Thése

Présenté par Lahsen Abouenour

UNIVERSITE MOHAMMED V DE RABAT DEPARTEMENT GENIE INFORMATIQUE

Three-levels Approach for Arabic Question Answering Systems

Octobre 2014

UNIVERSITE MOHAMMED V DE RABAT

ECOLE MOHAMMADIA D'INGENIEURS

THESE

Présentée pour l'obtention du

DOCTORAT EN SCIENCES

Au

DEPARTEMENT GENIE INFORMATIQUE

EQUIPE : Réseaux et Systèmes Intelligents

FORMATION DOCTORALE : Sciences et Techniques pour l'Ingénieur

Par

Lahsen Abouenour

Three-levels Approach for Arabic Question Answering Systems

Soutenue publiquement le **Samedi 11 octobre 2014** à **10h** devant le jury composé de :

Prof. Abdelkhaleq Cheddadi	PES, Ecole Mohammadia d'Ingénieurs, Université Mohamed V, Rabat	Président
Prof. Mohamed Benkhalifa	PES, Faculté de Sciences (FSR), Université Mohamed V, Rabat	Rapporteur
Prof. Horacio Rodriguez	Associate Professor, Université Polytechnique de Catalonia (UPC), Espagne	Rapporteur
Prof. Abdelilah Maach	Professeur Habilité, Ecole Mohammadia d'Ingénieurs, Université Mohamed V, Rabat	Rapporteur
Prof. Violetta Cavalli-Sforza	Assistant Professor, Al-Akhawayn University, Ifrane	Examinateur
Prof. Karim Bouzoubaa	Professeur Habilité, Ecole Mohammadia d'Ingénieurs, Université Mohamed V, Rabat	Directeur de thèse
Prof. Paolo Rosso	Associate Professor, Université Polytechnique de Valence (UPV), Espagne	Co-Directeur de thèse

Octobre 2014

DEDICATION

to my parents,

to my wife

to my children

to my sister and my brothers

to my friends

for their love, care and constant support,

with a sincere hope that this will make you proud of me.

Acknowledgments

I would like to express my deepest gratitude to **Dr. Karim Bouzoubaa**, first for supervising this thesis with enthusiasm, second for the innumerable meetings that we have held to discuss the main and detailed issues and third for his patience that were priceless encouragements for me. He kept motivating and supporting me along the way.

I owe an enormous debt of gratitude to my co-advisor, **Dr Paolo Rosso**, for many reasons. First, he put me in the right track saving me a lot of time and effort exploring uncountable paths. Second, he encouraged me to write papers and to submit them to the right conferences and journals. Third, I have learned from his long experience in the Natural Language Processing field. Fourth, for the different helps he gave me whenever I got in Valencia.

I am grateful to the committee members on my thesis, for all their insightful comments and remarks.

I would like to express my gratitude to **Dr Yassine Benajiba** and **Dr. Manuel Montes-y-Gomez** for their help and advises on the work I published in conferences and journals. **Dr José M. Gómez**, **Dr Jaouad Mousser**, **Emad Mohammed** and **Sandra Blasco** for making available their tools and resources.

I thank my family and friends for their incredible support; I could have never achieved so much without them.

This research has been supported by the grant of the **CODESRIA Council** (Conseil pour le Développement de la Recherche en Sciences Sociales en Afrique) under the reference number **SGRT. 38/T13**.

ملخص

موضوع البحث الذي أجري في إطار هذه الأطروحة هو تحسين فعالية أنظمة الإجابة الآلية على الأسئلة المعبر عنها كتابة باللغة العربية وذلك باستعمال الطرق السطحية والدلالية. هذه الأنظمة بالرغم من كونها ملائمة لتقليل حدة ظاهرة العبئ المعلوماتي المصاحب للكم الهائل من المحتوى الرقمي على الشبيكة إلا أن مجال البحث المرتبط بها والذي يندرج في إطار المعالجة الآلية للغة العربية لازال يتسم بندرة المحاولات وقلة الموارد.

الطريقة الهجينة المقترحة في هذا البحث تنقسم إلى ثلاثة مستويات: المستوى المبني على الكلمات وتقنية إغناء السؤال باستعمال العلاقات الدلالية بين الكلمات من الووردنت، المستوى المبني على بنية السؤال عن طريق نموذج مسافة الكثافة بين مقاطع "ن-غرام"، والمستوى المبني على التمثيل الدلالي باستعمال الرسوم البيانية المفاهيمية لمعنى السؤال من جهة ومعنى المقاطع النصية المرشحة من جهة أخرى بغرض عمل مقارنة آلية بحساب قياس التشابه الدلالي بين التمثيلين.

هذا البحث بيّن أهمية توسيع الموارد لتغطية أشمل للغة العربية وخاصة بالنسبة للقاعدة المعجمية ووردنت والتي اقترحنا تعزيزها بمحتوى جديد باستعمال موارد متاحة في لغات أخرى وكذلك بتطبيق تقنيات مختلفة من بينها تقنية المتراتبات المتواترة القصوى.

ومن أجل تبيان فعالية الطريقة الهجينة المقترحة وكذا تأثير التوسيع المعجمي، تم القيام بتجارب مختلفة على مجموعة من الأسئلة ذات تمثيلية مناسبة من حيث العدد، النوع، درجة تعقيد النصوص، الخ. هذه التجارب أظهرت تحسن أداء نظام سؤال/جواب بعد إدخال الطريقة الهجينة المقترحة ذات المستويات الثلاثة وذلك مقارنة مع نظام مرجعي مبني على محرك بحث وباستعمال قياسات معروفة في مجال بحثنا من بينها متوسط الرتبة التبادلية (MRR) و السي وان (c@1).

وتعتبر النتائج المحصل عليها بمثابة جواب إيجابي على السؤال البحثي الخاص بهذه الأطروحة حول مدى إمكانية بناء نظام فعال للإجابة الآلية على الأسئلة المعبر عنها كتابة باللغة العربية انطلاقا من الموارد والبرمجيات المتاحة ومدى قدرته على إظهار أداء مقبول للتعامل مع مختلف التحديات المطروحة في هذا المجال عامة وتلك المتعلقة باللغة العربية خاصة.

كلمات مفتاحية: أنظمة سؤال جواب، المعالجة الآلية للغة العربية، إغناء السؤال، شبكة الكلمات ووردنت، المعالجة الدلالية، التحليل النحوي، الموارد اللغوية.

Summary

The research conducted in this thesis presents an improvement of Arabic Question Answering (QA) systems performance through surface-based and deeper approaches. Although these systems are important to decrease the information overload problem, this Natural Language Processing (NLP) field witnesses just a few number of attempts and available resources.

The three-levels approach proposed in this work is composed of a keyword-based level relying on Query Expansion (QE) using Arabic WordNet (AWN) semantic relations, a structure-based level integrating the Distance Density N-gram (DDN) model and a semantic-based level considering the representation of meaning in both the question and the corresponding passages using the Conceptual Graphs (CGs) formalism and the comparison based on the semantic similarity score.

This research illustrated the importance of resource coverage enrichment, especially in the AWN lexical database that we extended using existing resources and various techniques, including the Maximal Frequent Sequences (MFS).

In order to show the effectiveness of this three-levels and hybrid approach, different experiments have been conducted considering question test-sets that are representative in terms of size, types, collection complexity, etc. The evaluation made shows an improvement of performance with the usage of the three-levels approach in comparison to the baseline system considering well-known measures such as the Mean Reciprocal Rank (MRR) and c@1.

The obtained results positively answer the research question of this thesis, i.e., the possibility of developing an Arabic QA system from existing resources and NLP tools with the ability to provide acceptable performance and to address the different QA challenges.

Keywords. Question Answering, Arabic Natural Language Processing, Query Expansion, Arabic WordNet, Semantic processing, Syntactic parsing, Linguistic resources.

Résumé

Les travaux de cette thèse présentent l'évaluation et l'amélioration des méthodes surfaciques et profondes dans le cadre des systèmes de Question Réponse pour la langue Arabe. Malgré l'importance de ces systèmes pour l'atténuation du problème de surcharge d'information, ce domaine du traitement des langages naturels connait une rareté au niveau des recherches associées ainsi qu'au niveau des ressources linguistiques utilisables.

La méthode à trois niveaux proposée comporte un premier niveau de traitement des questions selon les mots clés enrichis à travers les relations sémantiques de la ressource Arabic WordNet (AWN), un deuxième niveau portant sur la comparaison en tenant compte de la structure et de la densité des mots en utilisant le modèle Distance Density N-gram (DDN) et un troisième niveau basé sur la représentation en graphes conceptuelles dans un premier temps, et le calcul du score de similarité sémantique en passant par l'analyse syntaxique du texte dans un deuxième temps.

Un volet d'investigation a également concerné la proposition de méthodes semi-automatiques telles que Maximal Frequent Sequences (MFS) pour l'enrichissement de la ressource AWN et l'évaluation de l'impact de ce travail sur les performance.

Pour illustrer les performance de la méthode à trois niveaux proposée, plusieurs tests d'évaluation ont été effectués en utilisant des panels de questions présentant une bonne représentativité en termes de nombre, de type et de complexité, et ce en adoptant des mesures reconnues dans le domaine telles que la précision, le Mean Receprocal Rank (MRR) et le C@1.

Les résultats obtenus répondent positivement à la question de recherche qui est la possibilité de développer un système de Question Réponse pour la langue Arabe à partir des ressources et outils existants, d'atteindre des performance acceptables et de pallier aux différents défis de tels systèmes sur le plan général ainsi que sur le plan spécifique à cette langue.

Mots clés. Question Réponse, Traitement Automatique de la Langue Arabe, Extension de Requêtes, Arabic WordNet, Sémantique, Analyse Syntaxique, Ressources Linguistiques.

List of Tables

List of Figures

List of Abbreviations

Answered Questions (AQ)
Answer Extraction and Validation (AEV)
Arabic VerbNet (AVN)
Arabic WordNet (AWN)
Base Phrase Chunking (BPC)
Broken Plurals (BP)
Common Linguistic Categories (CLC)
Concept Confidence (CC)
Conceptual Graphs (CG)
Conditional Random Field (CRF)
Distance Density N-gram Model (DDNM)
Cross Language Evaluation Forum (CLEF)
General Architecture for Text Engineering (GATE)
Information and Data Reasoning for Answering Arabic Questions (IDRAAQ)
Information Retrieval (IR)
Iterative Query Expansion (IQE)
Java Information Retrieval System (JIRS)
Knowledge Base (KB)
Knowledge Representation (KR)
Local Context Analysis (LCA)
Machine Learning (ML)
Machine Translation (MT)
Maximal Frequent Sequences (MFS)
Mean Reciprocal Rank (MRR)
Modern Standard Arabic (MSA)
Named Entity Recognition (NER)
Natural Language Processing (NLP)
Part-of-Speech (PoS)
Query Expansion (QE)
Question Answering (QA)
Question Answering for Machine Reading (QA4MRE)
Question Analysis and Classification (QAC)
Passage Retrieval (PR)
Princeton WordNet (PWN)
Software Architecture For Arabic language pRocessing (SAFAR)
Search Engines (SEs)
Sound Plural (SP)
Suggested Upper Merged Ontology (SUMO)

Support Vector Machine (SVM)
Text REtrieval Conference (TREC)
Typed Dependencies-based Rule Confidence (TDRC)
Unified Verb Index (UVI)

Contents

Chapter 1

General Introduction

1.1 Background

Importance of Arabic

The role of the Arabic language has been prominent with respect to different perspectives. Historically, it was one of the main languages during the period of Arab science Golden Age, especially in Mathematics, Medicine, Astrology and Chemistry. Arabic has a tremendous religious significance in Islam considering that: (i) the Quran, one of the four Holy books, was revealed in Arabic, and that (ii) over 1.2 billion Muslims in the world pray five times a day using the Arabic language. Geographically, Arabic is an official language in 25 countries including the members of the Arab league. These countries are populated with more than 300 million people located between the Atlantic Ocean and the Persian Gulf including the Middle-east zone, making Arabic the fifth most commonly spoken language in the world[1].

Arabic requires research attention, especially in Natural Language Processing

The above elements show why the Arabic language was and is still important from a religion, social, economic and political angle. This also explains why it gets attention in various fields of research particularly in Natural Language Processing (NLP). Topics of interest in Arabic NLP include supporting usage of Internet in Arabian countries through efficient Search Engines (SEs), helping non-native Arabic speaking Muslims by providing language learning and Machine Translation (MT) tools, allowing companies targeting markets in the Arabic-speaking world to perform better data mining from opinion and sentiment detection tools over social media, etc.

Current state of Arabic NLP: maturity for some basic tasks

Work on Arabic NLP started in the 1970s, but the 2000s have witnessed an increase in Arabic-centered research due to its importance. As a result, some basic tasks such as Part-of-Speech (PoS) tagging, stemming and morphological analysis were well developed and practically reached maturity towards the late 2000s. They are no longer stumbling blocks holding back the development of higher-level Arabic NLP systems.

[1]According to the 2012 statistics of the *Ethnologue project*, available at: http://www.ethnologue.com/statistics/size

Arabic content is growing on the Web, classic SEs are not suitable, more sophisticated systems are needed

However, in today's world where the Web plays a key role in people's lives and companies' strategies, Arabic NLP is gaining momentum as the demand grows for new sophisticated systems as solutions to the increasing needs of users in terms of automatic text translation, information retrieval and extraction, etc. Recently, many surveys and statistical reports illustrate the impressive growth of online users as well as Arabic content[2]. For instance, this content has increased by around 2,501% between 2000 and 2012.

In this direction, the expansion of the Arabic language on the Web raises a classical problem, often referred to as information overload. Indeed, the availability of a huge amount of information written in Arabic cannot be efficiently exploited unless computers can make sense of the knowledge contained in this information and, in turn, help users with finding and/or extracting relevant content that better matches their queries or questions[3].

For other languages such as English and Spanish, this problem was reduced through the development of many NLP tools such as SEs that later were adopted worldwide. Unfortunately, the usefulness of these applications in the context of the Arabic language has shown some limitations due to various levels of differences between Arabic and those languages.

Complex IR systems can provide solutions beyond limitation of SEs for Arabic, but building them is a challenging task

Hence, providing effective solutions for the users of Arabic online content passes through the development of new Information Retrieval (IR) systems beyond the current used SEs. Unfortunately, the research on Arabic IR faces two main difficulties:

Firstly, Query Expansion (QE) and so on non-basic tasks that are useful for such systems are still lacking in maturity. Secondly, even though the second half of the last decade has known many efforts in developing new Arabic NLP resources, there are still some concerns about their availability, usability and coverage in particular annotated corpora, lexicons, ontologies, knowledge bases, etc.

Hence, the Arabic IR is still a field of opportunities

Consequently, researchers and professionals have still many opportunities in developing specific Arabic IR applications. For example, these applications can play an important role in

[2] Statistics of June 2012 from the Arabic Web Days initiative: http://www.arabicwebdays.com/front/initiative.aspx

[3] A q*uery* is a sequence of keywords (sometimes linked by boolean operators) which is used for querying an information retrieval system or a search engine, while a *question* is a precise query in natural language for asking a QA system.

some emerging trends such as the development of Arabic-based communication portals with the objective of targeting citizens by governments and customers by companies.

Motivated by the above, it makes sense to work on the case of Arabic Question Answering (QA) as complex IR systems, introducing QA systems, how they differ from classical IR systems and SEs, common challenges

The QA systems fall into this category of advanced IR applications that can bring valuable help for the users in their exploitation of the growing Arabic Web content or any other large document collection.

In comparison with classical IR applications and SEs, the idea behind these systems is allowing the computer to directly provide precise answers to natural language questions rather than lists of documents that require filtering efforts at the user end.

For instance, a user would need to know the name of the scientific capital of Morocco. Currently, the user can obtain the answer through the following tedious consecutive steps: (i) introducing the question "What is the scientific capital of Morocco?" (ما هي العاصمة العلمية للمغرب؟) to a classical SE, (ii) reviewing lists of returned snippets[4] in order to identify the documents that could potentially contain the right answer, (iii) accessing each potentially matching document and reading it in order to find the answer.

For the above simple question, step (ii) can be easy as usually the first list of snippets provided by SEs is sufficient to contain the answer. Therefore, step (iii) can be avoided if these snippets are relevant enough to contain the answer. However, for more complex questions, these two tasks (i.e. step ii and iii) may concern hundreds or thousands of snippets/documents. The role of an Arabic QA system is then to perform these two difficult steps instead of the users, saving their time and effort.

In the literature and similarly to what we have mentioned in the case of IR, it is reported that in order to be able to automatically process a question following the above pipeline of steps, a QA system requires not only basic NLP tasks such as PoS tagging and Base Phrase Chunking (BPC) but also more sophisticated tasks such as Named Entity Recognition (NER), Word Sense Disambiguation (WSD), Query Expansion, syntactic parsing, semantic representation and scoring, etc.

The users of Arabic QA systems would be interested in having the ability to automatically answer various types of questions. To design a system with such ability, researchers face the challenge of integrating most of the above NLP tasks.

[4] A snippet is short summary of a document, generally composed of two or three lines, which is displayed by SEs.

1.2 Problem description

To our knowledge, existing Arabic QA systems are limited either in terms of their scope as well as in terms of performance regarding the types of questions they are designed to answer. Today, the community of Arabic language users is still obliged to manually looking for precise answers to their questions which is a tedious task regarding the great amount of available Web information. This is due to the absence of a large scale and effective QA system.

To fill this gap, we first have to analyze the challenges that may be faced in the design and development of an Arabic QA system with such capabilities. These challenges are described in the following sub sections from the perspective of existing experiences.

1.2.1 Language challenge

Arabic is a highly inflected and derivational *language*. Its morphology and other particularities such as the absence of capital letters, the high ambiguity of undiacritized Arabic text and its syntactic flexibility have usually been reported as real challenges for PoS tagging, NER, QE, syntactic parsing as well as for other NLP tasks that are needed to develop Arabic QA systems.

According to the exiting works related to Arabic QA systems, the problem is how to deal with *language* challenges and particularities at different levels of question processing.

1.2.2 Web challenge

In the pipeline of processing a question, a QA system extracts answers from a *collection* of documents written in the targeted language. Recently, the great amount of information that is available online encouraged researchers for targeting the Web as a *collection*. Obviously, users of the online Arabic content are more interested in QA systems that can extract answers from the Web or any other huge document collection.

To our knowledge, there are currently no such systems for Arabic. The existing works only concern document *collections* that are not relevant for challenging QA systems. Indeed, we can identify here two kinds of works based on: (i) very small collections of documents (Hammo et al., 2004; Kanaan et al., 2009), or (ii) large collections with unique and formal sources such as Arabic Wikipedia (Benajiba et al., 2007a; 2007b).

1.2.3 Question and answer challenge

Another challenge concerns the *types of questions* and expected *answers*. The challenges are different depending on the complexity of the processed question. In the literature, these types range from the basic factoid questions where the user looks for a NE as answer (for instance, What is the name of the president of USA in the Second World War?) to HOW and WHY

questions (for instance, How did the financial crisis occurr at the end of the last decade?) where the user is looking for a more complex answer. Similarly to SEs, the popularity of a QA system within the users' community depends on its ability to process all types of questions.

The problem faced in most of the existing Arabic QA systems is two-fold: (i) in almost all these systems, the scope is limited to factoid questions. Moreover, despite this restriction, the overall performance remain unsatisfactory in comparison with other languages where maturity is already reached for these simple *types of questions*, and (ii) the processing of more challenging *questions* and the extraction of *answers* from different sources are still understudied.

1.2.4 Evaluation challenge

One of the key factors of success in the QA field is the organization of *evaluation* campaigns that helped researchers in benchmarking their systems according to standard measures and test-sets. The succession of yearly QA tracks such as TREC[5] and CLEF[6] allowed the improvement of performance to a mature extent for the considered languages and, recently, the development of more advanced QA tasks among which QA for Machine Reading (QA4MRE). Due to the absence of Arabic in the majorities of these tracks, the existing Arabic QA systems presented many drawbacks in terms of their evaluation process. [7]

Concerning test-sets, except the work on the ArabiQA system (Benajiba et al., 2007b) where the evaluation respected the same proportion of each NE class (PERSON, LOCATION, etc.) as in CLEF 2006, the other conducted experiments cannot be considered for comparison due to the nature and number of questions in their test-sets. These do not guarantee the coverage of challenging questions with representative percentages of different criteria (question length, NE classes, syntactical variations, etc.).

Unfortunately, the conducted experiments in the Arabic QA field are still lacking in: (i) deep experiences based on large question test-sets with relevant representativeness in terms of question types, length, domain, etc.; (ii) relevant results that can give a precise idea about the coverage and usability of used resources and tools in the context of Arabic QA; and (iii) **standard-de-facto** measures as well as nature of targeted collections that currently constrains any comparison with other similar systems.

[5] Text REtrieval Conference, http://trec.nist.gov/data/qa.html
[6] Cross Language Evaluation Forum, http://www.clef-campaign.org
[7] The QA4MRE campaign has been organized on behalf of the CLEF conference, representing an evolution of previous evaluation approaches. Machine Reading requires a deeper analysis and inference of text and in turn may need background knowledge acquisition. The Arabic language has been considered in this campaign since the 2012 edition.

1.2.5 Research question

From the above situation of Arabic QA, it is worth asking the following question that we will try to answer in the current research: Leveraging the current advances registered in different basic and non-basic Arabic NLP tasks, is it possible to build a QA system that can automatically answer different types of Arabic questions, deal with the above Arabic QA challenges and reach acceptable performance even in tricky contexts such as the Web?

1.3 Objectives

To be able to answer the above research question, the core of the work presented in this thesis has different objectives that try to face each of the previously mentioned challenges:

- *Language challenge*: introducing, in each stage of the Arabic QA pipeline, new NLP resources and tools that take into account one or many particularities of this language.
- *Web challenge*: designing an approach that makes it possible to automatically answer questions from the Web and from huge collections giving rise to various challenges for an Arabic QA system.
- *Question and Answer challenge*: considering different types of questions and expected answers aiming to deal with most users' needs.
- *Evaluation challenge*: conducting relevant experiments on Arabic QA with respect to the well-known measures and relevant sizes and natures of collections.

The above main objectives are broken down into detailed objectives, namely:

- Designing an effective approach for the improvement of the key module of Arabic QA systems. The objective here is to enhance passage ranking/scoring not only by considering surface-based approaches that are *language* independent but also by considering deeper approaches dealing with particularities of Arabic ;
- Identifying the main causes of failing in Arabic QA by analyzing the positive and negative examples with respect to the used approach and according to each *question type*;
- Analyzing the coverage and usability of the current Arabic resources and how they can be enriched for a better integration in the QA task;
- Investigating the usefulness of variety of tools related to different layers of Arabic NLP including stemming, POS tagging, NER, syntactical parsing, semantic representation, etc;

- Studying the extent to which *language* independent techniques can be useful for Arabic QA and its related tasks;
- Evaluating the impact on Arabic QA of each of those resources and tools following the trends of the well-known *evaluation* campaigns such as TREC and CLEF;
- Measuring the overall performance after applying the proposed approach on a large set of questions and within the context of the Web as a challenging *collection*;
- Comparing this approach with baseline SEs and other QA systems;
- Implementing the proposed approach as part of an integrated system where the other third-party NLP resources and tools can be combined, adapted and/or enriched for the sake of better performance of Arabic QA.

1.4 Document structure

The present document is structured as follows:

- Chapter 2 presents a study of existing works either related to the QA field as well as to its dependent tasks such as QE with a special focus on Arabic QA. This chapter is composed of four main sections. Section 2.1 introduces the QA task. Section 2.2 details the general architecture of a QA system and the evaluation of such systems. Section 2.3 highlights the main advances in Arabic QA; this section, first, introduces the main specific challenges of Arabic QA, and, second, emphasizes existing works in terms of either complete systems or QA modules for the Arabic language. Section 2.4 shows, for the purpose of benchmarking, the best systems developed for other languages. Section 2.5 draws a conclusion for the main issues of Chapter 2.
- Chapter 3 is devoted to the presentation of the three-level approach proposed for improving the Arabic PR module as part of a QA system for this language. This chapter is divided into two main sections: Section 3.1 introduces the chapter by recalling the status of the studied field (i.e., Arabic QA) and the objectives of this thesis. It also describes the methodology followed in this research. Section 3.2 provides, first, background information that presents the needs for Arabic PR modules, and, second, the three-levels approach. Indeed, a presentation at a glance gives an overview of the three levels before moving into details and describing experiments of the surface-based levels of this approach, namely, the keyword-based and structure-based levels. Section 3.3 summarizes this chapter.
- Chapter 4 goes through the investigation of the impact of resource enrichment on the performance of an Arabic QA using the surface-based levels described in Chapter 3. After the introduction of Chapter 4, Section 4.2 gives a theoretical analysis of Arabic

WordNet (AWN), the main resource used in the three-levels, according to three perspectives, especially its comparison with other WordNets. Section 4.3 presents another analysis of AWN coverage and usability through experiments. Section 4.4 provides details of the AWN extension we propose in order to overcome the shortcomings of this resource as shown in theoretical and experiment-based analysis. Section 4.5 gives an evaluation of this extension with respect to two different question test-sets. Section 4.6 draws the main conclusions of this chapter.

- Chapter 5 addresses the semantic-based level with the aim to process the types of questions requiring the understanding of meaning rather than the comparison of surface elements. After an introduction, Section 5.2 provides the necessary background related to the approaches based on similarity at a semantic level. Section 5.3 describes the ontology built for the purpose of this level. Section 5.4 presents and evaluates the implementation of this level using two experiments conducted with different test-sets of questions. Finally, a conclusion draws the main research results obtained in the semantic-based level.
- Chapter 6 is structured around four sections after an introduction. Section 6.2 introduces the integrated NLP platforms and their main objectives as well as their support for the Arabic language. Section 6.3 presents the SAFAR platform used in this work. Section 6.4 details the proposed architecture of the Arabic QA system called "IDRAAQ" as part of SAFAR platform and shows how the developed and separate modules of IDRAAQ can be used in similar applications.
- Chapter 7 draws the general conclusions of this research. Section 7.1 recalls the findings and research directions studied in this thesis, Section 7.2 highlights its main contributions and Section 7.3 discusses further challenges to be tackled in future works.
- Appendix A lists the papers where the results of the current research were published.
- Appendix B lists the 11 syntactic rules designed for the construction of Conceptual Graphs (CG).
- Appendix C provides the meaning of each tag used in the Stanford parser.

Chapter 2

Approaches and resources for Arabic Question Answering

2.1 Introduction

Interest in building QA systems started since the attempt made by Green et al. (1961) through the system called "BASEBALL". In 1965, the paper of Simmons (1965) addressed the efforts made by fifteen systems for automatically answering English questions. These first implemented systems focused on specific domains and questions. Then, the field moved to new trends thanks to the availability of online information and to the series of organized evaluation conferences and QA tracks. The interest of the most prominent companies such as Google, Yahoo!, Microsoft and IBM[1] in making such projects attests to the growing popularity of this field.

Typically, QA systems are built around a general architecture that combines many NLP components with the aim to automatically answer different types of user questions. The efficiency of these systems is measured through an evaluation process using relevant test-sets related to the targeted language.

In Arabic QA, challenges are not limited to those commonly faced by systems developed for other languages such as English or Spanish. Each integrated Arabic NLP component may positively or negatively impact the performance of the system unless it considers the particularities of this language such as its complex morphology and syntax, its high ambiguity, especially in undiacritized texts, the absence of capital letters, etc.

The objective of this chapter is fourfold: (i) introducing the main concepts of the QA task that will be used throughout the remaining chapters; (ii) drawing focus to Arabic QA specific challenges to help us understand the particularities of building such systems for Arabic; (iii) presenting existing systems and previous efforts in tackling Arabic QA; this overview will help us define priorities as well as measure the significance of our contribution; and (iv) reviewing the most important QA experiences in other languages as guidelines to be followed for the development of the Arabic QA task.

To address this fourfold objective, this chapter is structured as follows: Section 2.2 introduces some generalities about the QA task, the types of questions it tries to answer and the

[1] The IBM Watson is a computer with an integrated QA system developed in the context of the DeepQA project. IBM Research undertook a challenge to build a computer system that could compete at the human champion level in real time on the American TV quiz show, Jeopardy (Ferrucci et al., 2010).

evaluation process of such systems. Section 2.3 presents, through examples and existing experiences, the challenges faced and the advances made in terms of building systems for Arabic QA. Section 2.4 recalls some of the most successful works regarding other languages. Section 2.5 provides a synthesis of this chapter.

2.2 The Question Answering Task

2.2.1 Overview of the QA task

As a recall, a QA system, differently from a SE such as Google, tries to directly display the answer to user questions without presenting lists of candidate passages for further manual filtering. Figure 1 illustrates the difference between actual SEs and an ideal QA system.

As Figure 1 depicts, while both QA and SEs systems allow users to introduce inputs in natural language, each of them has a particular scope. Indeed, classical SEs can help users looking for information about a topic formulated through keywords (or logical expressions). They provide exhaustive results in terms of document lists related to the searched topic. On the other hand, QA systems are more suitable for users whose main need is getting a precise answer to a question without being requested to manually filter lists of documents related to the question keywords.

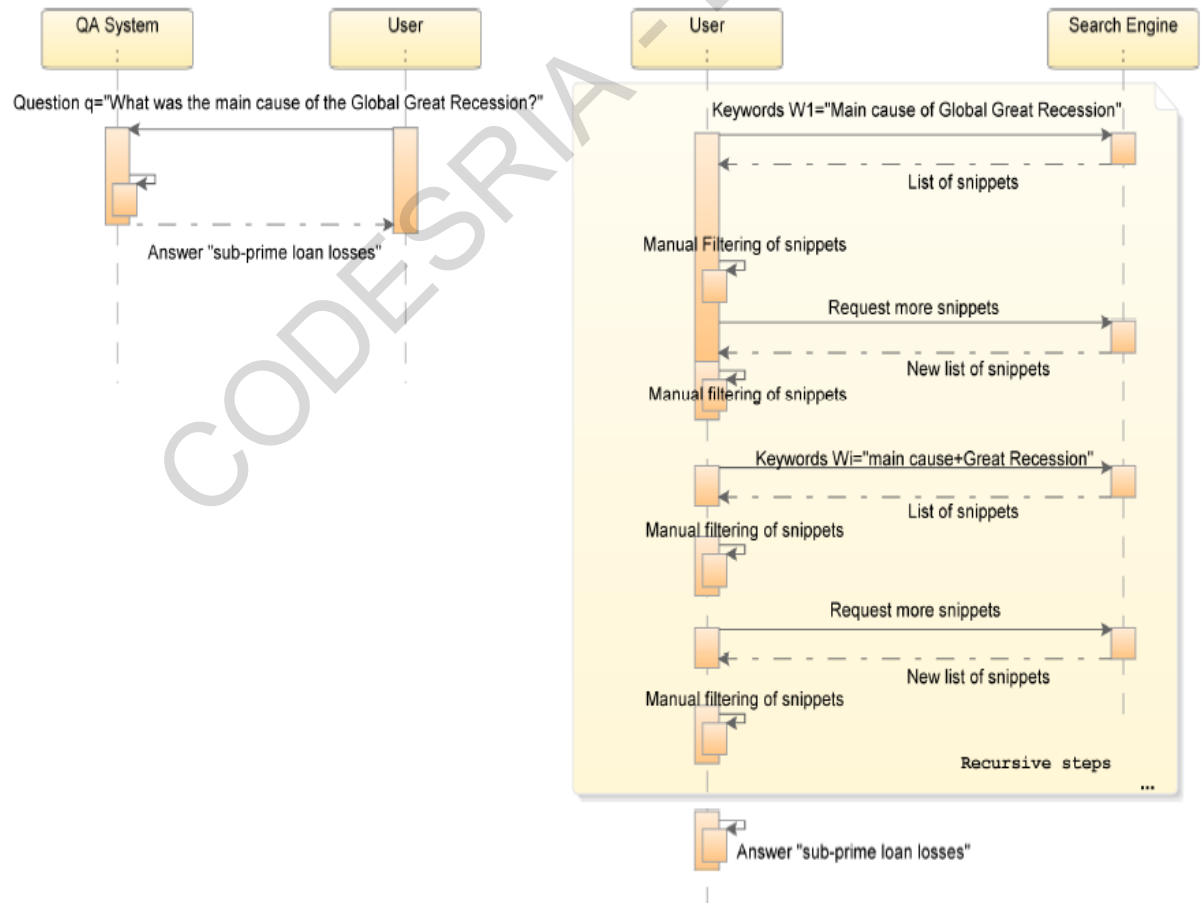


Figure 1. Difference between QA systems and SEs

Figure 1 shows the main steps performed by users in both cases (i.e., when using a QA system versus a SE such as Google). After a question (or a query) is introduced by the user (for example, what was the main cause of the Global Great Recession?), the QA system directly retruns the answer (i.e., sub-prime loan losses), whereas the SE outputs a list of snippets related to relevant documents. If the user tries to get the answer to his/her question through a SE, he/she is requested to manually make multiple steps in an iterative way to: (i) view the list of returned snippets, (ii) search for the needed answer, (iii) move to another list of snippets, and eventually (iii) change the query keywords.

In practice, a user can manually filter only a small number of snippets because the answer is usually in the first few documents returned by a SE (although it could be potentially in the middle or at the end). The QA system tries to carry out these steps for the end user. As a result, the user gains in terms of time and in terms of the amount of information that otherwise he/she has to process.

In the case of QA systems, two main types of data are processed: the question and a targeted collection. The targeted collection can be composed of documents written in the given natural language, Web pages that are accessed online by the system or any other information source.

The questions can be classified by their type and/or domain. For instance, when a user is looking for the answer to the question "Who is the Spanish football player who scored at the world cup 2010 final?", the expected answer is a name of a person and this question's type is called "factoid question", i.e. question for which the answer is a named entity (name of person, organization, place, etc.). Table 1 provides samples of question types that a QA system is requested to automatically answer.

Table 1. Summary of question types and challenges

Type	Expected answer	Examples	Challenges
Factoid	Getting a Named Entity (person, place, organization, etc.) related to a fact	Who is the first president of USA? *Answer: George Washington*	▪ Less challenging questions since question structure and keywords more likely occur in documents ▪ Evaluation of the answer is easier
List	A list of NE items	What are the most visited places in Morocco? *Answer: Marrakech, Agadir, Fez, Tangier*	▪ It is more probable that the expected list is scattered over different documents ▪ Evaluation of the answer is difficult
Definition	Information about a NE	Who is Ibn-Batutah? *Answer: he is a Moroccan explorer researcher and*	▪ Similarly to List questions, the answer can be collected from different documents ▪ Evaluation of the

		geographer	answer is difficult
Other	The answer can be of any type: Yes/No, facts, arguments, etc.	- Does Djibouti belong to the Arab Nations League? *Answer: YES* - What is the reason for decline in tourism activity in the world since 2009? *Answer: economic crisis*	▪ Inference and machine reading techniques are required ▪ Evaluation of the answer is difficult except for YES/NO questions and particularly for longer answers.

As shown in Table 1, the challenges of building a QA system differ from the question type perspective. Obviously, when the project only concerns factoid questions and/or list and definition questions about NEs, the task as well the used techniques may not require advanced processes. On the contrary, when the system is open to other allowed types of questions, it is necessary to integrate modules that allow obtaining an in-depth understanding of question and documents through different techniques among which we can cite deduction, text entailment, semantic reasoning, etc. The QA4MRE campaigns attempt to address also this kind of challenging questions.

Another challenging perspective is the domain to which questions belong. For instance, we can build a QA system devoted to all types of questions but only for a specific domain such as Biomedicine, Sport, etc. This domain limitation can help in reducing the language and question type challenges. Unlike open-domain QA systems, domain-specific QA systems can be more efficient since dedicated materials (resources, tools, question templates, etc.) can be built and ensure a high coverage of user terminology and needs. Nevertheless, approaches based on resources with high redundancy cannot be useful for restricted domain QA systems since related resources are small in size.

Beyond the type and domain of the processed question and the expected answer, there are two known approaches for answering user questions by a QA system:

(i) ***Surface approaches*** based on the comparison between strings of question and targeted documents. Generally, these techniques are language independent and, hence, are of limited interest especially when the gap between the question string and the answer string in documents is large. Techniques that help in reducing this gap will be explained later in this chapter.

(ii) ***Deep approaches*** trying to understand the user question and the knowledge in the available content. In this case, many other NLP tasks can be used such as language preprocessing (text segmentation, tokenization, etc.), processes from different NLP layers (morphology, syntax, semantics, etc.) and statistical and machine learning models. Actually,

these approaches are concerned with a deep analysis and language-dependent tools so that the most challenging types of questions (see Table 1 above) can be answered by the system. Nevertheless, these approaches are harder to implement due to the challenges of natural language processing.

Before presenting the state of the art of the existing QA systems with respect to the above approaches, we focus on the general architecture of QA systems and their evaluation.

2.2.2 General Architecture of QA systems

Whether using a surface or a deep approach, the existing QA systems follow a generic architecture. This architecture is a pipeline of three main modules: Question Analysis and Classification (QAC) Module, Passage Retrieval (PR) Module and Answer Extraction and Validation (AEV) Module (see Figure 2).

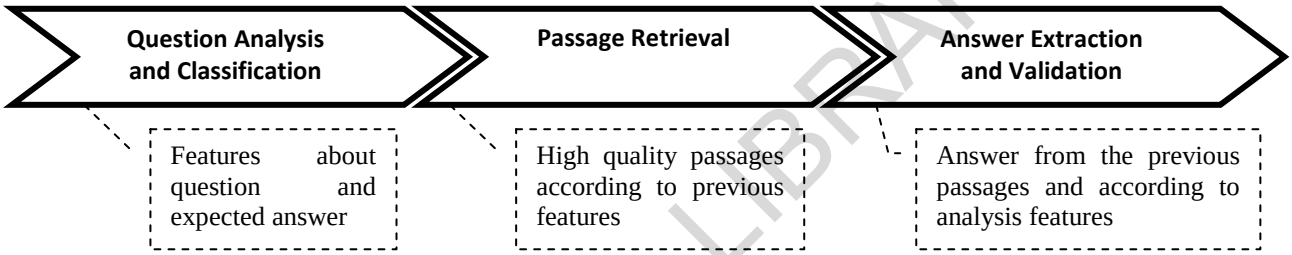


Figure 2. General architecture and modules of a QA system

As we can see, the purpose of each module can be summarized as follows:

(i) ***QAC module*:** In this module a question is analyzed in order to identify its type, extract its pattern and the structure of the expected answer, form and/or reformulate the query to be passed to the PR module, determine constraints on the expected answer, etc.

(ii) ***PR module***: This module is a core component of the QA system. Generally, an IR system (for instance a SE such as Google or Yahoo!) is used to retrieve documents and passages. Thereafter, this module has to perform a ranking process in order to improve the relevance of the candidate passages that better match the user question.

(iii) ***AEV module***: This module tries to extract the answer from the candidate passages provided by the previous module. In advanced QA systems, this module can be designed to construct the answer from one or many passages. Obviously, the AEV module will fail to return the correct answer if the candidate passages provided by the PR module are not relevant and do not contain the answer. There are some systems that also integrate answer validation in this module.

Let us present an example which illustrates how each module works and what kind of data it retruns. According to the above architecture, a QA system will process the question “What are the places in Morocco most visited by French tourists in the last decade?” as follows:

- The QAC module determines that the given question is of type “LIST” i.e., the expected answer is a list of items. In some cases, the module has a predefined list of question patterns and the question may be matched with these patterns. For instance, the module decides that the right pattern is “What are X [in P] [by Z] in Y?”. In this case, the module generates the structure of the expected answer under the format “X [in P] [by Z] in Y {are, is, ...}” so that the PR module makes a special emphasis on documents (or passages) containing this structure (in a superficial QA approach). The QAC module also defines the bag of words to be used in the queries. In our example, these words are for example: most; visited; places; tourists; last; decade. The module is concerned by the recognition of NEs (in the example: Morocco and French). It also identifies constraints on the answer (last decade: temporal constraint) in order to filter candidate answers according to these constraints.
- Basically, the PR module tries to extract from the document collection the best paragraph-sized fragments of text (i.e., passages) that are similar to the user question in terms of keywords and structure. In our example, let us assume such passages are as follows:
    - *Passage 0:* “the most visited places in Morocco by French tourists in the last decade are not the same for those coming from Germany”
    - *Passage 1:* “the most visited places in Morocco by French tourists in the last decade are Marrakech, Rabat and Fez”
    - *Passage 2:* “the most visited places in the Kingdom of Morocco by French tourists in the last decade are mainly those located in the cities of Marrakech, Rabat and Fez”
    - *Passage 3:* “Fez is one of the most visited cities in Morocco by French tourists in the last decade”
    - *Passage 4:* “In the last decade, French tourists have most visited Fez, Rabat and Marrakech”
    - Etc.

  Obviously, the PR module returns the first passage above when the collection contains a document with this passage. However, in real texts, the situation is often more complicated. The PR module may be challenged if:
    - none of the documents contain any passage related to the question;
    - there are documents with similar keywords but different structure (for instance passage 4 above) ;

- there are documents with similar structure but different terms (passage 2 and 3 above);
- etc.

Note that passages can be retrieved following two main approaches: (i) indexing each passage as separate document and retrieving it as such; (ii) retrieving relevant documents for a given question and then retrieving passages from these relevant documents (Khalid 2008). Both approaches require additional processes to face the above challenges. One of these additional processes is QE. In the case of the given example, a QE process can generate new terms for the NE "Morocco" such as "Kingdom of Morocco" or for the keyword "places" such as "locations", "cities" and "regions". These new terms can then be added in the question structure and be used at the retrieval stage. The consistence of existing QE techniques will be described later in this chapter.

- The AEV module usually integrates two sub modules for Answer Selection and Answer Validation respectively. The former concerns the pre-processing of passages (coming from the PR module) in order to extract sub content from them with its features; the first sub module returns a list of candidate answers. The latter sub module determines the correctness of this list on the basis of their features. The importance of the AEV module can be illustrated through "Passage 0" in the previous example. Indeed, even though this passage contains similar question keywords and structure, the pre-processing of the passage at the AEV module will result in the sub content "not the same for those coming from Germany"; from this content, the system can identify "Germany" as a candidate answer since the user expects a list of places and "Germany" is tagged with the NE feature "LOCATION". Nevertheless, the Answer Validation sub module will reject this candidate answer if it has additional information: Germany is not a place in Morocco.

As we can see from the above example, the performance of each module is impacted by the performance of its predecessors in the pipeline. Moldovan et al. (2003) reported that more than 36% of errors in QA are due to mistakes of question classification. (Llopis et al., 2002) assets that the quality of the results returned by the QA system depends mainly on the quality of the PR module it uses.

The previous example also shows two main findings:

- The three basic modules of a QA system have complementary roles. The information extracted and generated by each module can help in the other ones. Hence, their performance are highly dependent each to another;

- Features about question keywords are of great importance in the three modules and in turn for the whole system. For example, questions asking about NEs require information related to the expected NEs.

In order to obtain an efficient QA system, each module has requirements in terms of approaches and NLP components. Thus, measuring the impact of each used approach and component is of great interest at the development stage, and in turn for the usability of the system in real situations.

The performance of such systems can be measured at the module level and/or at the system level. Due to the importance of the evaluation in the process of building QA systems, we devote the following section to the presentation of the well-known evaluation campaigns, their trends in terms of evaluated systems and the resulting QA-oriented measures.

2.2.3 Evaluation in the QA field

2.2.3.1 Evaluation campaigns

Starting from 1987, an important trend has emerged in the NLP field: the organization of evaluation campaigns related to different tasks such as speech processing (Pallett 2003) and text understanding (Harman 1992). The Information Retrieval community kept track of this trend and witnessed the organization of the TREC (Voorhees and Harman 2005**)** in USA.

The field of IR was one of the most concerned by the so called "evaluation paradigm" (Adda et al., 1998). Indeed, the number of IR evaluation campaigns and the number of participants denote the importance of these campaigns for IR researchers.

Regardless of the nature of the evaluated task, these events succeeded in pushing forward the efforts made by researchers. They were, and still are, a framework for providing new data sets, developing methodologies for new topics of the concerned task, and bringing together all relevant actors to objectively compare their techniques:

In Europe, since 1994, the NLP and IR research communities have known an ongoing series of evaluation campaigns related to a variety of tasks including Morpholympics for German morphological analyzers (Hauser 1994**)**, Grace for French Part-Of-Speech taggers (Adda et al., 1998), Senseval and Semeval for lexical semantics (Edmonds and Kilgarriff 2003; Agirre et al., 2007) and the Conference and Labs of the Evaluation Forum (i.e., CLEF) for information access systems with an emphasis on multilingual and multimodal information with various levels of structure (Agosti 2007). In the 2012 edition of CLEF, even non-European languages such as Arabic have been introduced in the QA for Machine Reading track. Thus, the evaluation and benchmarking of Arabic QA systems can be encouraged and supported by this decision.

The significance of results presented in the above campaigns highlighted the issue of evaluation and its importance in the cycle of NLP projects. Thus, new terms are used within the community such as:

- Progress evaluation: in this evaluation, the current state of a system is assessed against a desired target state,
- Adequacy evaluation: in this evaluation, the adequacy of a system for some intended use is assessed,
- Diagnostic evaluation: in this evaluation, the assessment of the system is used to find where it fails and why,
- Hypothesis vs. reference data: hypothesis refers to data produced by the systems participating in an evaluation campaign while data created to represent the gold-standard are called “reference” (Mitkov 2005),
- etc.

Over the last two decades, the organized evaluation campaigns have followed a typical model which is composed of four phases as illustrated in Figure 3.

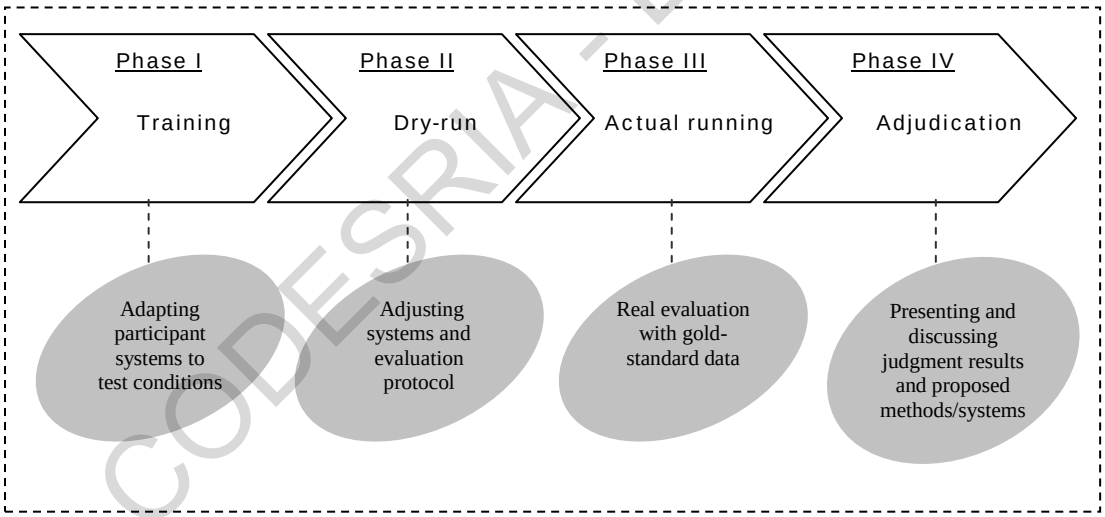


Figure 3. *Typical phases of an evaluation campaign*

As depicted in Figure 3, the two first phases of an evaluation campaign allow participants firstly to adapt their systems to the conditions of the final test (for instance taking into account test formats in terms of inputs and outputs) and secondly to perform any needed adjustment in terms of evaluation protocol or system functionalities using a small set of gold-standard data. The last two phases are simply those of the actual competition where participants process gold-standard data using their systems, send the output results for adjudication and get final results and ranking. Generally, a workshop is organized to reveal the final results, to present evaluated systems and methods and to have discussion between participants.

The QA evaluation series have always been held as a task under the different IR campaigns. This is the case of TREC for English, CLEF for European languages, the National Institute of Informatics Test Collection for IR systems (NTCIR[2]) for Japanese, the Russian Information Retrieval Evaluation Seminar (ROMIP[3]) for Russian language and the INititative for the Evaluation of XML Retrieval (INEX[4]). The Arabic language was rarely one of the languages in these QA tasks. To our knowledge, the only editions where this language was introduced are:

- TREC during 2001-2002 in the Cross-Language Information Retrieval track on Arabic CLIR[5],
- CLEF@2012 in the main task of the QA4MRE workshop[6].

2.2.3.2 QA measures

In general, in order to measure the performance and effectiveness of IR systems, a test collection is needed. This collection is composed of (Manning et al., 2008):

- a document collection which is a list of content to be indexed or formatted according to the system need,
- a set of queries expressed in natural language,
- a set of relevant judgements assessing a pair of document-query "relevant" or "nonrelevant". This is called the "gold standard" or "ground truth" judgment of relevance.

With respect to the evaluation of IR systems, two kinds of situations can occur: (i) unranked retrieval situations, and (ii) ranked retrieval situations. In the former, the system returns a set of documents for a query while in the latter, this set of document is ranked or restricted to the top k documents that better match the query (k is a number to be defined in the system).

In unranked retrieval situations, the evaluation is usually made through the following measures (Manning et al., 2008):

- ***Precision (P)*** which is the fraction of retrieved documents that are relevant to the question:

  $$P = \#(\text{relevant items retrieved}) / \#(\text{retrieved items})$$

- ***Recall (R)*** which is the fraction of relevant documents that are retrieved:

[2] http://research.nii.ac.jp/ntcir/
[3] http://romip.ru/en/
[4] https://inex.mmci.uni-saarland.de/
[5] http://terpconnect.umd.edu/~oard/research.html#trecclir
[6] In Chapter 4, we provide more details about this campaign in which we participated using our Arabic QA approach.

R = #(relevant items retrieved) / #(relevant items)

- ***F-Measure (F)*** which combines precision and recall and represents their weighted harmonic mean. The most used formula for F is:

$$F = 2*P*R / P + R$$

In the QA field, which is commonly referred to as a sub field of IR, the above measures are commonly used for unranked retrieval. Note that the evaluation of QA systems can be done for the whole system and/or for each module, especially the PR module and the AEV module. In both cases, the situation of ranked retrieval occurs. Hence, the above measures are not useful since they only inform about the effectiveness of the system in returning and in covering a high number of precise documents without highlighting the ability of the system to provide documents that are ranked according to their relevance.

As we have previously seen, the different QA evaluation campaigns use measures more suited to this task. The commonly used measures in the context of those campaigns are:

- ***Accuracy***: this measure is used to evaluate the quality of the overall QA system that provides one potential answer. Accuracy is a number between 0 and 1 that indicates the probability that the QA system will provide the correct answer on average. It is expressed as following:

  ```
  Accuracy = Number of correct answers / Number of questions
  ```

- ***Mean Reciprocal Rank (MRR):*** this measure is used to evaluate the quality of the overall QA system that provides a sorted list of 'n' potential answers. MRR is a number between 0 and 1 that indicates the quality of the sorted list of 'n' potential answers. The formula to compute MRR is the following:

$$MRR = \left(\frac{1}{|Q|}\right).\sum_{i=0}^{|Q|}\left(\frac{1}{rank(i)}\right)$$

- ***Answered Questions (AQ):*** is another measure for QA systems providing a sorted list of 'n' potential answers. AQ is a number between 0 and 1 that indicates the probability of the QA system to provide a correct answer in its sorted list of 'n' potential answers. AQ is expressed as follows:

$$AQ = n / |Q|$$

Where:

n is the number of answered questions. Note that a question is answered when the correct answer is contained in the list of answers returned by the system regardless the rank of the correct answer

|Q| is the number of considered questions

Since this measure does not penalize the system when it does not rank the correct answer as first, we can consider it as a relaxed version of both the accuracy and the MRR.

- ***c@1 measure:*** This measure, used in previous CLEF QA tracks since 2009, encourages systems to reduce the number of incorrect answers while maintaining the number of correct ones by leaving some questions unanswered. It is represented by the following formula:

  $$c@1 = (nr + nu * (nr/n)) / n$$

  where:

  nr: is the number of correctly answered questions

  n: is the total number of questions

  nu: is the number of unanswered questions

2.3 Advances in Arabic QA

The above characteristics of QA systems in terms of architecture and evaluation are commonly followed by researchers independently of the targeted language. Nevertheless, the core modules of these systems have to be developed and adapted to face the challenges specific to each language.

The high level of complexity of Arabic morphology and syntax are among the specific challenges that make the task of building an Arabic QA system tougher in comparison to other languages. In the next sub section, we present some of these challenges in the context of the QA task. Afterwards, we present existing Arabic QA works with a special focus on their level of resolution of such challenges.

2.3.1 Arabic QA challenges

Independently from the targeted language, each module of a QA system requires the integration of other basic and non basic NLP tasks. Figure 4 illustrates these requirements among the modules of a typical architecture of a QA system.

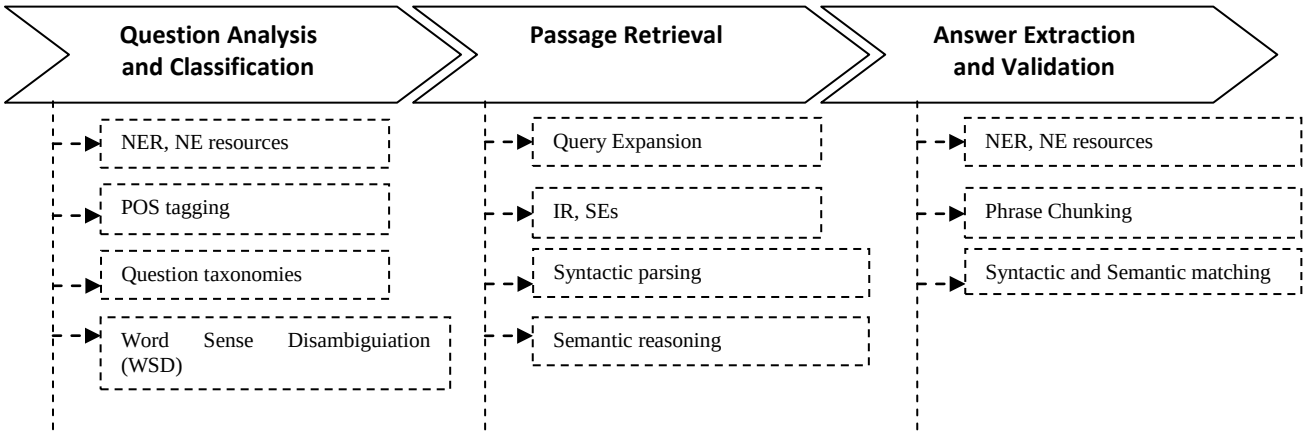


Figure 4. Sample of NLP tasks used in QA systems

Figure 4 gives an idea about some of the most useful tasks regarding a specific QA module. For example, many works related to the PR module have reported the use of QE, SEs/IR techniques, syntactic parsing and/or semantic reasoning. Also, PoS tagging, WSD and NER are among the tasks commonly used in the QAC module.

Obviously, performing each of the above related tasks is challenging because of the Arabic language particularities. In the following sub sections, we provide, through examples, the principles of these particularities that turn Arabic QA and its related NLP tasks into a challenge.

2.3.1.1 Arabic script

Arabic is written using a specific alphabet writing system called “Arabic script”. This system also used in other languages (for example Persian, Urdu, etc.) is different from the Latin system. Currently, the most available and efficient NLP tools are developed for languages such as English that uses Latin script. Researchers in the Arabic NLP community usually adopt some of these existing tools in their works by means of transliterating Arabic text using Latin characters (for example by adopting Buckwalter’s transliteration). [7]

2.3.1.2 Ambiguity in Arabic QA

Unlike Latin languages, Arabic is written using diacritics (i.e. *fatha, damma, kasra)* that play the role of vowels. Apart from children books and religious text, it is rare to find texts with full diacritization.

Beyond the classical ambiguity problem that is common to NLP for different languages, the non use of diacritics leads to additional challenges. This causes more ambiguous situations than any other language. The average is 19.2 per Arabic word versus 2.3 in other languages (Farghaly and Shaalan 2009).

[7] Throughout this document we use the Buckwalter table (see http://www.qamus.org/transliteration.htm) to transliterate Arabic words into Latin characters

Let us take some examples to show how this explosive ambiguity can present real challenges to Arabic QA:

- The first issue can be explained through the question " **من هو الشخص الذي قتل في جريمة قطار الشرق في الرواية الشهيرة لأجاثا كريستي؟** " (mn hw Al$xS Al*y qtl fy jrymp qTAr Al$rq fy AlrwAyp Al$hyrp l>jAvA krysty, i.e., "Who is the person who was killed in the Orient Express crime in the famous novel of Agatha Christie?"). The absence of diacritics in verb "**قتل**" presents at least two cases for the QA system: (i) "**قَتل**" with *fatha* above the letter "**ق**"which means that the question is "Who did kill in the famous story of Agatha Christie, *Murder on the Orient Express?*" so "**قتل**" here means "kill", and (ii) "**قتل**" with *damma* above the letter "**ق**"which means that the question is "Who is the person who was killed in the famous story of Agatha Christie, *Murder on the Orient Express?*" so "**قتل**" here means "was killed". Obviously, the AEV module will be negatively impacted by this ambiguity problem due to the non use of diacritics. Arabic QA systems that try to answer to questions from the Web will be more concerned by this challenge since the online content is usually undiacritized.
- The second issue rises when a PR module integrates a QE process. In our example, the verb of the question is one of the most interesting keywords for expansion. Trying to perform this expansion for the verb "**قتل**" (we suppose here that a WSD process implemented in the QAC module has already disambiguated the word "**قتل**" and eliminated all noun cases such as "**قتل**", i.e. "Killing" with *fatha* above the letter "**ق**" and *soukon* above the letter "**ت**") means to generate related terms such as "**إغتال**" i.e. to assassinate or "**إغتيل**" (i.e. to be assassinated). In the case the user looks for the person who "was killed", the former term is not relevant and will bring some noise to the PR module.

2.3.1.3 Complex morphology

Arabic is a highly agglutinative and derivational language. In Arabic, a word may replace a whole sentence in other languages. For instance, the sentence "*and with their return*" can be expressed in one Arabic word "**فبعودتهم**" which includes the stem "**عودة**" (i.e. return), the prefix "**فب**" (i.e., and with) and the pronoun ""**هم**" (i.e. plural pronoun). Thus, extracting keywords at the QAC module of an Arabic QA system will be more complex than any other language. Applying a light-stemmer (Khoja 1999) or a classical morphological analysis such as BAMA proposed by Buckwalter (2004) may be enough in some basic IR systems but not in advanced QA where the exact need of a user has to be caught by the system. In a question like " **من المخترعان الأمريكيان اللذان ينسب لهما صناعة أول طائرة؟** " (Who are the two American inventors that are known as the first creators of an aircraft?), the user looks for the name of two persons (i.e.

Orville Wright and Wilbur Wright). In English QA, the system catches this user need through the word "two". In Arabic, this information is embedded in the word "**المخترعان**" thanks to the suffix "**ان**". Actually, the above is just an example; the morphology of an Arabic word may contain large number of morpho-syntactic information (basic POS, gender, number, mood, case, etc.) that are important for each module of Arabic QA.

Concerning the derivational aspect of the Arabic language, it was reported that most of the Arabic words are derived from a three-letters root (sequence of three Arabic letters) and very few are from four or five letters roots. The effectiveness of IR and QA based on root, stem or word at the indexing or retrieval stage is still under research with different findings (Abu-Salem et al., 1999; Aljlayl and Frieder, 2002; Darwish and Oard, 2003; Larkey et al., 2007; Benajiba et al., 2007a; 2007b).

Similarly, discussion is still open about the usability of QE based on the generation of morphologically related words relying on root or stem. In fact, replacing a word with its morphologically related forms can completely change the meaning of a question. For instance, in the question "**متى كشف العالم النظرية؟**" (When did the scientist reveal the theory?), replacing the verb "**كشف**" (to reveal) by a verb with similar root such as "**اكتشف**" (to discover) results in changing the expected answer from the time of revealing the theory to the time of its discovery.

2.3.1.4 Syntax particularities

It was depicted in the previous sections that a QA system needs deeper analysis and understanding of the question at different levels especially syntax and semantic ones. In the Arabic language, the basic order of words is Verb-Subject-Object (V-S-O), but S-V-O, V-O-S, etc. are also possible (Green and Manning 2010). This may raise some issues in the Arabic PR module. Let us take the previous sample question "**متى كشف العالم النظرية؟**" (i.e. When did the scientist reveal the theory ?). The classical PR approach consists in retrieving the passages that contain the same word order as in "**كشف العالم النظرية**" (i.e. the scientist reveals the theory). However, the collection may contain the other possibilities listed in Table 2.

Table 2. Examples of the different word orders in Arabic sentences

VSO	**كشف العالم النظرية سنة 1994** *The scientist revealed the theorem*
SVO	**العالم كشف النظرية سنة 1994** *The scientist revealed the theorem*
VO	**كشف النظرية سنة 1994** *He revealed the theorem*
VOS	**كشف النظرية العالم سنة 1994** *The scientist revealed the theorem*

As we can see, it is also important that the PR module considers the different situations of the question keywords order. The syntactical analysis of Arabic is more challenging with respect to these word order possibilities, to ambiguity of undiacritized text and to the complexity of morphology for each word. Note that as we have seen, words may embed pronouns that replace the subject or the object in V-S-O, S-V-O, etc.

Regarding the different challenges of NLP in Arabic, the performance of syntactic parsing tools of this language is a constraint for their usability in the context of QA systems. Let us recall that at a semantic level, a QA system may use a technique that relies on syntactic parsing. Among these techniques we can cite the identification of semantic role labeling that was reported as promising for shallow semantic parsing (Gildea and Jurafsky 2002).

2.3.1.5 Named Entities in Arabic QA

As we have seen in the introduction of Section 2.3.1, NER is one of the most used tasks in QA systems, particularly in the analysis module. This task has been performed for the Arabic language based on algorithms such as in the work of Benajiba and Rosso (2008) based on supervised Machine Learning (ML) techniques namely Maximum Entropy, Support Vector Machine (SVM) and Conditional Random Field (CRF) or relying on resources and rules (Zaghouani 2012). Despite these efforts, the availability of public Arabic NER tools is still lacking.

For other languages such as English, NER systems commonly use capital letters as a main feature to identify NEs. These systems also rely on ML techniques to classify the NEs (Nadeau and Sekine 2007). With respect to the above techniques, the Arabic language presents some specific challenges. One of these challenges is that Arabic does not use capital letters. This constraint can be passed over for foreign NEs in Arabic text (for instance **جراهام بل** and **ألفريد نوبل**) by applying a morphological analyzer ; if the word cannot be analyzed, it is more likely that it is a NE. In undiacritized Arabic text, however, this technique is not efficient for Arabic NEs since they can be confused with adverbs or verbs. For example the question "**أين ولد عارف الطويل؟**" (Where is born Aref Tawel?) is asking for the place of birth of a Syrian actor (i.e. Aref Tawel). The two words composing this NE can also be interpreted by the morphological analyzer as adverb and adjective (**عارف الطويل** means the taller man who knows).

As an alternative to NER tools, a NE ontology such as YAGO (Suchanek et al., 2007) can be used and be integrated in an Arabic QA system.

2.3.1.6 Lacks of resources for semantic processing

The lacks of available resources has always been mentioned as an obstacle for Arabic NLP projects. It concerns almost all the tasks that can be integrated in a QA system. Resources are needed either for processing and evaluation.

A) Resources for processing

The Arabic NLP community needs in terms of resources range from the most basic ones such as electronic lexicon, corpora and dictionaries to the most advanced ones such as ontologies and knowledge bases. The last decade witnessed many efforts in the development of public resources especially those belonging to the first category. The needs of Arabic QA community are much more important regarding the nature and complexity of this task.

Ontologies can play a key role in a QA system. For instance, in the previous question "**من هو الشخص الذي قتل في جريمة قطار الشرق في الرواية الشهيرة لأجاثا كريستي؟**" (Who is the person who was killed in the famous story of Agatha Christie, *Murder on the Orient Express?*), humans can extract the right answer from the following passage:

ليكتشف في الصباح مقتل المسافر الأميركي راتشيت في المقصورة المجاورة له باثنتي عشرة طعنة متفاوتة القوة و التوصيف
Only to discover in the morning that the American Traveler Ratchet was killed in the nearer cabin with twelve stabs varying in terms of strength and characterization

An Arabic QA system can do the same if it has the information that "**مسافر**" (traveler) is a sub concept of "**شخص**" (person). Such information can be found in an ontology of the Arabic language such as the Arabic WordNet[8] (Felbaum 1998; Elkateb et al., 2006) or the one proposed by Jarrar (2011).

B) Resources for evaluation

As we have seen in Section 2.2.3, building a QA system passes through many experiments that require relevant test collections and data. For Arabic, there are just a few available resources for evaluation (Ezzeldin and Shaheen 2012):

- The first test set developed for Arabic IR and QA was the TREC 2001 and TREC 2002 text collections containing only 383 872 documents (some 800MB of data), the English TREC WT10g collection contains 1.6 million documents (10GB of data), and the English TREC GOV2 text collection contains 25 million documents (420GB of data) (Nwesri 2008).
- The second test set was proposed and made available for public by Benajiba et al. (2007a). This collection developed in the framework of the ArabiQA system contains 200 question/answer pairs and 11,000 documents from the Arabic Wikipedia in SGML format. Note that this is the format accepted by the CLEF campaign (see Section 2.2.3.1).

[8] In fact, AWN is a semantic network of Arabic words grouped into synsets rather than concepts.

In this section, we presented the most known particularities of the Arabic language with regard to QA systems. Despite these particularities turn into real challenges, the Arabic QA field has known various attempts to overpass them. The following section presents the systems developed in the context of these attempts.

2.3.2 Existing Arabic QA systems

The first works on Arabic IR started in 1990s with a limitation in terms of text collections size. The main focus of that works was the evaluation of the effectiveness of indexing by root, stem or surface words. In 2001, the TREC campaign has considered a track of 75 queries for testing Arabic retrieval as a monolingual and cross-lingual task (Nwesri 2008).

To our knowledge, the first built Arabic QA system is called "AQAS" (Mohammed *et al.,* 1993). The system presented some restrictions in terms of processed data that are mainly structured in the context of a knowledge base. Ten years after this attempt, the Arabic QA system called "QARAB" was proposed by Hammo *et al. (*2004). Unlike AQAS, this system is based on a set of rules for each question type excepting "WHY" and "HOW" questions that require more advanced processing. The performance of QARAB has not been tested following state-of-art evaluation methods. In fact, the only reported experiments are made by four native speakers who checked the correctness of QARAB answers for 113 questions. These experiments show that a recall and a precision of 97.3% were obtained.

Building new Arabic QA systems have gained much interest in the community of Arabic NLP, especially with the following works:

- ArabiQA (Benajiba et al., 2007b): authors of this work prepared an evaluation corpus on the basis of CLEF guidelines. Using this corpus, it was reported that light stemming has a positive impact over the PR module. An AEV module that obtained an accuracy of 83.3% has also been integrated in that system. This AEV system relies heavily on Named Entity Recognition.
- QASAL (Brini et al., 2009): is an attempt for building an Arabic Q/A which processes factoid questions (i.e., questions that have NE answers). Experiments have been conducted and showed that for a test data of 50 questions the system obtained 67.65% as precision, 91% as recall and 72.85 as F-measure.
- Kanaan *et al. (*2009) described a QA system for short Arabic questions relying on IR and NLP techniques. The authors used a text collection with 25 documents from Internet, 12 questions and some relevant documents. The reported performance obtained 100% precision for 0, 10 and 20% recall and 43% precision for 90 and 100% recall. According to the small size of the used collection and question set, their experiments cannot be considered as reliable results to compare with.

- AQuASys (Bekhti et al., 2011): is an attempt at building an Arabic QA system which is composed of three modules: A question analysis module, sentences filtering module and an answer extraction module. The system was evaluated with a set of 80 questions. The system obtains a precision of 66.25 %, a recall of 97.5 and an F-measure of 78.89.

A description of the modules integrated in these system-oriented works is provided in the next section. In each QA module, we also mention other attempts that are qualified as component-oriented, i.e., attempts where the main objective is to enhance a particular Arabic QA module or component regardless its integration in a full system.

2.3.3 Arabic QA modules

2.3.3.1 Question analysis classification

The QAC module of QARAB uses the Type-Finder and Proper Name-Finder system implemented by Abuleil and Evens (1998). In that module, the question type is identified by means of short list of stopwords (i.e.When, Who, etc.).

Authors of the ArabiQA system (Benajiba 2007b) built their own QAC module that integrates a developed component for NEs recognition and classification. Let us recall that NER is among the tasks that are most commonly used in a QAC module. The existing NER systems namely Siraj[9] by Sakhr, ClearTags[10] by ClearForest, NetOwlExtractor[11] by NetOwl and InxightSmartDiscoveryEntityExtractor[12] by Inxight are all for commercial ends.

The QASAL (Brini 2009) system integrates a QAC module that allows returning information about the processed question such as the NE representing the focus of the question, the keywords of the question, its class and its schemata.

2.3.3.2 Arabic passage retrieval

The PR module of QARAB (Hammo 2004) integrated a QE process based on root relatedness. This module generates new related terms having the same root of the original question keywords.

In ArabiQA (Benajiba 2007b), the PR module uses the Java Information Retrieval System (JIRS)[13] (Gomez et al., 2005) as a core component for passage scoring based on Distance Density n-gram Model. JIRS has been adapted for the Arabic language (Benajiba et al., 2007a) and used following the architecture illustrated in Figure 5.

[9] http://siraj.sakhr.com/
[10] http://www.clearforest.com/index.asp
[11] http://www.netowl.com/products/extractor.html
[12] http://www.inxight.com/products/smartdiscovery/ee/index.php
[13] JIRS is an Information Retrieval system developed by Gomez et al., (2005). Unlike the traditional search engines that are based on question keywords, JIRS retrieves passages that will most likely contain the answer. Its search technique is based on question n-grams using three different language independent models.

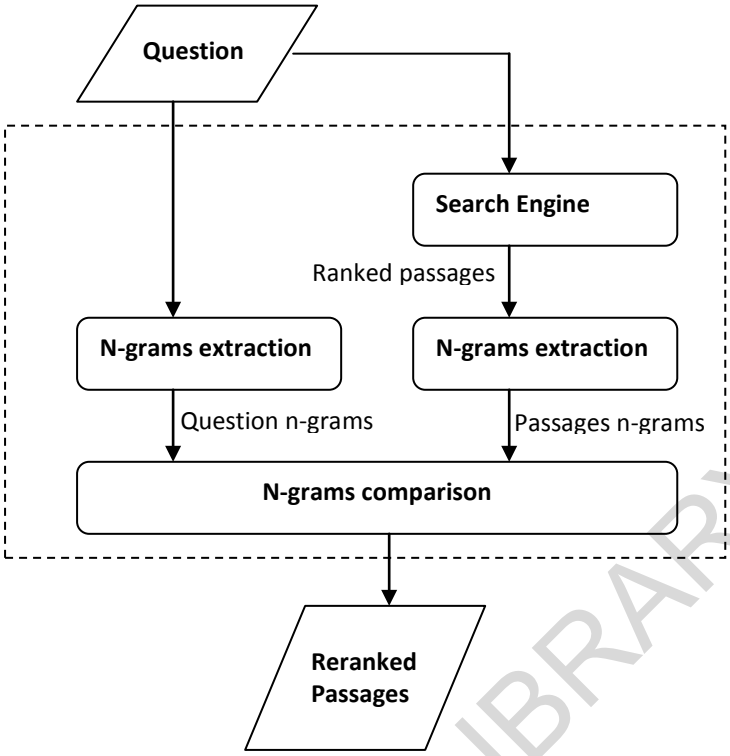


Figure 5. Usage of JIRS in ArabiQA
Source: (Benajiba 2007a)

Figure 5 shows that JIRS is used on top of a classical SE with the aim to improve the passage ranking. The experiments conducted by Benajiba et al. (2007a) are the first ones to consider the same questions proportion as in well-known evaluation track (CLEF 2006 in this case). The results with the use of JIRS and a NER system for Arabic helped to obtain a 83.3% precision.

In the QASAL system, the authors have used a passage based approach which consists in considering each passage as separate document. The retrieval of passages is made following two steps: (i) Step 1: retrieving passages related to the question focus (i.e. NE if factoid question, verb or noun otherwise), and (ii) Step 2: retrieving relevant passages by considering the other keywords of the question. In the second step, weight is assigned to each term.

As an improvement of the PR module of QASAL, the developers of the system cited the usage of Arabic WordNet as a QE resource in order to extract synonym terms.

2.3.3.3 Answer extraction and validation for Arabic

The importance of AEV is its ability to provide a precise answer to the user question instead of long passages or list of documents. For factoid questions, the extraction of a NE in the candidate passages can be enough to provide a right answer. For the Arabic QA systems, this is not a trivial task according to the particularities of Arabic previously described. The need of

an accurate NER for Arabic for this task rises. The NER developed in the framework of ArabiQA was reported to provide competitive F-measure (83.5 on the ACE 2003 BN data[14]).

For definition questions, this task is even more challenging for Arabic since the user needs definition passages about the question focus. Trigui et al. (2010) used manual lexical patterns of sequences at word, letter and punctuation levels as well as heuristic rules deduced from a set of correct and incorrect definitions. Preliminary experiments on 50 questions about organizations have been conducted using the above AEV method.

Other researchers have investigated N-gram matching method for Arabic AEV module. Abdelbaki et al. (2011) tested the usefulness of semantic similarity and N-gram matching between question's focus and candidate answers. The reported results are around 86% accuracy and 0.87 MRR. The experiments used a small test-set based on the ANERCorp[15] containing 316 articles and 240 questions.

2.4 Non Arabic QA systems experiences

In QA field, many research works are devoted to the English language. The existing systems consider different types of questions ranging from the simplest ones such as factoid questions to the most complicated ones (for instance why-questions, definition and opinion questions) requiring deeper approaches.

The future development of Arabic QA, either in a monolingual or in a cross-language context, can leverage the main lines and approaches investigated in the existing experiences related to other languages.

The next sub sections describe these existing experiences from two complementary perspectives: system-oriented and module or component-oriented. Section 2.4.1 shows the most important development and evaluation lines of whole QA systems. Section 2.4.2 highlights other works that only consider separate modules.

2.4.1 Development and Evaluation of system-oriented QA

The development of QA systems has known different trends in terms of targeted collections and used approach, within the communities of researchers attending the evaluation campaigns held annually or those working on Open Source projects.

2.4.1.1 Targeted collections

The early QA systems extracted answers from structured data. In 1993, a system called MURAX (Kupiec 1993) was among the first QA systems searching for answers in a document collection. In that work, the electronic version of an encyclopedia was used to

[14] http://www.nist.gov/speech/tests/ace/
[15] http://www.dsic.upv.es/%7Eybenajiba/resources/ANERCorp.zip

answer questions from the quiz game *Trivial Pursuit*. With the growth of available information on the Web, it was normal that many QA researchers thought using the online content to get the expected answer. The popularity of Web based QA systems can also be justified by the available Web search APIs that the common Search Engines such as Google, Bing and Yahoo! offer for developers. MULDER Kwok,(2001), NSIR (Radev 2002), ANSWERBUS (Zheng 2002) and START (Katz 2002) fall into this category of Web-based QA systems. Their performance, in particular LAMP (DellZhang 2002), is comparable to the best state-of-the-art question answering systems.

2.4.1.2 QA approaches

Initially, the most used approaches were surface-based (i.e., based on statistical techniques or symbolic/pattern matching). Thereafter, new approaches have emerged such as:

- *Rule-based* QA systems integrate heuristic rules that mainly rely on lexical and semantic features in the questions. This is the category of many existing projects such as Quarc (Riloff 2003) and Noisy channel Echihabi,(2000).
- *Knowledge representation-based* approaches, where the question and the passages are compared on the basis of their semantic representation.
- *Domain-oriented* approaches have the aim to reach higher accuracy by using a domain-specific Knowledge Base (KB) and/or a set of rules for the given domain. Obviously, the restriction means that a smaller amount of information is required in the built KB, which in turn means that the project is feasible in terms of budget (time, money and resources). This line of QA research was one of the lines followed by early built systems such as LUNAR (Woods 1972) for Geology domain and BASEBALL (Green 1961) which answers questions about the US baseball league. Other attempts in this category of QA systems are Geographic (Chung 2004), Biomedicine (Zweigenbaum 2003), WEBCOOP (Benamara and Saint-Dizier 2004).

2.4.1.3 Communities of QA research

A) Evaluation campaigns community

Over the editions, TREC and CLEF campaigns moved from traditional tracks of questions (short factoid and definition questions) and collections (especially newswires) to new challenges. The evolution of both campaigns is illustrated in Figure 6.

As illustrated in Figure 6, along the tracks of QA evaluation, each edition of CLEF and TREC advanced following one or many axes, in particular in terms of: (i) question types, (ii) evaluation (measures, answer validation, etc.), (iii) nature of the QA task (classical or more advanced such as machine reading and opinion detection), and (iv) content of the targeted collection.

For instance, Figure 6 shows that the edition 2002 of TREC was characterized by the use of a new measure, i.e. the confidence score, the CLEF 2011 was interested in a new collection types such as blogs and topic-oriented collections along with a new QA task, i.e. Machine Reading.

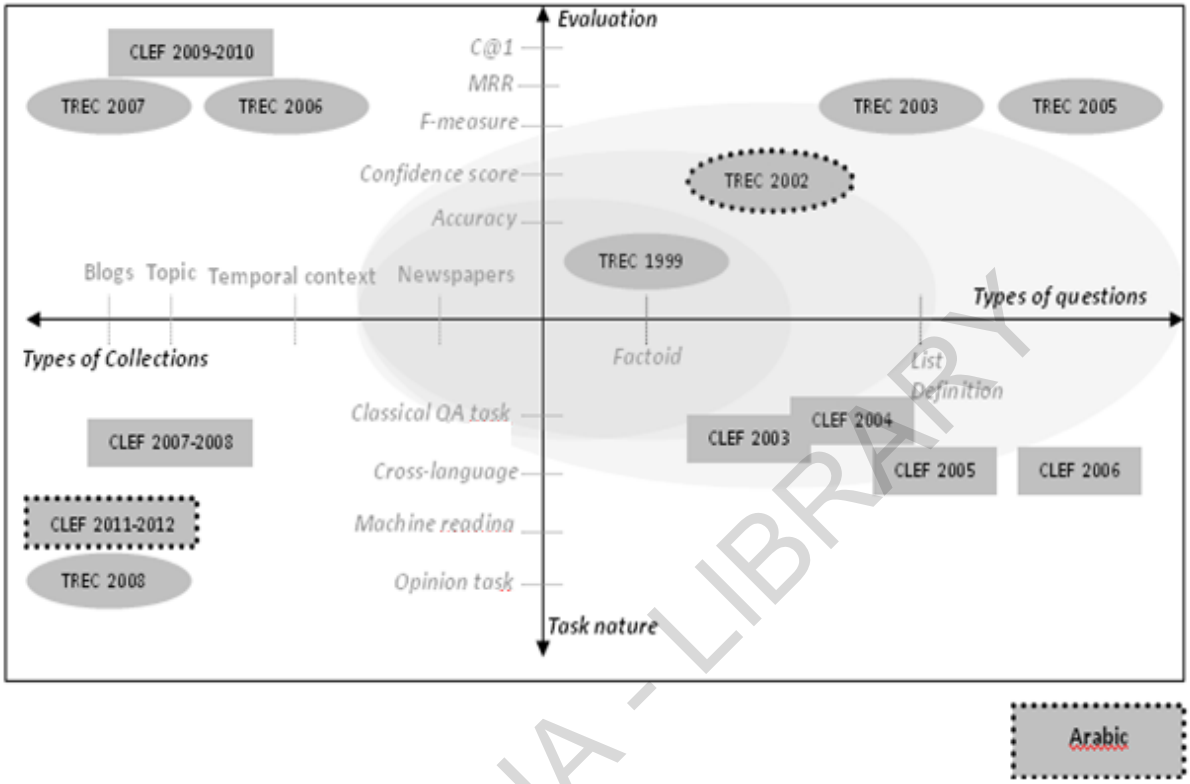


Figure 6. Evolution of the QA task in TREC and CLEF campaigns

Following, a summary of the main trends registered in these campaigns with respect to the four above axes:

- **TREC**
  - 1999 in TREC-8: factoid questions
  - 2002: the confidence-weighted score is used to assign confidence to answers
  - 2003: in addition to factoid questions, list and definition questions are also considered
  - 2005: events were added as possible targets
  - 2006: temporal context of the document collection is considered
  - 2007: further challenges to the QA by integrating blogs (less formal language) in the document collection. Best system for factoid questions (accuracy of 0.706), for other questions (F(β=3)=0.329)

- 2008: NIST organized QA track at Text Analysis Conference (TAC[16]). Factoid questions were removed and questions about opinions in blogs were highlighted. In this edition, definition and list questions were kept.

- **CLEF:** Unlike TREC, cross-language systems have been considered starting from the first edition (2003). IR engines were widely used by systems participating in CLEF over years. No real deeper understanding of documents was performed by these systems. Results did not go beyond 60%.
  - 2003: tasks to test monolingual (Dutch, Italian and Spanish) and cross-language (Dutch, French, German, Italian and Spanish source language queries to an English target document collection). Questions were generally short, factoid, unrelated to subjective opinions and not on definitions and multi-item answers (i.e. List questions). 10% of questions have "NIL" answers (the systems have to provide empty answers if no correct response in the document collection is found). The basic evaluation measure was the MRR (it will be described later in the next section). Average performance was 41% of correct answers in the monolingual task and 25% in the cross-language one. Accuracy reaches 29% and 17% respectively (Magnini et al., 2004).
  - 2004: questions mostly factoid, also definition and "how" questions introduced. Accuracy is the main measure. Average accuracy was 23.7% and 14.7 for bilingual runs (Magnini et al., 2004). The Confidence-weighted Score is used.
  - 2005: the number of target languages became 12 (Amharic, Bulgarian, Chinese, English, French, German, Hungarian, Indonesian, Italian, Portuguese, Russian, and Spanish).
  - 2006: closed (What are the two players who scored in the last match of Manchester United in 1999?) and open (Name schools in Rabat) list questions were considered.
  - 2007-2008: the track was focused on topic-related QA. Heterogeneous document collections were used.
  - 2009-2010: in ResPublicQA, a parallel document collection was used for multilingual QA. Focus on legal documents. Reason and Opinion questions are also considered. Systems were allowed to return either passages or exact answers. A new evaluation measure called `c@1` was used to reward systems

[16] http://www.nist.gov/tac/

that reduce the number of questions answered incorrectly without affecting systems accuracy.

- 2011-2013: QA4MRE was the new considered task. This task requires deep knowledge of text meaning. Systems used documents as well as background collection to extract the answer. The background collection is a variety of documents from different sources (newspapers, web, blogs, Wikipedia, etc.) on three topics, namely Aids, Climate change and Music and Society. Systems need a cognitive process with inferences, implications and presuppositions, etc. in order to extract the answer that can be implicit in the document.

B) Open source community

The community of Open Source QA systems has known significant works with projects such as Open Ephyra (Schlaefer et al., 2006) and Aranea (Lin 2007). Also, the community of Knowledge Representation (KR) has been interested in QA systems; therefore, the best performing systems integrate some kind of inference or reasoning (Peñas 2008).

2.4.2 Component-oriented QA

The objective of this section is to provide an overview of the main achievements in each module or component of the QA architecture. This is also the opportunity to show the requirement of each QA module in terms of resources and evaluation.

2.4.2.1 QAC Module

The Question Analysis and Classification module has been addressed in many works, among them we can cite Moldovan et al. (2000) who used the TREC-8 training to manually construct a question type hierarchy of about 25 types. Hovy et al. (2001) analyzed a set of 17,000 questions to come up with a question typology which is composed of 47 categories. Named Entities taxonomies can be useful in the classification of factoid questions. In this direction, it is reported that accuracy of classification is obtained with small NE taxonomies instead of large ones (Kurata et al., 2004).

Zhang and Zhao (2010) presented a Question Classification that uses words, named entities, PoS and semantics as classic features to classify the question.

The importance of question classification is its ability to provide a pattern or a structure of the expected answer. This is particularly useful in QA systems that use a surface-based approach where the candidate passages are defined by their similarity to the question or, better, to the expected answer. There are two main methods to question classification:

- *Statistical methods* using ML techniques: these techniques usually require large sets of annotated questions (van Zaanen 2002). Hence, lack of training

data can be a constraint to the use of these methods despite their effectiveness reported in many works (Li and Roth 2006);

- *Symbolic methods* based on pattern matching: this technique is the most used in recent evaluation campaigns. Since it is language-specific, this technique is based on pattern matching rules that are in most cases manually created (Dridan and Baldwin 2007). An alternative to this is learning patterns by querying the Web with question/answer pairs (Schlaefer et al., 2006).

In (Dridan and Baldwin 2007), a comparison between the two methods for the Japanese language shows that with less training data the former method can reach a higher accuracy. The identification of additional constraints such as grammatical relations between question terms was also among the issues studied in QAC (Scott and Gaizauskas 2001; Harabagiu et al., 2001).

One of the important pieces of information that a QAC module provides concerns NEs appearing in the question. We have seen above that the classification of the question can be performed on the basis of NE taxonomies. Nevertheless, this is not the unique way QA systems need information about NEs. This information is usually extracted by means of a NER system. Other methods use large NE ontologies such as YAGO[17] which contains 2 million entities. The positive impact of using a NER in QA is confirmed in many studies such as in (Noguera et al., 2005). Some researchers have used NER as a standalone component; this means that they use the NE classes taxonomy of the NER tool. Some others have proposed a QA-oriented NER in order to introduce a NE classes taxonomy which is more suitable to QAC. In (Moll et al., 2006), authors show that the use of multi-label QA-oriented NER system increases recall of named entities and benefits the task of QA.

As a summary to this part, we have seen that the works related to the QAC module were interested in analyzing and/or classifying questions at two levels: (i) *keyword level*: where POS taggers, NER systems and other tools and resources are required, and (ii) *question level*: in which the question is classified using different techniques such as question taxonomies, machine learning, pattern matching, etc.

2.4.2.2 PR Module

The number and diversity of research works related to PR revealed the importance of this module in the architecture of a QA system. In this section, we describe exiting experiments on three main sub tasks: (i) passage ranking as the core of basic PR modules, (ii) query expansion that can be used to support passage ranking and PR as well to consider shallow semantic features, and (iii) syntactic and semantic matching that also can support passage ranking with advanced semantic comparison between questions and passages.

[17] Yet an Other Great Ontology, available at http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html

A) Passage ranking

Usually, the main processed units in IR system are whole documents while QA systems only consider parts of text (passages) from these documents at the answer extraction. Passages are retrieved from the targeted document collection. This collection can be built from Web pages, wikis, blogs, databases or any other source of information that is not necessarily structured. From the TREC and CLEF campaigns, it turned out that with the huge amount of available information either on the Web or in databases, a collection of documents can not efficiently be queried by a QA system (mainly by the PR module) unless a pre-processing step is performed. This step helps in using the collection offline and getting an acceptable response time for users. The pre-processing step remains necessary for systems that use SEs such as Google or Yahoo! to get the candidate passages. This is the case of languages with a high Web redundancy such as English and Spanish. The existing approaches adopt this pre-processing stage in different ways, particularly in:

- document indexation and annotation (Prager 2001),
- logical representation of documents (Molla Aliod et al., 1998),
- shallow linguistic processing of document collection including tagging, NER and chunking (Milward and Thomas 2000).

Preparing a pre-processed document collection has shown great impact on the improvement of the AEV module as well as the PR module. However, adopting such solution is not always possible due to its cost in terms of time and money (linguistic experts have to get involved in such projects). This solution also does not meet the main objective of QA systems and NLP applications in general that is processing unstructured data.

In the context of PR, we distinguish between semantic, discourse and window-based passages. In each type of PR, boundaries are defined by identifying changes in topic, discourse, or byte spans, respectively. Beyond the type of used PR, two main approaches can be found in the literature for passage ranking. The first one consists in dividing documents into passages in the targeted collection and then ranking these passages. The second approach consists in indexing and ranking the documents at search time and then retrieving passages from the ranked documents (Callan 1994).

Ranking passages is the core process in the PR module. This process has to measure the distance between the question and each document/passage in the targeted collection. Weighting schemes help in assigning a relevance score to each document on the basis of:

- *Term frequency*: one of the most used algorithms in IR and PR module of QA systems in particular is tf*idf (Term Frequency * Inverse Document Frequency) technique (Murdock and Tesauro 2012). Let us recall that its main idea is the fact that a document is more relevant to a query term if the latter

occurs many times in it and less in all other documents. This technique decrease the noise of stopwords at document/passage ranking since obviously those stopwords appear in many documents and, therefore, idf will decrease TF*IDF. In the case of factoid questions the technique improves results with NE terms. For instance, if a document collection contains Wikipedia entries about NEs, the TF part increases the relevance of the Wikipedia document that describes the NEs since it will be cited in that document more times than any other document. The IBM's Watson QA system called "DeepQA" uses this technique in its PR component (Chu-Carroll et al., 2012).

- *Statistical analysis*: the most known family of statistical scoring functions is bm25 (Robertson et al., 2000). The function is based on a probabilistic information retrieval model that considers document features such as term frequencies, document frequencies, and document length.

Even though both scoring and ranking techniques come from the IR field, they are widely used in QA systems after being adapted to work on passages instead of whole documents. Hence, these techniques are not effective when they deal with some QA specific challenges. For instance, these classical techniques are not sensitive to relations between question terms and cannot help in identifying passages where these relations occur in addition to that terms. Typically, a PR algorithm comes after a document retrieval one. Generally, only topic based document retrieval is performed before PR.

There are many attempts to propose PR algorithms with higher effectiveness in the context of the QA task. Tellex et al. (2003) compare eight PR algorithms, namely:

- Alicante (Llopis and Vicedo *2001*): relies on the number of appearances and *idf* values of the term in the query and passage;
- *bm25* (Robertson *et al., TREC 4*): the previously mentioned algorithm based on a probabilistic information retrieval model;
- IBM (Ittycheriah *et al., TREC 9*): based on weighted sum of various distance measures (matching words, thesaurus match, miss-match words, dispersion, cluster words);
- ISI (Hovy *et al., TREC 10*): it is also based on weighted sum of various features such as matching of NEs, question terms and their stems;
- MITRE (Light *et al., J. of Natural., Lang. Eng., Special Issue on QA 2001*): this is a baseline algorithm which counts the number of query terms that appear in the sentence;

- MultiText (Clarke *et al.*, *TREC 9*): it also uses weights of terms through *idf* with the particularity of assigning more importance to short passages with many query terms;
- SiteQ (Lee *et al.*, *TREC 10*): this algorithm weights passages (3 sentences window) based on the density of question terms.

Figure 7 illustrates the main findings of that work. In fact, the eight considered algorithms have been trained on TREC-9 and tested on TREC-10. The comparison between those algorithms was based on the MRR and percent of unanswered questions.

The first finding is that the used document retrieval system can differently impact each algorithm. This difference is statistically significant in the case of PRISE and Oracle retriever. The former is more suitable for PR and confidence ranking.

The second finding is that the Density-based scoring performs the best passage retrieval algorithms for factoid questions. This is the case of IBM, ISI, and SiteQ algorithms.

The Density-based models have gained attention due to their usability in the processing of factoid questions. Gomez et al. (2007a) have proposed the language-independent system called JIRS that implements the Distance Density N-gram Model. This model considers sequence of 'n' adjacent words (n-gram) extracted from a sentence or a question. All possible n-grams of the question are searched in the collection. It also assigns them a score according to the n-grams and weight that appear in the retrieved passages. This system does not use any language knowledge, lexicon or syntax (Gomez et al., 2005), but just shallow adaptations are needed for its use in the context of a given language (for instance, the list of stopwords has to be adapted to consider the one corresponding to the targeted language).

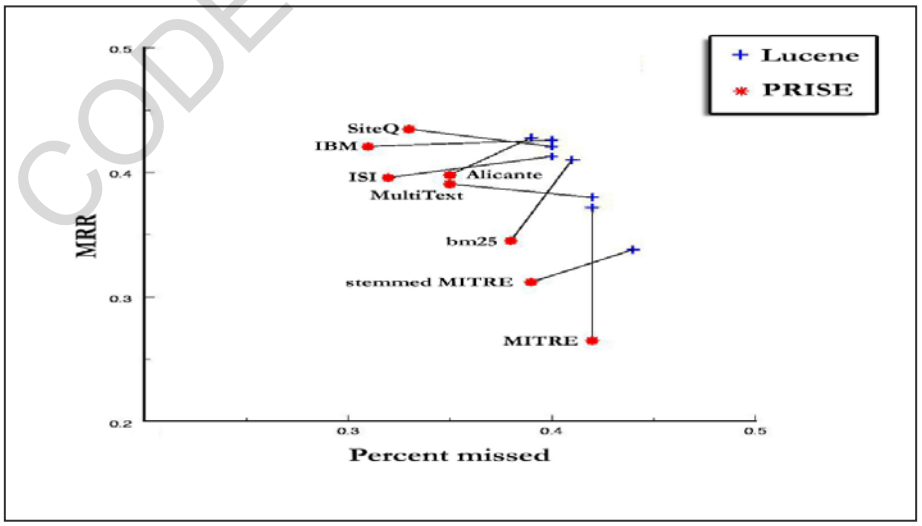


Figure 7. Experiments on eight PR algorithms

Source: Tellex et al. (2003)

It was reported in (Gomez et al., 2007a) that JIRS improves the coverage measure (Roberts and Gaizauskas 2004). In (Gomez et al., 2007b), it was shown that the JIRS system increases the MRR. Both works have used a Spanish collection and 200 CLEF questions. Balahur et al. (2010) also used the system in increasing the precision of a process that combines Opinion Analysis with other challenges, such as the ones related to English QA.

The above mentioned works, especially the one of Tellex et al. (2003) and those using JIRS highlight the usefulness of density-based algorithms. Nevertheless, they also mention that in addition to question terms and their density in passages, there is a need of recognizing syntactic relations. This issue was pointed out by Katz and Lin (2003) proving the importance of relationship analysis rather than only considering term density. Cui et al. (2005) also compared a dependency relation matching method with a density based one. The former is performed by examining the grammatical dependency relations between query terms and key terms within passages. This comparison results in a significant increase in performance of up to 77.83% in MRR using the relation-based method as compared to the density-based passage retrieval module.

To sum up, passage ranking is a core component of the PR module. As we have seen above, the related research proposes different algorithms and techniques that were evaluated and compared in real text situations. Although, most of them were designed for IR applications, their use in PR modules of QA systems was reported to be useful.

B) Query expansion in PR modules

Many works investigated the enrichment of the query itself for the improvement of the performance of the PR module. This is usually done via QE by adding additional related terms to the original terms of the question.

In the IR field, many QE techniques have been investigated by researchers. The basic ones focus on fixing spelling errors by searching for the corrected form of the words (João Pinto 2008). Other QE techniques rely on morphological relations and reformulate the user query by adding the different variations that are generated from keywords stems (Larkey et al., 2008). Although these QE techniques obtain higher recall (Monz 2003; Bilotti et al., 2004), it is difficult to assert that they improve the precision. This is why researchers have investigated other QE research directions especially the use of semantic relations. Generally, a semantic QE technique is performed by considering the synonyms of the query keywords. A thesaurus can be used as a base for such a process (Nanba 2007). However, the use of a thesaurus, which is generally built based on statistical techniques, presents many disadvantages. Indeed, building a thesaurus is time consuming since a great amount of data has to be processed. Moreover, the precision of thesaurus based QE in terms of semantic distance has to be proved.

In the QA field, QE has been widely used and tested as a promising technique for the enhancement of PR modules. Different research directions have been followed with respect to

the use or not of stemming. Bilotti (2004), quantitatively compared two different approaches to handling term variation: applying a stemming algorithm at indexing time, and performing morphological query expansion at retrieval time. The results of the conducted experiments show that morphological expansion yields higher recall. In early works on QE for QA (Grefenstette 1992; Grefenstette 1994), it is mentioned that gain in terms of performance is more important when applying stemming and a second loop of expansions of words rather than restricting QE on the nearest neighbours of a term.

Another research direction is using statistical methods for QE. Local Context Analysis (LCA) was one of these methods whose idea is incorporating additional contextual terms for enhancing passage retrieval., Term co-occurrence statistics helps in defining these added terms. In this direction, Sun et al. (2006) presented an interesting work where two new QE methods have been proposed: (i) dependency relation-based term expansion (DRQET) which was used in a density-based passage retrieval system (Croft and Harper 1979; Cui et al., 2004), and (ii) dependency relation-based path expansion (DRQER) which is integrated in a relation based passage retrieval.

Both DRQET and DRQER techniques relies on a Web corpus (built from a set of extracted snippets) as a training resource to score relations between terms. Thereafter, the table of scores is used to weight expanded terms by DRQET and expanded paths by DRQER. The conducted experiments considered factoid questions from the TREC-12 QA task. They showed that in terms of MRR scores, the first method (i.e., DRQET) combined with density-based score outperforms the LCA by 9.81%, while the second method (i.e. DRQER) obtained 17.49% improvement over a corresponding relation-based passage retrieval system without query expansion.

Moldovan et al. (2003) proved that keyword expansions based on lexico-semantic alternations from WordNet (Fellbaum, 1998) increments the scores by 15% when they are used in the PR component of the QA system. Pasça and Harabagiu (2001) conducted experiments on a set of TREC-8 and TREC-9 questions with lexical and semantic QE. These experiments show that the precision score is higher (73.7%) when combining lexical and semantic alternations (against 67.6% if only lexical alternations are applied and 55.3% if any alternations are considered). On the other hand, it also was reported that QE-based on WordNet synsets and gloss (Yang and Chua 2003) brings more noise than information to the query and that using EuroWordNet synonyms as related terms in the query degrades results (Voorhees 1993).

WordNet-like resources were not the only resources used in QE, we can also cite ConceptNet which is developed by MIT Media Laboratory (Liu and Singh 2004). Comparison between the WordNet and ConceptNet tested on TREC-6, TREC-7 and TREC-8 datasets shows that the WordNet enhances the precision while the ConceptNet improves the recall (Hsu et al., 2006). The use of WordNet or ConceptNet in QE presents some drawbacks:

- The coverage of WordNet or ConceptNet regarding the processed language is not always sufficient to guarantee the application of QE on question terms;
- WordNet or ConceptNet are linguistic resources developed by lexicographers. This means that the relatedness between words in these resources is what we should have in theory. However, in real-world text, these words could be used differently with respect to these two resources (Peetz and Lopatka 2008).

Recently there has been an increasing interest in using online resources such as Wikipedia in the generation of related terms within a QE process. In (Peetz and Lopatka 2008), Wikipedia disambiguation pages was used in a promising QE process. Experiments in that work show that this resource presents a more user-friendly distribution of terminology and topics than Wordnet.

A QE process based on WordNet can enrich the user question with a high number of related terms. This can result in two drawbacks: challenging the PR module with many sets of passages retrieved for the queries generated from QE and bringing some noise to the AEV module since this module works on top of what the PR module retruns. In order to overcome this problem, QE is applied on just important terms of the question. In the Raposa system (Sarmento, 2008), factoid questions in Portuguese are answered thanks to three types of queries: keyword, pseudo-stemming and verb expansion queries. Verbs are usually chosen as head of questions.

In both IR and QA fields, there has been a particular interest in expanding the query/question on the basis of user feed-back. This family of QE techniques is called Iterative Query Reformulation (IQR). It consists in modifying the query by adding, replacing or removing terms using previous user experience. For instance, a QA system may integrate an IQR module which allows users to inform the relevance of retrieved passages for their questions. Rocchio's algorithm (Rocchio, 1971) is a standard algorithm for relevance feedback. This algorithm uses a Vector Space Model. In the literature, Liebeskind et al. (2013) cite Iterative Query Expansion (IQE) because it consists in only adding terms to improve the query starting from the relevance of the previous ones.

This section focused on the Query Expansion task that has widely been used in PR module. Hence, we showed the different approaches applied to queries in IR and to questions in QA. Statistical-based and resource-based QE approaches remain the most commonly adopted within the community of IR and QA. Usefulness of lexical resources such as WordNet and Wikipedia was reported in many experiments.

C) Syntactic and semantic matching in PR modules

The interest of researchers in more advanced PR modules has increased with the need of processing new types of questions such as those asking for the reason of a fact. We have

previously seen that classical PR modules focus on comparing the question and passages relying on the number and density of occurence of original question keywords in the targeted documents/passages. Besides these classical PR approach, the QA field has known new trends especially with the introduction of syntactic tree matching (Cui et al., 2005) as well as semantic processing and reasoning.

The idea behind syntactical processing in PR is that the answer is more likely surrounded by the same tree or subtree representing the user question (Wu et al., 2005). Note that syntactic information can be extracted from tagging and parsing tools (Manning and Schutze 1999). The former help in labeling each word in a sentence with the appropriate PoS such as "verb", "noun" and "adjective". The latters provide deeper syntactic analysis of the sentence by representing the relationships between sentence constituents through a syntactic tree. This tree is of great interest since it also helps in recognizing phrases (noun phrases, verb phrases, prepositional phrases, etc.) and hence in identifiying blocks of words that need to be considered as a whole, and not just as a bag of words. In (Li and Croft 2001), it is shown through experiments on the TREC-9 track questions that a combination of syntactic information with heuristics for ranking potential passages can perform about 10% better than the ranking based on heuristics.

Syntactic matching was mentioned as a solution for many challenges in QA such as paraphrasing. Let us recall that this concerns the fact that many passages may contain answers to the processed question but cannot be well-ranked by the PR module due to their formulation is different from the original question. For example, when a user asks for "Who scored for Manchester United in the 1999 champions league final?", a classical approach would highlight passages such as "Ole Gunnar Solkskaer and Teddy Sherringham scored for Manchester Untied in the Champions League Final of 1999 against Bayern Munich" but not "One of the two goals of Manchester Untied in the Champions League Final of 1999 has been scored by Ole Gunnar Solkskaer". This issue has been widely discussed in the QA community (Mollá and Van Zaanen 2005; James et al., 2003; France et al., 2003).

Syntactic parsing has also been used as an intermediate step towards semantic processing in QA. The two PR modules of the AQUA (Vargas-Vera and Motta 2004) and the QUANDA systems (Breck 1999) make use of knowledge representation on the basis of syntactic parsing as an intermediate step. The work of Salloum (2009) is an example of these PR modules that represent text (in the question and the passages as well) through a formalism such as conceptual graphs[18] (Sowa 1984) and then compare the two representations (i.e., the

[18] Conceptual Graphs (CGs) are a kind of semantic networks introduced by John Sowa in 1984 on the basis of Peirce's existential graphs. Each CG contains Concepts which can be related by conceptual relations. A Conceptual Graph is a directed graph of nodes that correspond to concepts, connected by labeled and oriented arcs that represent conceptual relations. Conceptual nodes represent entities, attributes or events. They are

representation of the question and the one of the passage) in order to rank the candidate passages. Hensman and Dunnion (2005) proposed an approach for the use of CGs in an automatic representation of text based on syntactic parsing, WordNet and VerbNet (Kipper-Schuler 2006). The general idea of these works can be summarized in three steps: (i) syntactic parsing of the given text (question or passage); (ii) generating the CG on the basis of VerbNet frames and semantic roles; and (iii) performing operations and similarity scores between the CG of the question and the CG of the candidate passage.

In order to apply this technique for a given language, there are many requirements that have to be guaranteed:

- Syntactic tools for the considered language;
- Ontologies containing the concepts that will occur in the built conceptual graphs. Thus, these ontologies have to cover a high number of concepts that can occur in questions and passages. Generally, since these resources are budget and time consuming, researchers prefer to focus on a specific domain such as biology. In this direction, Graesser et al. (1991) used a conceptual graph representation for QA systems for stories;
- Artificial Intelligence plateform for CG operations and semantic reasoning;

The comparison between the CG of the question (CG-Q) and the CG of a given passage (CG-P) was made through similarity scores. For calculating the semantic score Montes-y-Gómez (2001) proposed the following formula:

$$SemanticScore(P) = \sum_{c_i \in C} (weight(c_i) * \beta(c_i, \pi(c_i))) / \sum_{c_i \in C} (weight(c_i)$$

Where C is the collection of the concepts of the passage P, weight(ci) is a weight assigned to the word related to the concept ci of the graph CG-P and β(ci,π(ci)) is the distance between ci and its projection in the generalization graph between CG-Q and CG-P, it is defined as follows:

$$\beta(c_i, \pi(c_i)) = 1 \quad [\text{if } type(c_i) < type(\pi(c_i))]$$

$$\beta(c_i, \pi(c_i)) = 1 - \min(\delta(type(c_i), type(\pi(c_i))), 5)/5$$

$$[\text{if } type(c_i) \geq type(\pi(c_i))]$$

Where $\delta(type(c_i), type(\pi(c_i)))$ is the number of nodes between $type(c_i)$ and $type(\pi(c_i))$ in the ontology.

---

denoted with square brackets. Relational nodes determine the type of the relation between two conceptual nodes (Sowa, 1984).

Stephane (2003) adopted similar formula with different weights for PoS (1 for verbs, 0.8 for nouns and 0.16 for adjectives and adverbs).

As a conclusion to this part, the approaches based on syntactic and semantic matching in PR have gained interest due to their ability to support advanced QA systems. For some complicated types of questions (such as *why* and *how* questions), classical methods are reported to be unsuitable.

The main sub tasks of PR presented in this section, namely passage ranking, QE and syntactic/semantic matching try to bring solutions to the various challenges faced in this module. According to the existing works, statistical approaches to passage ranking is limited in terms of usefulness when processing complex questions. Considering shallow semantic features through QE or syntactic and semantic matching can be helpful to process such complex questions.

2.4.2.3 AEV Module

Answer extraction is also important in a QA system. Actually, even when a PR module reaches a high level of accuracy, the provided candidate passages cannot be useful for the end user until the answer extraction and validation module exploits them efficiently. Many works with various levels of processing and different approaches concerned this module. As we have seen in the previous section, layers such as syntax are of great interest for PR. Syntactic structure matching has also been applied to answer extraction (Shen and Klakow, 2006).

The QA system described in (Shen et al., 2006) integrates an AEV module which extracts exact answers from the processed candidate passages. Two main strategies are used in this module: (i) surface text pattern matching and (ii) correlation of dependency relation path. In order to implement these two strategies, authors considered different tools such as LingPipe[19] for named entity recognition, Abney's chunker (Abney 1989) for NP chunking and MINIPAR (Lin 1994) for dependency parsing.

In CLEF 2006, answer validation exercise which is added in the third QA module (i.e., Answer Extraction) has been considered as a pilot task. The reason for adding this step is due to the problem of error propagation in the traditional QA pipeline. No matter in 2005 more than 80% of the questions were answered by at least one participant, the upper bound of accuracy in systems performance was 60%. The basic idea is that once a pair [answer + passage] is returned by a QA system, an hypothesis is built by turning the pair [question + answer] into the affirmative form. If the related text (a passage or a document) semantically entails this hypothesis, then the answer is expected to be correct (Rodrigo et al., 2010).

[19] http://www.alias-i.com/lingpipe/

This section presented some of the AEV research works. It figured out the trend of using NER and dependency parsing for this QA module. Also, it highlighted the objective of the validation stage once a list of candidate answers is extracted.

The study of the approaches used in existing works related to QA modules allowed understanding some challenges faced during the different parts of a QA pipeline. This study also shows that for building an Arabic QA system that has the ability to process simple and complex questions, we are requested to move towards more advanced techniques based not only on statistical or surface processing but also on syntactic and semantic features.

2.5 Chapter summary

In this chapter, we have seen that QA systems analyze a question expressed in natural language, retrieve passages from a collection of documents (for instance the Web), rank these passages according to their relevance to the expected answer and extract/validate this answer for the end user. These systems gained researchers' attention for their ability to save time and effort for the user when long lists of Web snippets and document passages are provided by the classical SEs and IR systems.

Generally, a QA system is built following a three-module architecture: the QAC module, the PR module and the AVE module. These systems are evaluated by means of IR measures such as recall and precision as well as QA-specific measures such as accuracy, MRR and C@1.

Even though this architecture and evaluation process are language-independent, the core components of an Arabic QA module have to be developed to tackle some challenges that are specific to this language. These challenges (e.g. high ambiguity level, complex morphology and syntax, lacks for resources and tools) have been presented in this chapter with examples from the QA perspective.

The particularities of the Arabic language such as its complex morphology and syntax, as well as the high ambiguity of unvoweled text, explain the delay registered in the efforts and results of Arabic QA research. Indeed, we observed that:

- The early works on Arabic QA do not report results of complete experiments nor do they use standard measures such as accuracy and MRR commonly used for other languages in evaluation campaigns. This can also be due to the fact that Arabic was not considered in the existing campaigns (Arabic was introduced only in TREC 2002, CLEF 2012 and CLEF 2013). Recently, the standard metrics of QA evaluation are considered in the case of Arabic works but still only handle small sized test collections;

- The scope of these Arabic QA systems is restricted in terms of collection type (structured texts, children book, etc.), question type (factoid questions in most cases, definition questions in others), QA module (especially Passage Retrieval), etc.;
- The reported works do not integrate deeper QA approaches in order to understand the meaning of the question and compare this meaning with the knowledge existing in the targeted collections;
- Open-domain Arabic QA, especially systems querying the Web as a targeted collection, is scarcely cited as future work, though it is an important line of research in QA considering the substantial amount of information on the Web and Social Media.

In this chapter, we reviewed the most prominent systems for other languages with the aim to define the new lines of research to explore in Arabic QA research. In this review we first described system-oriented works and results from some evaluation campaigns, namely CLEF and TREC, and moved to a more detailed review of component-oriented attempts.

This review has lead us to the conclusion that the QA field has witnessed a notable evolution in the last decade due to two main reasons:

- Organization of yearly QA campaigns (TREC, CLEF, etc.) that made available new evaluation metrics and resources such as test collections, baselines, surveys, etc. As a result, a typical architecture for QA systems took shape and conclusions about the use of certain approaches were drawn in particular for statistical ones,
- Development of mature resources and tools for the basic layers of NLP (morphological analyzers, syntactical parsers, etc.) as well as for other tasks such as NER and QE (for instance NE gazetteers, Ontologies, WordNets and VerbNets). Building a QA system depends on the availability of such resources and tools.

The insights of the evaluation campaigns show that researchers in this field followed one of (or a combination of) two main kinds of approaches: (i) deep level approaches that are language-dependent and (ii) surface level approaches that can be applied for different languages. Table 3 makes a comparison of these approaches in terms of complexity, performance, etc.

Table 3. Differences between deep and surface approaches

Surface approaches	Deep approaches
Use language-independent tools	Require language dependent resources and tools
Quick implementation	Implementation is more complex
Limited performance	Higher performance if the required resources and tools are available
There are reported open-domain works	Most works are domain-specific
Useful mainly for factoid questions	Suitable for different question types

According to the high requirements of deep approaches in terms of language-dependent resources and tools, they are not used in simple QA systems especially in the case of factoid questions. These types of questions are usually formulated using the same structure and keywords that can be found in the document collections. The system based on the surface approach makes a better matching between the question and passages by comparing their surface elements (i.e., keywords and structure). For other types of questions, the reported works adopt deep approaches that also help in reaching high performance when the quality of linguistic resources and tools is guaranteed. Otherwise, some restrictions have to be applied for instance to limit the considered domain and hence to prepare adequate requirements in terms of resources and tools.

In terms of performance, QA systems for languages such as English and Spanish have matured, especially when processing factoid questions. Even though there are many reported experiments on different types of QA systems, it is difficult to show what contributes to the performance of a system and what does not, due to the complexity and number of components of such systems.

Finally, the development of Arabic QA systems is highly concerned by: (i) the availability of resources as well as tools related to other tasks such as QE, NER, syntactic parsing, etc., and (ii) the feasibility of exploiting some techniques reported as effective for other languages.

Chapter 3

The three-levels approach for Arabic QA

3.1 Introduction

3.1.1 Problem and objectives

As summarized in Chapter 2, the state-of-the-art shows that Arabic QA research is limited to a basic processing of questions. Experiments were focused on surface-based approaches that did not integrate semantic understanding. Building an Arabic QA system with the ability of answering and processing different types of questions and document collections is still unachieved. This requires new approaches, leveraging existing NLP tools and resources and developing new needed ones.

Between 2001 and 2010, efforts in the field of Arabic NLP have resulted in the development of more available resources and tools, especially for handling basic tasks such as POS-tagging, morphological analysis, phrase chunking, syntactic parsing, etc. The availability of these materials is, at the time being of this research, an opportunity for the development of more advanced Arabic QA systems.

On the other hand, we have seen in the previous chapter that the same decade has witnessed the emergence of QA devoted to other languages. Consequently, different approaches were proposed and tested in the framework of evaluation campaigns. From these approaches, the simplest are surface-based using IR or SE applications as a starting point and then extracting text similar to the question in terms of keywords and/or structure. Beyond this simple approach, the evaluation campaigns, held regularly, show a trend towards the introduction of new approaches based on semantic comparison rather than surface comparison.

The need for these new approaches is mainly motivated by the objective to answer new types of questions beyond the factoid questions and to overcome the challenges of extracting answers from challenging Arabic texts (such as the content available on the Web). Reaching this objective would increase the popularity of QA systems among users more interested in the Web and Social Media's content.

Henceforth, we will present our approach towards building an efficient Arabic QA system with the ability to deal with the specific challenges related to: (i) the processing of the Arabic language, (ii) the difficulty of extracting answers from large document collections such as the Web, (iii) the users' expectations to automatically answer different types of questions and (iv) the significant evaluation of the system.

3.1.2 Methodology

From the study of the advances of the QA field, we have learned the importance of passage retrieval in the pipeline of a QA system. Once enhanced, this module would provide passages with a high quality regarding the user question. These passages can then be exploited in the next module for answer extraction and validation.

The methodology of our research is experiment-oriented and thus composed of the following main steps:

- Step 1: designing a three-level approach to support keyword-based, structure-based and semantic-based processing of Arabic questions by means of integrating exiting resources and tools developed for Arabic NLP;
- Step 2: evaluating the keyword-based and structure-based levels of this approach on well-known test collections. The objective of this step is observing the performance obtained for each question type, especially factoid questions that can effectively be processed at keyword and structure levels;
- Step 3: experimenting the impact of resource enrichment on Arabic QA with respect to the previously evaluated levels;
- Step 4: evaluating the effectiveness of semantic-based level especially with the aim to process more complex questions, beyond the factoid questions, such as definition and why-questions.

In the next two sections, we present details about Step 1 and Step 2. Chapter 4 will describe the work conducted in Step 3, while Chapter 5 is devoted to Step 4.

3.2 Three-level approach for Arabic PR

3.2.1 Background

Almost all existing Arabic QA systems and attempts are surface-based approaches. That is, the retrieval and ranking of candidate passages rely on their surface similarity with the given

question. Thus, keywords of the question are searched within these passages. Generally, a passage containing the highest number of keywords is considered to be relevant.

More refined versions of this approach are based not only on the existence of question keywords in passages but also on their density. Indeed, this is especially useful when the target is the Web as a document collection. Snippets returned by SEs are usually small in size (two to three lines) and most likely do not contain all question keywords (even if the document corresponding to this snippet contains all question keywords). Such approaches allow the improvement of the ranking based on the existence of a maximum number of keywords together in the passage. This was reported to be useful for Arabic PR (Benajiba et al., 2007a).

By definition, in factoid questions the expected answer is a Named Entity (name of place, person, etc.). Usually, such answer appears in the passage together with the question keywords. In this case, surface approaches are effective. However, for other types of questions or even for long and complex factoid questions, such approaches fail.

A deeper approach based on semantic matching was reported as being effective to improve the precision especially for those types of questions where surface approaches fail (Peng et al., 2005). Therefore, we propose a three-level approach combining both families of approaches with the aim to leverage the advantages of each of them.

3.2.2 Approach at a glance

The three-level approach proposed for Arabic PR is composed of three levels to achieve two main objectives as illustrated in Figure 8.

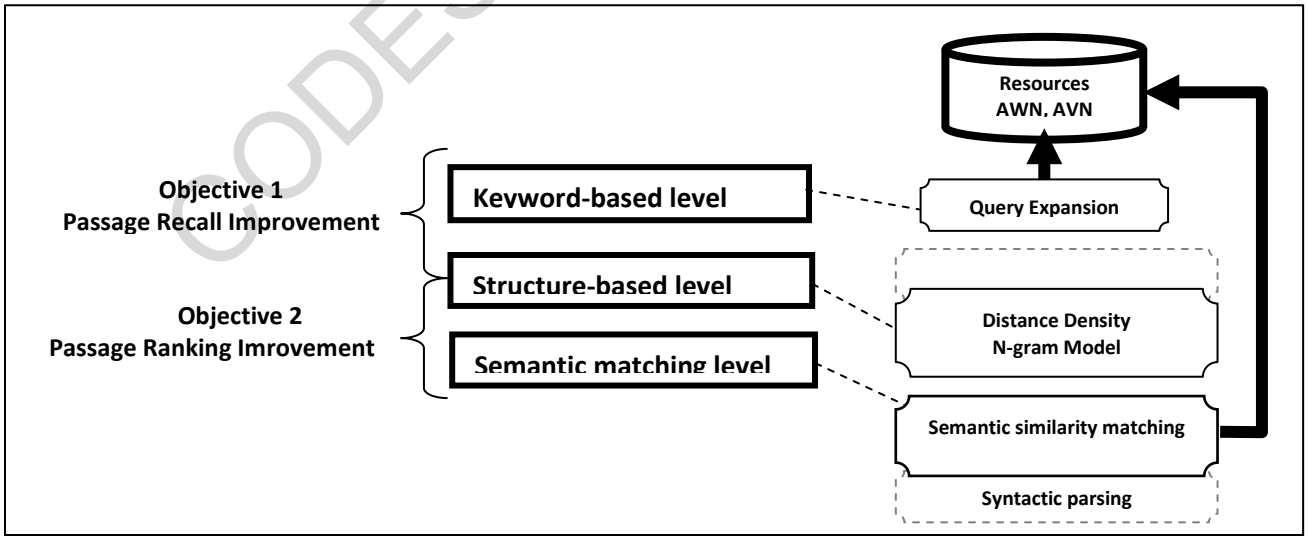


Figure 8. Tasks and levels of the proposed approach

The first objective is concerned with the improvement of the passage recall. The main aim is retrieving a high number of passages containing question keywords and their related terms and, therefore, increasing the ability of the system to retrieve relevant documents. This is made through the "Keyword-based level". In this level, the question keywords are extended to their semantically related terms through the Arabic WordNet resource. Indeed, since there are many ways to formulate a question in natural language, a QE process can be used in order to overcome the situations where the PR process eliminates relevant passages containing other forms of the question keywords or words related to them. For instance, if the question contains the keyword طريق (Tryq : a way) the query used by the PR process can be expanded to include its other morphological forms like طرق (Trq : broken plural of Tryq) or طرقات (TrqAt : plural of Tryq). A more advanced QE process relies also on semantic relations. For example, we can include keywords like ممر (mmr : path) or مسار (msAr : trajectory) since they are similar in meaning with respect to the original keyword.

As mentioned in Chapter 2, some QE techniques using light-stemming can improve recall, while others improve precision. Generally, QE increases the recall at the expense of precision. Thus, the second objective of our approach is improving passage ranking after the improvement of passage recall through QE.

The second objective is related to two levels: (i) The Structure-based level that reduces the possible noise generated by the extraction of a large number of passages in Step 1. It is based on the Distance Density N-gram model; and (ii) The Semantic matching level which increases precision through a deeper matching approach. In this level, the semantic representation of the question and passages to be ranked allows to measure their semantic similarity.

Towards reaching both objectives, each level makes use of existing Arabic NLP resources and tools. The key resource used in our approach is AWN that is exploited in the keyword and semantic-based levels. Consequently, we are also interested, in Chapter 4, not only in evaluating its coverage and usability but also in enriching its content according to the possible shortcomings registered in the experiments.

The following sub sections provide details about the keyword-based and structure-based level that represent the surface side of our approach, while Chapter 5 is devoted to the deeper semantic-based level.

3.2.3 Passage recall improvement

The keyword-based level starts from questions keywords and try to search for their semantically related terms in addition to their morphological variations. The QE process uses relations between words existing in the AWN lexical resource (Elkateb et al., 2006).

3.2.3.1 Keyword-based level

To understand why these relations are exploited and why they help in a QE process, let us, first, introduce the AWN lexical resource. A WordNet (WN) is a lexical database of a given language that focuses on common-class words: nouns, verbs, adjectives, adverbs and adverbials (the latter being mostly nouns, adjectives and participles used in an adverbial role, e.g. 'willingly'). Independently of the concerned language, WNs allow the user to explore the relationship of words to each other. They are also useful in a number of language processing tasks requiring the understanding of the meaning of language. Such tasks include information retrieval (Rila et al., 1998), word sense disambiguation (Navigli 2009), automatic text classification (Elberrichi 2008), automatic text summarization (Dang and Luo 2008), question answering (Clark et al., 2008) and machine translation (Anwarus Salam et al., 2009), among others.

In terms of WN structure, words are grouped into synsets. The members (i.e., words) of a synset are synonyms and can be used in a sentence without changing its meaning. Generally, they express a concept which is distinct from all the other WN concepts.

Synsets are interlinked by means of conceptual-semantic and lexical relations such as hyponymy, meronymy and antonymy (Table 4 lists examples of these relations). The first WordNet was built for the English language (named Princeton WordNet)[1].

Regarding the Arabic language, the AWN was released in 2007 (Black et al., 2006; Elkateb et al., 2006; Rodriguez et al., 2008) and followed the development process of English WordNet and Euro WordNet[2] (Vossen 1998). It utilized the Suggested Upper Merged Ontology (SUMO)[3] as an Interlingua to link AWN to previously developed WNs.

Table 4. Examples of AWN semantic relations

Relation	Synset #1	Synset #2	Relation meaning
Hyponymy	كَبَحَ (verb to restrain)	مَنَعَ (verb to prevent)	كَبَحَ *نوع من* مَنَعَ to restrain *is kind of* to prevent
Meronymy	شَقّة (apartment, flat)	بناء (building)	شَقّة *جزء من* بناء apartment, flat *is part of* building
Antonymy	صَلاح، خَيْر (goodness)	السُّوء (badness)	السُّوء *بخلاف* صَلاح، خَيْر badness *opposite of* goodness

In each relation, we have two members: Synset #1 and Synset #2. For instance, the synset "كَبَحَ" (to restrain) plays the role of hyponym of synset "مَنَعَ" (to prevent) in the first relation type. Table

[1] http://wordnet.princeton.edu/

[2] http://www.illc.uva.nl/EuroWordNet/

[3] http://www.ontologyportal.org/

4 also shows the meaning of the relation between the two synsets in each case (column 4, the underlined italic expression represents the meaning of each relation).

Our QE approach is based on AWN due to the following advantages it offers:

- The AWN lexical database is a free resource for MSA;
- It is based on a nearly standard design (i.e., Princeton WordNet);
- AWN has a structure that is similar to WordNets existing for approximately 40 languages.[4] Therefore, cross-language processes could be considered later as an enhancement of the present work;
- It is also connected to SUMO ontology. Let us recall briefly that SUMO is an upper level ontology which provides definitions for general-purpose terms and acts as a foundation for more specific domain ontologies. It contains about 2000 concepts.

AWN offers the possibility to export its content and structure onto many formats so that researchers can use it in their context. Figure 9 illustrates the structure of AWN database and its mapping onto the English WN.

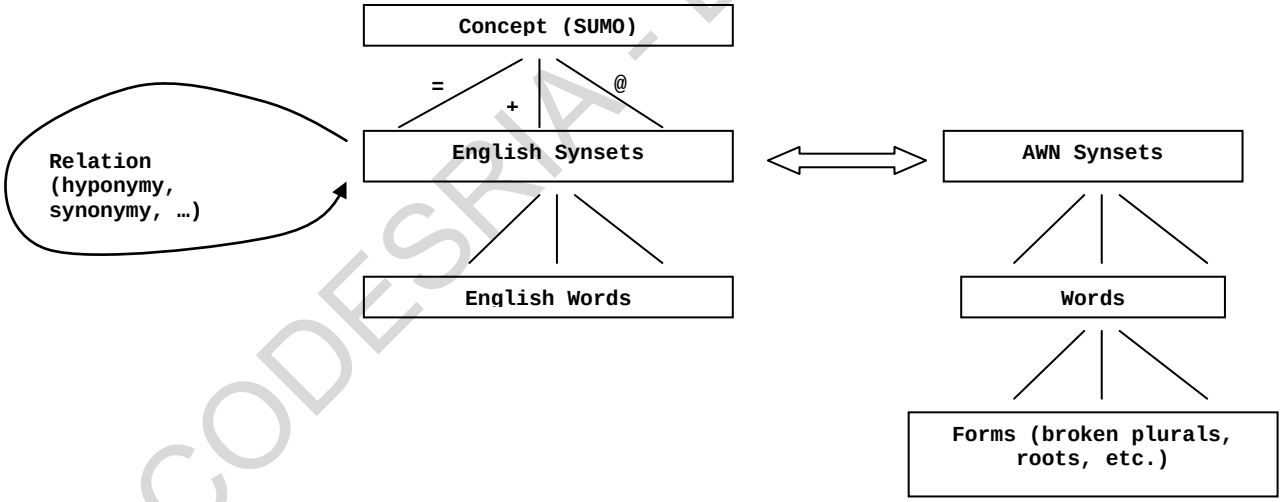


Figure 9. The structure of the AWN lexical database

The AWN data are divided into four entities:

[4] Including English, Italian, Spanish, French, Basque, Bulgarian, Estonian, Hebrew, Icelandic, Latvian, Persian, Romanian, Sanskrit, Tamil, Thai, Turkish, etc.

- Items which are conceptual entities, including synsets (a set of words with the same part of speech that can be inter-changed in a certain context), ontology classes and instances. Besides a unique identifier, an item has descriptive information such as a gloss. Items lexicalized in different languages are distinct;
- Word entity is a word sense, where the citation form of the word is associated with an item via its identifier;
- A form is an additional form of a word. It is considered as a dictionary information (not merely an inflectional variant). The forms of Arabic words that go in this entity are the root and/or the broken plural form, where applicable;
- A link relates two items, and has a type such as "equivalence," "subsuming," etc. Links connect sense items to other sense items, e.g. a PWN synset to an AWN synset, a synset to a SUMO concept, etc. Note that the "@", "+" and "=" symbols in the figure above refer to the INSTANCE_OF, MORE_GENERAL and EQUIVALENT mapping types respectively.

The QE process uses, in addition to the morphological QE, four semantic relations connecting AWN synsets: synonymy, hyponymy, hypernymy and AWN synset-SUMO concept mapping. Each question keyword is substituted by its semantically related terms extracted from the AWN. Figure 10 is an illustration of the QE process.

A given question is composed of stopwords and keywords. Our process only concerns keywords and is performed, for each of them, as follows:

i. Derivational forms of W_i using "AL KHALIL" system[5]. This system analyzes words and provides their root. The root is searched within the AWN lexical database to get the forms sharing the same root. Here, potentially a lot of irrelevant forms could be generated. This problem will be reduced by the structure-based level described later in this chapter.

ii. If the given keyword is matched with at least one corresponding synset in AWN, then the QE process is performed over the corresponding synsets in a recursive way in order to extract:

   - Terms that share the same AWN synsets (Synonyms (S));
   - Terms that share the AWN synset (Synset S_{hyper}) that is more general than S (Hypernym (S));

[5] http://sourceforge.net/projects/alkhalil/

- Terms that share the AWN synset (Synset S_{hypo}) that is more specific than S (Hyponym(S));
- Terms related to AWN synsets (Synsets S_{SUMO}) provided by the SUMO concepts appearing in the formal definition of the SUMO concept which is equivalent to each S.

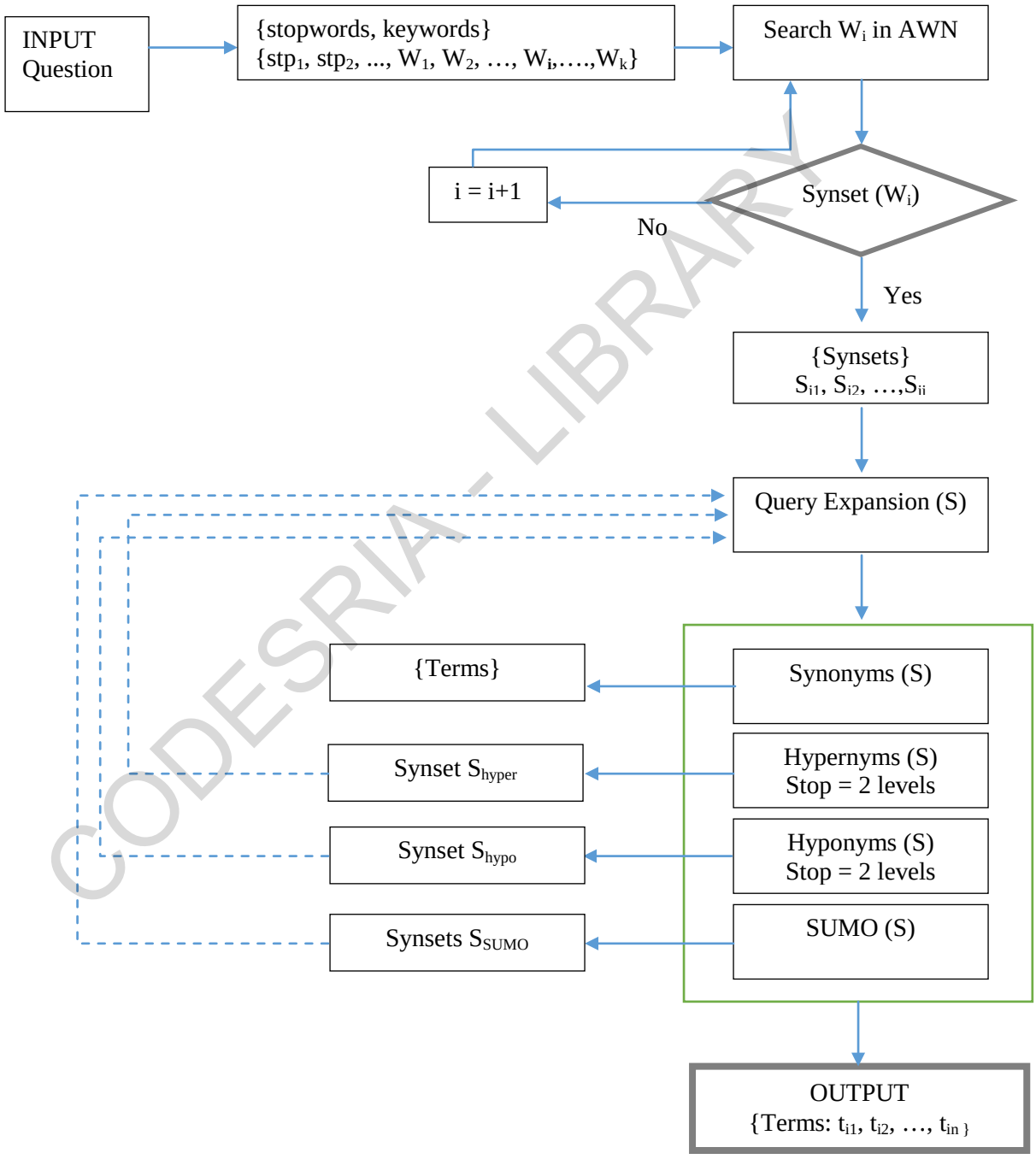


Figure 10. Design of the AWN-based QE process

The threshold of the recursive process is set to levels for the Hyponymy and Hypernymy relations. Due to the quality of the AWN synonymy relation and the few SUMO-AWN mappings relations, we did not set a threshold for these two relations.

At the end of this process, we have a list of terms that are semantically related to the question keywords. The synonymy, hyponymy and hypernymy relations extract related terms from the direct neighborhood of the keyword in AWN, while the SUMO relation explores other contexts of the keyword in AWN.

The generated terms are used to form new queries by substituting a keyword in the question by its related terms. Note that in the case of Named Entities keywords, we substitute the keyword just by its synonyms. The hypernyms are just added before the keyword in the question. This is due to the fact that a hypernym of a NE is usually its category (for instance person, country, etc.).

3.2.3.2 Example of passage recall improvement

To show the effectiveness of the described QE process, we provide an example starting from the question: "ما هي المناصب التي تقلدها سيلفيو برلسكوني؟" (i.e., What positions did Silvio Berlusconi hold?). The proposed QE process is applied only on the question keywords excepting the stopwords: ما (mA : what), هو (hw : he) and الذي (Al*y : that).

For example, let us apply the QE process on the keyword "المناصب" (AlmnASb : positions). This keyword is the broken plural of the noun "المنصب" (manoSib : position). In AWN there is a synset "مَنْصِب - وَظِيفَة" (manoSib : position – wZyfp : job) which contains the keyword "المناصب" among the possible forms of the word "مَنْصِب" (manoSib : position) which is a member to this synset. Figure 11 illustrates the entry of this synset in the AWN lexical database. It shows the information displayed in the AWN browser provided in the released version. In the right side of Figure 11, we have the gloss of the selected item, i.e., the synset "مَنْصِب - وَظِيفَة". Below this gloss, we have a snapshot of the AWN hierarchy based on the hyponymy relation with a special focus on the selected synset (the given synset is highlighted). The other side of the figure illustrates other kinds of information about this synset, including the English counterpart of the AWN hierarchy, the corresponding SUMO hierarchy, the synonyms of this synset, etc.

Our QE process exploits this information to generate new related terms. As we can see, one of such a terms (which is not included in the original question keywords) is the synonym "wZyfp : job". By considering the AWN hierarchy, the considered entry has two direct supertypes: مهنة (mhnp : job) and نشاط (n$AT : activity). It has only one subtype "مَنْصِبسِكْرتِير" (manoSib_sikortiyr : secretaryship).

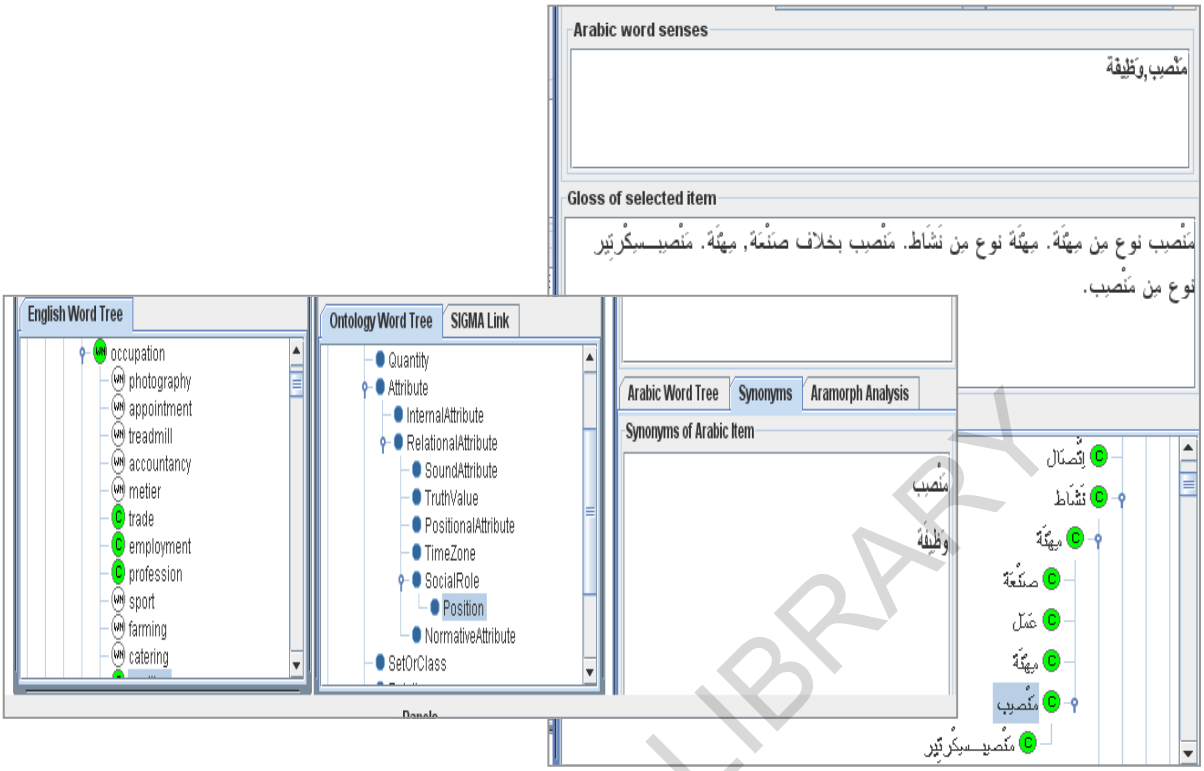


Figure 11. The neighborhood of the synset "مَنْصِب - وَظِيفَة" in the AWN hierarchy

Regarding the SUMO relation, the given synset corresponds to the concept "POSITION" (see the highlighted concept in Figure 11). The definition of this concept in the SUMO ontology is as follows:

"A formal position of responsibility within an ***&%Organization****. Examples of Positions include president, laboratory director, senior researcher, sales representative, etc."*

Given that the SUMO concepts are preceded by the symbols "&%**"** and **"?"**, we can identify the SUMO concept "ORGANIZATION" (written in bold) as being related to the "POSITION" concept. By following the links set by Niles and Pease (2003) between SUMO concepts and AWN synsets, the concept "ORGANIZATION" corresponds to the synset "جَمْعِيَّة" (jamoEiy~ap : association). In this stage, we exploit the information provided by AWN for this synset as we did in the case of the starting synset "مَنْصِب - وَظِيفَة".

The neighborhood (supertypes and subtypes) of this new synset allows us to reach new terms such as: "مُنَظَّمَة" (munaZ~amap : organization), "جَمَاعَة" (jamaAEap : community), "حكومة" (Hkwmp : government) and "نِظَام سِيَاسِي" (niZaAm siyaAsiy : political system). The SUMO concept "ORGANIZATION" is also linked to the synset "رَئِيس" (ra}iys : Chairman). Recursively,

new terms could be reached in the neighborhood of this synset such as مَلِك (malik : king), رَئِيس الوِزَرَاء (ra}iys AlwizaraA' : prime minister) and رَئِيس الدَّوْلَة (ra}iys Ald~awolap : head of nation). Figure 12 illustrates the result of the recursive QE process that we perform starting from the question keyword "المنصب". Note that boxes with labels 1, 2, 3 and 4 refer respectively to the QE by synonyms, definition, subtypes and supertypes. Note that the non expanded boxes refer to a non existing AWN entry (synonym, definition, subtype or supertype).

Our QE process generates three groups of new terms:

- Terms reached by the hyponymy (subtypes) and hypernymy (supertypes) relations: "مهنة" (mhnp : profession), "تَفَاوَضَ" (tafaAwaDa : negotiation), "قِيَادَة" (qiyaAdap : command), "ضَبْط" (DaboT : control), "صَنْعَة" (SanoEap : workmanship), "عَمَل" (Eamal : work) and "نشاط" (n$AT : activity). These terms represent the direct neighborhood of the given question keyword.

- Terms such as رَئِيس (ra}iys : president) and جَمْعِيَّة (jamoEiy~ap : association) that do not exist in the direct neighborhood of the considered AWN synset but can be reached through the definition of the SUMO concept equivalent to that synset.

- Terms not existing in the direct neighborhood of the considered AWN synset but can be reached through the SUMO concept "IntentionalProcess" equivalent to the second supertype of the given synset. The definition of this concept uses two other SUMO concepts: "CognitiveAgent" and "Process". The former is equivalent to the synset represented by the terms "شَخْصِيَّة" ($axoSiy~ap : personality) and "ذات" (aAt* : self). The latter is equivalent to the synset symbolized by "حَدَث" (Hadav : evant) and "وُقُوع" (wuquwE : occurring).

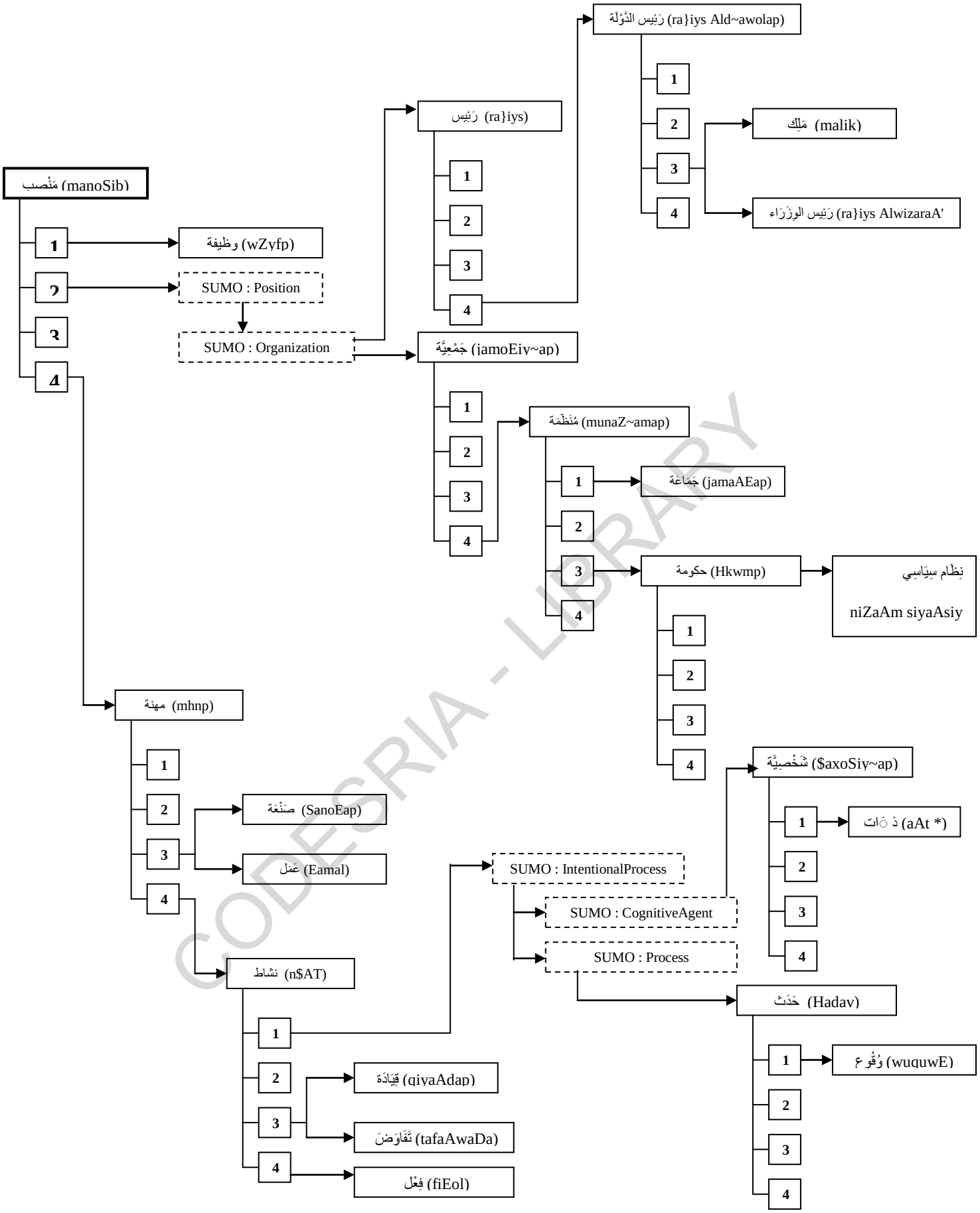


Figure 12. The QE process applied on the keyword “manoSib” using AWN and SUMO relations

Using the QE process based on AWN and SUMO relations for the question keyword "المناصب", we generate new semantically related terms such as "حكومة" (Hkwmp : government) and رَئِيس الوزَرَاء (ra}iys AlwizaraA' : prime minister). Such terms, integrated in the enriched queries, would help in retrieving answers like "رئيس الوزراء الإيطالي" (the Italian prime minister), "رئيس الحكومة الإيطالية" (the president of the Italian government) or " رئيس نادي أس ميلان الإيطالي" (the president of the AS Milan club).

Obviously, the QE process can also generate terms such "حَدَث" that are irrelevant for the question. This is due to the recursive aspect of the process which allows moving from the neighborhood of the keyword synset in AWN to another synset. Thus, setting a threshold to avoid such undesired terms is necessary. Therefore, by experiments this threshold is set by considering only two levels of supertypes and subtypes. For the synonyms and terms produced by the SUMO relation, we do not set any limit due to the importance of the former relation and the low number of connections between AWN synsets and SUMO concepts in the latter one.

As we have seen, improving passage recall is obtained through a recursive QE process based on the synonymy, the hypernymy/hyponymy (stops at level 2) and the SUMO connections in AWN. This process results in a high number of related terms that would enrich queries for a better retrieval of relevant passages. The next section explains how the noise produced by irrelevant QE terms is filtered by considering the structure-based level on top of the retrieved passages, and at what extent relevant terms can improve the PR module.

3.2.4 Structure-based level and passage ranking improvement

In a PR module of a QA system, it is worth improving passage ranking after we obtained a good passage recall. The performance of the answer extraction and validation module highly depends on the relevance of the top ranked passage since, usually, only limited number of passage are considered by the AEV module.

The keyword-based level (described in Section 3.2.3.2) allowed us to guarantee a list of passages with a high recall performance. The ranking at this level is based on the occurrence of question keywords and their related terms in the given passage. To improve passage ranking in the next two levels, we follow two methods: (i) ranking passages according to the Distance Density N-gram between keywords and related terms appearing in the same passage; this is a structure-based comparison which is more effective in the case of factoid questions, and (ii) ranking passages that have similarity in meaning with the processed question; this similarity is computed on the basis of the comparison between the semantic representation of the passages and the question. The latter method has the aim to improve ranking for more complicated questions beyond the case of factoid questions. Also, one of the advantages of the Distance Density N-gram model that can turn into a drawback in complicated questions is

the fact that it assigns the same weight to passages that are reformulations of the question, i.e., those containing the maximum N-gram composed of question terms.

3.2.4.1 Distance Density N-gram Model ranking

We have seen in Chapter 2 that among many passage ranking algorithms, density-based ones are effective, especially for factoid questions. Therefore, we integrate a density-based algorithm in order to deal with factoid questions (Later in Chapter 5, we show how the third level, i.e., the semantic-based level, deals with the other types of questions). This model finds question structures in the passages and gives a higher similarity value to those passages that contain more grouped structures. This similarity depends on the density of question terms in the passage. It is calculated as the sum of all N-gram weights, multiplied by the distance factor and divided by the sum of all term weights of the question.

We focus on the Distance Density N-gram Model (DDNM) implemented in the JIRS system (described in Chapter 2 Section 2.3.3.2) due to its usefulness in the context of QA systems as reported in many works (see Chapter 2). Also, there is an adaptation of the JIRS system made by Benajiba et al. (2007a) in order to take into account the particularities of the Arabic language such as:

- *Text normalization*: it consists of bringing all variants of a character, especially "أ" and "و" to a one "normalized" form;
- *Stopwords*: a list of stopwords that is built for the Arabic language;
- *Diacritization*: the meaning of a word can change if diacritics are removed or added;
- *Agglutinatitive words*: a word can embed a full sentence.

To show how the similarity score is calculated with an example in Arabic, we take the question "من هو القاتل في رواية جريمة قطار الشرق السريع؟" (Who is the killer in the novel *Murder on the Orient Express*?) and the following passages retrieved using Google SE:

Table 5. Sample passages for the given question

ID	Passage
p1	في رواية أغاثا كريستي التي تدور أحداثها في قطار الشرق السريع، جريمة تحدث داخل حافلة القطار، ومجموعة من الناس يسعون لمعرفة من هو القاتل، في ..
p2	نشرت هذه الروايه بالدول العربيه بأسم اخر "كالمعتاد" هو القضيه الكبرى وهي النسخه التي وجدتها ..... 19 -جريمه في قطار الشرق السريع (1934). بعد أنتهاءه ...
p3	قطار الشرق السريع الذي خلدته الكاتبة البوليسية البريطانية أجاثا كريستي عام 1943 ... وأكثر ما يعيق وصول هذا المشروع السياحي والترفيهي إلى دنيا العرب هو مراكز «جريمة في بلاد الرافدين» وميزة روايات أجاثا إنك لا تستطيع تخمين القاتل مهما كانت ..
p4	تسردُ الكاتبة أجاثا كريستي في قصّة مثيرة عنوانها: "جريمة قتل في قطار الشرق السريع" ... إلى أن أتاه أحدُ المسافرين المشهورين, هو المحقّق البلجيكي الشهير ...Murder on the) وبعد أكثر من نصف قرن من نشر هذه الرواية المثيرة, (التي صوّرت فيلما ... موت واحد هو: هل كانت كل دوافع جريمة قتل المبحوح من قبل 11 قاتل وقاتلة, ...
p5	كتبت أغاثا كريستي من روايات وقصص الجريمة سبعا وستين رواية طويلة، وعشرات من القصص ... وحينما سافرت بقطار الشرق السريع خرجت بقصة مشهورة من قصصها وهي (جريمة في قطار الشرق السريع) ... من خلال تكرار ترديدهم لأنشودة الأطفال , أدركوا جميعاً أن القاتل ليس فقط أحدهم , ... (11)شركاء في الجريمة (تومي و توبنس)

As we can see, the right answer to the sample question which is "11 قاتل" or "11 شركاء في الجريمة" exists in the passages ranked by the SE in the 4th and 5th position respectively. Now, let us apply the first step of the structure-based level by assigning weights to each passage.

Considering passage 5 (p5) from the above list, the weight of this passage can be expressed relying on the formula:

$$W_k = 1 - \frac{\log(n_k)}{1 + \log(N)}$$

Where:

- W_k is the weight of the question term t_k, this weight ranges from 0 to 1
- n_k is the number of passages containing the question term t_k
- N is the number of passages considered by the system

The overall score of a passage is then calculated through the following formula:

$$Score_{Passage} = \sum_i w_i$$

In our case, p5 contains all the terms of the question except "هو". For example, the term "رواية" appears in the five considered passages. Therefore, the weight of the term "رواية" is:

$$W_{\text{رواية}} = 1 - \log(5)/(1+\log(5)) = 0.59$$

Similarly, we have calculated the weights of the terms in order to compare passage 2 (p2) and passage 5 (p5). Table 6 lists the calculated weights. The weights assigned allow the identification of relevant passages. In our example, passage 5 is assigned a weight of 4.18 versus 3.53 for passage 2, and this means that passage 5 is more relevant than passage 2. In a similar way, we calculated the weights of the other three passages (see Table 7 and Figure 12).

As detailed in Table 7, the passages containing the right answer, i.e., p4 and p5 have been re-ranked better when their weights have been taken into account (henceforth, we call this the evident weight ranking).

Table 6. Term weights in passage 2 and passage 5

k	P_2	n_k	N	W_k
1	$W_{\text{في}}$	5	5	0.59
2	$W_{\text{رواية}}$	5	5	0.59
3	$W_{\text{جريمة}}$	5	5	0.59
4	$W_{\text{قطار}}$	5	5	0.59
5	$W_{\text{الشرق}}$	5	5	0.59
6	$W_{\text{السريع}}$	5	5	0.59
			$W_{p2} = 3.53$	

k	P_5	n_k	N	W_k
1	W القاتل	4	5	0.65
2	W فى	5	5	0.59
3	W رواية	5	5	0.59
4	W جريمة	5	5	0.59
5	W قطار	5	5	0.59
6	W الشرق	5	5	0.59
7	W السريع	5	5	0.59
			W_{p5} = 4.18	

Table 7. Passage weights for the given example

Passage	SE Rank	Weight	Evident weight Rank	Human Rank
P1	1	4.82	1	3
P2	2	3.53	5	5
P3	3	4.18	4	2
P4	4	4.82	2	1
P5	5	4.18	3	1

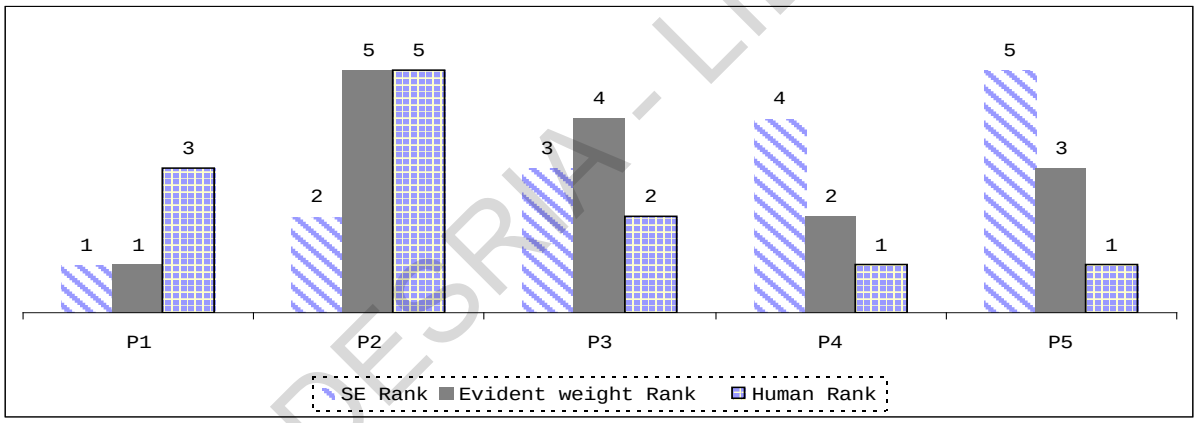


Figure 12. Passage ranking before and after weight assignment

Figure 12 shows that the rank of both passages gained two positions, moving from the 4th and 5th positions to the 2nd and 3rd positions respectively after assigning weights to terms.

The next step now is applying the DDNM in order to re-rank the passages according to this distance and calculate the score of their similarity to the processed question. This score is given by the formula:

$$Sim(p,q) = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{\forall x \in P} h(x) \frac{1}{d(x, x_{max})}$$

Where x is an N-gram of p formed by q (i.e., question) terms, w_i are the weights previously defined, h(x) is the sum of weights of all question terms appearing in the N-gram x, and

$d(x, x_{max})$ is the expression of the distance between the N-gram x and the N-gram with the maximum weight x_{max}, the formula expressing this factor is:

$$d(x, x_{max}) = 1 + k.ln(1 + D)$$

Where D is the number of terms (i.e., terms not appearing in the question) between the N-gram x and the N-gram with the maximum weight x_{max}. The factor k gives an importance to the distance factor in the similarity score. We use the value $k=0.1$ since it was reported that this is the best value according to conducted experiments (Gómez et al., 2007b).

Table 7 shows the new ranking after calculating the similarity score for the five sample passages.

Table 7. Similarity scores after applying the DDNM

	x(1-grams)	x(2-grams)	x(3-grams)	x(4-grams)	x(5-grams)	Sim(p,q)
Passage 1	0.55	1.02	0.00	2.35	0.00	0.81
Passage 2	1.00	0.00	0.00	0.00	2.94	0.82
Passage 3	1.41	0.89	1.77	0.00	0.00	0.84
Passage 4	1.99	0.00	0.00	2.35	0.00	0.90
Passage 5	0.97	0.00	0.00	0.00	2.94	0.81

For each passage, we can see the score showing its similarity with respect to the given question. Also, details are provided about the sub scores of each N-gram in these passages. For example, passage 4 has been assigned the highest similarity score (0.90) which was originated from four 1-grams (i.e., جريمة , هو, الرواية and قاتل) and one 4-gram (i.e., في قطار الشرق السريع). To show the gain in terms of passage ranking improvement, we have illustrated the four ranking methods in Figure 13.

The ranking based on the DDNM performs better since it succeeded in bringing the most relevant passage (from a human perspective) to the first position. Later, this will allow the extraction of the right answer (i.e., قاتل وقاتلة 11).

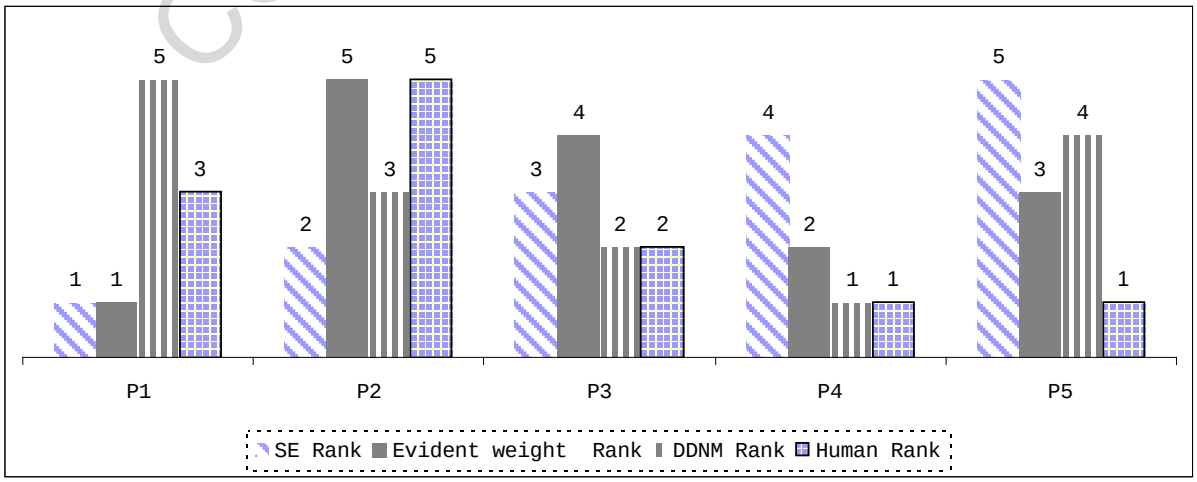


Figure 13. Passage ranking improvement with the DDNM

Another point that deserves to be mentioned is ability of the DDNM ranking to automatically recover the human ranking for passage 3. However, regarding passage 5, which is also one of the most relevant from a human perspective, the improvement of its rank gained through the evident weight rank (i.e., 3rd position) has been lost after applying the DDNM-based ranking (i.e., 4th position). This may be due to the behavior of JIRS when processing long passages such as passage 5.

3.2.4.2 Query Expansion injection in DDNM

The example described in Section 3.2.3.2 has shown the effectiveness of QE in improving the passage recall while the second one has illustrated the gain in terms of passage ranking enhancement. In the latter example, the DDNM has a nearly human ranking for the five passages. However, there are still some drawbacks since there is a relevant passage which is re-ranked wrongly.

The passage recall obtained using QE allowed the structure-based level to re-rank higher number of passages containing not only original question keywords but also semantically related terms. The idea now is considering these terms also in the passage ranking. Henceforth, we calculate the similarity score on the basis of N-grams composed not only from question terms but also from semantically related terms generated from QE.

The new measures are calculated with the following assumptions:

1. The weights of the terms coming from QE are reduced by multiplying the weight of the corresponding term (i.e., original question term) by a factor set heuristically to `0.9` (on the basis of a `0.05` decrease according to the number of considered levels in the hypernymy/hyponymy relation, considering the maximum levels is 20, in our case we have a two-level decrease); so if `x` is an N-gram containing a question term t_1 extended by QE by the term t_{11} then the weight of t_{11} (W_{t11}) in `h(x)` is $0.9 * W_{t1}$;
2. The distance `d(x,`x_{max}`)`is calculated with the same formula, where `D` now is the number of terms between `x` and x_{max};
3. Two N-grams `x` and `y` are completely different if they have no term in common and there is no term in `x` that is a semantic extension of another term in `y`.

The first assumption is motivated by the requirement of reducing the possible noise generated by QE since irrelevant terms could also be brought to the considered queries. The second assumption makes sense after the injection of QE terms in the DDNM model. In fact, we are interested here to assign higher scores to passages containing a higher density of both question keywords and QE terms. Finally, the third assumption is a consequence of the second one.

With these assumptions, we take the same example as in Section 3.2.4.1. We calculate the new similarity scores after injecting one of the terms that are semantically related to the question keyword "رواية" (a novel). From a human perspective, "قصة" is one of these related words. Table 8 lists the new similarity scores after considering the new term "قصة".

Table 8. Similarity scores after QE injection in DDNM

	x(1-grams)	x(2-grams)	x(3-grams)	x(4-grams)	x(5-grams)	Sim(p,q)
Passage 1	0.55	1.02	0.00	2.35	0.00	0.81
Passage 2	1.00	0.00	0.00	0.00	2.94	0.82
Passage 3	1.41	0.89	1.77	0.00	0.00	0.84
Passage 4	**2.02**	0.00	0.00	2.35	0.00	**0.91**
Passage 5	**0.99**	0.00	0.00	0.00	2.94	**0.82**

The unique observed changes regarding the similarity scores are marked in bold. The term "قصة" (a story) exists in two passages: p4 and p5. Note that these two passages contain either the term "قصة" or one of its forms, especially "قصص" (stories) which is its broken plural form. We consider all its forms since the QE would theoretically generate all these forms. Considering this new term has lead to the following changes with respect to our previously mentioned assumptions:

- In passage 4, the 1-gram "الرواية" (the novel) is not considered since, according to *Assumption #3* there is another such 1-gram, i.e., "قصّة" (story); the distance $d(x,x_{max})$ is now reduced to (1+0.1*LN(3)) instead of (1+0.1*LN(24)) previously registered with the original term "رواية" (*Assumption #2* is applied here); this results in a 0.91 similarity score (versus 0.90 without QE injection in DDNM);
- Similarly, in passage 5, the 1-gram "قصصها" (her stories) replaces the 1-gram "رواية" (novel) bringing the similarity score based on DDNM up to 0.82 instead of 0.81.

These changes have positively influenced the passage ranking. Indeed, with the injection of QE terms and according to the described assumptions, it was possible to maintain the relevance of passage 4 and to deal with the drawback registered for passage 5 (relevant according to human ranking) after applying the DDNM without QE injection. That is, the new rank of passage 5 is more relevant that the baseline (i.e., the rank assigned by the SE).

Table 9. Passage ranking improvement with the QE injection in DDNM

Passage	SE Rank	Weight	Evident weight Rank	DDNM Rank	DDNM+QE	Human Rank
P1	1	4.82	1	5	5	3
P2	2	3.53	5	3	3	5
P3	3	4.18	4	2	2	2
P4	4	4.82	2	1	1	1
P5	5	4.18	3	4	3	1

Note that this performance was obtained when injecting the term "قصة" (story) that is semantically related to the keyword "رواية" (novel). Now let us see if the AWN-based QE is able to provide such term. In the AWN database, the keyword "رواية" (novel) corresponds to

two synsets with the AWN IDs “riwaAyap_n1AR” and “riwaAy~p_n1AR”. In terms of synonyms we can generate new terms such as “تقرير شفوي” (oral report) and “تقرير” (report). Regarding the hypernyms and hyponyms, we can get terms including “عمل أدبي” (literary work), “كتابة” (writing) or “رسم” (drawing). Using the SUMO-related terms, we have again the term “كتابة” (writing). Unfortunately, the AWN lexical database does not contain the term “قصة” (story), and this is a real limitation to the use of this resource as a support of our multi-levels approach.

Another example showing the effectiveness of the passage ranking using QE injection in DDNM is the case of NEs. Generally, more importance is devoted to NEs in factoid questions. Rather than replacing NEs with their related NEs, it is worth enriching the question with its class (i.e., hypernym). For example, if the question was “ من هو القاتل في رواية أجاثا كريستي جريمة قطار الشرق السريع؟” (Who is the killer in the story of Agatha Christie *Murder on the Orient Express*?) then it makes sense to inject the hypernym of the NE keyword “أجاثا كريستي” (Agatha Christie) when applying the DDNM model. Therefore, the 3-gram “الكاتبة أجاثا كريستي” (The author Agatha Christie) would be considered in passage 4, and in turn the similarity score will be increased. Once again, the released version of AWN only contains just a few number of NEs.

As previously mentioned, our research methodology is based on experiments conducted using large collections of questions in order to evaluate performance as follows:

- Before and after applying the surface side of our approach, i.e., keyword-based and structure-based levels;
- Before and after performing AWN enrichment (NEs, nouns, etc.) to show the impact of its coverage in the improvement of our approach;
- Before and after applying the deeper side of our approach, i.e., the semantic-based level.

The next section presents the obtained results regarding the first evaluation. The second and third evaluations will be presented in Chapter 4 and Chapter 5 respectively.

3.2.5 Surface-side evaluation

The current section presents the experiments conducted to evaluate the effectiveness of the surface-based levels (i.e., keyword-based and structure-based levels) of our Arabic PR approach. The first sub section presents the tasks performed in the evaluation process and how performance are measured. The second sub section describes the test-set of questions. The third sub section provides the obtained results for each conducted experiment. The fourth sub section discusses these results and highlights the significance of the conducted experiments.

3.2.5.1 Evaluation process and measures

Since we are interested in the PR module, the evaluation process does not use a full Arabic QA system. Rather, it integrates a pipeline of the three modules of a typical QA system (see Chapter 2 Section 2.2.2) as illustrated in Figure 14.

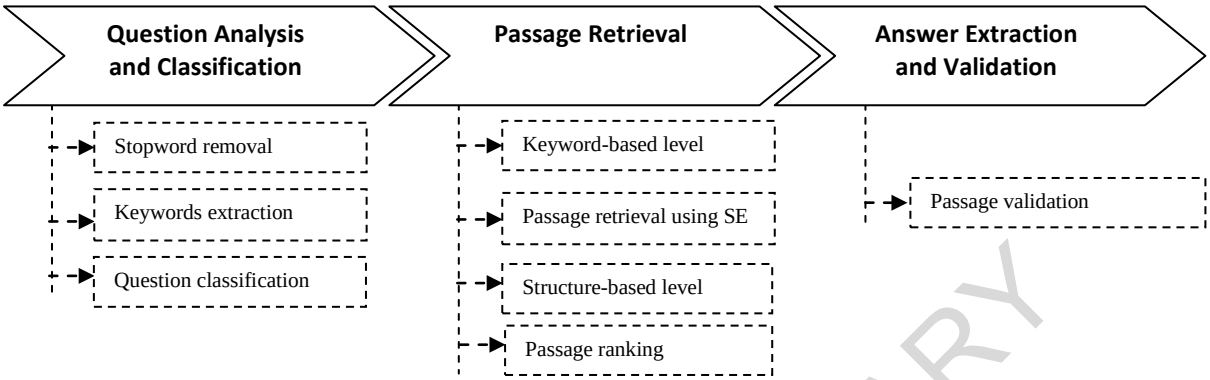


Figure 14. Modules of the QA process

The above figure illustrates the tasks performed in each module with respect to the general architecture of a QA system as follows:

- *Shallow question analysis and classification module*: In this module, a question is analyzed by: (i) removing stopwords using the list provided in JIRS after a slight enrichment; (ii) extracting question keywords; (iii) classifying the question on the basis of the test-set information.
- *Passage Retrieval module* that performs: (i) the keyword-based level providing a list of semantically related terms and enriched queries using the AWN-based QE process (see Section 3.2.3.2), (ii) the retrieval of passages from the Web by using the enriched queries as inputs in the Yahoo! API;[6] (iii) the structure-based level based on the indexation and searching processes of the JIRS system; and (iv) the ranking of passages on the basis of the keyword-based and structure-based levels to obtain the five most relevant passages with respect to the given question.
- *Answer Extraction and Validation module:* This module validates each of the five passages from the list provided by the PR module. The validation is based on the right answer related to the given question. A passage is marked as "valid" if it contains the right answer.

In order to measure the performance of our approach, we did not adopt recall, precision and F-measure since they are most suitable to unranked retrieval situations (see Chapter 2, Section 2.2.3.2). In this part, our main objective is evaluating the passage ranking improvement which

[6] http://www.yahoo.com

we evaluate by adopting the other well-known measures introduced in Chapter 2 (see Section 3.2.3.2):

- The Accuracy registered for a question set S, calculated according to the formula:

$$\mathbf{Acc} = \frac{\mathbf{1}}{\mathbf{Ns}} \sum_{k \in s} V_{k,1}$$

  Where N_s is the number of questions of the question set S (two question sets are considered, CLEF set and TREC set, their description is presented in the next section).

  $V_{k,j}$ is a value assigned by the AEV module to the passage j related to question q_k from the list of the five passages provided by the PR module. This value is equal to 1 if the answer to the question q_k is found in the passage having the rank j (j is between 1 and 5), and it is equal to 0 otherwise. In the accuracy measure, we are only interested in the best-ranked passage (j=1).

- The Mean Reciprocal Rank registered for a question set S, its formula is:

$$MRR = Avg_{k \in s}\left(\frac{1}{5}\sum_{j=1}^{5}\frac{V_{k,j}}{j}\right)$$

  Where k is the index of a question q_k belonging to the set S (i.e., CLEF and TREC), j is the rank of a passage. Unlike the accuracy, the MRR is interested in the first five passages (j={1,2,3,4,5}).

- The number of Answered Questions, the number of questions in a set S (CLEF or TREC) for which we find the answer in at least one of the passages ranked in the first five positions. It is calculated according to the formula:

$$AQ = \frac{1}{Ns}\sum_{k \in s} \max(Vk, j)$$

  Where k is the index of a question q_k belonging to the set S, N is the number of question contained in the set S and $V_{k,j}$ the value assigned to the five passages returned in response to the question q_k.

3.2.5.2 Test-set questions

As we have seen in Chapter 2, the evaluation campaigns organized on QA provide material such as collections of questions with their right answers as well as collections of documents from which answers should be searched and extracted. Among this material, we have been interested in those provided by the CLEF and the TREC tracks.

In order to achieve the main objectives of this research regarding the investigation of our QA system from the perspective of the different previously mentioned challenges (*language*, *Web collection*, *questions* and *evaluation*), we consider questions available from different yearly editions of CLEF (between 2003 and 2010) and TREC (between 1999 and 2008). These editions provided questions and document collections that let the participating QA systems to address the above challenges.

Using these two test data sets allows us to conduct experiments with the same distribution of questions in terms of covered topics, question categories, nature of the expected answer, etc. This helps to compare the performance obtained for Arabic QA with the baseline system (Yahoo! API) and the surface-based levels of our approach.

The test data provided by the two competitions covers a considerable variety of languages (English, French, Spanish, Italian, Dutch, etc.). Unfortunately, the Arabic language is not among them. Therefore, there is a need of translating the questions and documents into Arabic. In the context of the current work, we manually translated all the considered TREC and CLEF questions available in English and French.

The number of translated questions[7] is: 1500 for the TREC set and 764 for the CLEF set. These questions are classified into different domains (sport, geography, politic, etc.) and different types. The types are identified on the basis of the expected answer. The considered types are:

- MEASURE: for instance "What distance does the Granada-Dakar rally cover?" " ما هي المسافة التي يغطيها رالي غرناطة - دكار؟"
- ABREVIATION: for instance "What is NASA?" "ما هي ناسا ؟"
- COUNT: for example "How many people are killed by landmines every year?" " كم عدد الاشخاص الذين يقتلون سنويا من جراء الألغام الأرضية ؟"
- PERSON: "What is the name of the Queen of the Netherlands?" " ما هو اسم ملكة هولندا ؟"
- OBJECT: "What is exhibited in the Vitra Design Museum?" " ما الذي يعرض في متحف فيترا للتصميم ؟"
- LOCATION: for instance "What is the capital of Chechnya?" "ما هي عاصمة الشيشان ؟"
- ORGANIZATION: "Which organization does Vanessa Redgrave support?" " ما هي المنظمة التي تدعمها فانيسا ريدجريف ؟"
- TIME: "When was the Universal Declaration of Human Rights approved?" " متى تمت المصادقة على الإعلان العالمي لحقوق الإنسان ؟"
- LIST: "Tell me names of robots." " أعطي أسماء روبوت."

[7] Available for download from the Web site of the Ibtikarat Team at: http://sibawayh.emi.ac.ma/web/i/?q=projects or alternatively from the NLE Lab Web site at: http://www.dsic.upv.es/grupos/nle/downloads.html

- OTHER: " What is a risk factor for cardiovascular diseases? " " ما هو أحد عوامل الخطر لأمراض القلب والأوعية الدموية ؟"

Tables 10 and Table 11 show, for each set, the number of questions belonging to the different question types.

Table 10. CLEF questions per types

TYPE	#Q	Percent.
PERSON	183	24%
LOCATION	123	16%
TIME	93	12%
COUNT	89	12%
ORGANIZATION	63	8%
ABREVIATION	34	4%
OBJECT	26	3%
LIST	22	3%
MEASURE	16	2%
OTHER	115	15%
TOTAL	**764**	**100%**

Table 11. TREC questions per types

TYPE	#Q	Percent.
LOCATION	307	20%
PERSON	258	17%
TIME	208	14%
ABREVIATION	133	9%
COUNT	106	7%
ORGANIZATION	57	4%
MEASURE	56	4%
OBJECT	29	2%
LIST	6	0.4%
OTHER	340	22.7%
TOTAL	**1,500**	**100%**

In both sets, questions mainly belong to factoid types (for which the expected answer is a NE). Indeed, roughly 60% of the CLEF and 55% of the TREC questions ask about PERSON, ORGANIZATION, TIME and LOCATION. The percentage of unclassified types (OTHER) of questions is more important in TREC where it represents 23% versus 15% in the CLEF set. Generally, these are the questions that have higher complexity to answer by QA systems.

Another feature that deserves to be mentioned is the number of words per question. This is important as complexity of processing also depends on the question length (the longest ones are often the more complicated to analyze at different levels including morphology level, syntax, semantic, etc.). We illustrate in Figure 15 the distribution of questions according to three length ranges: (i) questions containing more than 10 words; (ii) questions containing between 5 and 9 words; and (iii) questions with less than 5 words.

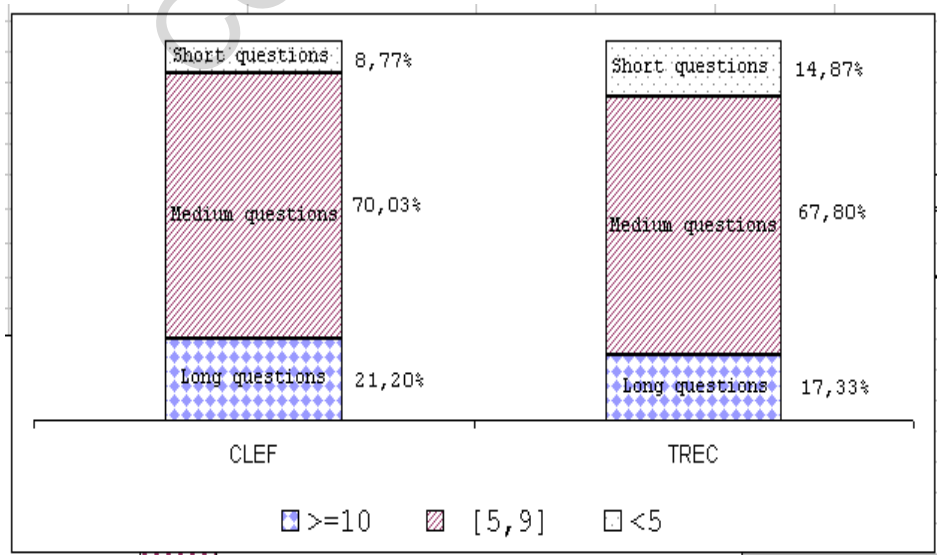


Figure 15. Distribution of CLEF and TREC questions according to the length feature

Both sets of questions are quite similar in their length distribution. The percentage of long questions is slightly higher in the CLEF set as 21.2% of its questions contain more than 10 words (versus 17.33% in TREC) and 70.03% are formed by 5 to 9 words (while 67.8% of TREC belongs to this range). Note that the overall average of words in both sets is quite similar (around 7.26 words per question). The difference in length would help us to re-evaluate the behavior of JIRS when processing long Arabic questions and passages (previously mentioned at the end of section 3.2.4.1).

In terms of content, 75% of the questions in the CLEF and TREC sets could be manually classified under 6 different topics as illustrated in Figure 16.

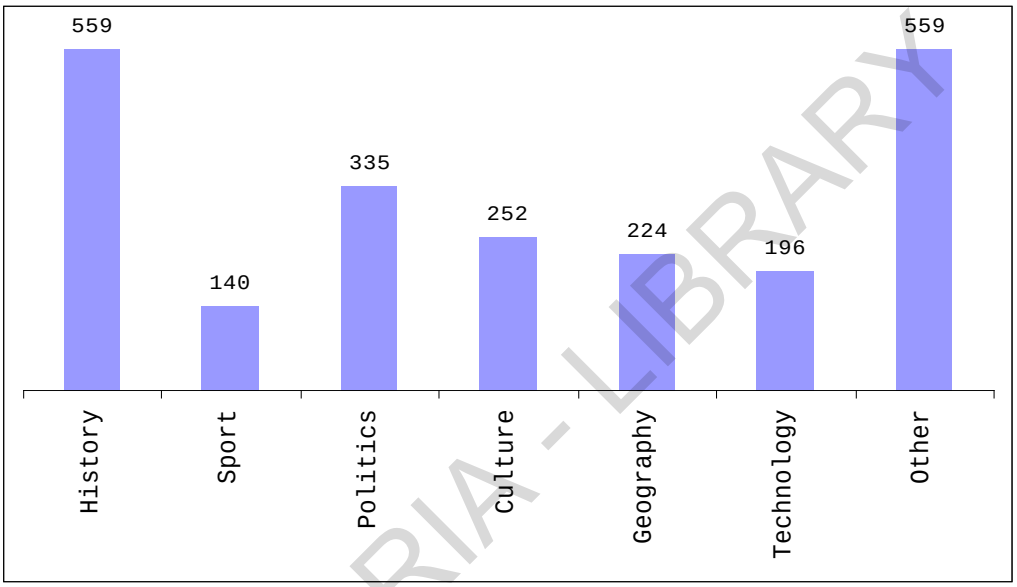


Figure 16. Distribution of CLEF and TREC questions according to the topic

We can see that "History" is among the most frequent topics in these questions with 559 questions (nearly 25% out of the 2,264 questions of the CLEF and TREC sets). This raises another complexity level, since processing questions about "History" deals with the issue of temporal information, especially in a dynamic collection of documents such as the Web.[8]

3.2.5.3 Results

A) Query Expansion

For the TREC questions, the QE has been performed for 858 questions (out of 1,500). This means that AWN contains corresponding entries for 57.2% of the TREC questions. This percentage is higher in the case of the CLEF questions reaching 80.10% (612 questions out of 764). The overall coverage of AWN with respect to the two question sets is 64.93%. Table 12

[8] In the CLEF and TREC 1999-2008 competitions, it was given static collections.

and Table 13 show the AWN coverage for the two question sets with respect to the type of QE.

TABLE 12
AWN semantic relations coverage for the CLEF Questions

RELATION TYPE	#Q	%
Synonyms	608	99.35
Supertypes	143	23.37
Subtypes	102	16.67
SUMO-Definitions	36	5.88

TABLE 13
AWN semantic relations coverage for the TREC Questions

RELATION TYPE	#Q	%
Synonyms	850	99.07
Supertypes	179	20.86
Subtypes	132	15.38
SUMO-Definitions	26	3.03

The coverage for the two question sets has the same trend in the four considered semantic relations. Indeed, for 99.35% of the CLEF questions covered by AWN (versus 99.07% of the TREC set) there is at least one keyword that can be expanded by its synonyms in the AWN. That is, for almost all the questions, at least one query can be formed by replacing the keyword in the question by one of its synonyms. However, the average of the generated queries from the synonymy relation does not exceed 3.65 queries per question for the CLEF set and 4.26 for the TREC set. The coverage of AWN in terms of hyponymy (subtypes) and hypernymy (supertypes) relations is under 25% (the best percentage is registered for CLEF questions by around 17% for subtypes and 23% for supertypes). For the SUMO-definition relation, the coverage is very low and is close to 6% for the CLEF questions.

With respect to the considered question sets, the above statistics show that AWN is more developed regarding the synonymy relation. However, the hierarchy of synsets in terms of hyponymy/hypernymy relation needs more efforts. Also, the connection existing between AWN synsets and the SUMO concepts did not allow for generating a higher number of terms from this relation.

By considering only the subset of questions that can be expanded (65% of the CLEF and the TREC questions), we can evaluate the impact of enriching queries by new generated terms in the context of Arabic PR. Note that in this evaluation, we do not take into account yet the passage ranking based on the question structure. Table 14 shows the results obtained in this experiment.

Table 14
Keyword-based performance using QE for the CLEF and the TREC questions (Strict Validation)

	CLEF		TREC	
MEASURES	Without QE	With QE	Without QE	With QE
Acc	5.07%	8.35%	3.38%	5.24%
MRR*	1.66	3.12	1.21	2.04
AQ	12.09%	17.97%	7.58%	12.82%

* Note that the MRR has been multiplied by 100 in order to have a better readability.

We can see that with QE we obtain a better performance in terms of Accuracy, MRR and number of Answered Questions. Indeed, by using the QE based on AWN we gain 3.28%, 1.46 and 5.88%, for the CLEF set and 1.86%, 0.83 and 5.24% for the TREC questions, respectively.

In the case of the CLEF questions, the number of answered questions represents nearly 18% of the 764 questions of this set. For the remaining 82%, there are two cases: (i) the answer is not in the first five passages; or (ii) the answer appears in one of the five passage but it could not be identified by our automatic Answer Validation process.

The latter case can occur due to different factors. Generally, this is due to the multi-word answers, i.e., answers with more than one word. For instance, if the question is " متى ولد توماس مان ؟" (When was Tomas Mann born?) and the answer is "6 يونيو 1875" (6th June 1875), our process fails to extract the answer in a passage containing just the month and the year or the year only. Therefore, we investigate the introduction of relaxations for a lenient validation as follows:

- For the date answers, if the process fails to extract them we try then to search only the year;
- In the date answers, we search also with the Arabic corresponding months such as "أيلول" (September) ;
- If the process fails to identify the answer in a passage, we try identifying its stem instead of the entire word.

In addition to those relaxations, we perform, for multi-word answers, a sub process which allows identifying passages that contain at least one word of the answer. For instance, the question "من الذي اخترع الهاتف ؟" (Who did invent the telephone?) has the answer " الكسندر غراهام بيل" (Alexander Graham Bell), so if the word "الكسندر" (Alexander) appears in a passage then it is listed for a manual validation. Therefore, we obtain a list where each row contains the question, its answer, and the passage containing parts of the answer (e.g., Alexander). This list is then manually checked in order to confirm whether the concerned passages are relevant or not. This manual validation is only done in the evaluation step and allows us to avoid any impact on the results obtained due to the mentioned relaxations. After this lenient validation, we obtained the results listed in Table 15.

Table 15
Keyword-based performance using semantic QE for the CLEF and the TREC questions (Lenient Validation)

	CLEF		TREC	
MEASURES	Without QE	With QE	Without QE	With QE
Acc	11.76%	14.40%	8.16%	12.35%
MRR	3.85	5.59	3.1	5.05
AQ	25.16%	29.74%	16.78%	23.43%

The results presented in Table 15 show that the performance in term of accuracy, MRR and number of Answered Questions has been improved after the lenient validation. In the next section, we conduct new experiments by injecting terms generated using QE (four terms per question on average) in the Density Distance N-gram Model.

B) QE Injection in DDNM

As described in Section 3.2.5.2, the structure-based level considered in this evaluation is based on the JIRS system that implements the DDNM. An amount of m passages (snippets) is extracted from the Web using the queries enriched by the QE process (evaluated in the previous section). The value of m is equal to 1,000 since previous experiments (Gomez et al., 2007), with the same system, concluded that the optimal value of m is between 800 and 1000 for the Spanish CLEF document collection. The results of this experiment are shown in Table 16 by distinguishing performance according to the use of JIRS (i) without QE and (ii) with QE.

Table 16
Structure-based performance using JIRS for the CLEF and the TREC questions (Strict Validation)

	CLEF		TREC	
MEASURES	Without QE	Using QE	Without QE	Using QE
Acc	8.77 %	**11.60%**	6.41%	**8.51%**
MRR	3.99	**5.26**	2.78	**3.7**
AQ	12.09 %	**16.01%**	6.99%	**10.60%**

Injecting QE terms in DDNM improves the performance in both sets of questions for the three considered measures. Accuracy and MRR have been improved in CLEF (11.6% Accuracy, 5.26 MRR) and in TREC (8.51% Accuracy, 3.7 MRR) if compared to what we had obtained with the keyword-based evaluation in the strict validation process (8.35% Accuracy, 3.12 MRR for CLEF and 5.24% Accuracy, 2.04 MRR for TREC). However, the number of

answered questions has slightly decreased. Considering the lenient validation allows for increasing the reached performance. Table 17 lists these performance.

Table 17
Structure-based performance using JIRS for the CLEF and the TREC questions
(Lenient Validation)

	CLEF		TREC	
MEASURES	Without QE	Using QE	Without QE	Using QE
Acc	19.89 %	**21.90 %**	13.64%	**18.99%**
MRR	9.12	**10.08**	6.06	**8.61**
AQ	27.45 %	**29.90%**	15.50%	**24.48%**

According to the above performance, the combination of QE and DDNM show the better results for both sets of questions. The obtained Accuracy is 21.90% in CLEF (versus 14.40% before applying the structure-based level) and 18.99% (versus 12.35%). Similarly, the MRR is 10.08 and 8.61 respectively (versus 5.59 and 5.05 respectively with only the keyword-based level). The number of answered CLEF questions moved from 29.74% before considering DDNM to 29.74% (versus 23.43% to 24.48% for TREC).

3.2.5.4 Discussion

The conducted experiments confirm the potential effectiveness of the combination made by introducing the terms generated in the keyword-based level into DDNM implemented in the structure-based level.

The experiments showed that, regardless of the question set, the performance in terms of accuracy, MRR, and number of Answered Questions improves when we include separately our QE process based on AWN and then JIRS as a structure-based PR system.

The highest performance is obtained when we include JIRS together with QE. Indeed, for the TREC questions the accuracy shows a significant increase from 8.16% to 18.99%, the MRR from 3.1 to 8.61 and the percentage of answered questions from 16.78% to 24.48% with different relaxations included in the lenient validation process.

The effectiveness of using JIRS together with QE is even better in the case of the CLEF questions. Indeed, Accuracy is close to 22% instead of 12%, MRR is 10.08 rather than 3.85. The use of JIRS and QE allows obtaining the answer in one of the first five returned passages for about 30% of the questions (versus 25.16% before JIRS+QE). Table 18 shows a detailed analysis of the obtained results in terms of the number of answered questions per type of question.

Table 18 Types of the answered questions per question set (Lenient Validation)

TYPES	CLEF		TREC	
	Without JIRS+QE	With JIRS+QE	Without JIRS+QE	With JIRS+QE
ABREVIATION	6.49%	1.64%	2.78%	5.24%
COUNT	7.14%	8.74%	9.03%	5.71%
LIST	2.60%	2.73%	0.69%	0.95%
LOCATION	19.48%	21.86%	21.53%	22.38%
MEASURE	2.60%	1.64%	7.64%	5.24%
OBJECT	2.60%	2.19%	2.78%	4.76%
ORGANIZATION	5.19%	9.29%	6.94%	7.14%
OTHER	13.64%	12.57%	17.36%	13.33%
PERSON	29.87%	25.68%	14.58%	23.81%
TIME	10.39%	13.66%	16.67%	11.43%

For instance, 25.68% of the answered CLEF questions are of the type PERSON and 21.86% are of the type LOCATION. For the same questions, 80.87% of the answered questions are factoid ones while this percentage is 75.71% for the TREC set. Let us now consider the overall performance among all the 1,470 questions (those covered by the AWN-based QE) in both sets as listed in Table 19.

Table 19
The overall performance before and after using the semantic QE with JIRS
(Lenient Validation)

1,470 CLEF+TREC questions		
MEASURES	Without JIRS and QE	With JIRS+QE
Acc	9.66%	**20.20%**
MRR	3.41	**9.22**
AQ	20.27%	**26.74%**

The improvement of Arabic PR is significant as Accuracy is nearly 20.20% (versus 9.66%), MRR is 9.22 (versus 3.41) and the percentage of answered questions is 26.47% (versus 20.27%). The gain obtained in terms of MRR was the best with respect to the two other measures. This means that our approach increases the probability of having the expected answer in the first five ranked passages.

To show the significance of the obtained results, we use the Student's paired t-test. Therefore, for each measure (Accuracy, MRR and number of Answered Questions), we consider the directional hypothesis that “performance becomes better when we inject QE terms in DDNM”. The null hypothesis is:

H_0 = the performance (acc, MRR or #answered questions) is not positively impacted by the use of QE and DDNM combined.

Our alternative unilateral hypothesis is:

H_1 = the performance (Acc, MRR or #answered questions) is positively impacted by the use of QE with DDNM combined.

The t-test value is calculated using the following formula:

$$t = \frac{\left|\bar{x}_{no} - \bar{x}_{jq}\right|}{\sqrt{\frac{s^2(no)}{n(no)} + \frac{s^2(jq)}{n(jq)}}}$$

Where:

$\bar{x}_{no}$ is the mean (in terms of the considered measure) of the sample processed without using QE and DDNM.

$\bar{x}_{jq}$ is the mean (in terms of the considered measure) of the sample processed using QE and DDNM.

s^2 is the variance of the sample

$n(S)$ is the number of observations in the sample S. In our case, we take into consideration four observations related to the different question collections used previously (858 TREC questions, 612 CLEF questions, 82 TREC questions and 82 CLEF questions).

The degree of freedom is: $df = 7$

The calculated t-test values are:

- In the case of accuracy: $t=3.42$
- In the case of MRR: $t=1.45$
- In the case of the Number of answered questions: $t=2.23$

According to the t-test values above, we can reject the null hypothesis in the case of the accuracy ($t=3.42, df=7, p<0.05$) and the number of Answered Questions ($t=2.23, df=7, p<0.05$). For the MRR, the difference in performance between the two samples is not significant.

3.3 Chapter summary

In this chapter, we presented our surface-based approach for the Passage Retrieval module. It consists in combining the QE process based on AWN semantic relations with the Distance Density N-gram Model that relies on structure similarity. To test its effectiveness, we highlighted, through the processing of a sample question and the retrieval of corresponding passages, the difference in terms of passage ranking after applying the keyword-based level and the structure-based level, separately and then as a combined level. The passage ranking

obtained by the combined level (i.e., the surface-based approach) was the closest to the human ranking. It was also better than the ranking provided by the baseline system (i.e., the considered Search Engine) for the given question.

Thereafter, we presented the experiments that we conducted on a set of 2,264 translated TREC and CLEF questions provided by both campaigns over ten years (between 1999 and 2008). The surface-based level was applied to each question in the set. This allowed us to measure the performance in terms of Accuracy, MRR and Number of Answered Questions. The obtained performance regarding the three measures was better than the one registered by the baseline system, i.e., the Yahoo! API (20.20% versus 9.66%, 9.22 versus 3.41 and 26.47% versus 20.27% respectively). The statistical t-test also showed the significance of these results.

The conducted experiments indicate encouraging performance results in light of the following elements:

- The experiments were conducted in an open domain (the Web). This means that the content of passages is not always written in a formal style like the one used in the questions;
- The passages (snippets) returned by the baseline system (i.e., Yahoo! API) are usually so small that it is difficult to have both the question terms and the expected answer in the same passage;
- Questions are not about the Arabic culture. Indeed, the CLEF and TREC questions used in the test are translated from the European and the American cultures respectively to the Arabic language. Hence, we are not sure that the available Arabic content in the Web will cover the questions topics or not. This will cause a low redundancy level. Unfortunately, the DDN model works better when redundancy is high in which case it is more likely to retrieve at least one relevant passage in this case.
- Most of the answers are NEs that are transliterated from English or French to the Arabic language. Therefore, answers could not be found in Arabic texts and the performance can be affected by spelling errors.

As mentioned in the examples presented in this chapter, the enrichment of AWN is required for a better usability of this resource in a complex application such as QA systems. The next chapter investigates this enrichment by focusing on the type of content that occurs in questions not covered by AWN as well as those not answered by the keyword-based and structure-based levels.

Chapter 4

AWN resource enrichment

4.1 Introduction

The proposed approach for Arabic PR uses the Arabic WordNet, as described in the previous chapter, for different needs: (i) extraction, in the keyword-based level, of semantically related terms with respect to a given question keyword, (ii) injection of these terms to calculate the similarity score based on the Distance Density N-gram model, and (iii) use of AWN hierarchy as a support of ontological resources with the aim, first, to allow the representation of the question and passages in terms of Conceptual Graphs and, second, to make their semantic comparison. The issue latter will be described in Chapter 5.

The use of AWN was motivated by its nearly-standard design leveraging the experiences registered within the last decade in building over 40 WordNets. However, as we showed in the examples presented in the previous chapter, this resource has to be enriched and adapted for the objective of covering a higher number of common classes words, including nouns and verbs, and being linked to NEs since factoid questions are important in any QA system.

In this part of our research, we do not try to build a new release of AWN with a full support of Modern Standard Arabic (MSA). Rather, we are aiming for a more enriched release that can be evaluated and compared to the standard release in terms of coverage and usability in the context of Arabic QA. To achieve this goal, we have been inspired by existing experiences, either for the extension of WNs or their use in different applications, especially in Information Retrieval and QA.

The remainder of this Chapter is structured as follows: Section 4.2 provides a theoritical analysis of the AWN resource according to two lines: (i) content and (ii) usability. Section 4.3 presents an experience-based analysis of this resource starting from the evaluation conducted in the previous chapter. Section 4.4 describes different methods used in terms of the AWN enrichment and the amount of entries added by means of these methods. Section 4.5 provides an evaluation of the impact of the new added content on the effectiveness of our Arabic PR approach. Finally, we conclude this chapter with a synthesis of the main issues investigated in this part of the research.

4.2 Theoretical analysis of Arabic WordNet

Generally, the existing experiences related to the construction of WordNets followed the trend aiming for better coverage of main concepts and semantic relations. These experiences have given rise to many development methods to overcome several known WordNet challenges. These challenges became more conspicuous when dealing with languages less commonly addressed by NLP research. The latter case includes, among others, Arabic and Hebrew, the most prominent members of the Semitic language family.

The construction of AWN followed the general trend, leveraging the methods developed for Princeton WordNet (PWN) (Fellbaum 1998) and EuroWordNet (Vossen 1998). The result was a linguistic and semantic resource that complies with the WN structure while considering some specificities of Arabic such as entry vocalization, Broken Plurals (BP) (i.e., irregular plural) and roots. The first release of this resource may well be viewed as a valuable step in terms of the following findings:

- The most common concepts and word-senses in PWN 2.0 have been considered in AWN;
- AWN provides some culture-specific senses. For instance, the word sense أرض الكنانة (the land of Egypt), which is commonly used in Arabic to refer to the country "Egypt", belongs to the synset "جُمْهُورِيَّة" (republic);
- AWN is designed and linked to PWN synsets so that its use in a cross-language context is possible;
- Similarly to other WordNets, AWN is connected to the SUMO ontology (Niles and Pease, 2001; Niles and Pease, 2003; Black et al., 2006). A significant number of AWN synsets was, indeed, linked to their corresponding concepts in SUMO. Statistics show that 6,556 synsets in AWN (65.56% of the synsets) are linked to 659 concepts in SUMO (65.9% out of 1,000 concepts). These links complements synset information related to synsets with the formal definitions provided by SUMO. Note that on behalf of the SUMO project, an ontology, among others, was developed around the concepts that are specific to the Arabic culture.

Before releasing AWN, the lack of linguistic resources had always been an obstacle to the development of efficient and large scale Arabic NLP systems. Once released, AWN quickly gained attention and became known in the Arabic NLP community as one of the rare freely available lexical and semantic resources.

In the absence of any study about the AWN project after it was launched, it is interesting to evaluate the resource in terms of two aspects: *coverage* and *usability*. Concerning AWN *coverage*, it seems logical to begin by comparing AWN contents with those of a lexicon covering modern standard Arabic and with other WordNets.

4.2.1 Comparison to existing WordNets

AWN contains around 18,925 Arabic word-senses[1] belonging to roughly 9,698 synsets,[2] very poor content indeed in comparison to other WordNets. Table 20 presents a comparison among Arabic, Spanish[3] and English[4] WordNets contents, as well as the estimated ratio of the number of word lemmas in each Wordnet to the number of words in large lexical resources corresponding to each language.[5]

Table 20 Comparison of AWN content to the English and Spanish WNs

	Arabic	Spanish	English
WN Synsets	9,698	57,424	117,659
WN Word-Senses	18,925	106,566	206,941
WN Word Lemmas (WL)	11,634	67,273	155,287
Language Lemmas (LL)	119,693	104,000	230,000
Ratio lemmas (WL/LL)	9.7%	64.7%	67.5%
Ratio Word-lemmas (WN/English WN)	7.5%	43.3%	100.0%
Ratio Synsets (WN/English WN)	8.2%	48.8%	100.0%
Ratio Word-senses (WN/English WN)	9.1%	51.5%	100.0%

Table 20 shows that: (i) on the one hand, the released AWN contains only 9.7% of the estimated number of word lemmas in the Arabic lexicon considered (versus 67.5% for the English WN and 64.7% for the Spanish WN), which in turn represent roughly 7.5% of those existing in English WN; and (ii) on the other hand, the number of synsets in AWN represents only 8.2% of the English WN synsets. For the Spanish WN, this number represents 48.8% of the English WN synsets.

The link between word lemmas and synsets is established through word-sense pairs that represent 9.1% of what exists in English WN (51.5% in the case of Spanish WN). Furthermore, AWN synsets are mainly linked by only two kinds of relations hyponymy and synonymy, versus the seven semantic relations used in English WN (which also include antonymy and meronymy, among others).

[1] In WordNet, a word lemma that appears in *n* synsets has *n* word-senses.
[2] AWN statistics are extracted from the AWN browser and database available at: http://www.globalWordNet.org/AWN/AWNBrowser.html
[3] Spanish WN 1.6 statistics are extracted from the MultiWordNet project, see: http://multiWordNet.fbk.eu/online/multiWordNet-report.php
[4] English WordNet 3.0 statistics are extracted from: http://WordNet.princeton.edu/WordNet/man/wnstats.7WN.html
[5] The considered lexical resources are: DIINAR.1 lexicon for Arabic (http://diinar.univ-lyon2.fr/) which presents the advantage of containing voweled and lemmatized entries that exist in the language, the Spanish lexicon and the British English Source Lexicon (BESL) for English (both are large and contain morphological information). The three resources are published by ELRA (statistics are extracted from http://catalog.elra.info).

4.2.2 AWN compared to existing MSA lexicon

To make the AWN coverage described in Table 20 more precise, detailed figures about the number of AWN synsets and words are presented in Table 21 with an emphasis on the following three elements:

- Nouns and verbs, as the main Common Linguistic Categories (CLC);
- Named Entities, as one of the most important types of dynamic information to link with the AWN resource, since AWN is designed for various Arabic NLP applications and domains, including the Web, where NEs are widely used; Also, we are interested in NEs since the injection of their hypernym is effective in the structure-based level as shown in the exemple provided in Chapter 3, Section 3.2.4.2;
- Broken plurals, as a linguistic characteristic mainly specific to Arabic, which are formed by changing the word pattern, not by using regular suffixation. AWN can be used in different NLP applications, particularly, in Information Retrieval, but the Arabic light stemming algorithms that are reported to be effective in this field do not extract the correct stem for BP (Goweder and De Roeck 2001). The use of lexical resources that integrate these BP forms can resolve such problems. Therefore, it makes sense to devote more attention to the enrichment of AWN in terms of BP forms.

Table 21 Detailed AWN statistics

Statistics	CLC		Dynamic information	Arabic-specific characteristic
	Nouns	Verbs	Named Entities	Broken Plurals
No. AWN Synsets	7,162	2,536	1,155	126
No. AWN Word-senses	13,330	5,595	1,426	405
No. AWN Distinct Lemmas	9,059	2,575	1,426	120
No. Baseline Lexicon Lemmas (BLL)	100,236	19,457	11,403	9,565
Percentage AWN Lemmas/BLL	9.0%	13.2%	12.5%	1.3%

In Table 21, we compare the number of lemmas in AWN with DIINAAR.1 as a baseline lexicon (Abbès et al., 2004). This comparison shows that, with respect to the three elements under consideration (CLC, Dynamic Information, etc.), the gap between the two lexical resources is significant. In fact, lemmas in AWN account for only around 9% of nouns and 13.2% of verbs in the baseline lexicon. For dynamic information, this percentage is about 12.5%. The BP forms, peculiar to Arabic, are hardly covered in AWN: it only contains 1.25% of similar forms in the baseline lexicon.

In previous work (Alotaiby et al., 2009), experiments conducted on nearly 600 million tokens from the Arabic Gigaword corpus (Graff 2007) and the English Gigaword corpus (Graff et al., 2007) showed that the total number of Arabic word types needed in any application is 1.76 times greater than that of English word types required for the same application. On the basis of the foregoing statistics, it is clear that AWN *coverage* is limited compared to the DIINAR.1 lexicon for Arabic and to other WNs. Therefore, one may question the usefulness of the resource and its response to the needs in different applications.

4.2.3 AWN in NLP applications

As mentioned above, another point that deserves to be addressed is AWN *usability*. While the efficiency of other WNs (e.g., English and Spanish) in different NLP applications has been proven through several research efforts and experimental results (Kim et al., 2006; Wagner 2005), AWN was considered in just a few applications. In fact, AWN was only used and cited as:

- A comparative resource to evaluate a Web-based technique for building a lexicon from hypernymy relations with hierarchical structure for Arabic (Elghamry 2008);
- A resource for Query Expansion (El Amine 2009);
- A resource to be linked to the PanLex 2.5 which is a database that represents assertions about the meanings of expressions (Baldwin et al., 2010); [6]
- A source of information for building an Arabic lexicon by incorporating traditional works on Qur'anic vocabulary (Sharaf 2009);
- A promising resource that (i) allows the exploration of the impact of semantic features on the Arabic NER task (Benajiba et al., 2009a; 2009b) and (ii) improves the question analysis module in the Arabic QA system called QASAL (Brini et al., 2009a; Brini et al., 2009b).

To sum up, from a theoretical perspective AWN presents many advantages, including WN structure compliance, mapping to other ontologies and consideration of some Arabic specificities; nevertheless, its patent coverage weaknesses explain its use in just a few projects. Currently, world-wide interest in the development of WNs is increasing. For example, this is shown from the 2012 edition of the Global WordNet conference[7] that revealed around 55 projects related to new WN construction, existing WNs enrichment, WNs and lexical resources integration, WN applications and other WN efforts. The AWN project will have to keep up with such dynamism.

[6] http://utilika.org/info/panlex-db-design.pdf
[7] The conference has been held every two years since 2004. The Web site of the 2012 edition can be accessed from: http://lang.cs.tut.ac.jp/gwc2012/

As we showed in Chapter 3, a semantic QE process based on AWN could improve the passage recall as well as passage ranking in an Arabic QA system, though the resource is requested to be enriched and adapted. To achieve this goal, we follow a three-steps approach:

- Step 1 focuses on analyzing, from an experience-based perspective, the *usability* and identifying the shortcomings of the current AWN lexical database in the context of Arabic QA;
- Step 2 enriches the AWN *coverage* following the lines identified from the shortcomings raised in step 1;
- Step 3 uses both the standard and the enriched release of AWN as part of an ontology adapting its lexical design and combining its coverage with the semantic and syntactic information integrated in Arabic VerbNet. This step will allow the implementation of the semantic-reasoning based level (in order to improve passage ranking on top of the two first levels (i.e., keyword-based and structure-based).

Jointly, the three steps aim to explore different possibilities for extending and using AWN coverage in order to increase the usefulness of AWN for Arabic NLP in general, while satisfying the specific need to achieve the best performance possible for Arabic QA.

The next sections present the results obtained in the three-steps approach for the enrichment and adaptation of AWN.

4.3 Experiment-based analysis of AWN

In order to address the main lines to be followed in extending AWN *coverage* for promoting its *usability*, a detailed analysis of the AWN content is required. There is also a need to identify the gap between this content and what is required by NLP applications, such as Arabic QA, in terms of resource *coverage*.

The current experiment uses the keyword-based and structure-based levels of our approach that aims at improving passage recall and ranking. In Chapter 3, we were interested in the *usability* of AWN for Arabic QA systems. AWN help us to improve the quality of passage ranking. For each user question, the underlying process tries to retrieve passages from the Web most likely to contain the expected answer. Our process is mainly based on the AWN-based semantic QE process previously described and evaluated (the examples and experiments related to this process were presented in Chapter 3). In the current experiment, the process is applied to all question keywords. As raised in the examples and experiments of Chapter 3, the overall performance of the AWN-based approach will be impacted by two main factors: (i) non-coverage of question keywords by AWN, so that the QE process cannot be applied to the given question, and (ii) extraction of a limited number of related terms; let us recall that, from the example presented in Chapter 3, the passage ranking will provide better results if a higher number of related terms are injected in the DDN model.

In order to evaluate AWN in relation to these two factors, we analyzed 2,264 translated questions extracted from CLEF and TREC. The results obtained are given in Table 22. Note that the statistics of the last four rows of the table were manually calculated. The results presented in Table 22 show that we were able to apply the AWN-based QE process to only around 65% of the questions considered in that study—the remaining 35% contained keywords that were not covered by AWN—and that the keywords covered can be expanded by, on average, 4 corresponding synonyms from AWN.

Table 22 Analysis of the AWN coverage for the CLEF and the TREC questions

Measures	CLEF	TREC	Overall	%
No. Questions	764	1,500	2,264	-
No. Questions covered by AWN	612	858	1,470	64.93%
Avg. AWN word lemmas per question	3.65	4.26	4	-
No. Questions Not Covered (QNC) by AWN	152	642	794	35.07%
QNC with NE keywords	127	420	547	68.89%
QNC with Verb keywords	44	262	306	38.54%
QNC with Noun keywords	81	508	589	74.18%
QNC with Broken Plural keywords	0	18	18	2.27%

A more in-depth analysis of the results in Table 22 reveals that over 74% of the questions not covered by AWN contain noun word lemmas, around 69% include NEs and roughly 39% are composed of at least one verb. We can also notice that BP forms (the irregular form of plural) are present in over 2% of these questions (only 120 such forms exist in AWN: this represents around 1.71% of the well-known existing BP lists). For example, the TREC question "متى وقعت حرائق الرايخستاغ ؟" (When did the Reichstag fires happen?) is formulated with three keywords: the verb "وقع" (happen), the BP "حرائق" (fires) and the NE "الرايخستاغ" (Reichstag). Since none of these keywords exists in AWN, the question can not be extended using the QE process.

The experiment-based analysis displays the AWN weaknesses previously pointed out and highlights the need to expand its coverage. To extend AWN content, particular interest is attached to semi-automatic methods among the most commonly used by researchers when enriching WordNets. These methods help to avoid the limitations of: (i) the manual approach, which consumes time and effort and tends to result in low coverage resources; and (ii) the automatic approach, which raises the coverage at the expense of accuracy and confidence.

In the following sub sections, we propose two types of AWN extension: (i) *resource-based extension of NEs and verbs* using existing English resources, and (ii) *process-based extension of nouns* using a hyponymy pattern recognition process.

4.4 Semi-automatic enrichment of AWN

4.4.1 Resource-based enrichment

Diab (2004) already proposed a resource-based AWN extension by means of Arabic English parallel corpora and English WordNet. In this subsection, we also extend AWN on the basis of existing English resources. Rather than using parallel corpora in recovering the Arabic side, we have explored using the Google Translation tool which can provide good results when processing unique entries (NEs or verbs).

4.4.1.1 Named Entities Extension using the YAGO Ontology

Various research efforts have aimed at extending WordNets with NEs. Indeed, adding new NEs synsets to WN is of paramount importance in the field of NLP because it allows using this unique resource for NE recognition and other tasks. Toral et al. (2008) automatically extended PWN 2.1 with NEs using Wikipedia. NEs in Wikipedia are identified and integrated in a resource called Named Entity WordNet, after a mapping performed between the is-a hierarchy in PWN and the Wikipedia categories. Al Khalifa and Rodriguez (2009) also demonstrated that it is possible to enrich NEs in AWN by using the Arabic Wikipedia: in that work, experiments showed that 93.3% of automatically recovered NE synsets were correct. However, due to the small size of the Arabic Wikipedia, only 3,854 Arabic NEs could be added.

One way to tackle monolingual resource scarcity problems is to use available resources in one language to extend existing WordNet in another one, as was done by Sagot and Fiser (2008) for the French WN. In this direction, we have been interested in using the YAGO ontology[8] (Suchanek et al., 2007) for the following reasons:

- It covers a great amount of individuals (2 millions NEs);
- It has a near-human accuracy around 95%;
- It is built from WordNet and Wikipedia;
- It is connected with the SUMO ontology;
- It exists in many formats (XML, SQL, RDF, Notation 3, etc.) and it is available via tools[9] which facilitate exporting and querying it;
- Its usage avoids the challenges of performing NE identification as was done by Al Khalifa and Rodriguez (2009).

[8] Yet Another Great Ontology: available at http://www.mpi-inf.mpg.de/YAGO-naga/YAGO/downloads.html
[9] http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html

The YAGO ontology contains two types of information: *entities* and *facts*. The former are NE instances (from Wikipedia) and concepts (from WordNet), whereas the latter are facts which set a relation between these entities. The YAGO ontology was already used as a semantic resource in the context of IR systems (Pound et al., 2009).

In order to enrich the NE content in AWN, we perform an automatic mapping process as illustrated in Figure 17.

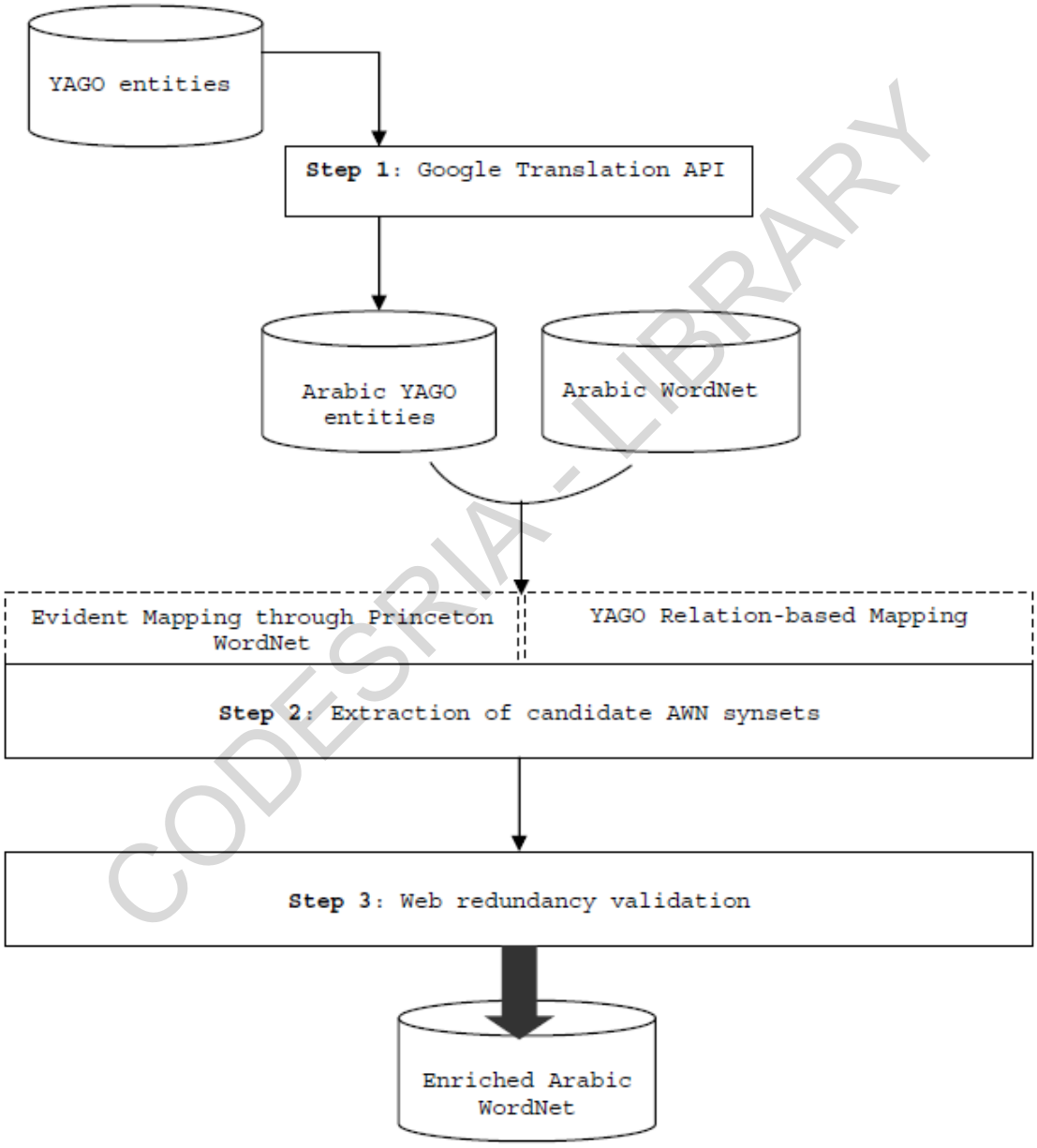


Figure 17. Automatic mapping process between YAGO NEs and AWN Synsets

The process is composed of three main steps that are descibed as follows:

- Step 1: translation of YAGO entities into Arabic instances by means of Google Translation API (GTA).[10] Based on the manual checking of 1,000 translated NEs, we have observed that this automatic translation has attained an accuracy of 98.2% when applied to a one or two-word NE.
- Step 2: extraction of candidate AWN synsets to be associated with the created instances. It was possible to add the translated YAGO entities to AWN through two kinds of mappings:
    - *Evident mapping*: the WordNet synsets corresponding to a given YAGO entity are extracted using the facts involving the YAGO "TYPE" relation (in YAGO, there are 16 million facts for this relation); the AWN synsets corresponding to the identified WordNet synsets are then connected with the given entity. For example, the YAGO entity "Abraham_Lincoln" appears in three facts for the YAGO "TYPE" relation; from these facts, the three English WN synsets "president", "lawyer" and "person" are extracted. Hence, the YAGO entity "ابراهام لينكولن" (i.e., Abraham Lincoln) can be added as an instance corresponding respectively to AWN synsets identified by "**رَئِيس**" (president), "**مُحَام، مُحَامِي، وَكِيل**" (lawyer, attorney) and "**شَخْص، إِنْسَان**" (person, human);
    - *YAGO relation-based mapping*: it consists in supposing that the arguments of some YAGO relations can be systematically added to AWN as instances of specific synsets. For example, the second argument of the YAGO relation "bornIn" is likely to be an instance of the AWN synset "مدينة" (city : identified by madiynap_n1AR in AWN). Following this idea, we have specified for a set of 19 YAGO relations (out of 99) whether the first or the second argument of the relation should be used and which AWN synset should be linked to it. Table 23 shows the mapping made for the 19 relations. Using this mapping, 331,851 candidate NEs have been extracted and passed on to the validation process.
- Step 3: consists of the automatic validation of the links between YAGO entities and corresponding AWN synsets (using both mapping types). This step aims at eliminating incorrect mappings as well as wrongly translated entities by means of Web redundancy. For instance, in YAGO facts, the entity "Association_for_Computing_Machinery" is present in the second argument of the relation "isLeaderOf". Therefore, with respect to the mapping listed in Table 23, this entity is a candidate for being an instance of the synset بلد (country : balad_n1AR). Using the Yahoo! API, we extract the Web snippets that strictly match the expression

[10] http://code.google.com/p/google-api-translate-java/

"بلد جمعية الآلات الحاسبة" (Association for Computing Machinery country). The given entity is then not added in the AWN extension under the synset بلد (country : balad_n1AR) since the number of extracted snippets does not exceed a specific threshold (set heuristically to 100), meaning that the given candidate NE is most likely not an instance of the considered synset. Unlike step 1, we use the Yahoo! API instead of the Google API in order not to have a biased validation of NEs (translated in one API and validated by another one). After applying this validation step, we were able to eliminate over 13% of the candidate mappings. Table 24 gives the detailed percentage of eliminated entities per YAGO relation.

Table 23 Mapping between YAGO relation and AWN synsets

YAGO relation	AWN synset	AWN synset id
actedIn	إبداع (creation : AibodaAE)	ibodaAE_n1AR
bornIn	مدينة (city : mdynp)	madiynap_n1AR
diedIn	مدينة (city : mdynp)	madiynap_n1AR
hasCapital	مدينة (city : mdynp)	madiynap_n1AR
hasCurrency	بلد (country : balad)	balad_n1AR
hasNumberOfPeople	بلد (country : balad)	balad_n1AR
hasPopulation	بلد (country : balad)	balad_n1AR
hasPopulationDensity	بلد (country : balad)	balad_n1AR
hasUnemployment	بلد (country : balad)	balad_n1AR
inTimeZone	منطقة-مقاطعة (region : mnTqp)	minoTaqap_n1AR
isCitizenOf	مدينة (city : mdynp)	madiynap_n1AR
isLeaderOf	بلد (country : balad)	balad_n1AR
isMarriedTo	زوج-زوجة (married : zwj)	zawoj_n1AR
livesIn	مدينة (city : mdynp)	madiynap_n1AR
locatedIn	مدينة (city : mdynp)	madiynap_n1AR
originatesFrom	منطقة (region : mnTqp)	minoTaqap_n1AR
politicianOf	بلد (country : balad)	balad_n1AR
worksAt	مُؤَسَّسَة (institution/establishment: u&as~asap)	mu&as~asap_n1AR
wrote	كاتب (writer/author : kAtb)	kaAtib_n1AR

Table 24. YAGO and AWN evident mapping statistics

YAGO relation	# entities	Eliminated entities
actedIn	28,836	35.09%
bornIn	36,189	20.59%
diedIn	13,618	12.92%
hasCapital	1,368	6.78%
hasCurrency	367	0.00%
hasNumberOfPeople	6,171	0.00%
hasPopulation	77,928	9.78%
hasPopulationDensity	44,628	0.00%
hasUnemployment	41	0.00%

inTimeZone	2	0.00%
isCitizenOf	4,865	0.00%
isLeaderOf	2,886	0.00%
isMarriedTo	8,416	0.00%
livesIn	14,710	11.11%
locatedIn	60,261	14.03%
originatesFrom	11,497	26.67%
politicianOf	6,198	0.00%
worksAt	1,401	3.45%
wrote	12,469	27.27%
Total	**318,612**	**13,24%**

Four YAGO relations, namely “actedIn”, “wrote”, “originatesFrom” and “bornIn” cover the major part of the eliminated entities (35.09%, 27.27%, 26.67% and 20.59% of the entities candidate in each relation were eliminated respectively). For example, in the case of the relation “wrote”, many cases are due to translation errors (the automatic translation is not effective when it processes long titles of books and stories). Another example of these eliminated entities is the names of countries such as Morocco linked to the synset مدينة (city : mdynp) using the YAGO relation-based mapping for the “bornIn” or “diedIn” relation.

The three-step process described was performed for three million YAGO entities. We found out that it was possible to keep 433,339 instances (145,135 NEs thanks to the first mapping in Step 2 and 288,204 NEs from the second mapping) that were connected with 2,366 corresponding AWN synsets. Let us recall that in the original AWN release, there are 1,067 synsets having instances (i.e., NEs). The new numbers represent an increase of nearly 205%. Also, the high number of instances allows an acceptable coverage of real-world NEs. These instances belong to different categories as listed in Table 25.

Table 25 Statistics of NE classes augmented in AWN

Cat. ID	NE categories	Number	%
1	PERSON	163,534	37.7%
2	LOCATION	73,342	16.9%
3	EVENT	14,258	3.3%
4	PRODUCT	14,148	3.3%
5	NATURAL OBJECT	8,512	2,0%
6	ORGANIZATION	8,371	1.9%
7	FACILITY	4,312	1,0%
8	UNIT	3,513	0.8%
	Sub Total	**289,990**	**66.9%**
9	OTHER	143,348	33.1%
	Total	**433,339**	**100%**

The major part (66.9%) of NEs that were linked to AWN synsets can be classified under 8 categories. The most frequent ones are PERSON (37.7%) and LOCATION (16.9%). The remaining NEs (33.1%) are grouped under the “OTHER” category.

Most of the added PERSON entities are foreign names; however, this will not impact the experimental process (reconducted and presented later in this chapter) using CLEF and TREC questions containing the same nature of names. Also, we did not investigate using an Arabic NER system as alternative to the resource-based approach in order to avoid any eventual inaccuracy of such a system.

The feasibility of enriching AWN coverage by NEs coming from YAGO was investigated. Nevertheless, we understand that building an Arabic YAGO linked to the English one could presumably be the most suitable option for dynamic information such as NEs (rather than adding these NEs directly in AWN). The interesting amount of NEs that we have linked to AWN synsets will at least help in considering their mapping to already existing PWN NEs and also to deal with issues related to irregular spelling of Arabic NEs.

4.4.1.2 Extension using VerbNet and Unified Verb Index

Rodriguez et al. (2008a) have investigated two possible approaches for extending AWN. In both cases, the purpose was just to show the potential usefulness of such approaches for semi-automatic extension of the resource. In both works, it was reported that the results were very encouraging, especially when compared with the results of applying the eight EuroWordNet heuristics (Vossen 1998). However, further experiments are needed in order to add a number of words to AWN synsets. The first approach deals with lexical and morphological rules, while the second considers Bayesian Networks as an inferencing mechanism for scoring the set of candidate associations (Rodriguez et al., 2008b). The Bayesian Network (BN) doubles the number of candidates of the previous heuristics approach (554 candidate words using BN versus 272).

In our own work, in order to enrich the verb content in AWN, we have followed a two-step approach inspired by what was proposed by Rodriguez et al. (2008a). The first step consists in proposing new verbs to add to AWN; the second step aims at attaching these newly proposed verbs to corresponding AWN synsets.

Considering the first step, while Rodriguez and his colleagues made use of a very limited but highly productive set of lexical rules in order to produce regular verbal derivative forms, we obtained these forms by translating the current content of VerbNet (Kipper-Schuler 2006) into the Arabic language. Our reasons are two-fold:

(i) To avoid the validation step where we need to filter the noise caused by overgeneration of derivative verb forms (unused forms can be generated);

(ii) To allow advanced AWN-based NLP applications to use the syntactic and semantic information about verb classes in VerbNet and their mappings to other resources such as FrameNet (Baker et al., 2003) and PropBank (Palmer et al., 2005).

The translation concerns the 4,826 VerbNet[11] verbs distributed into 313 classes and sub classes. After the process of translating every single verb using the Google Translation Web page (note that, unlike the Google Translation API, this translation Web page can provide more than one possible translation for a unique verb entry), a manual validation was performed to check the correctness of the translation, as well as to select the verb lemmas to be added to AWN. Thanks to this semi-automatic process, we were able to obtain 6,654 verbs for the next step. The same process was applied on verbs covered by the Unified Verb Index (UVI).[12]

In the second step, the attachment of Arabic verbs with AWN synsets was done by constructing a graph which connects each Arabic verb with the corresponding English verbs that are present in PWN. Figure 18 illustrates this step: A stands for the Arabic verb, E_j for the English verb number j, S_i for PWN synset number i and S_{ai} for AWN synset number i.

As Figure 18 shows, each English verb can be connected to different PWN synsets. Then they are connected with their equivalent synsets in AWN. After building the graph connecting each Arabic verb with the corresponding PWN synsets through English verbs, the relevant connections were selected by applying 3 of the 5 graph heuristics adopted in (Rodriguez et al., 2008a). We set the limit at the third heuristic because the percentage of noise attachment increases starting from the fourth heuristic and even more after applying the fifth one.

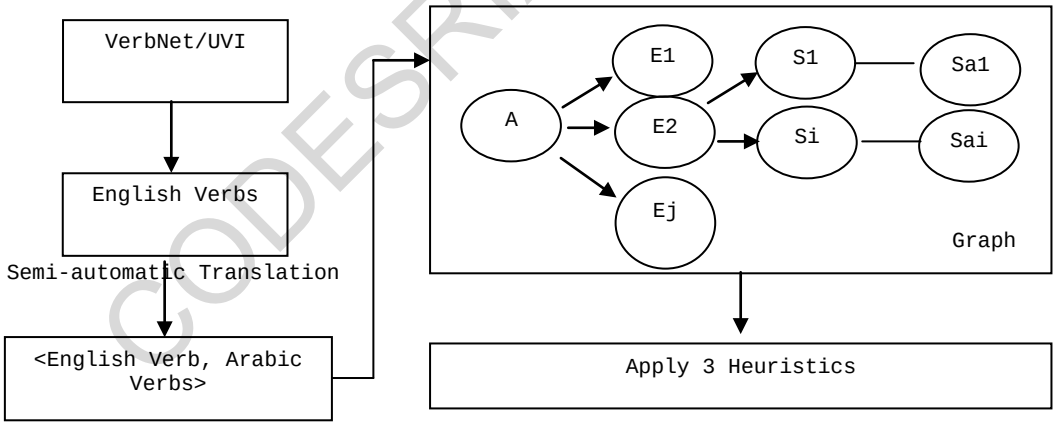


Figure 18**.** Enrichment of verbs in AWN and their attachment to synsets

Let us recall the definition of each heuristic as described in that work:

[11] VerbNet is a lexicon classifying verbs into classes with descriptions of these classes in terms of members, syntactic and semantic frames, etc. This lexicon is described with more details in chapter 5

[12] The Unified Verb Index is a system which merges links and Web pages from four different natural language processing projects: VerbNet, PropBank, FrameNet and OntoNotes Sense Groupings (http://verbs.colorado.edu/verb-index/)

- Heuristic 1: If a unique path Arabic-English-Synset (AES) exists (i.e., A is only translated as E), and E is monosemous (i.e., it is associated with a single synset), then the output tuple <A, S> is tagged as 1;
- Heuristic 2: If multiple paths AE1S and AE2S exist (i.e., A is translated as E_1 or E_2 and both E_1 and E_2 are associated with S among other possible associations) then the output tuple <A,S> is tagged as 2;
- Heuristic 3: If S in AES has a semantic relation to one or more synsets, S_1, S_2 … that have already been associated with an Arabic word on the basis of either Heuristic 1 or Heuristic 2, then the output tuple <A, S> is tagged as 3;
- Heuristic 4: If S in AES has some semantic relation with S_1, S_2 … where S_1, S_2 … belong to the set of synsets that have already been associated with related Arabic words, then the output tuple <A, S> is tagged as 4;
- Heuristic 5: Heuristic 5 is the same as Heuristic 4 except that there are multiple translations E_1, E_2, … of A and, for each translation Ei there are possibly multiple associated synsets S_{i1}, S_{i2}, …. In this case the output tuple <A, S> is tagged as 5.

Note that tags 1, 2 and 3 help in identifying the <A, S> tuple generated by the first, second and third heuristic respectively. Table 26 presents the results obtained using the described verb extension process.

Table 26. Results of the AWN verb extension process

	VerbNet		UVI		Total
	Number	Percentage	Number	Percentage	
Considered Arabic verbs	6,654	-	3,431	-	10,085
Connected Arabic verbs	5,329	80.09%	1,115	31,13%	6,444
Verbs existing in AWN	2,760	41.48%	542	15,80%	3,302
Newly Added Verbs (NAV)	2,569	38.61%	573	16,70%	3,142
- NAV with Heuristic 1	184	2,77%	129	3,76%	313
- NAV with Heuristic 2	158	2,37%	43	1,25%	201
- NAV with Heuristic 3	2,227	33,47%	401	11,69%	2,628
Connected AWN synsets	1,361	-	1,906	-	3,267

We succeeded in connecting 5,329 of the Arabic verbs translated from VerbNet with the corresponding AWN synsets (1,361 distinct synsets). Even though around 41.5% of these verbs (2,760 verbs) already existed in the current release of AWN, the process added new synset attachments for them. The remaining 2,569 verbs were not in AWN and could be added. Heuristic 1 allowed the generation of a few but accurate verbs and attachments (2.77%), whereas Heuristic 3 succeeded in coming up with a higher number of less relevant verbs (33.47%). With respect to the verbs generated from UVI, the overall newly connected verbs were 6,444, 3,142 of which were new additions.

4.4.2 Process-based enrichment

4.4.2.1 Background

Relying on resource-based extension is not the only line of investigation for enriching WordNets. Process-based semi-automatic techniques have also been adopted by researchers in order to refine the hyponymy relation in WordNets, as well as to add new noun and verb synsets (Hearst 1992; Costa and Seco 2008; Tjong Kim Sang and Hofmann 2007). Hyponymy discovery is another useful direction for WordNet enrichment that allows the automatic extraction of hyponym/hypernym pairs from text resources such as the Web. For instance, A and B form a hyponym/hypernym pair if the meaning of B covers the meaning of A and is broader (Tjong Kim Sang and Hofmann 2007). There have been many attempts with the aim of the automatic acquisition of such hyponymy pairs. Hearst (1992) was among the first researchers to have proposed and investigated a pattern-based approach in order to resolve this problem. This approach consists mainly in using a set of lexical and syntactic patterns to generate a list of concepts linked using the considered semantic relation. For instance, in English, the pattern "X including Y_1 (, Y_2, ..., and |or Yn)" helps to identify the nouns Y_1, ..., Y_n as candidate hyponyms of the noun X. For example, "cinema" and "drawing" can be extracted as hyponyms of "arts" from the text "The institute focuses on different arts including cinema and drawing". It was reported that adopting these kinds of pattern-based approaches allows the harvesting of semantic relations in general and hyponymy particularly in languages such as English (Pantel et al., 2006; Snow et al., 2005), Spanish (Ortega-Mendoza et al., 2007) and Dutch (Tjong Kim Sang and Hofmann 2007).

As for Arabic, there have been few such attempts in comparison to other languages like English. The work of Elghamry (2008), which proposed an unsupervised method to create a corpus-based hypernym/hyponym lexicon with partial hierarchical structure, is one of these few attempts. In that work, the acquisition process was bootstrapped relying on the lexico-syntactic pattern "بعض X مثل Y1…Yn" (some X such as Y1,…Yn). The effectiveness of the suggested method was demonstrated through a comparison between the extracted entries with those of AWN, but a single lexico-syntactic pattern ("بعض X مثل Y1…Yn") was used. This limitation had two causes: (i) it was reported that Arabic patterns which are equivalent to those proposed in (Hearst 1992) do not give significant results and (ii) there was no Arabic parser available to facilitate the detection of noun phrases in the context of the other patterns. With the availability of Open Source Arabic syntactic parsers like the Stanford Arabic Parser,[13] the latter reason is no longer valid: such syntactic parsers can reduce the noise generated by a long list of Arabic lexico-syntactic patterns.

[13] http://nlp.stanford.edu/software/lex-parser.shtml

4.4.2.2 Enriching hypernymy relation in AWN

In line with the above-mentioned research efforts for Arabic and other languages, our aim is to augment the coverage of AWN noun synsets (currently there are 7,162 noun synsets versus 82,115 in the English WN) while simultaneously enriching the hyponymy (is-a) relation between these synsets. The two-step method proposed by Ortega-Mendoza et al. (2007) and García-Blasco et al. (2010) was adapted to achieve the target enrichment. Figure 19 illustrates the general architecture of our approach.

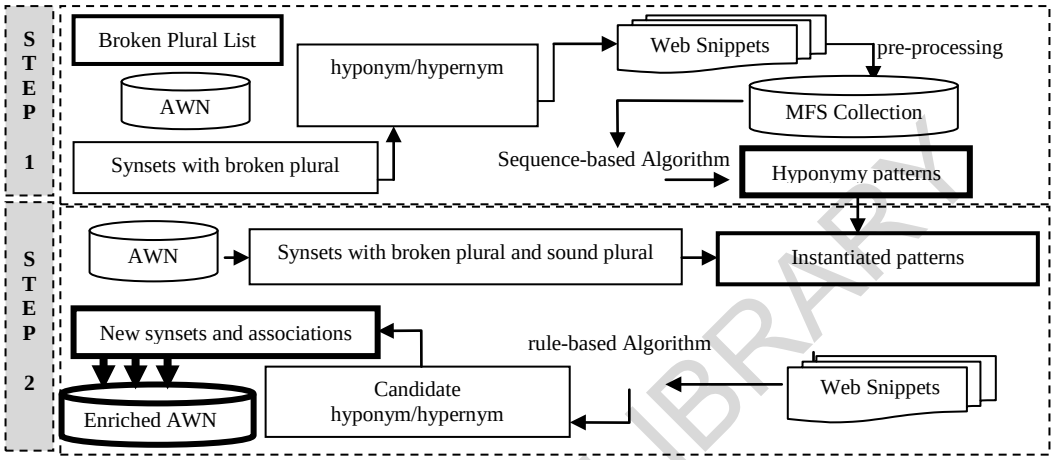


Figure 19 General architecture for Arabic Hyponym/Hypernym pairs detection

Figure 19 depicts the two-step method. It can be summarized as follows:

• Step 1: It identifies hyponymy patterns over snippets retrieved from the Web. These snippets match a set of queries formed by hypernym/hyponym pairs;

• Step 2: It instantiates the identified patterns. The instantiation is performed by searching for hypernym/hyponym pairs that match the given pattern.

The following sub sections explain more in detail the two steps illustrated in the previous figure, presenting how these steps are implemented for the Arabic language and highlighting the main results obtained after this implementation.

A) Identifying lexico-syntactic patterns

According to Ortega-Mendoza et al. (2007), we need a seed list of hypernym/hyponym pairs to be used as queries. In our case, we have built this list from the synsets existing in AWN. For instance, the synset (fan~ / art) فنّ is described by the following synonyms: (<inotaAj_fan~iy : artistic production) إِنْتاج فنّيٍ, (AibodaAE_fan~iy : artistic innovation) اِبْداع فنِّي and (fan~ / art) فنّ. Figure 20 shows the context of this synset in the AWN hierarchy using the hyponymy relation.

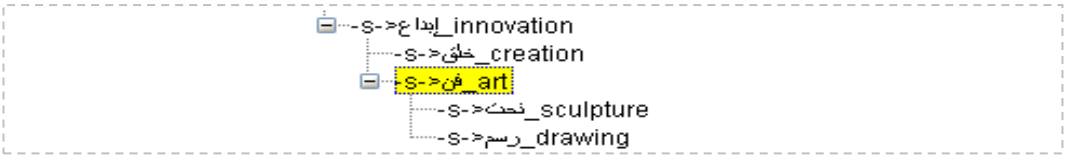


Figure 20. Context of the synset fan~ in the hierarchy of AWN

Only two hyponyms of the synset فنّ (fan~ : art) are present in the current version of AWN, namely “sculpture” and “drawing”. In the English WordNet 3.0, 13 hyponyms (gastronomy, perfumery, origami, etc.) exist under the equivalent synset (art).

To know about how this synset appears together with its hyponyms in a text, we have queried the Web with a set of hand-coded hyponymy patterns instantiated using the given synset and its hyponyms. Table 27 shows the used queries and sample snippets obtained as results.

Table 27 Sample snippets obtained using instantiated patterns as queries

Instantiated pattern (in Arabic)	Instantiated pattern (in English)	Sample of obtained snippets (in Arabic)	Sample of obtained snippets (in English)
نحت و غير ذلك من **فن**	**sculpture** and other **arts**	فن النحت من أقدم الفنون وأكثرها انتشارًا وتنوعًا في العالم...	Sculpture is one of the oldest arts, the most widespread and diverse in the world...
فن الأخرى خاصة **نحت**	other **art** in particular **sculpture**	الفنون عامة وفن النحت خاصة يعتبر من أهم المجالات التي تعكس بصدق بالغ تفاعلات...	Generally, the arts and in particular sculpture, are one of the most important areas that truly reflect deep interactions...
فن الأخرى على غرار **نحت**	other **arts** such as **sculpture**	قواعد الفن على غرار الفن الاغريقى أو الرومانى...	The rules of art such as Greek or Roman arts...
من أهم هذه **فن** هناك **رسم**	**drawing** is one of the most important **arts**	هناك تقنيات مختلفة للفنون التشكيلية والرسم التي تجعل الاختلافات في الرسم سواء...	There are different techniques of Fine Arts and painting that make the differences ...

From the above example, the hypernym is usually used in its plural form which can be generated by adding specific suffixes (for instance –arts- فنون is the sound plural of فن –art-). This is similar to other languages such as English. According to some research on large Arabic corpora (Goweder and De Roeck 2001; Boudelaa and Gaskell 2002), BP forms constitute around 10% of texts, and BP forms account for 41% of the different plural forms used in texts. Therefore, we used BP forms to automatically extract patterns and built a list of seed hypernym/hyponym pairs starting from the AWN synsets which have a BP form.

Since the current version of AWN contains only a few BP forms, we decided to begin enriching AWN by connecting its synsets and words with such new forms. To perform this

task we relied on 3,000 Arabic BP forms extracted from Emad Mohamed's list[14] that we automatically connected these forms to the corresponding AWN words using the singular entry existing in that list. The content of the list as well as the connections so-created were manually validated. In all, we connected 1,934 synsets with the corresponding BP form (nearly 24.3% of the AWN noun synsets), using 1,696 hypernym/hyponym pairs to identify lexical patterns (the other synsets do not appear in relevant number of snippets). A description of the procedure used is oulined below.

For each seed pair, we extract from the Web the first 20 distinct snippets corresponding to the results returned by the Yahoo! API when using the following request forms: "HYPONYM+HYPERNYM" and "HYPERNYM+HYPONYM". The next challenge was to retrieve the relevant lexical patterns from the previously mentioned collection of snippets. Currently, different techniques are suitable for such a task. One of these techniques is based on the retrieval of the Maximal Frequent Sequences (MFS) of words. In fact, many research works (Denicia-Carrel et al., 2006; Ortega-Mendoza et al., 2007; García-Blasco et al., 2010; García-Hernández et al., 2010) highlighted the usefulness of this technique for pattern discovery over text.

Following Ahonen-Myka (2002), a sequence is defined as a set of ordered elements (for instance, words). The frequency of a sequence of words is determined by the number of sentences that contain this sequence. A sequence is maximal if it is not a subsequence of any other. That is, if it does not appear in any other sequence in the same order. MFS are all the sequences that appear in β sentences (where β is the defined frequency threshold) and are not subsequences of any other MFS. To make these maximal frequent sequences more flexible, García-Hernández (2007) has introduced the concept of gap which is defined as the maximum distance that is allowed between two words in a MFS. Following this, if we set the gap to 0, the words in the MFS will be adjacent words in the original text. For example, $<w_{i0}, \ldots, w_{in}>$, with $i_j \in 1 \ldots k$, is a maximal frequent sequence of k words, $i_j = i_{j-1}+1$, $j > 1$, when $gap = 0$, and $i_j \leq i_{j-1+\eta} + 1$, when $gap = \eta$.

In our work, we adopted MFS for two main reasons: (i) it has achieved a higher performance for languages such as English and Spanish (Denicia-Carrel et al., 2006; Ortega-Mendoza et al., 2007; García-Blasco et al., 2010; García-Hernández et al., 2010), and (ii) it is language-independent, which allows us to leverage for Arabic tools that have been developed for the aforementioned languages.

Specifically, we used the MFS-algorithm proposed by García-Blasco et al. (2010). It allows the processing of a document collection (that must be just plain text, divided into lines) and searches for the MFS on the basis of three parameters introduced before running it:

[14] http://jones.ling.indiana.edu/~emadnawfal/arabicPlural.txt

- Minimal Frequency (MF): It is the minimum number of times the sequence must appear. If a sequence appears twice in the same sentence, it will only count as 1 for the frequency;
- Minimal Length (ML): It is the minimum number of words that must compose the sequence;
- Maximal Gap (MG): It is the maximum distance allowed between two consecutive words in the maximal frequent sequence. The greater this value is, the more flexible the extracted patterns will be.

Extracting a high number of hyponymy patterns depends on the coverage of the document collection used. In this work, we built a collection from 102,900 snippets corresponding to 1,696 Web queries (a query is formed from AWN hyponym/hypernym pairs). In order to guarantee the correctness of the extracted patterns, we manually evaluated the patterns that resulted from applying the MFS-algorithm on a small subset of the collection (5,145 snippets, which represent 5% of the collection). We used different parameter values while considering the following constraints: (i) since a `MF>20` only generates 2 candidate patterns and a `MF<5` generates an excessive number of patterns, we considered a range between 5 and 20 for this parameter, (ii) according to the lengths observed in a manually built list of hyponymy patterns, a range between 3 and 7 was set for MG. Table 28 shows the results of the MFS-algorithm on the small subset of the collection.

As we can see, when the parameters are `MF=20, ML=2 and MG=7`, the algorithm (which is applied on the small subset of the collection) is able to generate 27 candidate patterns of which 5 patterns (18.52%) are manually qualified as correct hyponymy patterns. This percentage is the highest among the different runs corresponding to the different MFS parameters values.

Table 28 Results of MFS parameter setting in the context of the Arabic language

	Run #1	Run #2	Run #3	Run #4	Run #5	Run #6
Minimal Frequency (MF)	20	20	20	15	10	5
Maximal GAP (MG)	3	5	7	7	7	7
Minimal Length (ML)	2	2	2	2	2	2
#Patterns	19	26	27	46	113	1,019
#Hyponymy Patterns	2	3	5	7	17	135
%Hyponymy Patterns	10.53%	11.54%	18.52%	15.22%	15.04%	13.25%

To apply the MFS-algorithm on the whole collection, it makes sense to maintain the same ML and MG parameters values, as they are collection-coverage independent. However, the MF has to be changed to 400. Indeed, unlike ML and MG, the MF depends on the collection coverage and in our case MF is calculated accordingly (MF=102,900*20/5,145). With these parameter values, we succeeded in extracting 23 relevant hyponym patterns from the whole

snippet collection. These patterns, after manual validation, were used in the pattern instantiation step (Step 2).

B) Instantiating Patterns

The main objective of the pattern instantiation step is to retrieve candidate hyponym/hypernym pairs with which to enrich the current AWN hierarchy. Generally, a pattern has one of the two following forms: "<Phrase> HYPONYM <Phrase> HYPERNYM" or "HYPERNYM <Phrase> HYPONYM <Phrase>". Instantiating these patterns means that we replace the HYPERNYM part by the synset names from AWN and the other parts by a wild character (such as *). For instance, the pattern " العديد من HYPR مثل HYPO" (many HYPR such as HYPO) is instantiated with the synset الأسلحة (Al>slHp : weapons) which is the BP of سلاح (silAH : weapon). The query resulting from this instantiation is: " العديد من مثل الأسلحة *". This query is passed on to the search engine (i.e., the Yahoo! API) in order to retrieve the most relevant and matching snippets. Table 29 lists samples of the extracted snippets.

Table 29 Sample snippets obtained using the pattern " العديد من HYPR مثل HYPO"

Snippets (in Arabic)	Translation (in English)
وله **العديد من** الأسلحة **مثل**: العصا, السيف... أحد الاسباب التي حدت من انتشار الفن هو التحفظ من المعلمين واختيار التلاميذ بحذر حتى لا تنتقل الاسرار للمنافسين...	...have **many** weapons **such as** stick, sword ...
أي معلومات غير موثقة يمكن التشكيك بها وإزالتها. وسم هذا القالب منذ: نوفمبر 2010 ... 1957 حيث تم من خلال تصميمة تطوير وإنتاج **العديد من** الأسلحة **مثل** إم 240...	... developing and producing **many** weapons **such as** ***M240*** ...
تستخدم بعض الأسلحة الكيماوية الغير قاتلة، مثل *الغاز المسيل للدموع ورذاذ* ... فإن **العديد من** الحروب هي جزئيا أو كليا مستندة إلى أسباب اقتصادية، **مثل** الأزمة...	... several chemical weapons such as *tear gaz* ... **many** wars are completely or partially triggered by economic causes, **such as** *crisis* …
لعام 1939-1945م تجد في اللعبة **العديد من** الاسلحة **مثل** *الدبابات والصواريخ ومدفع الهاون والكثير من* الاسلحة لعبة مثيرة واكشن جداً جداً...	... you'll find in this game **many** weapons **such as** big *tanks, rockets and mortars* and lot of other weapons ...
هناك **العديد من** الأسلحة **مثل**: *السيوف والخناجر والفأس والقوس والسهم والمسدس وتو بليد وألخ*..... آعذروني لعدم وجود صور للاسلحه آستعراض بسيط للاسطورهـ...	... There are **many** weapons **such as** *swords, daggers, ax, bow and arrow, pistol* ...
لم تكن تستطيع الوصول إليه من قبل, القتال سيكون باستخدام ... **العديد من** الأسلحة **مثل** *السوط والسيف والخناجر والفؤوس والسحر* والكثير من الأسلحة الأخرى المتنوعة...	...using **many** weapons **such as** the *whip and the sword, daggers, axes and magic* ...

The words of the pattern are in bold, the synset used for its instantiation is underlined while the candidate hyponyms are both underlined and in italic. As we can see, in the above

example, the left side of the pattern contains the targeted hyponyms. Therefore, a rule-based algorithm was applied in order to analyze the left side and extract from it nouns that could be added as hyponyms of the synset الأسلحة (Al>slHp : weapons).

The list of the 23 hyponymy patterns identified in the previous step was instantiated using both 700 AWN synsets (hypernyms) that have BP forms and then using 700 other AWN synsets with their Sound Plural (SP)[15] form. Let us recall that only BP forms have been used as seed pairs of the hyponymy relation while we used both forms in the instantiation phase. This should allow us to determine whether the patterns discovered using a plural form (in our case BP) can be useful in identifying hyponyms for the other form (i.e., SP). Table 30 presents the results obtained.

Table 30 Experimental results of the AWN noun hyponymy extension

	Using BP	Using SP	Overall/Total (distinct)
#AWN hypernym synsets	700	700	1,400
#Successful patterns	17 (73.91%)	9 (39.13%)	17 (73.91%)
#Candidate hyponyms	1,426	828	2,254
Avg. candidate hyponyms per AWN synset	2.04	1.22	1.61
#Correct hyponyms	458 (32.12%)	415 (50.12%)	832 (36.91%)
#AWN hypernym synset with correct hyponyms	94 (13.43%)	191 (27.29%)	284 (40.57%)
#New correct hyponyms (not existing in AWN)	265 (57.86%)	205 (49.40%)	459 (55.17%)
#New AWN associations(hypernym/hyponyms)	193	196	359

As depicted in Table 30, instantiating the 23 patterns with BP forms opens up the possibility of getting an average of around two candidate hyponyms per AWN hypernym synset (versus 1.22 using the sound plural form). Note that candidate hyponyms are extracted using a set of automatic rules. These candidate hyponyms are then manually validated in order to identify correct hyponyms (Two persons validated around 2,300 hyponyms within approximately two days). With regard to BP forms, around 74% of the patterns considered succeeded in generating correct hyponyms. The list of these patterns also includes all the patterns that succeeded with SP forms (9 patterns). The difference in pattern accuracy can be explained by the following fact: when using the SP form in the query, snippets often contain the singular instead of the plural stem. Therefore, such snippets will not be relevant and hardly match the pattern considered. This confusion does not occur with the BP having a different pattern and stem.

[15] A Sound Plural (SP) is formed by adding a suffix without changing the pattern of consonants and vowels inside the singular form as in the case of Broken Plural (BP)

The results listed in Table 30 also show that 832 correct hyponyms were identified (roughly 37% of the candidate hyponyms). About 60% of these could be added to AWN as new synsets. Even though the remaining hyponyms already existed in AWN, new hypernym/hyponym associations in which they participate could still be added.

According to Table 30, our process succeeded in generating hyponyms for approximately 41% of the 1,400 hypernym synsets considered. The number of hyponyms per hypernym ranges from 1 to 29. Figure 21 illustrates the distribution of the number of hyponyms per hypernym.

Figure 21 Distribution of the number of hyponyms per hypernym

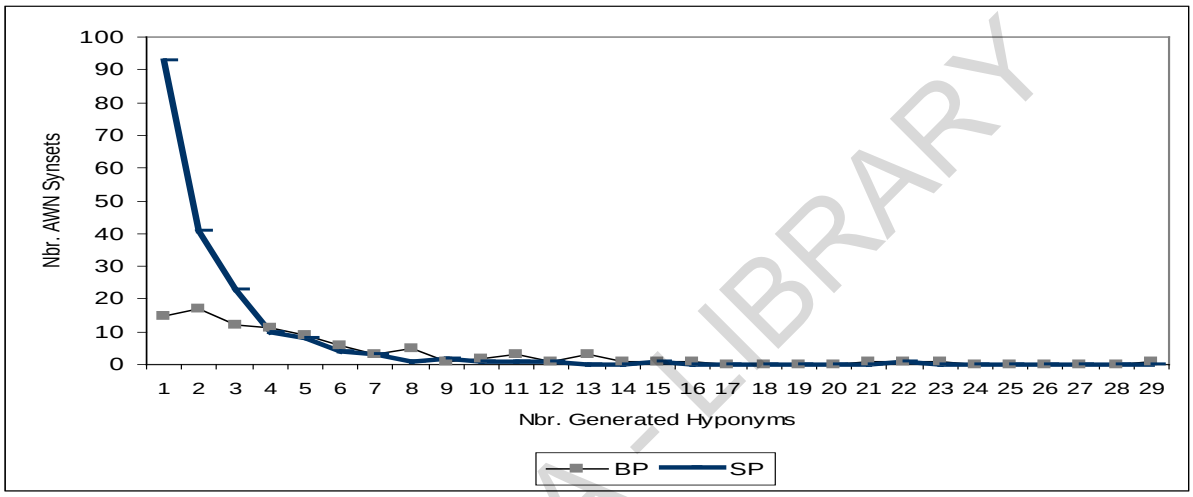


Figure 21 contains two curves, corresponding to BP and SP hyponym generation respectively. The first curve reveals that with the BP form, for instance, only one hyponym is extracted for 15 AWN hypernym synsets. While Table 30 shows that SP forms help in generating correct hyponyms for a higher number of AWN synsets (191 vs 94 with BP forms), Figure 21 depicts an unbalanced distribution of these hyponyms over these synsets. In fact, for around 54% of the BP forms the process succeeded in generating at least 4 correct hyponyms, whereas this percentage did not exceed 17.5% for SP forms that can be confused with singular nouns embedding the suffix of SP, i.e., "ات". To sum up, using both forms as hypernyms guarantees that more AWN synsets will acquire hyponyms, but not with the same accuracy. Table 31 lists the patterns that generate a high average of hyponyms per synset.

Table 31 Top relevant hyponymy patterns

Pattern	English translation	Avg. hyponyms per synset
HYPO مثل HYPR من العديد	Many HYPR such as HYPO	1.32
HYPO ك HYPR من العديد	Many HYPR for instance HYPO	1.30
HYPO مثل HYPR بعض	Some HYPR such as HYPO	1.13
HYPO مثل الأخرى HYPR	Other HYPR such as HYPO	1.10
HYPO ك الأخرى HYPR	Other HYPR for instance HYPO	0.89
HYPR من ذلك وغير HYPO	HYPO and other HYPR	0.88

The best hyponym patterns contain the hypernym part in the middle or at the beginning. The experimental results show that we have fulfilled our aim, i.e. to enrich the noun content and hierarchy of the AWN. Indeed, thanks to the use of a set of automatically discovered patterns (via an MFS-based algorithm), it was possible to add 459 new synsets (which account for 7.53% of the number of existing noun synsets) and 359 new associations between synsets using the hyponymy relation (around 2% of the existing associations).

The proposed technique is promising since it allows suggesting candidate hyponyms that can be validated and integrated under AWN synsets. In principle, this way is faster than adding these hyponyms from scratch, especially if we consider the following further possibilities:

- Extracting new patterns by setting other values for MFS parameters (these patterns can help in generating new hyponyms);
- Using a recursive process in which generated hyponyms play the role of hypernyms.

Since the technique is relation-independent, it could also be used for enriching AWN by adding new relations between synsets such as the meronymy (part of) relation.

4.4.3 Extension coverage

As described above, it is possible to semi-automatically extend the content of NEs, verbs and nouns in AWN. For each case, we made use of an adapted existing approach and/or resources developed for other languages. Thanks to this extension process, we obtained the results summarized in Table 32 and Table 33.

Table 32 Nouns, verbs and NEs *Coverage* improvement

Figures	Common Linguistic Categories			Dynamic Information		
	Nouns and Verbs			Named Entities		
	Original	Extended	Added	Original	Extended	Added
No. AWN Synsets	9,698	10,198	5.2%	1,155	2,366	205%
No. AWN Word-senses	18,925	37,463	98.0%	1,426	433,339	30,288%
No. AWN Distinct Lemmas	11,634	15,005	29.0%	1,426	433,339	30,288%
No. Baseline Lexicon Lemmas (BLL)	119,693	-	-	11,403	-	-
Percentage of AWN Lemmas/BLL	9.7%	12.5%	2.8%	12.5%	3,800%	3,788%

Table 33 BP Coverage improvement

Figures	Arabic specific characteristic		
	Broken Plurals		
	Original	Extended	Added
No. AWN Synsets	126	1,934	1,435%
No. AWN Word-senses	405	2,682	562.2%
No. AWN Distinct Lemmas	120	1,395	1,062%
No. Baseline Lexicon Lemmas (BLL)	9,565	-	-
Percentage AWN Lemmas/BLL	1.3%	14.6%	13.3%

The above results show not only the usefulness of the different AWN extension techniques, but also the significance and the extent of the new content. The most successful outcomes were the addition of the equivalent of roughly 38 thousand times the original number of NE (the number of AWN synsets having instances increased by 205%), as well as the large number of new noun and verb lemmas (15,005 vs. 11,634 in the original version) and new BP forms (1,395 vs. 120 in the original version).

A low coverage improvement was registered for synsets extension (+5.2%). This low increment can be justified as follows: (i) the process used for the automatic extraction of hyponyms was not recursively applied in the current work. Indeed, the hyponyms identified by this process could be used as hypernyms on which we apply the same process again to extract new hyponyms; (ii) the number of extracted snippets was limited to 20 and served as a text collection from which new hyponyms were extracted. Considering a higher number of snippets could increase the number of candidate hyponyms and, therefore, new AWN candidate synsets too. Note that the technique is quite similar to the one used by Snow et al. (2005) where AWN entries have been extended with hyponyms on the type level. However, this approach does not consider all possible senses for a word type.

With respect to the statistics of the proposed AWN release, the previously highlighted gap (see Table 20 in Section 4.2.1) relative to the Arabic lexicon (e.g. DIINAR.1) and other WNs considered is now reduced. Table 34 shows the new comparison.

Table 34 Comparison of the extended release of AWN with the English WN 3.0 and the Spanish WN

	Arabic		Spanish	English
	Original	Extended		
WN Synsets	9,698	10,198	57,424	117,659
WN Word-Senses	18,925	37,463	106,566	206,941
WN Word Lemmas (WL)	11,634	15,005	67,273	155,287
Language Lemmas (LL)	119,693	-	104,000	230,000
Ratio lemmas (WL/LL)	9.7%	12.5%	64.7%	67.5%
Ratio Word-lemmas (WN/English WN)	7.5%	9.7%	43.3%	100.0%
Ratio Synsets (WN/English WN)	8.2%	8.7%	48.8%	100.0%
Ratio Word-senses (WN/English WN)	9.1%	18.1%	51.5%	100.0%

We can see that the extension of AWN now covers around 12.5% of the estimated number of word lemmas in the baseline Arabic lexicon (versus 9.7% without extension). Moreover, after the AWN extension, word senses represent 18.1% of what already exists in the English WN (versus 8.2% before the extension).

This enriched content in terms of nouns (including BP forms), verbs and NEs has been manually validated with the collaboration of three lexicographers. The judgement rate differs from a content type to another. In fact, 41% of the hyponymy relations between noun synsets,

75% of the verbs added as synonyms in the AWN synsets and 91% of the BP forms were judged as true.

For the time being, we have developed a Web interface[16] that presents both the original and the extended content of AWN in order to allow researchers to explore and/or validate the results of the proposed extension. The interface we developed allows:

- Navigating within the AWN hirerachy (synsets tree);
- Consulting the general information of a selected synset (words, part-of-speech, etc.);
- Identifying the source of information (original or extension) using labels (for instance, NS for identifying new synsets, NI for new instances, etc.).

4.5 Impact of the extension on Arabic PR

Following the experimental process described in Section 4.2.5.1 of Chapter 3, we re-conducte an evaluation in order to see whether the performance of the AWN-based PR approach is improved after extending the content of AWN.

In the current section, we present the two runs of this new evaluation: (i) the first run using the same CLEF and TREC questions as in the previous evaluation (see Section 4.2.5.2 of Chapter 3). Note that these questions were analyzed to show the AWN coverage shortcomings; and (ii) the second run using the collection of questions prepared in the framework of the Question Answering for Machine Reading task of CLEF 2012. For both sets of questions, we are interested in comparing the performance before and after the AWN enrichement.

4.5.1 Evaluation using the original test set

The current section presents and discusses the results obtained with the first run executed with the test set of 2,264 CLEF and TREC questions that were previously considered before AWN enrichment. Table 35 presents the results of the new experiment. The same table also recalls the results that were obtained in the first evaluation.

Table 35. Results before and after AWN enrichment

Measures	Baseline PR System	PR using the Keyword-based and Structure-based levels (based on AWN)	
		Original AWN	Enriched AWN
Accuracy	9.66%	17.49%	26.76%
MRR	3.41	7.98	11.58
Nr. AQ	20.27%	23.15%	35.94%

[16] The Web interface can be viewed at: http://sibawayh.emi.ac.ma/awn_extension.

From the results above, we can see that accuracy, MRR and number of correctly answered questions were significantly improved after using our approach in comparison with the baseline PR system (i.e., Yahoo! API). Furthermore, our AWN-based approach obtained a higher performance when it was based on the enriched content of AWN. Indeed, while the original content allows the application of the approach on 1,470 questions (64.93% of the CLEF and the TREC collection), the extended content raises this number to 1,622 (71.64% of the collection). This brought about an increase in accuracy from 17.49% to 26.76% (both are higher than the 9.66% registered with the baseline PR system).

MRR also increased from 7.98 to 11.58 and the percentage of answered questions (for which the answer is found in the first five positions) went up from 23.15% to 35.94%. The improvement was also observed when considering each of the CLEF and the TREC sub collections separately with the different types of AWN extension.

Figures 22, 23 and 24 illustrate the gain in terms of Accuracy, MRR and AQ respectively, before and after enriching AWN with each type of content (NEs, verbs and nouns). These figures also recall the performance with the baseline PR system.

The first finding that deserves to be mentioned in these detailed results is the fact that generating new terms by the QE process in the keyword-based level does not decrease the performance due to the use of the structure-based level. This is true independently of the type of the enriched content.

The second finding is the noticeable performance improvement (MRR is doubled, 35% of questions were answered) observed when using the AWN extended with NEs. This can be explained by the significant percentage of questions containing NE keywords (see Table 22). Thus, the high number of NEs added to the AWN synsets helped us to obtain this improvement.

By analyzing the runs corresponding to these results, we find that the increase in performance (also in the case of verb and noun enrichement) is not only due to the possibility of applying the AWN-based approach to a higher number of questions, but also to the fact that for each keyword in the question a higher number of related terms are now generated thanks to the extension of AWN.

For instance, in the TREC question "من هو الدكتاتور الكوبي الذي أطاح به فيدل كاسترو خارج السلطة في عام 1958؟" (Who is the Cuban dictator who was overthrown by Fidel Castro out of power in 1958?), thanks to the AWN extension it was possible to apply the QE process on the verb "أطاح" (overthrow) which was newly added in AWN under the synset ">asoqaTa_v1AR / أسْقط". This helped us to get the right answer "باتيستا" (Batista) in the first 10 snippets returned by the Yahoo! API. Applying the DDN model on top of this QE process allows drawing this answer to the first 5 snippets.

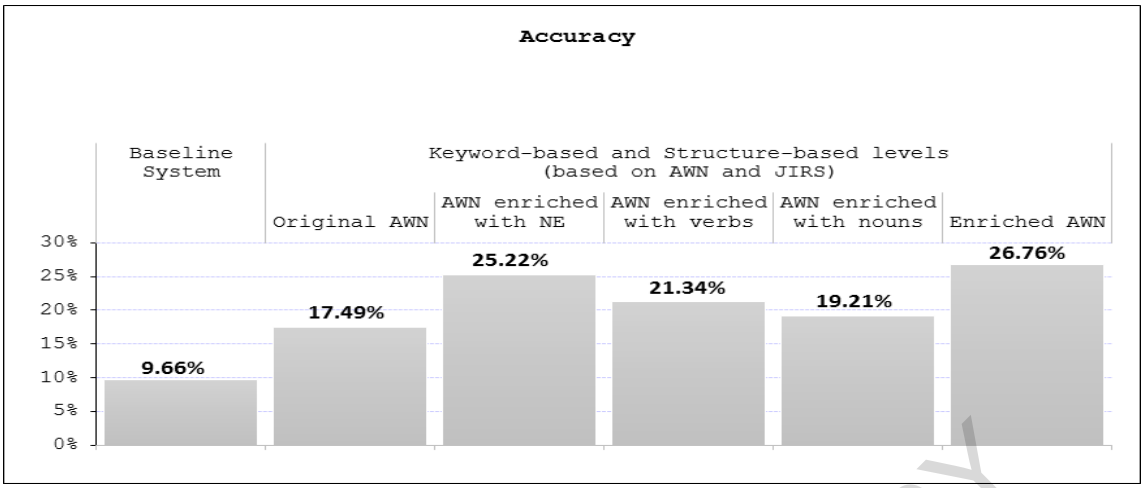


Figure 22. Details of Accuracy improvement

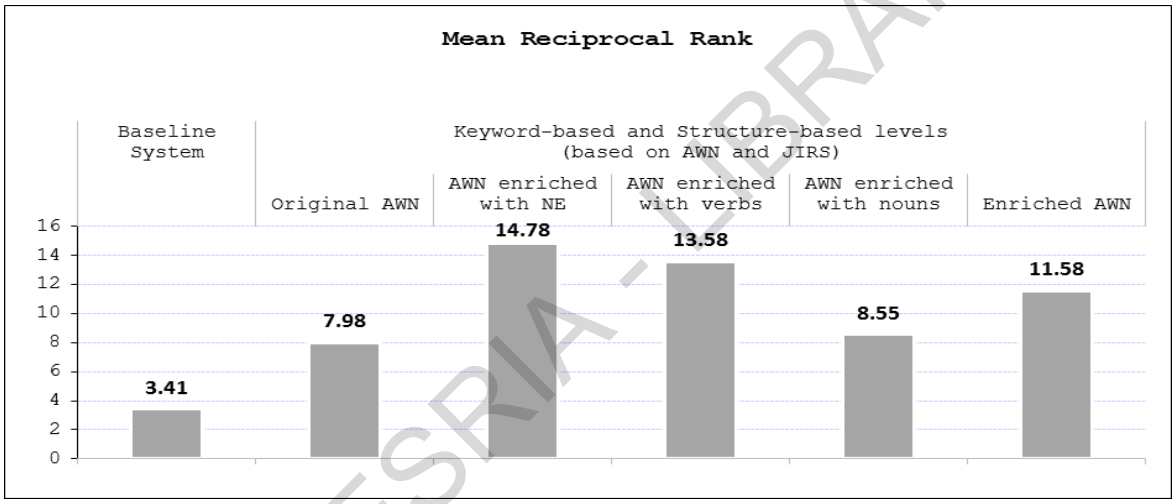


Figure 23. Details of MRR improvement

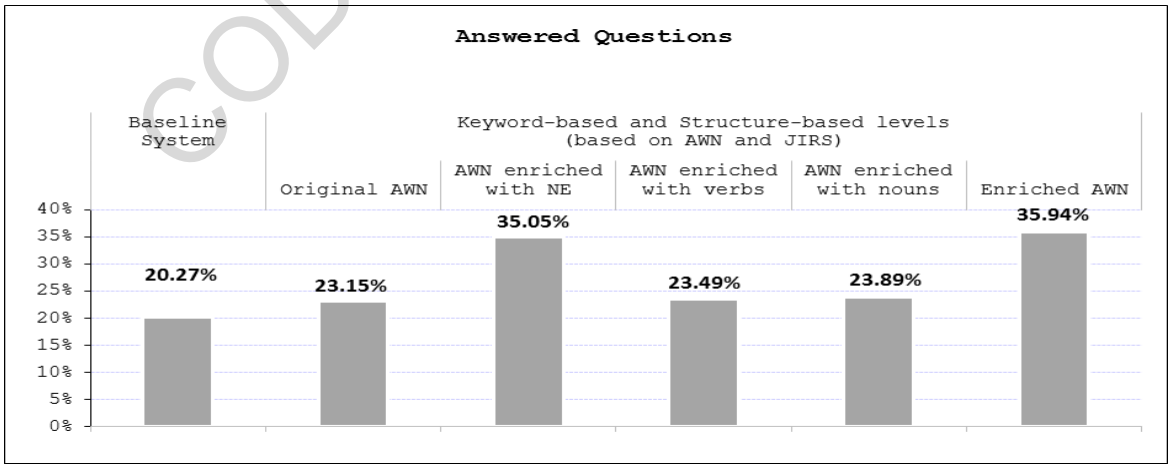


Figure 24. Details of Answered Questions improvement

To summarize, within the scope of the first run of the experiment just described, we were able to show an improvement using the extended content of AWN instead of the original content.

This is a concrete example of the usability of the AWN extension. In the next section, we present the results obtained using another test set of questions different from the one that served to analyze AWN shortcomings.

4.5.2 Evaluation using the QA4MRE test set

Let us recall that the first evaluation, described in Chapter 3, allowed us to show that our approach succeded in improving the three measures with respect to the values obtained using the baseline PR system (i.e., the Yahoo! API); thereafter, the coverage of the AWN lexical database was semi-automatically extended in order to deal with the shortcomings of this resource for the questions of the test set. Note that these shortcomings result in a poor QE process (generating just a few number of related terms) or in a non ability of applying this process. Once extended, the goal of this second evaluation is to measure the gain in performance after using the AWN enrichment.

This new evaluation contains two runs: (i) the first run with the same questions analyzed for the AWN enrichment; this run showed an improvement of accuracy, MRR and AQ measures, and (ii) the second run with a different test set of questions prepared in the framework of the QA4MRE Task of CLEF 2012. We participated in this competition with the aim to evaluate the keyword-based and structure-based approach in such a specific task and also to compare its performance with other systems.

The 2012 test set is composed of 4 topics; each topic includes 4 reading tests. Each reading test consists of one document, accompanied by 10 questions, each with a set of 5 answer options per question. Therefore, for each language task, there are in total:

- 16 test documents (4 documents for each of the four topics);
- 160 questions (10 questions for each document);
- 800 choices/options (5 for each question with one correct answer and four incorrect answers).

Questions have the following characteristics:

- They are in the form of multiple choice, where for each question, 5 possible answers are given;
- They are designed to focus on testing the comprehension of one single document;
- They test the reasoning capabilities of systems, which means that inferences, relative clauses, elliptic expressions, meronymy, metonymy, temporal and spatial reasoning, and reasoning on quantities may be exploited;

- They may involve background knowledge, i.e., information that is not present in the test document given. In such cases, information from the background collections is needed to fill in the knowledge gap to answer the question.

The distribution of these questions over the different types is presented in Table 36 (Peñas et al., 2012).

Table 36. Distribution of question types

Question type	Total number of questions	Percentage
PURPOSE	27	16.88%
METHOD	30	18.75%
CAUSAL	36	22.50%
FACTOID	36	22.50%
WHICH-IS-TRUE	31	19.38%
TOTAL # of QUESTIONS	160	100.00%

The distribution of the 160 questions is quite similar over the 5 considered categories. This shows how complex are the questions of this test set in comparison with the previous test set composed of 2,264 CLEF and TREC questions where factoid questions represented over 50% of the set (versus 22.5% in the QA4MRE set). Below we give some examples for the other categories contained in this test set:

- FACTOID: Where or When or By--Whom
- CAUSAL: What was the cause/result of Event X?
- METHOD: How did X do Y? Or: In what way did X come about?
- PURPOSE: Why was X brought about? Or: What was the reason for doing X?
- WHICH IS TRUE: Here one must select the correct alternative from a number of statements, e.g. What can a 14 year old girl do?

We apply on each question the keyword-based and the structure-based levels. The answer checking process matches candidate answers with returned passages. The first run that we have submitted uses the original AWN while the second run uses the enriched AWN.

Each test receives an evaluation score between 0 and 1 in order to calculate the c@1 measure (see the description of this measure in Chapter 2, Section 2.2.3.2) and accuracy. As previously mentioned, the c@1 measure encourages systems to reduce the number of incorrect answers while maintaining the number of correct ones by leaving some questions unanswered. This measure is considered as a relaxed form between accuracy and MRR.

Systems receive evaluation scores from two different perspectives: (i) at the question-answering level: correct answers are counted individually without grouping them; and (ii) at the reading-test level: figures are given both for each reading test as a whole and for each separate topic.

Obtained results also present number of unanswered questions with right and wrong candidate answers. However, in both runs, we did not consider this possibility in the submitted outputs. Table 37 and Table 38 present the obtained results in terms of: (i) overall accuracy among the test set of questions and (ii) overall and detailed c@1 performance.

Table 37. Overall accuracy over the two runs

RUNS	OVERALL ACCURACY	ANSWERED		UNANSWERED		
		RIGHT	WRONG	EMPTY	RIGHT	WRONG
run #1 (Original AWN)	8%	12	21	127	-	-
run #2 (Enriched AWN)	13%	21	49	90	-	-

The overall accuracy reaches 13% with run #2 using the enriched AWN lexical database which represents an increase of 5% in comparison with the 8% accuracy obtained in run #1 using the original AWN. This confirms the results obtained in the previous evaluation.

Accuracy was calculated over the 160 questions, including the unanswered questions (i.e., questions for which our approach does not provide any answer). If we only consider the 75 questions that are mentioned by CLEF as being answerable without any extra knowledge (in our two runs we did not use such knowledge which also includes the background collection of CLEF 2012), the accuracy in run #2 becomes 28% which is slightly higher than the 26.76% accuracy registered in the previous experiments (using the 2,264 CLEF and TREC questions). This confirms the effectiveness of our approach based on an enriched AWN even in the context of complex questions (let us recall that factoid questions that are more simple to answer only represent 22.5% of the used test set).

Table 38. Overall and detailed c@1 over the two runs

RUNS	c@1 measure				
	Overall	Topic #1	Topic #2	Topic #3	Topic #4
run #1 (Original AWN)	0.13	0.25	0.18	0.05	0.05
run #2 (Enriched AWN)	0.21	0.36	0.19	0.08	0.17

Regarding the c@1 measure, Table 38 shows the overall of 0.21 for the second run (versus 0.13 for the first run). With respect to this measure, our approach registered a different performance over the four topics (AIDS, Climate Change, Music and Society, and Alzheimer's disease). Indeed, from Table 38 the maximum score was obtained for Topic #1 (i.e., AIDS) in the two runs (0.25 in run #1 versus 0.36 in run #2). Moreover, this score is higher than the mean score (0.32) over all best runs registered in this topic by all the participating systems for different languages including English.

At reading-test level, our system obtained its best score of c@1 measure when answering questions belonging to Topic #1 (i.e., AIDS). Figure 25 illustrates a comparison between the

best c@1 measures obtained in reading-tests over the four topics. Topic #3 is the one for which lower performance have been reached.

Let us analyze questions for which our system succeeds and those for which it fails, i.e., questions belonging to the above topics (i.e. topic #1 and #3). Most of the answered questions are factoid ones (When, Who, What, etc.). This shows that using Arabic WordNet mapped with YAGO (which contains high number of Named Entities) has a positive impact on system performance especially when processing factoid questions.

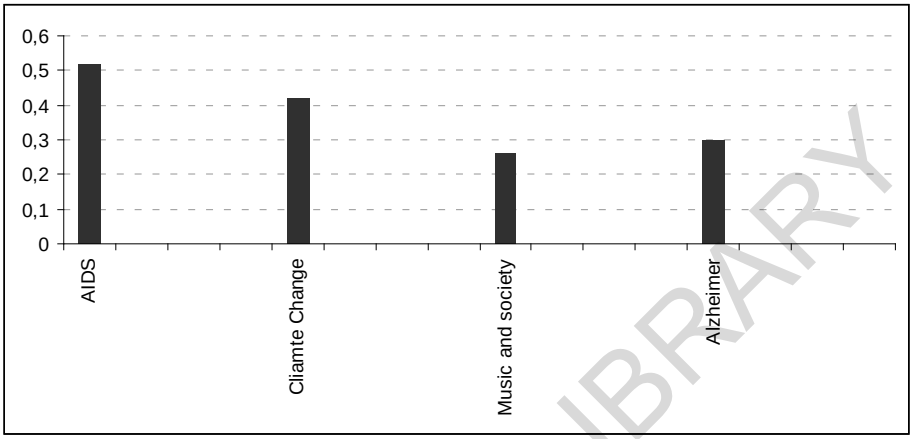


Figure 25. Best c@1 obtained in reading tests over topics

On the other hand, the questions where the system fails to extract a correct answer fall into five categories:

- Questions that are not factoid such as LIST questions (questions starting with Give a list of ...) and REASON questions (questions starting with Why ...);
- Questions with translation errors. For instance, in reading-test #4 question #4 the translation of "What is the mechanism by which HIV-positive Brazilians receive free ARV drugs?" is " ماهي الأليات المستعملة لإعطاء البرازيليين المصابين بداء نقصان المناعة البشرية المضادة للفيروسات القهقرية مجاناً؟" which is not an understandable Arabic question. This remark can also be applied on reading-test documents.
- Questions not starting with question stopword (such as What, When, etc.). For example, reading-test #6 question #3 "ووفقا للحكومة البرازيلية، ما هي الأسباب الرئيسية لتغير المناخ؟" (According to the Brazilian government, what is one of the main reasons for climate change?)
- Questions with long candidate answers. For instance, questions #3 and #4 in reading-test #13 "ما هو النظام الغذائي الذي يمكن أن يخفض من خطر الإصابة بمرض الزهايمر؟" (What type of diet may reduce the risk of Alzheimer's disease?) and " لماذا لا يوصى باستعمال أنابيب التغذية لمرضى الزهايمر الذين لديهم صعوبات في البلع؟" (Why are feeding tubes not always recommended for Alzheimer's patients who have difficulties with swallowing?).

Following, let us discuss the comparison of our approach to the baseline of the competition and to other systems (for Arabic, English, Dutch, etc.). The baseline has five possibilities when trying to answer a question: it can select the correct answer to the question, or it can select one of the four incorrect answers. Then, the overall result of this random baseline is 0.2 (both for accuracy and for c@1). Systems applying a certain kind of processing and reasoning should be able to outperform this baseline (Peñas et al., 2012).

The c@1 obtained by our approach for run #2 (0.21) using the surface-side (the keyword-based and structure-based levels) of our approach is higher than the 0.2 of the baseline system. This allows run #2 to be ranked at the 27th position in a list of 40 submitted runs. It outperforms the unique other competitor system for Arabic (its best corresponding run scored 0.19). With respect to the overall results at reading-test level, run #2 was ranked at the 7th position (out of 40 runs).

According to CLEF 2012 results, the average c@1 among questions that can be answered without using any extra knowledge is 0.30 (see Table 39). By considering only the total number of these questions (75), we find out that with run #2 we obtain a c@1 of 0.48 which is significantly over the mentioned average.

Table 39. Classification according to the knowledge required to answer questions

Source: (Peñas et al., 2012)

Types of question	#of questions	c@1
NO EXTRA KNOWLEDGE REQUIRED	75	0.30
BACKGROUND KNOWLEDGE REQUIRED	46	0.28
INFERENCE REQUIRED	21	0.20
INFORMATION NEEDS TO BE GATHRED FROM DIFFERENT SENTENCES or PARAGRAPHS	20	0.27

Table 39 also shows that, unsurprisingly, among the participating systems, the highest average c@1 was registered in the case of questions that need no extra knowledge to be answered. The least c@1 average was obtained for questions requiring inference to answer them. The semantic-based level (the third level of our approach) will help in answering this kind of questions and, therefore, in improving even more the performance of our Arabic QA approach. The next chapter describes the research undertaken to consider semantic-based reasoning on top of the surface-side of our approach (i.e., keyword-based and structure-based levels).

4.6 Chapter summary

In the first part of the present chapter, we highlighted the main *coverage* shortcomings in AWN from both: (i) a theoretical perspective by comparing its content to a representative Arabic lexicon and to WordNets in other languages, and from (ii) an experience-based perspective by analyzing the keywords of non extended and non answered Arabic questions

from the first evaluation presented in Chapter 3. We also explained how these shortcomings impact the *usability* of this resource and have been the reasons behind its limited use in Arabic NLP applications.

In the second part of this chapter, we started from the above analysis to identify the lines of investigation for the release of an enriched AWN with respect to the needs in the context of Arabic QA. The targeted contents were: (i) nouns and verbs, as the main common linguistic categories, (ii) Instances or NEs, as one of the most important types of dynamic information to link with the AWN resource, taking into account our interest in answering questions from the Web, where NEs are widely used; also, we are interested in NEs since the injection of their hypernym is effective in the structure-based level (as shown in the exemple provided in Chapter 3) especially for factoid questions; and (iii) broken plurals, as a linguistic characteristic mainly specific to Arabic and widely used in the analyzed test set of CLEF and TREC questions.

Once identified, we proposed semi-automatic techniques based either on other resources such as YAGO for NEs, VerbNet and UVI for verbs, manually prepared lists for BP or on process-based methods including MFS for hypernymy/hyponymy enrichement. From a theoritical perspective, these techniques allowed us to achieve an enrichment of AWN by suggesting new NEs, verbs and nouns (including BP forms).

The content in terms of NEs represents the best improvement since 433,339 instances were linked to their corresponding AWN synsets. This number is nearly 38 thousand times more than the number of NEs in the current release of AWN. Furthermore, a significant amount of verbs (+122% with respect to the original AWN) was linked to AWN verb synsets. A semi-automatic extraction of noun hyponyms also allowed extracting new AWN synsets and associations. As a comparison, the content of the enriched version of AWN exceeds that of the Spanish WN.

In the third part of this chapter, we were interested in showing from an experiment-based perspective, the usability of the AWN resource after its enrichment. For a more significant evaluation, new experiments were conducted by considering two different sets of questions: (i) the same set of the 2,264 CLEF-TREC questions considered in the first evaluation and that were analyzed to identfy AWN shortcomings; note that 55% of these questions are factoid and can be effectively processed by surface-based approaches, and (ii) a set of 160 questions from the main task of QA4MRE organized in CLEF 2012; the latter set contains a different distribution of question types with a lower percentage of factoid questions (22.5%) and a higher number of complex questions requiring deeper approaches rather than surface-based ones.

Both experiments showed improvement when applying the surface-based levels (keyword-based and structure-based levels) of our approach after the enrichment of AWN. In the first

experiment, we obtained an accuracy of approx. 27% while in the second experiment yielded 13% accuracy. The lower performance from the latter is due to the second set containing a higher number of complex questions requiring semantic and knowledge-based approaches. Nevertheless, this performance remains promising if we only consider the amount of questions that can be answered without extra knowledge (75 questions as mentioned by CLEF organizers). In this case, the accuracy is 28% which is on par with the one registered for the first set.

Another measure, c@1, allows analyzing the performance in comparison with other systems for Arabic and other languages. With respect to this measure, the obtained performance was about 0.21, higher than the baseline performance, and allowed for obtaining an acceptable ranking among the participating systems.

To conclude this part of the research, it is shown how the surface-based levels can improve the Arabic PR, especially when the AWN is used with a high coverage of the Modern Standard Arabic. The need of a semantic-based level is also highlighted for a better processing of complex questions beyond factoid ones. The next chapter presents the semantic-based level and a discussion of the performance achieved from its application in Arabic QA.

Chapter 5
Semantic-based Level

5.1 Introduction

In Chapter 3, we have evaluated the effectiveness of the surface-based levels (i.e., the keyword-based and structure-based levels) of our PR approach in comparison with a baseline system (i.e., the Yahoo API) and using the Web content as a target collection. This evaluation allowed us to confirm the improvement in performance through measuring the number of answered questions, the accuracy and the MRR. However, the analysis of the obtained results also revealed the shortcomings of the Arabic WordNet resource in terms of coverage. Then, in Chapter 4, we proposed an enrichment of AWN based on semi-automatic techniques and leveraging other resources such as YAGO for NEs, VerbNet and UVI for verbs, etc. The surface-based levels performed better after this enrichment. This has also been proved by conducting new experiments using another test-set of questions devoted to the QA4MRE task. The QA4MRE evaluation allowed us to make a comparison with systems for other languages such as English. It highlighted the importance of processing the other types of questions using semantic-based approaches, beyond the surface-based ones.

In the present Chapter, we focus on (i) implementing the levels of our approach addressing the semantic-based level with the aim to process the types of questions requiring the understanding of meaning rather than the comparison of surface elements (i.e., keywords and structure) and (ii) improving in general the performance of the system. This chapter is structured as follows: Section 5.2 provides a background related to the approaches based on similarity at a semantic level. Section 5.3 describes the ontology we have built for the purpose of semantic-based level reasoning and experiments. Section 5.4 shows the approach we propose for the representation of questions and passages in CGs and their comparison as well as its evaluation using two question test-sets. Section 5.5 draws the main conclusions of this chapter.

5.2 Background

At this point, we turn our attention to ranking passages with respect to their semantic

similarity to the question, and not simply by surface similarity. Among the approaches used in this direction, we can cite the work of Hensman (2005). Authors follow two steps: (i) Step 1: representing the text (question or passages) in term of CGs; and (ii) Step 2: comparing both representations on the basis of a CG operation (Maximal Joint, Generalization, Projection, etc.). Let us recall that the generated CGs are directed graphs of nodes that correspond to concepts, connected by labeled and oriented arcs that represent conceptual relations (Sowa 1983). The CG formalism has the advantage of being close to both natural and computers languages.

Step 1 is the most challenging in these approaches, involving different resources (WordNet, VerbNet, WordNet domains, etc.) and NLP tools (morphological analyzer, syntactic parser, etc.). This step mainly relies on the VerbNet (VN) (Kipper-Schuler 2006) resource. VerbNet is organized into verb classes extending Levin (1993) classes through refinement and addition of subclasses to achieve syntactic and semantic coherence among members of a class. By identifying the syntactic frame that better matches the processed text, it is then possible to use the semantic frame of the verb as a basis to construct a CG for the text. Figure 26 illustrates an example provided by Hensman and Dunnion (2004) regarding the use of the different resources in this step.

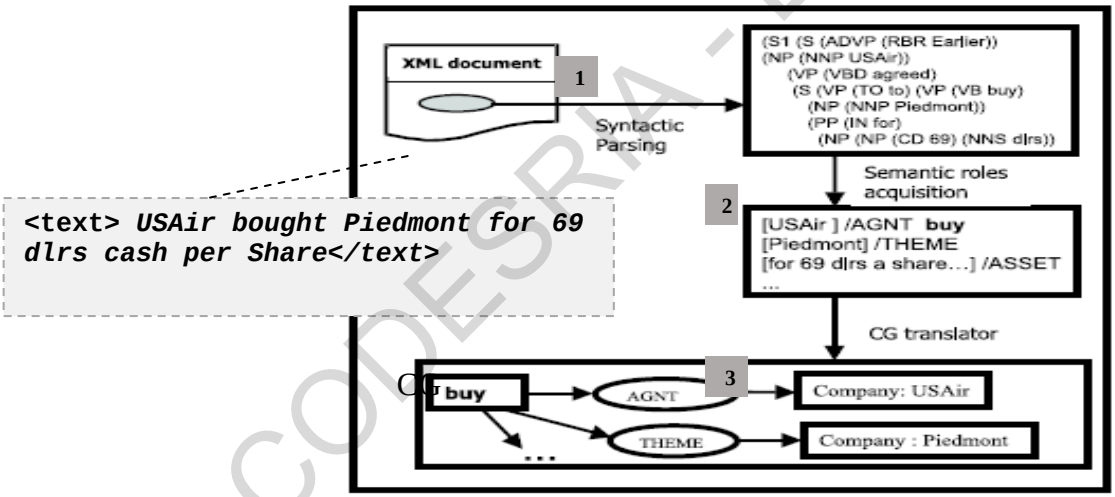


Figure 26. Example illustrating the step of text representation in CGs
Source: Hensman and Dunnion (2004)

This step is performed as follows:

1- *Syntactic parsing* by (i) Parsing the text and getting its syntactic tree (Figure 26 presents an example of a syntactic tree in a linear form); (ii) Identifying the verb in the text using a PoS tagger; and (iii) Recognizing the syntactic frame of VN that better matches the syntactic parsing of the text (parsing performed in Step 1) and the given verb (identified in Step 2);

2- *Semantic role acquisition* from the text with respect to the syntactic frame identified in step 3 (Figure 26 illustrates that *USAir* has the role *AGNT*, *Piedmont* has the role *THEME*, etc.);
3- *Translation of text into CG* using the semantic frame that corresponds to the syntactic frame recognized in Step 3;

This part of our research has the objective to evaluate the effectiveness of the semantic-based level on top of the surface-based levels. In order to implement the semantic-based level, the approach we adopt is hybrid: (i) it uses the same resources adopted by Hensman (2005), i.e., WordNet and VerbNet, for the translation of text into CG and (ii) it is based on the formula of the semantic similarity proposed by Montes-y-Gómez et al. (2001). Our approach also proposes some new adaptations specific to the Arabic language. Before moving to details about the semantic-based level, the next section describes the ontology that we constructed from AWN and AVN to be used in the two steps of this level.

5.3 Ontology construction

To develop the above two-steps method for the semantic-based level, we have two main requirements: (i) an ontology containing concepts and relations that can be used to construct question and passage CGs, and (ii) operations over CGs in the framework of the same ontology. Let us briefly recall that an ontology presents the knowledge about a domain with formal definitions about concepts and relations between these concepts (Gruber 1993).

In our work, we took the decision to construct such an ontology since, to our knowledge, there is no available one. This new ontology, called "AWN-AVN ontology" is mainly built from: (i) the AWN resource as its design is quite similar to an ontology, and (ii) the AVN resource as it provides formal representation of verb syntax and semantics. Hence, the design of our AWN-AVN ontology is structured around these elements as follows:

- Concepts and their hierarchy are extracted from AWN synsets and hyponymy relations especially QA;
- Concepts are assigned lexical information such as synonyms and situations about these concepts. Situations formalize the syntactic and semantic frames (this part is detailed later in this chapter) in terms of CGs.

5.3.1 Concepts and hierarchy

Figure 27 illustrates the global design of this AWN-AVN ontology. Note that the AWN-AVN ontology contains not only static information (concepts, lexicon and situations) but also

dynamic information, for example NEs i.e., instances (or individuals) such as names of persons and places that are important to be recognized by applications especially by QA systems. This is particularly important in the case of factoid questions where the expected answer is a NE. In the following subsections, we highlight the process performed to populate our ontology with respect to the above elements.

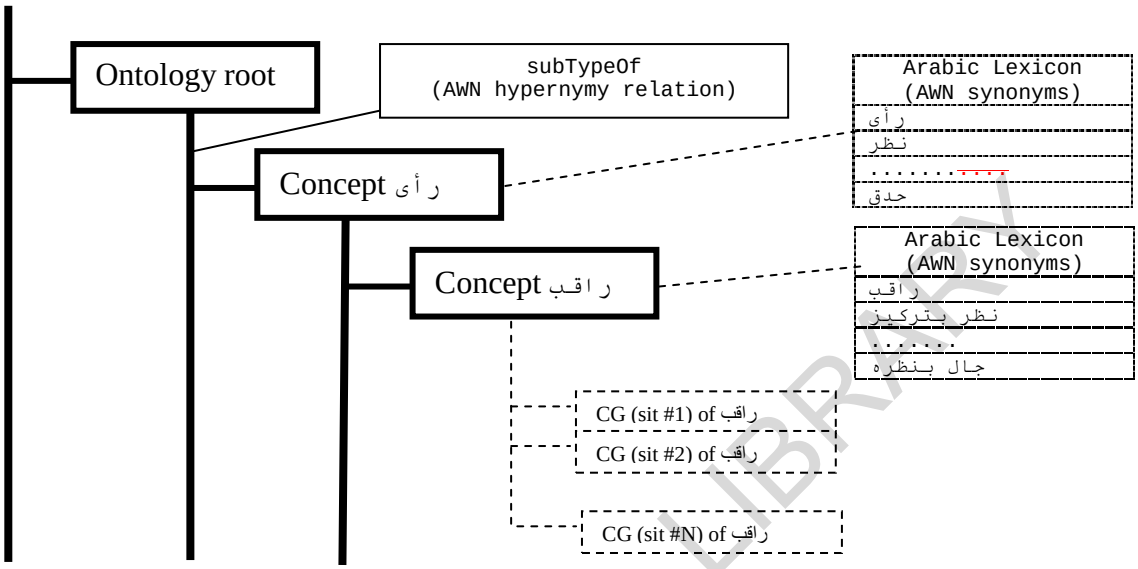


Figure 27. Design of the AWN-AVN ontology

In Figure 27, boxes with bold lines refer to concepts, while boxes with dashed lines refer to additional information about concepts. The root is the most general concept of the ontology. Under this general concept we can find the other concepts extracted from AWN synsets. Each concept can have hyponym concepts i.e., concepts that are more specialized (their meanings are more specific). The lexicon (illustrated by tables in Figure 27) is the natural language counterpart of the concept, i.e., the words that refer to this concept in the considered language (Arabic in our case). For example the concept "**رأى**" (to see) can be expressed in the Arabic lexicon by one of the following words: **رأى, نظر, حدق**, etc. The concept itself has another subconcept which is a specialization of "**رأى**", namely "**راقب**" (to supervise) expressed in natural language by:**راقب , جال بنظره,نظر بتركيز** , etc. According to the different expressions, in natural language, of the same concept, we can have syntactic-semantic situations extracted from AVN that can be applied to a given concept: for example, situation where the syntax contains `V+Agent+Patient` (for instance, **راقب النظار الهلال**) with a specific meaning, another where the syntax contains `V+Agent+Patient+PP` (for instance, **رأى الشرطي اللص في الليل**), etc. The situations are simply use cases of the concept with respect to two perspectives: syntax and

semantic. Each situation refers to a syntax case together with the corresponding semantic meaning. These situations are translated into CGs as described in the following section.

5.3.2 CG situations

5.3.2.1 Arabic VerbNet

The Arabic VerbNet resource covers a large number of Arabic verbs exploiting Levin's classes (Levin 1993) and the basic development procedure of Kipper-Schuler (2005). The current version of AVN has 336 classes populating 7,744 verbs and 1,399 frames[1]. Figure 28 shows an example from AVN related to class raOaY-1 (i.e., رأى, to see).

Each class contains information about (i) class members (i.e., verbs belonging to the class), for instance رأى (to see), لاحظ (to observe), etc. (ii) themroles and frames that represent syntactic-semantic situations of its members (for example, V Experiencer Stimulus), and eventually (iii) its subclasses and sibling classes (in the above example, the subclass is identified by *raOaY-1.1* and there is no Sibling class).

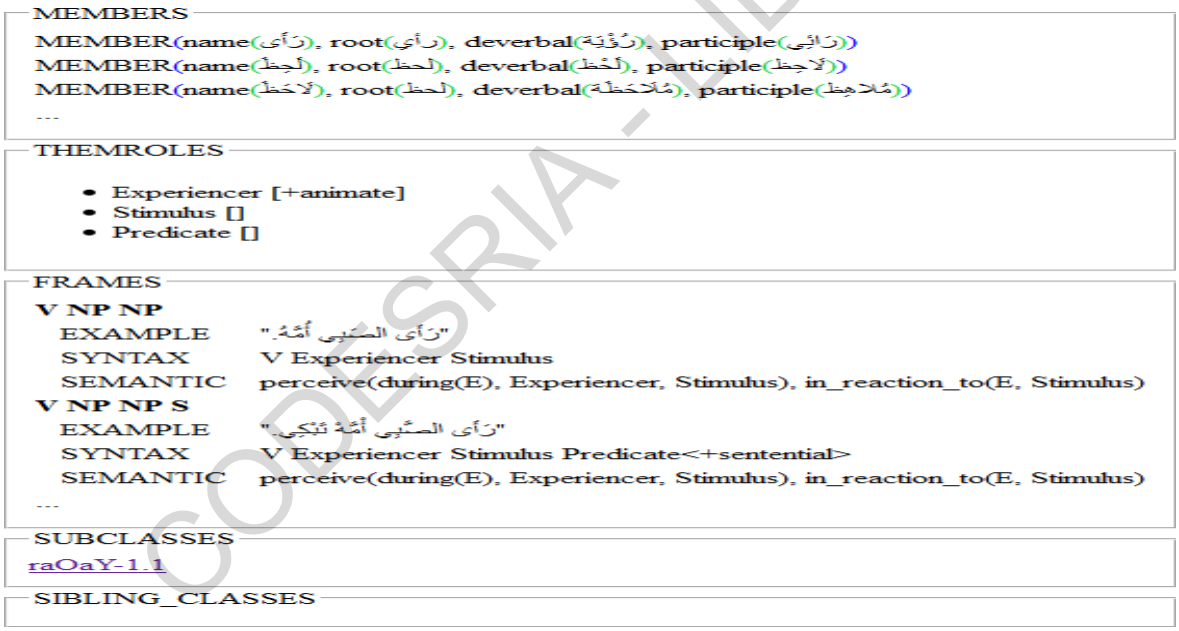


Figure 28. A snapshot of the AVN class raOaY-1

The top level of each class shows the verbs that are members of the given class. Each verb member is identified by the verb itself (e.g. رأى), its root form (e.g. رأى), its deverbal form (e.g. رؤية) and its participle (e.g. رائي). Also, the thematic roles and their restrictions are encoded at

[1] http://ling.uni-konstanz.de/pages/home/mousser/files/Arabic_verbnet.php

the top level of classes; restrictions are lists of selectional constraints on semantic roles. Some frames define local restrictions that are specific to the given frame and are combined with the common restrictions (i.e., those appearing at the top level of a class).

Frames related to a given class are presented with an example sentence (for instance, رَأَى الصَبِي أُمَّهُ.), a syntactic and a semantic structure. The latter structure contains semantic predicates including arguments and temporal information similarly to that proposed by (Moens and Steedman 1988).

Subclasses (for instance raOaY-1.1) have a similar structure as the main classes (i.e., raOaY-1). Obviously, subclasses can also have subclasses in a recursive way. A subclass inherits all properties of the main class. Therefore, verbs appearing in these subclasses have new syntactic and semantic frames in addition to those of the main class. On the other hand, sibling classes are specific to the Arabic language and are detailed in the work proposed by Mousser (2010). Briefly, a sibling class is created to populate the verbs resulting from alternations requiring morphological changes.

5.3.2.2 Transformation of AVN frames into CGs

The structure and content of AVN classes is an interesting starting point to enrich the verb nodes of our ontology using semantic and syntactic information. To achieve this enrichment, we perform a two-steps technique (Figure 29):

- Step 1: The first step is concerned by the extraction, from AVN, of verbs together with corresponding frames content. A given verb can appear as member of different classes. Therefore, we extract the frames from all these classes as well as from their super classes (considering the principle of frame inheritance).
- Step 2: we generate CGs based on the extracted semantic information and integrate them in the ontology as situations of each concept (corresponding to the concerned verb members).

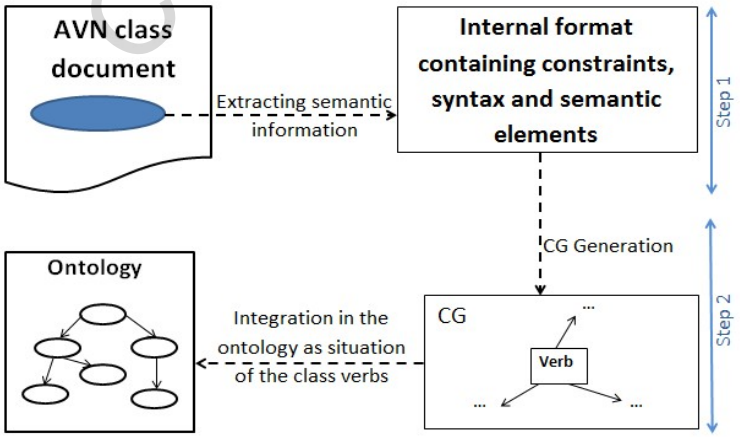


Figure 29.

General process for the semantic extraction

The aim of step 2 is to generate a global CG which is composed of three subCGs (depicted in Figure 30): (i) “SyntaxCG” for the syntactic frame that can be applied to a given verb, (ii) “SemanticCG” for the meaning of the verb by means of themeroles and predicates, and (iii) “ConstraintCG” for the constraints existing on themeroles used in the first and second subCGs.

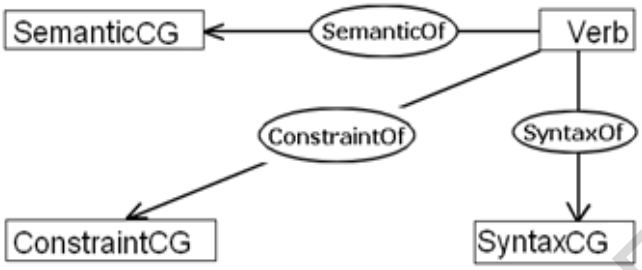


Figure 30. Form of the situation CG corresponding to the AVN frame

The global CG is formed around a verb concept linked to the other subCGs through three ontology relations, respectively “SyntaxOf”, “SemanticOf” and “ConstraintOf” (illustrated in elliptical shape). Figure 31 illustrates the steps performed to generate the global CG.

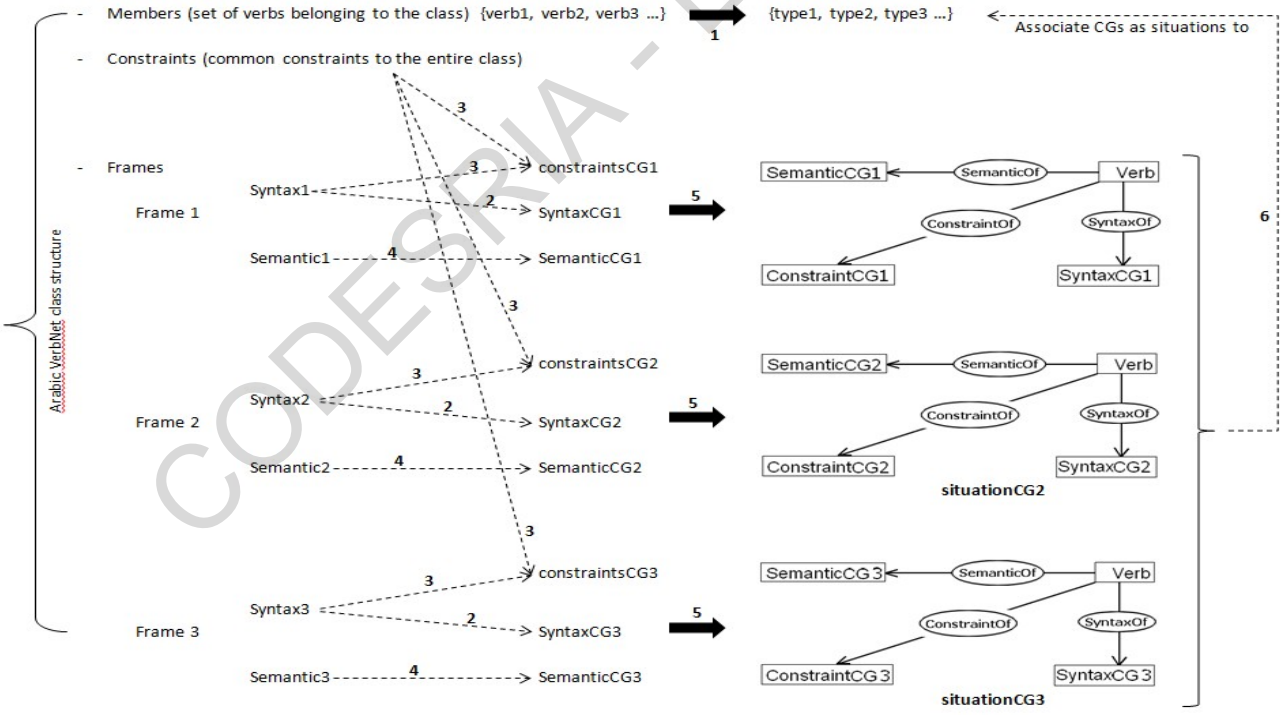


Figure 31. Process of AVN frames transformation into CGs

The step “CG generation” is performed through the following five substeps:

- *Step 2.1*: For given verb in AVN, we locate the corresponding concept (i.e., AWN synset). A verb can be associated with different possible concepts. To disambiguate these possibilities, we consider the concept having an ontology lexicon that contains the highest number of verbs sharing the same class of the given verb.

- *Step 2.2*: For each syntactic frame extracted in step1, the succession of syntactic constituents such as Noun Phrases (NP) and Prepositional Phrase (PP) are represented in the "SyntaxCG" using general concepts (for instance the concept "np**"** connected through the ontology relation "followedBy"). Examples of resulting Syntactic CGs are provided below:

```
Syntactic CG1:
 [np : *c2 ] -
       -followedBy->[np : *c3 ],
       <-followedBy-[verb : *c1 ]

Syntactic CG2:
 [np : *c2 ] -
       -followedBy->[np : *c3 ]-
        followedBy->[np: *c4],
       <-followedBy-[verb : *c1 ]
```

- *Step 2.3*: We construct the subCG "ConstraintCG" from class restrictions (must be applied to all the verbs of a class) and specific restrictions (those that are specific to the given frame). The following CG is the "ConstraintCG" generated for the class illustrated in Figure 28 :

```
Constraint CG:
[list : "[?c2(animate)]"]
```

As can be noticed, the above CG represents the restriction related to the second syntactic constituent of Frame 1, i.e., `c2` (see Figure 28) which has the themerole Experiencer and does not consider the restrictions on the other themeroles since they are not used in the frames of the given class. The resulting CG shows that the constraint on the concept of type "np" and identified by "c2" in the syntactic CGs (CG1 and CG2) must be "animate".

- *Step 2.4*: The CGs corresponding to the semantic frames are constructed by means of a semi-automatic process. Let us take the same AVN class illustrated in Figure 28. The first semantic frame shows two issues:

    - During the event related to verbs that are members of the given class, the syntactic constituent "Experiencer" (i.e., the second NP referenced by "c2" in SyntaticCG1) perceives the syntactic constituent "Stimulus" (i.e., the third NP referenced by "c3");

- This event is in reaction to the syntactic constituent “Stimulus”.

Hence, the two above issues of the semantic frame are represented in the semantic CG as follows:

```
Semantic CG:
 [event : *p1 ]-
 -duringOf->[cg:[perceive:*p2 ]-
 -experiencerOf->[np : ?c2 ],
 -stimulusOf->[np : ?c3 ]],
 -inReactionTo->[np : ?c3 ]
```

In the above semantic CG, the references used in the syntactic and constraint CG are reused for the same constituents in order to make a connection between parts of the global CG (illustrated in Figure 30). As shown in the semantic CG, the two AVN predicates “*perceive*” and “*in_reaction_to*” are represented differently: the former becomes the concept “*perceive*” whereas the latter becomes the relation “*in_reaction_to*”. The decision of which representation form should be used (concept or relation) is made manually. Thereafter, many types of automatic transformation generate the resulting CG. This was applied to the 146 different predicates contained in AVN as shown in Table 40.

Table 40. Transformation of AVN predicates into semantic CGs

AVN predicate groups	Example	No. predicates	%	No. transformation types
group 1	adopt,allow, attempt,contact	87	60%	1
group 2	free,depend, meet	39	27%	39
group 3	together-apart, harmed-disconfort	8	5%	4
group 4	-	3	2%	1

Table 40 shows that group 1 is composed of 87 (about 60%) of the available predicates are mapped using the same semi-automatic algorithm (i.e., process allowing the transformation of the frames where these predicates appear into CGs). The remaining predicates can be classified under 3 groups: group 2 contains 39 predicates (about 27%) that are mapped using 39 different algorithms (the manual task in this case is repeated 39 times); group 3 only concerns 8 predicates with 4 different algorithms (one per predicate pair); finally, group 4 contains 3 other predicates requiring a different algorithm.

- *Step 2.5*: We construct the global CG as explained above (Figure 30).

- *Step 2.6*: The resulting global CG is associated with concept extracted after Step 1.1. The general concept "verb" is substituted in this global CG by each associated concept. Here are the two CGs corresponding to the two frames of the previous example (class raOaY-1):

Global CG 1:

```
[verb : *c1 ] -
        -syntaxOf->[cg : [np : *c2 ] -
                        -followedBy->[np : *c3 ],
                        <-followedBy-[verb: ?c1 ]
                    ],
        -constraintOf->[list :"[?c2(animate)]" ],
        -semanticOf->[cg : [event : *p1 ] -
                            -duringOf->[cg : [perceive : *p2 ] -
                                                -experiencerOf->[np : ?c2 ],
                                                -stimulusOf->[np : ?c3 ]
                                            ],
                            -inReactionTo->[np : ?c3 ]
                    ]
```

Global CG 2:

```
[verb  : *c1 ] -
        -syntaxOf->[cg : [np : *c2 ] -
                        -followedBy->[np : *c3 ]-followedBy->[np],
                        <-followedBy-[verb : ?c1 ]
                    ],
        -constraintOf->[list :"[?c2(animate), ?c4(sentential)]" ],
        -semanticOf->[cg : [event : *p1 ] -
                            -duringOf->[cg : [perceive : *p2 ] -
                                                -experiencerOf->[np : ?c2 ],
                                                -stimulusOf->[np : ?c3 ]
                                            ],
                            -inReactionTo->[np : ?c3 ]
                    ]
```


5.4 Implementation and evaluation of the semantic level

The present section details the implementation of the semantic level on the basis of the AWN-AVN ontology. It shows the adaptations we have made for the needs of processing Arabic text (questions or passages) with respect to similar approaches. As we previously mentioned, the used approach considers two main steps: (i) Step 1 which is devoted to represent text (the question and candidate passages) using the CG formalism and (ii) Step 2 that compares each passage CG with the question CG to measure the semantic similarity between both representations. Thus, the current section is divided into two subsections describing each step.

In both subsections, we consider the following question extracted from the 2013 QA4MRE test-set related to the first topic "Alzhaimer": "ما هو الجزء من جسم الإنسان الذي يتم فيه تشكيل اللويحات؟" (What is the part of the human body where the formation of plaques occurs?). For this question, there is a list of five candidate answers (from which our semantic level is asked to decide the correct one). For instance, the answers given for the above sample question are (the right answer in the gold standard is underlined and written in italic):

- الرئتين (Lungs)
- الأكتاف (Shoulders)
- *الرأس (Head)*
- الأيادي (Hands)
- لا شيء مما سبق (None of these answers)

Using the surface-based levels and considering the proposed enrichment of AWN, we were able to extract and rank the passages according to the similarity score based on injection of QE terms in the Distance Density N-gram Model, let us call it the Surface Similarity (SFSim) (described in Chapter 3). Table 41 lists the top eight passages with the best `SFSim`.

As we can see, none of these passages have a `SFSim` which can allow to consider it for the extraction of the right answer. The best score was 0.44, since the structure of the question as well as the keywords used to formulate it are not significantly present in these passages. From a human perspective, seven of these passages show the right answer (see the sentences written in bold in Table 41). Now, let us see how our semantic-based approach can be used to improve the Semantic Similarity (SSim). To achieve this goal, we start by performing the first step which consists in representing the question and the eight passages in CGs. The first CG is labelled `CG-Q` while the CG of a passage P_i is labelled $CG\text{-}P_i$ (where `i` is the rank of the passage in Table 41).

Table 41. Passages retrieved and ranked using the Surface-based levels

Rank	Passage	Surface Similarity
1	عندما يتم التعامل مع الأجسام المضادة لمريض الزهايمر التي ترتبط باللويحات، هذه اللوائح تصبح المغلفة للأجسام المضادة .الاستجابة المناعية اللاحقة **لمسح دماغ المريض من هذه اللوحات** يحتمل أن تكون خطرة.رائع، نستطيع الآن استخدام هذه الأجسام المضادة لمساعدة المرضى؟ ليس جيدا تماما ...**في حين أن الأجسام المضادة تمسح بكفاءة اللويحات من الدماغ**، إلا أنه قد يخفض من نجاحها الآثار الجانبية.لإزالة اللويحات من الأجسام المضادة المغلفة من الضروري تنشيط خلايا من الجهاز المناعي وهذا يمكن أن يسبب التهاب في الدماغ .في حين أن هذا الالتهاب هو في الواقع مفيد في إزالة البكتيريا مثل السل، إلا أن الالتهاب في الدماغ له عواقب وخيمة .	0.44
2	وعلى الرغم من الكثير من البحث والتجارب المحتملة من العقاقير لا يوجد علاج واحدا يمكن أن يبطئ المعدل الذي تموت به الخلايا العصبية.فما هو تعوقنا من تطوير هذه العلاجات الجديدة؟ هناك العديد من المشاكل ولكن ربما أهمها هو أننا لا نفهم تماما كيف يعمل هذا المرض، وكيفية تصميم نموذج المرض في المختبر.و قد يكون السبب في ذلك أن العديد من الأدوية قد فشلت في التجارب السريرية – فيتوضح أن ما نحتاج إليه هو نهج جديد للعلاج .من خلال النظر في الدم أو أدمغة المرضى يمكن أن نحصل على فكرة عن ما يسبب الخطأ في مرض الزهايمر وبالتالي تصميم علاجات جديدة محتملة.من خلال الدراسات **نحن نعلم الآن أن بروتين يسمى Aß و لسبب غير معروف يتراكم في دماغ مرضى الزهايمر، ويشكل كتل كبيرة تسمى لويحات**. فكرة واحدة تشير إلى أن تراكم هذه اللوحات يتسبب في تدمير الخلايا العصبية، وبالتالي إزالة هذه اللوحات يمكن أن تحمي الخلايا العصبية من الموت.نتيجة لهذه الملاحظة، يجري حاليا وضع العلاجات لتفريق هذه الودائع أو تقليل كمية Aß المنتجة.كان واحدا من أكثر الأفكار ثورية و فعالية هو "تلقيح" مرضى الزهايمر ضد هذه اللوحات.هذه عملية مشابهة جدا لما يحدث عندما يتم التطعيم ضد مرض مثل السل TB .لقاح السل يعلم الجهاز المناعي على التعرف وتذكر نسخة من البكتيريا الميتة التي تسبب هذا المرض.هذا يعني أنه عندما مواجهة البكتيريا الحقيقية، الخلايا المناعية يمكنها محاربة العدوى .	0.38

3	مكافحة مرض الزهايمر؟ الحصول على الجهاز المناعي آمنJames FullerJames Fuller يناقش أبحاثه في تطوير أجسام مضادة ضد مرض الزهايمر في المادة الثالثة والأخيرة من المقالة للحصول على جائزة Max Perutz الكتابة العلوم عام 2012.تخيل العيش مع العلم انه خلال العقد القادم سيتم تدمير دماغك ببطء عن طريق جسدك .كما تنطفء الخلايا العصبية مثل الشموع، ماذا ستخسر بعد ذلك؟ هل ستكون الذكريات الثمينة؟ القدرة على أداء مهمة كل يوم؟ ربما جانب من شخصيتك؟ عائلتك وأصدقائك سوف يشاهدون ببطء عجز وتآكل الشخص الذي يحبون.تخيل مع تقدم البحث الآن إلى أنه مع كل خبراتنا الطبية ليس هناك شيء يمكننا القيام به .لا يوجد علاج واحد يبطئ هذا التدهور .	0.36
4	لقاح السل يعلم الجهاز المناعي على التعرف وتذكر نسخة من البكتيريا الميتة التي تسبب هذا المرض.هذا يعني أنه عندما مواجهة البكتيريا الحقيقية، الخلايا المناعية يمكنها محاربة العدوى .أثناء هذه العملية، يجري انتاج الأجسام المضادة التي تربط بالبكتيريا؛ ثم يتم تنشيط الجهاز المناعي لابتلاع و تدمير الأجسام المضادة للبكتيريا المغلفة ومنعكم من الحصول على المرض.عند هذه النقطة هل يمكن أن نتساءل، ما لا يجب أن نفعل مع مرض الزهايمر؟يمكن للعلماء إنتاج الأجسام المضادة في المختبر التي تربط أي شيء تقريبا : فيروس، خلية سرطانية **أو حتى لويحات توجد في أدمغة مرضى الزهايمر** .عندما يتم التعامل مع الأجسام المضادة لمريض الزهايمر التي ترتبط باللويحات، هذه اللوائح تصبح المغلفة للأجسام المضادة .	0.35
5	بواسطة إجراء تعديلات صغيرة لهيكل الأجسام المضادة يمكن أن تتحكم في كيفية استجابة الجهاز المناعي إلى العلاج .نأمل أن هذه الأجسام المضادة الجديدة **تزيل اللويحات دون إحداث مزيد من الضرر في الدماغ.التقدم في مجال الرعاية الصحية يزيد في طول الفترة الزمنية التي نعيش،** ولكن نوعية الحياة لدينا في السنوات المتقدمة لم يزد بنفس المعدل .الخرف هو واحد من أكبر التحديات التي تواجه العلم والمشكلة سوف تزداد سوءا إذا لم يتم العثور علاجات جديدة قريبا.اختبار قوة الجهاز المناعي باستخدام الأجسام المضادة تكون واحدة من هذه العلاجات، ولقد كان هذا بالفعل استراتيجية فعالة في علاج الالتهابات البكتيرية، والسرطان والتهاب المفاصل الروماتيكي .	0.32
6	ن**أمل أن هذه الأجسام المضادة الجديدة تزيل اللويحات دون إحداث مزيد من الضرر في الدماغ.**التقدم في مجال الرعاية الصحية يزيد في طول الفترة الزمنية التي نعيش، ولكن نوعية الحياة لدينا في السنوات المتقدمة لم يزد بنفس المعدل .الخرف هو واحد من أكبر التحديات التي تواجه العلم والمشكلة سوف تزداد سوءا إذا لم يتم العثور علاجات جديدة قريبا.اختبار قوة الجهاز المناعي باستخدام الأجسام المضادة تكون واحدة من هذه العلاجات، ولقد كان هذا بالفعل استراتيجية فعالة في علاج الالتهابات البكتيرية، والسرطان والتهاب المفاصل الروماتيكي .لو تمكنا من جعل هذا النوع من العلاج آمن لعلاج مرض الزهايمر، فإنه سيكون خطوة جيدة إلى الأمام للحد من معاناة الملايين من المرضى في جميع أنحاء العالم .	0.26
7	أثناء هذه العملية، يجري انتاج الأجسام المضادة التي تربط بالبكتيريا؛ ثم يتم تنشيط الجهاز المناعي لابتلاع و تدمير الأجسام المضادة للبكتيريا المغلفة ومنعكم من الحصول على المرض.عند هذه النقطة هل يمكن أن نتساءل، ما لا يجب أن نفعل مع مرض الزهايمر؟يمكن للعلماء إنتاج الأجسام المضادة في المختبر التي تربط أي شيء تقريبا : فيروس، خلية سرطانية **أو حتى لويحات توجد في أدمغة مرضى الزهايمر** .عندما يتم التعامل مع الأجسام المضادة لمريض الزهايمر التي ترتبط باللويحات، هذه اللوائح تصبح المغلفة للأجسام المضادة .**الاستجابة المناعية اللاحقة لمسح دماغ المريض من هذه اللوحات يحتمل أن تكون خطرة.**رائع، نستطيع الآن استخدام هذه الأجسام المضادة لمساعدة المرضى؟ ليس جيدا تماما ...في حين أن الأجسام المضادة تمسح بكفاءة اللويحات من الدماغ، إلا أنه قد يخفض من نجاحها الآثار الجانبية.لإزالة اللويحات من الأجسام المضادة المغلفة من الضروري تنشيط خلايا من الجهاز المناعي وهذا يمكن أن يسبب التهاب في الدماغ .	0.20
8	**الاستجابة المناعية اللاحقة لمسح دماغ المريض من هذه اللوحات يحتمل أن تكون خطرة.**رائع، نستطيع الآن استخدام هذه الأجسام المضادة لمساعدة المرضى؟ ليس جيدا تماما ...في حين أن الأجسام المضادة **تمسح بكفاءة اللويحات من الدماغ،** إلا أنه قد يخفض من نجاحها الآثار الجانبية.لإزالة اللويحات من الأجسام المضادة المغلفة من الضروري تنشيط خلايا من الجهاز المناعي وهذا يمكن أن يسبب التهاب في الدماغ .في حين أن هذا الالتهاب هو في الواقع مفيد في إزالة البكتيريا مثل السل، إلا أن الالتهاب في الدماغ له عواقب وخيمة .الالتهاب في مرضى الزهايمر يحدث حول الأوعية الدموية في الدماغ، يسبب الضرر والنزيف، الأمر الذي يحتمل أن يؤدى إلى مزيد من التدهور وفقدان الذاكرة.بالنسبة لمشروع الدكتوراه فأنا أنمي أجسام مضادة جديدة لمنع هذه الآثار الجانبية المؤذية .	0.19

5.4.1 Approach at a glance

This section presents the three-steps process used to construct a CG representation from a given text (a question or a passage text) in order to perform semantic comparison. These steps are illustrated in Figure 32.

The process contains three steps as shown in Figure 32. These steps are preceded by splitting the text of the question or the given passage into sentences.

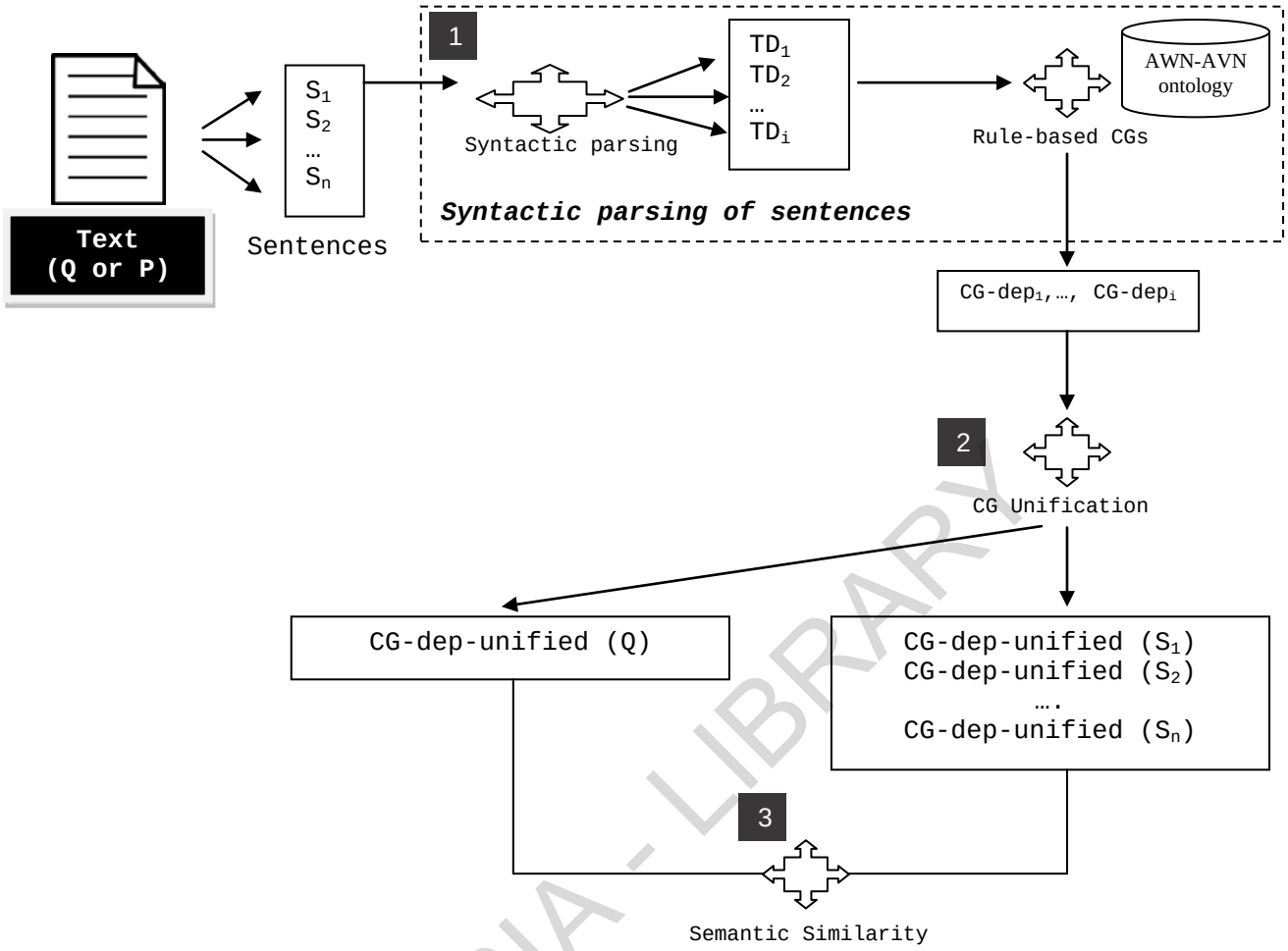


Figure 32. Representation of text in CG

5.4.1.1 Syntactic dependencies CGs

Each sentence is syntactically parsed using the Stanford parser (Manning and Jurafsky, 2012) which is an open source syntactic tool for English, Arabic and Chinese. It provides the parsing tree, the word tags as well as the Typed Dependencies (TD) [2] (see sample TDs in Table 42).

For example, the pair of words in bold {جسم, من} has the relation « *pobj* ». This means that the *dependant* (i.e., جسم) is a preposition object of the *head* (i.e., من). The parser also provides the tag of each word. For instance, the word "يتم" has the tag "VBP" which refers to a "*Verb, non-3rd person singular present*" in the present tense, "الإنسان" has the tag "DTNN" which refers to

[2] The Stanford typed dependencies representation was designed to provide a simple description of the grammatical relationships in a sentence that can easily be understood and effectively used by people without linguistic expertise who want to extract textual relations. The meaning of the tag dependencies adopted by the Stanford Arabic Parser is given in Appendix C.

a “Noun, singular or mass with Determiner”, اللويحات for “Proper noun, plural with Determiner”, etc.

Table 42. Dependencies provided by the Stanford Arabic Parser applied on the sample question

dependant	head	type_dep	tag_dependant	tag_head
ما	Root	Root	VBP	
هو	ما	iobj	PRP	VBP
الجزء	ما	dobj	DTNN	VBP
من	الجزء	prep	IN	DTNN
جسم	**من**	**pobj**	**NN**	**IN**
الإنسان	جسم	dep	DTNN	NN
الذي	يتم	nsubj	WP	VBP
يتم	الإنسان	rcmod	VBP	DTNN
فيه	يتم	dobj	NN	VBP
تشكيل	فيه	dep	NN	NN
اللويحات	تشكيل	dep	DTNNS	NN

We are interested in the transformation of these dependencies into subCGs that we call CG-dep_i (dep_i refers to the CG of the dependency i in the parsing result). This transformation is performed using a rule-based technique similar to the one proposed by Hensman and Dunnion (2004) and adapted for the Arabic language. Table 45 shows some examples of the 11 rules set (see Appendix B for the full description of these rules) for this purpose. For example, rule #4 states that if the Gouvernor Tag (GTag) is a tag of a verb (such as VBP) and the Dependant Tag (DTag) is equal to “NN” and the dependency type returned by the Stanford parser is “dobj” (Direct object), then two cases occur:

- The dependent tag is neither “NNP” nor “NNPS”: in this case the conceptual graph of the dendency is constructed following the pattern:

  CG-dep = [cg : [Conc(G)]<-objOf-[Conc(D)]]

- The dependent tag is “NNP” or “NNPS”: in this case the dependent is tagged by the Stanford parser as a singular or plural proper noun respectively, therefore, we follow the pattern:

  CG-dep = [cg : [SupConc(D) : D]<-objOf-[Conc(G)]]

  Where Conc(G)is the concept related to the word having the role “Gouvernor”, Conc(G)is the concept related to the word having the role “Dependant” and SupConc(D)is the super concept of the NE corresponding to the dependant in the AWN-AVN ontology. Let us recall that almost all NEs were extracted from YAGO (see Chapter 4).

The concepts related to a word are matched using its morphological analysis[3] that has the same PoS tag provided by the Stanford parser. For instance, the CG related to TD_{11} in Table 43 has the pattern:

$$CG\text{-}dep_{11} = [Conc(تشكيل)]<\text{-}attributeOf\text{-}[Conc(اللويحات)]$$

Where Conc(تشكيل) is the concept "إعداد_id380" and Conc(اللويحات) does not match any concept in the ontology. In this case, we set Conc(اللويحات) to "اسم" which is the general concept of all nouns in the ontology. The real CG becomes:

$$CG\text{-}dep_{11} = [إعداد\_id380]<\text{-}attributeOf\text{-}[اسم]$$

Table 43. Rule-based subCG generation applied on the sample question

ID dep	dependant	head	type_dep	tag_dependant	tag_head	Applied rule
1	ما	Root	Root	VBP		None
2	هو	ما	iobj	PRP	VBP	None
3	الجزء	ما	dobj	DTNN	VBP	**Rule #4** CG-dep = [Conc(ما)]<-objOf-[Conc(الجزء)]
4	من	الجزء	prep	IN	DTNN	**Rule #9** CG-dep = [prep : *p_1 "من"]
5	جسم	من	pobj	NN	IN	None
6	الإنسان	جسم	dep	DTNN	NN	**Rule #3** CG-dep = [Conc(جسم)]<-attributeOf-[Conc(الإنسان)]
7	الذي	يتم	nsubj	WP	VBP	**Rule #5** CG-dep = [Conc(الذي)]<-agentOf-[Conc(يتم)]
8	يتم	الإنسان	rcmod	VBP	DTNN	**Rule #10** CG-dep = [Conc(يتم)]-attributeOf->[cg : Conc(الإنسان)]
9	فيه	يتم	dobj	NN	VBP	**Rule #4** CG-dep = [Conc(يتم)]<-objOf-[Conc(فيه)]
10	تشكيل	فيه	dep	NN	NN	**Rule #6** CG-dep = [Conc(فيه)]<-is-[Conc(تشكيل)]
11	اللويحات	تشكيل	dep	DTNNS	NN	**Rule #3** CG-dep = [Conc(تشكيل)]<-attributeOf-[Conc(اللويحات)]

5.4.1.2 CG unification

Once all the $CG\text{-}dep_i$ are constructed, we need a unification of these CGs to generate a unique CG representing the given sentence. For this purpose, a "Join" operation over these

[3] Can be downloaded from http://sourceforge.net/projects/alkhalil/

CGs is processed. To show an example of the result provided by this CG operation, let us take two sample CGs[4]:

```
CG cg1 :
  [Drive]-
       -obj->[Car],
       -agnt->[Human :{Bob, Andre}]

CG cg2 :
  [Drive]-
      -manr->[Fast],
      -agnt->[Boy :{Bob, John, Sam}]
```

The result of the "GetMaximalJoin" operation between cg1 and cg2 is the following CG:

```
[Drive]-
       -obj->[Car],
       -agnt->[Boy :Bob],
       -manr->[Fast]
```

The resulting CG is the union of the two CGs and is composed of: (i) the subgraph that is common to cg1 and cg2 on matched concepts (i.e., concepts Car and also Human since it is a generalization of the concept Boy), (ii) a copy of parts that are specific to cg1 (none in this case), and (ii) a copy of parts that are specific to cg2 (i.e., -manr->[Fast] in this case).

Once the unification step is performed, we have now a unified CG (`CG-dep-unified (Q)` and `CG-dep-unified(S)` in figure 32) for the question and unified CGs for each sentence in the candidate passages. The next section details the used semantic similarity score and the adaptations introduced.

5.4.1.3 Semantic similarity score

The comparison between the CG of the question `CG-dep-unified(Q)` and that of each sentence in candidate passages `CG-dep-unified(S)` is based on the measure of the semantic similarity using the formula proposed by Montes-y-Gómez et al. (2001) considering some adaptations. Therefore, in order to compare two CGs, we need to identify the concepts and the relations that are common among these two CGs. These concepts and relations appear in the result of the generalization of the two CGs. Hence, before measuring this semantic

[4] Example from the Amine Platform Web Site: http://amine-platform.sourceforge.net/

similarity, we perform the *Generalization* operation between CG-dep-unified(Q) and CG-dep-unified(S) to get the *generalized* CG G_c.

To measure the similarity between two CGs, say CG-Q and CG-P, we first calculate the Conceptual Similarity (S_c) and Relational Similarity (S_r) between each graph to the *generalized* graph G_c, and, second, we calculate the overall similarity between CG-Q and CG-P based on these two similarities. The formula of S_c and S_r are analogous to the well-known Dice coefficient (Montes-y-Gómez et al. 2001):

$$s_c = \frac{2n(G_c)}{n(G_1)+n(G_2)}$$

Where $n(G_c)$is the number of concepts in the graph G. In our case G_1=CG-Q and G_2=CG-P. This conceptual similarity ranges from 0 if $n(G_c)=0$ and 1 if $n(G_1) = n(G_2)$.

As for the S_r similarity, it calculates the degree of connection between concepts nodes in the *generalized* CG (G_c) and the degree of connection between the same concepts in the original CGs (i.e., CG-Q and CG-P). This similarity is calculated through the formula:

$$s_r = \frac{2m(G_c)}{m_{G_c}(G_1)+m_{G_c}(G_2)}$$

Where m(Gc) is the number of relational nodes in the *generalized* CG, and m_{Gc} (Gi) is the number of the relational nodes in the immediate neighborhood of the graph G_c in the graph Gi (i.e., G1=CG-Q or G2=CG-P) which is the subgraph of Gi directly linked to a concept from G_c.

The combined semantic similarity (i.e., the combination of both the conceptual and the relational similarities) is then calculated using the formula:

$$S = S_c \times S_r$$

Montes-y-Gómez et al. (2001) proposed a modified version of this formula to give more importance to the conceptual similarity at the expense of the relational one. In our implementation, we consider another type of adaptation for the combined semantic similarity. The objective behind this adaptation is the introduction of the factor that we call Typed Dependencies-based Rule Confidence (TDRC). This factor is based on the idea that each of the 11 rules previously used to construct the CG-Q(dep) and CG-P(dep) is assigned a Rule Confidence (RC). This is set starting from the confidence of the syntactic parsing for

Arabic text as observed in the training questions (i.e., the 2012 QA4MRE questions that allowed us to deduce these rules). Thus, the confidences are given in Table 44.

As we can see, Table 44 provides a sorted list of the previously described rules according to the assigned confidence. There are 8 rules (out of 11) that have a confidence higher than 0.9, the remaining rules have lower confidences (0.833, 0.750 and 0.667 respectively).

Table 44. Confidences assigned to typed dependencies-based rules

Rule	TD Condition	Rule Confidence
Rule #10	"dependency-type={prep}"	0.990
Rule #3	"GTag ={NN} and DTag ={DTNN,DTNNS}"	0.986
Rule #11	"dependency-type={rcmod} and DTag={V*}"	0.974
Rule #4	"GTag ={V* } and DTag ={NN} and dependency-type=dobj"	0.973
Rule #9	"DTag ={JJ } and dependency-type={amod}"	0.971
Rule #8	"DTag ={CD }"	0.970
Rule #1	"GTag=JJ and DTag=NN"	0.941
Rule #6	"GTag ={NN } and DTag ={NN}"	0.917
Rule #7	"GTag ={CD }"	0.833
Rule #5	"GTag ={V* } and dependency-type={iobj, nsubj, dep, xcomp}"	0.750
Rule #2	"GTag ={NN,NNS} and DTag ={NNP,NNPS}"	0.667

Since a unified CG-Qu(dep) or a CG-Pu(dep) is constructed from many typed dependencies-based rules, we set the TDRC factor of the unified CG-dep as the average of the TDRC of each $CG\text{-}dep_i$. The latter is calculated using the fomula:

$$TDRC(CG\text{-}dep_i) = RC(r)*CC$$

Where $CG\text{-}dep_i$ is the CG constructed from a TD i using the rule r that has the confidence RC; CC is a concept confidence factor which is equal to:

$$CC = 1-(N_c * 0.1 / N)$$

Where N is the number of concepts used to represent the $CG\text{-}dep_i$ and N_c is the number of "اسم" (noun) and "فعل" (verb) concepts (fictive concepts) that we used in this CG, replacing the unmatched concepts in the ontology (i.e., when there is no concept matching the given head or dependent word in the typed dependency). This confidence ranges from 0.9 (in case the two concepts are fictive, i.e., N_c=2) to 1 (in case the two concepts are successfully matched in the ontology, i.e., N_c=0). The overall Semantic Similarity (SSim) is, therefore, expressed as follows:

$$SSim(CG) = S \times Avg(TDRC_{depi})$$

Where $Avg(TDRC_{depi})$ is the average TDRC among the graphs $CG\text{-}dep_i$.

5.4.2 Experiments

5.4.2.1 QA4MRE@2013 test-set

A) Questions

The present section is devoted to the presentation of the obtained results after applying the semantic-based level described above on a test-set of questions with a challenging complexity requiring this kind of semantic processing. Let us recall that the surface-based levels were evaluated using two test-sets: (i) the CLEF-TREC 1999-2008 questions containing 2,264 questions among which a large number are factoid questions, and (ii) the QA4MRE test-set edition 2012 that contains 160 questions with a more challenging complexity for surface-based approaches. Thus, we used the latter test-set as training set to design the rule-based CG construction. In order to conduct significant experiments for the semantic-based level, we adopted a similar test-set, i.e., the 2013 version of the QA4MRE test-set that contains 284.

As in the 2012 campaign, the task focuses on the reading of single documents and the identification of the answers to a set of questions about information that is stated or implied in the text. Questions are in the form of multiple choice, where a significant portion of questions have no correct answer among the given alternatives proposed. While the principal answer is to be found among the facts contained in the test documents provided, systems may use knowledge from additional given texts (the 'Background Corpus') to assist them with answering the questions. In our experiments, we did not consider such background corpus. Some questions also test a system's ability to understand certain propositional aspects of meaning such as modality and negation. Such aspects are not considered in our semantic-based level. Figure 33 illustrates the distribution of questions among the considered topics.

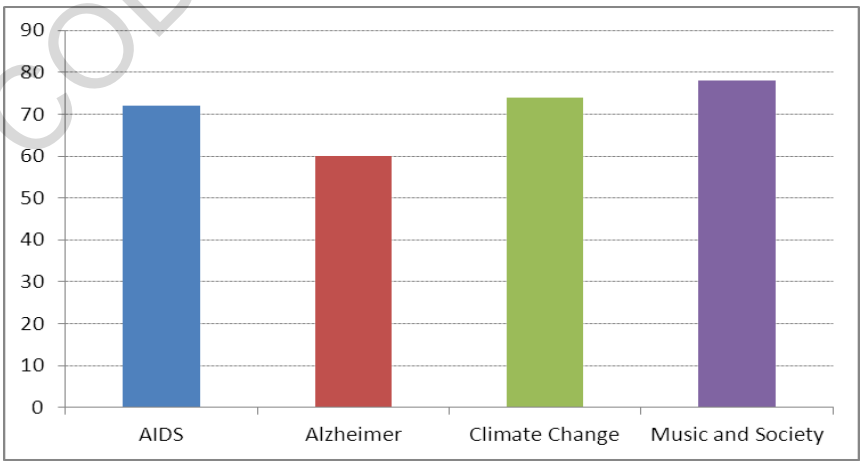


Figure 33. Distribution of the QA4MRE@2013 questions over topics

Similarly to the 2012 test-set, the 2013 set is composed of 4 topics, namely “Aids”, “Climate change” and “Music and Society” and “Alzheimer”. Each topic will include 4 reading tests. Each reading test will consist of one single document, with at least 15 questions and a set of five choices per question. There are 44 auxiliary questions that are duplicates of the main questions, but without required inference, allowing to test the ability of systems to use inference and its impact on the question treatment.

B) Results

For each question in the test-set, we perform the surface-based level. From the set of the resulting passages we extract a subset of 15 passages that are assigned the best surface similarity score. Thereafter, we perform, either for the question and the considered passages, the two steps of the semantic-based level, i.e., representing text in CG and comparing the CG-Q with CG-P.

B.1) Surface-based evaluation

The considered questions contain on average 9 words. The keyword-based level tries the generation of new related terms for each word based on the QE process relying on the AWN semantic relations. In this experiment, we also considered the new content after the AWN enrichment proposed in Chapter 4. The process generates an average of 14 new terms for each word. This allowed the extension of 234 questions (i.e., 82.39% of the overall test-set). After performing the structure-based level, the surface similarity score allowed ranking the passages to have the best 15 passages. The five possible answers for the question are then validated against the first five passages. Table 45 shows these results.

Table 45. Results of the surface-based evaluation for the 2013 QA4MRE test-set

	Number	Percent.	Remark
Questions	284	-	
Questions Extended by extended AWN	234	82.39%	out of 284
Avg words extended	2		
Avg new words generated by question	102	-	
Questions ANSWERED by Surface-based level	164	57.75%	out of 284
Questions UNANSWERED by Surface-based level	120	42.25%	out of 284
Questions Correctly ANSWERED	21	7.39%	out of 284
"		12.80%	out of 164
c@1	0.11		

As we can see, only 12.8% of the answered questions are correctly answered. If we also consider the unanswered questions, this percentage decreases to 7.39%. The obtained c@1 measure is around 0.11. These results show the limits of the surface-based regarding the processing of non-factoid questions. Indeed, the used test-set only contains 15 factoid

questions. Another challenging point is the fact that 217 of the 284 questions (76.4%) have a multi-word answer according the goldstandard. This may have an impact on the answer validation step. On the other hand, the correctly answered questions have a quite similar distribution over the four considered topics as illustrated in Figure 34.

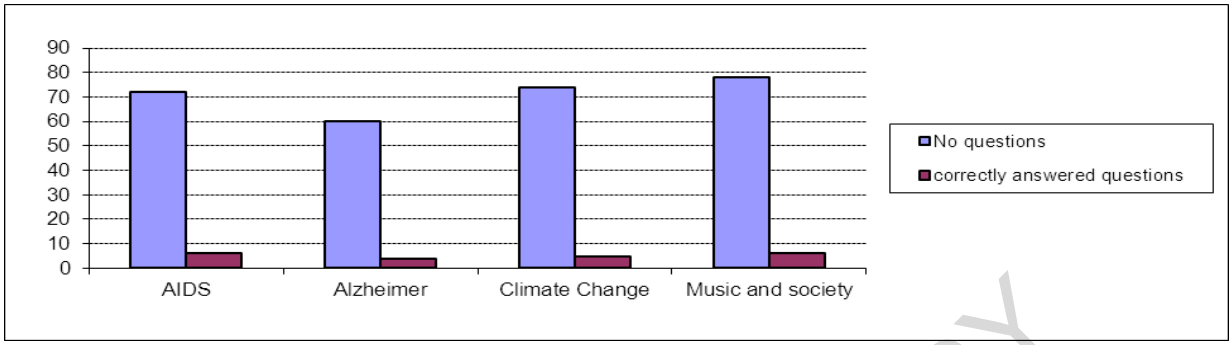


Figure 34. Correctly answered questions over topics

The runs figure out that among the 2,388 extracted passages (there are many questions for which we could not extract 15 passages as expected before experiments), only 9 were assigned a surface similarity score over 0.9, 60 were assigned a similarity higher than 0.7 as shown in Figure 35.

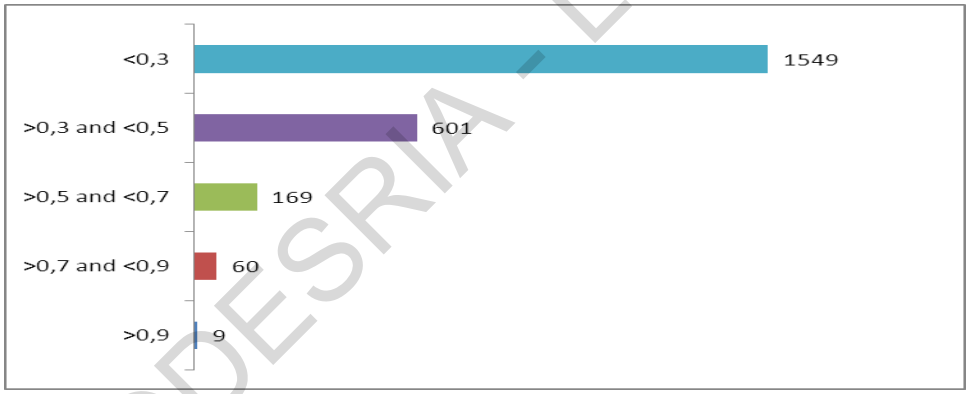


Figure 35. Distribution of the obtained Surface Similarity Score

This shows the huge difference between questions and passages in terms of keywords and structure which is challenging for a surface-based approach. The adoption of a semantic-based level as previously described in this chapter has the aim to overcome such challenges. The results obtained are described and discussed in the following section.

B.2) Semantic-based evaluation

For each question, we extract the best 15 passages according to the surface similarity score (for 35% of the questions we could not extract more than 8 passages, the average of the extracted passages per question is 10).

We performed syntactic parsing by means of the Stanford parser for the set of 284 questions and their corresponding 2,734 passages. The parsing of the passages was preceded by splitting them into phrases in order to increase the accuracy of parsing. The statistics of the questions and passages that were matched by our typed dependencies rules are listed and illustrated below.

Table 46 shows the high coverage of the Stanford parser that allowed getting parsing solutions for around 98.6% of the questions and 83.1% of the passages. For the remaining questions and passages, the parser could not process the text due mainly to the limit reached in terms of text length despite the splitting of passages into phrases.

Table 46. Applied typed dependencies rules for questions and passages

	Questions (Q)		Passages (P)	
	Number	%	Number	%
Set	284	-	2,734	-
- Q or P matching rules	280	**98.59%**	2,272	**83.10%**
Typed Dependencies (TD)	2,632	-	25,008	-
- TDs matching rules	1,473	**55.97%**	15,156	60.60%
-> Rule #1	69	4.68%	6,471	42.70%
-> Rule #2	152	10.32%	14,229	93.88%
-> Rule #3	222	15.07%	7,786	51.37%
-> Rule #4	196	13.31%	7,734	51.03%
-> Rule #5	255	**17.31%**	14,274	**94.18%**
-> Rule #6	174	11.81%	7,215	47.60%
-> Rule #7	9	0.61%	355	2.34%
-> Rule #8	40	2.72%	1,837	12.12%
-> Rule #9	72	4.89%	7,346	48.47%
-> Rule #10	222	15.07%	12,357	81.53%
-> Rule #11	62	4.21%	3,432	22.64%
- Rules overlap rate	4.82%		8.76%	

Figure 36. Distribution of question' typed dependencies over rules

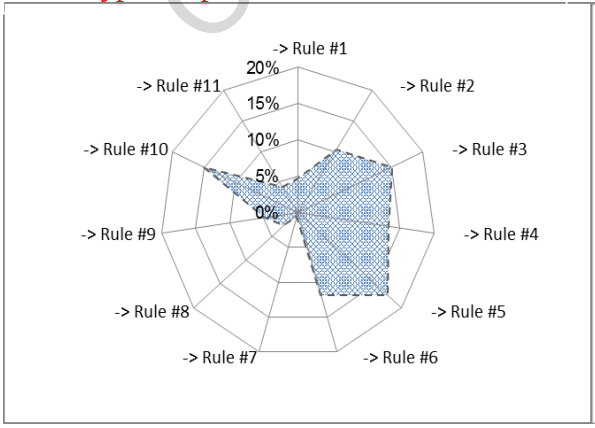
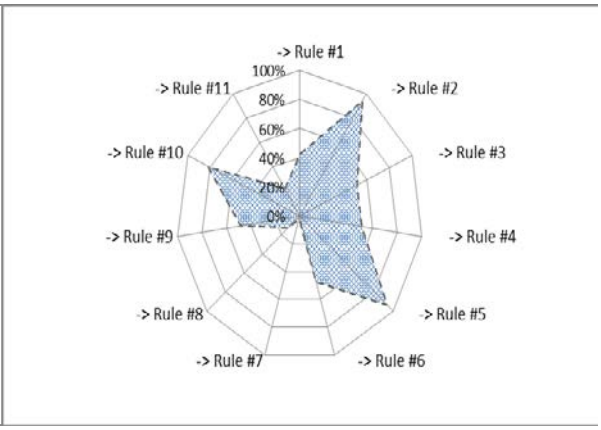


Figure 37. Distribution of passages' typed dependencies over rules



For both questions and passages, all the 11 rules were applied at least once. As illustrated in Figure 36 and Figure 37, the most applied one is Rule #5 in both sets (17.31% of the matched TD in question and 94.18% in passages). The ranking of 4 rules (Rule #1, Rule #7, Rule #8 and Rule #11) is also the same in both sets. Note that for around 5% of question TD and 43% of question TD, more than one rule was applied for the same TD. This is due to the fact that in some cases, the two rule conditions match the given TD. This mainly concerns Rule 1# and Rule #9. To match words (dependent and governor words) with their corresponding concepts, the morphological analysis of the 2,734 extracted passages provided 600,399 possible solutions. The distribution of these solutions over PoS (64.6% are nouns and 32.9% are verbs) is quite similar to the one registered in the questions. The number of distinct stems in these solutions is 4,306. The matching process recognized 322 question stems in the Arabic VerbNet resource and 1,252 stems in the corresponding passages. The details of this matching are presented in Table 47.

Table 47. Cross resource matching statistics – AVN matching

	Questions		Passages	
	Number	%	Number	%
Distinct stems	873	-	4,306	-
Matched in AVN	322	**37%**	1,252	**29%**
-> verb-matching	139	43%	511	41%
-> deverbal-matching	147	46%	547	44%
-> participle-matching	36	11%	194	15%

Around 43% of the recognized stems in the questions were matched using the verb-matching, 46% approx. using the deverbal-matching and only 11% using the participle-matching. As for passages, there is a number of 1,252 matched stems which is lower in percentage (29%) than that registered for questions (37%). Nevertheless, the distribution of this number over the different types of matching is quite similar (41% using verb-matching, 44% using deverbal-matching and 15% using participle-matching).

The second part of the matched words were recognized using the AWN content in our ontology. This consists of considering the Standard and Enriched versions of AWN as shown and illustrated in Table 48 and Figure 38.

Table 48. Cross resource matching statistics – AWN matching using Standard and Enriched versions

		Covered by AWN			
	No Distinct Stems	Standard AWN	%	Enriched AWN	%
Questions	873	399	45.70%	568	65.06%
Passages	4,308	1,559	36.19%	2324	53.95%

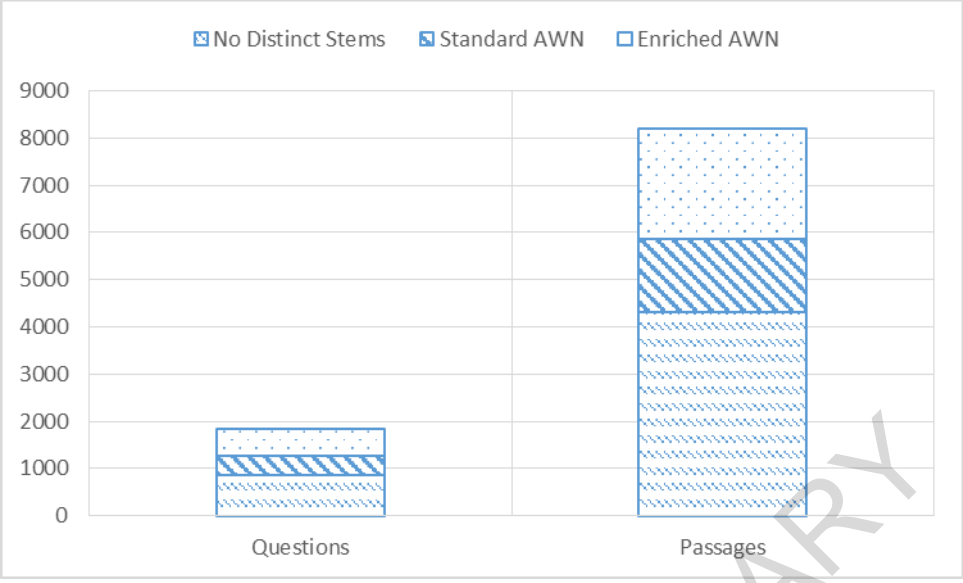


Figure 38. Comparison between stem coverage in Standard and Enriched AWN

The Standard AWN covers around 46% of the question stems and 36% of passage stems. This percentage is even better with the version of AWN that were enriched by nouns (including Broken Plurals), verbs and NEs (the description of this enrichment is provided in Chapter 4) reaching 65% of question stems and roughly 54% for passages. This shows the effectiveness of enriching this resource, not only for QE as shown in the evaluation presented in Chapter 4, but also for the application of the different steps of our semantic-based approach.

B.3) Three-levels performance

After representing the question and candidate passages in terms of CGs, we performed the semantic similarity score proposed by Montes-y-Gómez et al. (2001) between both CGs. Thereafter, we measure the performance of the system using the surface-based approach described in Chapter 3 and, then, after using the semantic approach based on this semantic similarity. Table 49 and Table 50 display the obtained results.

As we can see, the semantic-based approach that uses the CG representation (constructed through the built ontology) improves the performance in terms of the percentage of correctly answered questions from 7.39% of the 284 questions to 16.2%.

Another aspect that deserves to be mentioned is the high percentage of questions that were given an answer by the system (77.11% versus 57.75%). The improvement was also registered regarding the c@1 measure which penalizes systems providing wrong answers. The semantic-based approach obtained 0.20 c@1 (versus 0.11 with the surface-based approach).

Table 49. System performance using the surface-based levels on CLEF 2013

	Number	Percent. / Accuracy	Remark
Questions	284	-	
Questions ANSWERED by IDRAAQ (based on AWN)	164	57.75%	out of 284
Questions UNANSWERED by IDRAAQ (based on AWN)	120	42.25%	out of 284
Questions Correctly ANSWERED	21	7.39%	out of 284
		12.80%	out of 164
c@1	0,11	-	-

Table 50. System performance using also the semantic-based level on CLEF 2013

	Number	Percent. / Accuracy	Remark
Questions	284	-	
Questions ANSWERED by IDRAAQ (based on AWN)	219	77.11%	out of 284
Questions UNANSWERED by IDRAAQ (based on AWN)	65	22.89%	out of 284
Questions Correctly ANSWERED	46	16.20%	out of 284
		21.00%	out of 219
c@1	0.20	-	-

5.4.2.2 TREC and CLEF 1999-2008 test-set

We conducted the same experiment using the first set of questions described in Chapter 3 and Chapter 4. This set contains 2,264 CLEF and TREC questions from various editions (1999 to 2008). Table 51 recalls the results of the surface-based levels.

Table 51. System performance using the surface-based levels on CLEF-TREC 1999-2008

Measures	Baseline PR System	PR using the Keyword-based and Structure-based levels (based on AWN)	
		Original AWN	Enriched AWN
Accuracy	9.66%	17.49%	26.76%
MRR	3.41	7.98	11.58
Nr. AQ	20.27%	23.15%	35.94%

After performing the semantic level, the system was able to correctly answer around 38% of the questions versus approx. 36% answered with the surface-based levels with respect to the enriched version of AWN. The details of this performance is given in Table 52.

Table 52. System performance using also the semantic-based level on CLEF-TREC 1999-2008

	Number	Percent. / Accuracy	Remark
Questions	2,264	-	
Questions ANSWERED by IDRAAQ (based on AWN)	1,489	65.77%	out of 2264
Questions UNANSWERED by IDRAAQ (based on AWN)	775	34.23%	out of 2264
Questions Correctly ANSWERED	861	38.03%	out of 2264
		57.82%	out of 1489
c@1	0.51	-	-

The main gain was in terms of c@1 which is significantly improved from 0.2 in the previous test-set to 0.51 in the CLEF-TREC 1999-2008 test-set. This can be explained by two facts:

- The current test-set contains factoid questions that are simpler to process than the complex questions contained in the CLEF 2013 test-set;
- The Web passages used in this experiment are smaller in size than the passages used in the previous test-set, so the syntactic parsing is more accurate in this case and helps in generating precise CGs;
- The CLEF-TREC 1999-2008 test-set is the set used to analyze AWN shortcomings and to enrich this resource. The new content added in AWN better matches the words existing in this test-set and, therefore, CGs are constructed with concrete concepts instead of fictive ones (i.e., اسم and فعل). This avoid us to introduce the concept confidence which decreases the semantic similarity (see Section 5.4.1.3).

5.4.2.3 Comparison with Arabic QA systems

In terms of c@1, our three-levels approach performs better than existing systems, such as the one proposed by Trigui et al. (2012)[5] that was designed for factoid questions and shallow Arabic QA. This system obtained a 0.19 c@1 in the QA4MRE 2012 edition that is under the performance that we obtained: 0.20 obtained with the QA4MRE 2013, 0.21 with the QA4MRE 2012 and 0.51 with the CLEF-TREC 1999-2008 testset containing a higher percentage of factoid questions.

[5] The only system that used the same test-set, i.e., CLEF. For the other Arabic QA systems, experiments were conducted using a specific and small question test-set that cannot be representative for such evaluation.

5.5 Chapter summary

In this chapter, the semantic-based level was implemented on top of the surface-based level in order to address the shortcomings registered in the experiments when processing complex questions. Thus, we constructed a new ontology from the AWN and AVN resources with the aim to support the semantic-based level. Indeed, in this level (i) we construct the conceptual graph of both the question and the candidate passages and (ii) we rank these passages according to the semantic similarity score calculated between CGs.

The construction of CGs follows the syntactic-based technique relying on a set of typed dependencies rules that we designed for the Arabic language starting from a training test-set of questions (i.e., CLEF QA4MRE 2013 test-set) and using the syntactic parsing provided by the Stanford parser.

The semantic similarity score is based on the formula proposed by Montes-y-Gómez (2001) with the introduction of a TDRC confidence composed of two factors: (i) the RC factor that considers the confidence of a given rule as observed in the training test-set, and (ii) the CC factor that introduces the concept confidence, i.e., the confidence of the CG is decreased when it is constructed using fictive concepts instead of real concepts matched in the AWN-AVN ontology.

In order to show the effectiveness of this semantic-based level, two experiments were conducted using different sets of questions:

- The CLEF 2013 test-set that contains complex types of questions and just a few number of factoid questions (22%). The answers are searched in the document collection provided by the QA4MRE workshop.
- The CLEF-TREC 1999-2008 test-set that have a large representativeness in terms of size and for which the answer is searched in the Web as a target collection. This allows for testing the semantic-level with real-world text.

The results obtained for both test-sets show an improvement of system performance in terms of the number of answered questions (+8.81% improvement in the CLEF 2013 test-set and +2.09% in the CLEF-TREC 1999-2008 test-set in comparison to the performance with the surface-based levels).

With respect to these experiments, the system provides answers to a high number of questions (65.77% and 77.11% respectively). However, the obtained c@1 measure (which penalizes systems providing wrong answers) is higher with the CLEF-TREC 1999-2008 test-set (0.51) than the CLEF 2013 test-set (0.2).

The performance of the three-levels approach proposed in this research is better with the factoid questions represented by the CLEF-TREC 1999-2008 test-set, with 38.03% of correctly answered questions (versus 16.20% in the CLEF 2013 test-set containing complex questions).

Finally, we can conclude that the three-level approach which is an hybrid combination of the surface and deeper approaches allows for obtaining a better performance than the baseline system. Indeed, the percentage of correctly answered questions registered a gain of +17.76%, moving from 20.27% using the baseline system to 38.03% using the three-level approach.

Chapter 6

The IDRAAQ system as an integrated application in SAFAR Platform

6.1 Introduction

In Chapters 3, 4 and 5, we have proposed and evaluated a new hybrid approach for effective passage retrieval in the context of Arabic QA. This approach combines the advantages of the surface-based techniques that provide better results with factoid questions and deeper techniques with semantic-based processing over questions and passages. The keyword-based, the structure-based and the semantic-based levels of our approach make use of different and heterogeneous resources and NLP tools. The integration of this material in an Arabic QA system or other sophisticated applications requires a suitable architecture and platform to reduce the complexity of use and to optimize the response time.

In this chapter, we propose a new Arabic QA system called "IDRAAQ" which is constructed on top of an integrated Arabic NLP platform. The objective behind this work is discussing four important issues that can affect such a project: (i) the importance of using an integrated NLP platform to reduce the complexity and time for developers since we need to use different NLP components for third parties, (ii) the impact of system architecture on the overall performance, especially the response time, (iii) the possibility of analyzing results considering an integrated environment and (iv) the contribution in further Arabic QA research or similar applications by providing IDRAAQ separate modules to the research community.

This chapter is structured as follows. Section 6.2 introduces the integrated NLP platforms and their main objectives as well as their support for the Arabic language. Section 6.3 presents the SAFAR platform used in this work. Section 6.4 details the proposed architecture of the IDRAAQ system as part of SAFAR platform and shows how the developed and separate modules of IDRAAQ can be later constituents of similar applications.

6.2 Background

In the context of NLP in general and Arabic NLP in particular, the following issues can be mentioned as example of trends that can help researchers in the development of more sophisticated applications:

- Unification of researchers effort in the different NLP communities;

- Development and making available of open source NLP programs and allowing the reuse of already programmed modules;
- Standardization of information representation formalisms for a better sharing of resources;
- Benchmarking systems following the evaluation campaigns guidelines, test-sets and measures.

6.2.1 Examples of NLP platforms

Following the above trends, there have been many propositions of integrated platforms for NLP developers. In the next sub sections, we briefly describe examples that are reported by (Ezzeldin and Shaheen 2012) as being usable for QA systems.

6.2.1.1 GATE

The General Architecture for Text Engineering (GATE)[1] project started in 1995 at the University of Sheffield with the proposition of a suite of developed java tools. This platform presents the advantage of handling various languages, including English, Spanish, Chinese, Arabic, Bulgarian, French, German, Hindi, Italian, Cebuano, Romanian and Russian. It also provides preprocessing tools for many document formats (such as TXT, HTML, XML, DOC, PDF) and databases.

GATE as a platform includes an information extraction system called ANNIE (A Nearly-New Information Extraction System) having the form of set of modules for English comprising: a tokenizer, a gazetteer, a sentence splitter, a PoS tagger, a named entities transducer and a co-reference tagger. The GATE platform also provides plugins for machine learning with Weka, RASP, MAXENT, SVM Light, and a fast LibSVM integration.

Regarding its use for the development of QA systems, GATE offers an implementation that can help in the querying of the Princeton WordNet lexical database as well as various search engines such as Google, Yahoo and Lucene.

6.2.1.2 Open NLP

Apache OpenNLP[2] is a machine learning based library for the processing of natural language text that supports many NLP tasks like tokenization, sentence segmentation, PoS tagging, NER, chunking, parsing, maximum entropy, perceptron based machine learning, and co-reference resolution.

[1] GATE Official Website: http://gate.ac.uk
[2] OpenNLP Official Website: http://opennlp.apache.org

OpenNLP consists in the proposition of many modules: Sentence Detector, Tokenizer, NER module, Document Categorizer, PoS Tagger, Chunker, Parser, Co-reference Resolution module, Corpora processor, Machine Learning (Maximum Entropy) module.

6.2.1.3 Stanford NLP Toolkit

The Stanford NLP Toolkit[3] is a group of libraries that cover the most common tasks of NLP, especially those needed by the QA task. Among the libraries included in Stanford NLP Toolkit:

- Stanford Parser implementing probabilistic natural language parsers, a PCFG and dependency parsers, and a lexicalized PCFG parser;
- Stanford PoS Tagger which is a maximum-entropy PoS tagger for English, Arabic, Chinese, French, and German;
- Stanford NE Recognizer which is a CRF sequence model with a list of features for NER in English and German;
- Stanford Word Segmenter which is a CRF-based word segmenter also supporting Arabic and Chinese;
- Stanford Classifier which is a machine learning classifier for text categorization, a maximum entropy and multi-class logistic regression model;
- Phrasal: a phrase-based machine translation system.

6.2.1.4 NooJ Platform

NooJ[4] is a freeware linguistic engineering development environment. It can process various text formats with the ability to annotation using the XML language. This platform also uses PERL-type regular expressions, NooJ regular expressions and NooJ grammars so that any morphological, lexical, syntactic or semantic information annotated in the text can be used inside NooJ expressions and grammars. NooJ provides a module (i.e., the Context-Free Grammars that are Recursive Transition Networks) that can help developers or users to recognize and annotate certain sequences of texts.

NooJ features some Arabic language resources. These resources are a sample text, a dictionary of 10,000+ verbs, their inflection in the form of a NooJ inflectional grammar, and a group of morphological grammars for verb prefixes and suffixes. Brini et al. (2009) used these platform grammars in the development of the Arabic QA system called “QASAL”.

6.2.2 Support of Arabic QA

As we have seen in the previous chapters, the development and evaluation of a QA system involves various NLP components (resources and tools). The usability of an integrated NLP

[3] StanfordNLP Group Official Website: http://nlp.stanford.edu/software/index.shtml
[4] NooJ Official Website: http://www.nooj4nlp.net

platform for such a development can be studied according to many criterion. The main criterion is the ability of the platform to support various and complicated pipelines of processing. Let us recall here that in the framework of our research, we adopted the general pipeline of a QA system architecture that is composed of three QA modules (Question Analysis and Classification, Passage Retrieval and Answer Extraction and Validation). In addition, we proposed a three-level approach in the context of the PR module. Such a complex architecture is hardly supported by the above-described platforms.

Another criterion that might be studied is the ability of these platforms to support the different particularities of the Arabic language. As we have seen, there are many existing NLP components that were useful for the development of our three-level approach for Arabic PR, starting by the QE process based on AWN, the DDN model implementation in the JIRS system and ending by the syntactic and semantic processing of questions and passages relying on the Stanford parser and the built AWN-AVN ontology. The integration of these components in the existing NLP platforms requires much efforts in the case of NooJ (already used for Arabic QA) since it is written in .NET and most of those components are written in java.

The architecture of GATE, OpenNLP and the Stanford NLP toolkit does not provide clear services that can be directly used by Arabic QA researchers. For example, to use the Stanford parser in our system, we dedicated much time for adapting the output of its processes in order to be passed to the other modules of our approach despite the parser and our program are both written in java. Furthermore, we were obliged to optimize the loading of the grammar used by this parser to have an acceptable response time when processing large number of sentences in the context of the conducted experiments.

In addition, there is a problem of high dependency between processes and resources that makes these platforms less flexible to support the development of an Arabic QA system with the involvement of different resources and tools.

6.3 SAFAR platform project

Due to the limitations presented above regarding the use of existing NLP platforms, a new research project has been initiated in 2012 to come up with an integrated Arabic NLP platform called "SAFAR" (Software Architecture For Arabic language pRocessing)[5] that has the following main objectives:

- to integrate resources and tools available in the community of Arabic NLP;
- to help developers of Arabic-oriented applications by reducing the time and efforts needed to learn and use existing NLP components;

[5] In the Arabic language, the word « SAFAR » refers to « a long travel » which is the suitable description of such long way project. This project is conducted by the Ibtikarat team in the Mohammadia School of Engineers, University Mohamed V Rabat, Morocco.

- to facilitate the evaluation and benchmarking process, especially for applications designed around a complex architecture such as QA systems;
- to consolidate the separate works conducted in a given Arabic NLP field;
- to guide the standardization of resource presentation and tools outputs.

To be able to achieve the above objectives, SAFAR is initiated as a modular platform providing an integrated development environment (Souteh and Bouzoubaa 2011). It includes various layers as depicted in Figure 39.

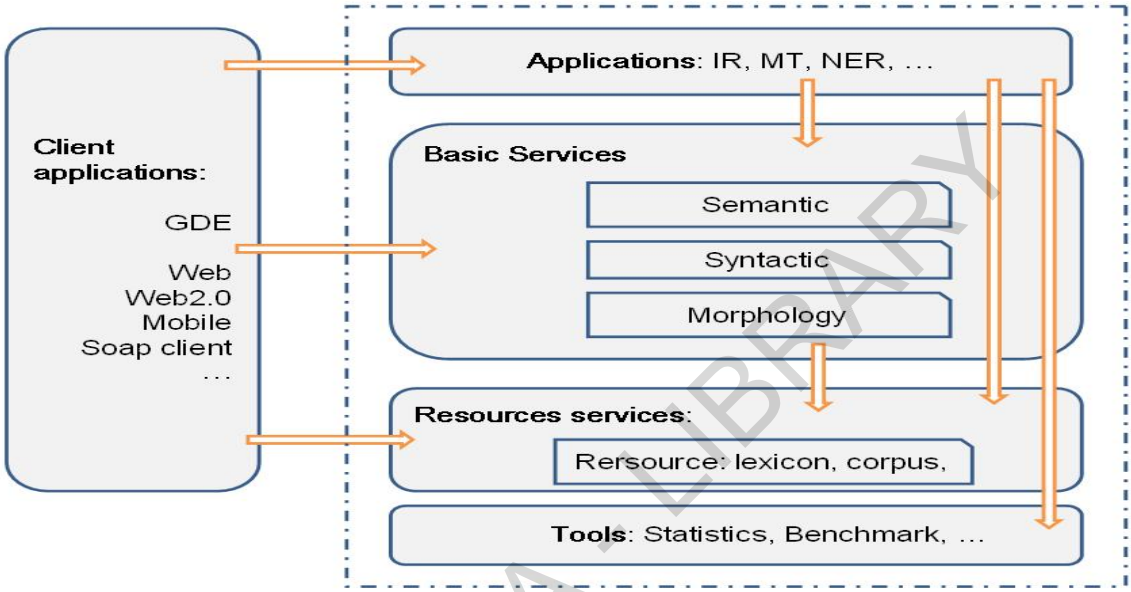


Figure 39. General architecture of SAFAR
Source: (Souteh and Bouzoubaa, 2011)

The architecture of SAFAR clearly defines the main layers needed in any Arabic NLP project. The core of this platform is constituted by the Basic Services Layer (BSL) containing the basic tasks of NLP such as morphology analysis, syntactic parsing, semantic processing, etc. In addition to this layer, resources such as lexicon, WordNets and corpora can be called from the Resource Services Layer (RSL).

Regarding the other layers proposed by SAFAR, we can cite: (i) the Tools Layer (TL) containing the different language independent material facilitating for example statistics, evaluation and benchmarking, (ii) the Client Application Layer (CAL) offering the different interfaces that can be used either by humans (e.g. demos, mobile interfaces, etc.) or machines (e.g. Web services, EDI and file exchange, etc.), and (iii) the Applications Layer (AL) which is the repository of Arabic NLP applications that are and will be developed by the research community using the other layers of SAFAR (i.e., BSL, RSL, etc.).

The SAFAR platform answers the mains needs of the IDRAAQ system for Arabic QA. Its BSL and TL layers provide the necessary Arabic morphological and syntactic analyzers and parser and language-independent (N-gram based) tools used in the three levels approach inte-

grated in IDRAAQ. The RSL layer contains the main resources considered in this research such as AWN and AVN.

6.4 Integrated architecture of IDRAAQ

In the current research, we addressed the main question whether it is possible to build an Arabic QA system from the existing NLP components or not. We believe that the current research proved that we succeeded in enhancing performance of the key module of a QA system for Arabic, i.e., the PR module. Our three-level approach based on existing Arabic NLP resources and tools and, also, on new processes (i.e., semantic representation and similarity scoring) that we have developed, provides improved results especially with the enriched AWN resource.

To consolidate this work and keep on building the whole Arabic system, we participated in the 2012 CLEF track with a new system called "IDRAAQ"[6] (Information and Data Reasoning for Answering Arabic Questions). This system integrates the three levels of our Arabic PR approach. The goal of this section is to show the development of IDRAAQ around the services of SAFAR and how the resulting modules and resources can be used by other Arabic QA developers to improve IDRAAQ or another system.

6.4.1 Architecture at a glance

The IDRAAQ system is designed around the general pipeline of QA modules, i.e., the question analysis module, the passage retrieval module and the answer extraction and validation module. Figure 40 illustrates this pipeline architecture.

In the application layer of SAFAR, only processes that are directly related to the QA task are implemented. These processes are divided into the three common QA modules as follows:

- *QAC module*: question analysis and classification using a simple question classifier based on some keywords showing the type of the question. For instance, after the preprocessing of a given question (keyword extraction, tokenization and sentence segmentation using the corresponding classes in SAFAR), if the first word is, for instance, "من" (who), the question classifier assigns the type "factoid".
- *PR module*: the three levels of our proposed approach are implemented under this module in the SAFAR-AL. In this layer, we implement processes such as QE based on the semantic relations of AWN (SAFAR-RSL), the DDN model implemented through the JIRS system (SAFAR-TL) and the semantic-based level relying on the SAFAR-BSL (morphology through the implementation of Alkhalil analyser provided by SAFAR, syntax through the implementation of the Stanford parser and the semantic

[6] In Arabic, the word "IDRAAQ" has the following meanings and senses: to understand, to recognize, to reach an objective, knowledge, intelligence, etc.

representation of a text in CG that we developed in this research). This module also calls some other tools from SAFAR-TL such as the Yahoo API to extract Web snippets.

- *AEV module*: this module is a simple process that helps in extracting or validating the right answer from a list of possible answers. The output of this module follows the same xml format used for instance in the CLEF campaign.

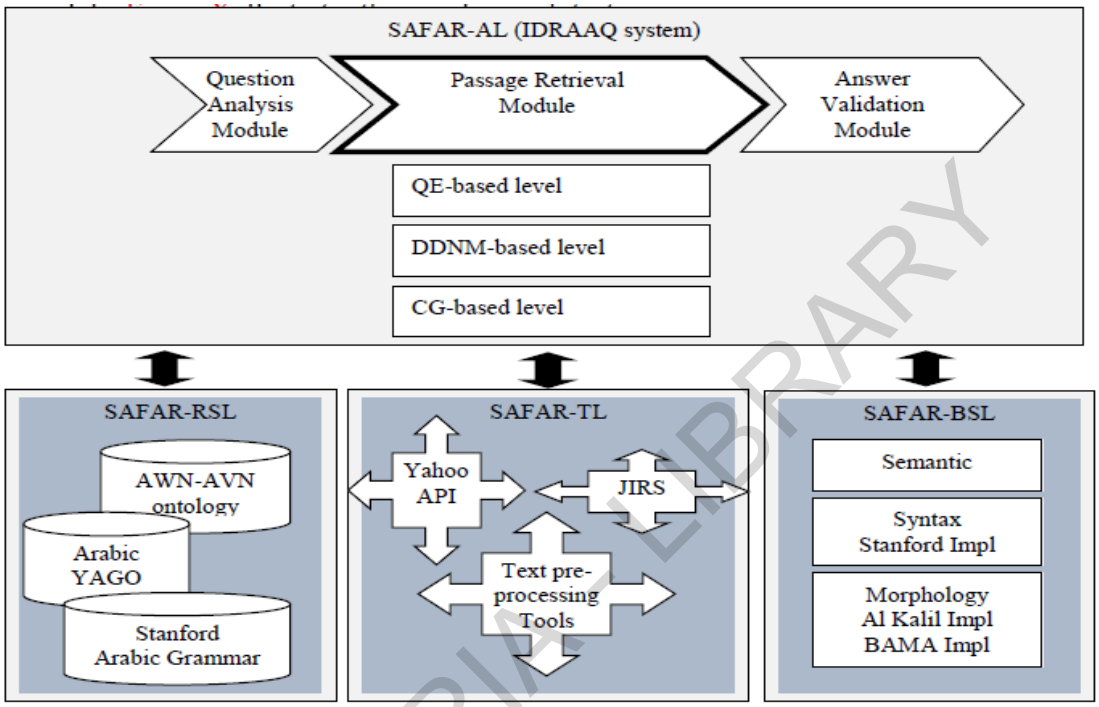


Figure 40. Three modules of the IDRAAQ system

6.4.2 IDRAAQ and SAFAR layers

In this section, we go through the details of the development of IDRAAQ as part of the SAFAR platform and the impact of this integration on the satisfaction of Arabic QA developers' needs.

6.4.2.1 Application Layer

Figure 41 illustrates the package "safar.applications" in the java project related to the IDRAAQ system. This package provides three main sub packages:

- *Evaluation package*: it contains a java class that can run an evaluation over a test-set of questions (e.g. ClefEvaluation). The main inputs and outputs of this evaluation are saved in the model class related to the given test-set (i.e., clef, trec, etc.). However, the classes allowing the reading of the test-set question and the measuring of performance are located in the SAFAR-TL layer. In the evaluation package,

we load the needed SAFAR resources (e.g. AWN-AVN ontology, Stanford Grammar for Arabic) once in order to optimize the response time.

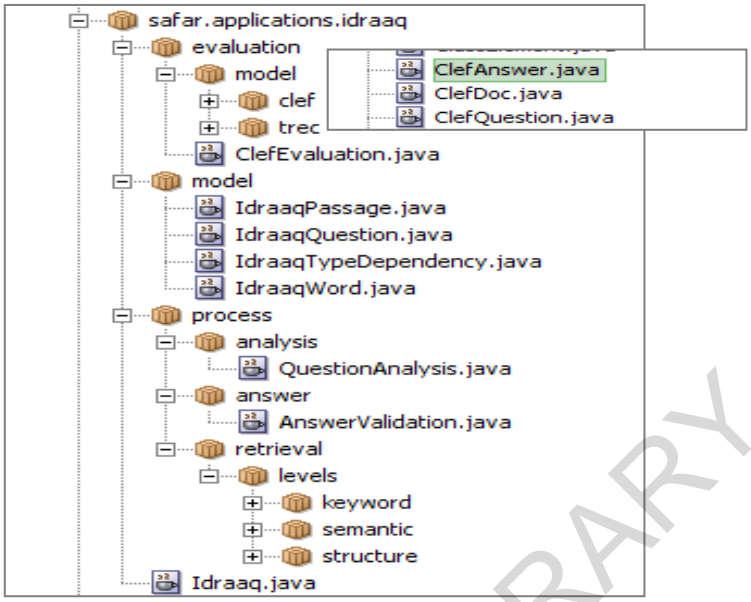


Figure 41. IDRAAQ in the SAFAR-AL layer

- *Model package*: it contains the needed classes to store the IDRAAQ specific objects. For instance, to store the result of passage retrieval, we use the class "IdraaqPassage".
- *Process package*: it is composed of the core modules of our Arabic QA system. As we can see, the passage retrieval module is divided into three sub packages related to the keyword-based, structure-based and semantic-based levels respectively.

6.4.2.2 Resource Services Layer

Figure 42 illustrates the package "safar.resource" which contains a linguistic resource loader class allowing to keep in memory the different needed resources. The IDRAAQ modules call these resources on-demand without being obliged to re-load them every time, so we avoid any heap space memory errors, especially at evaluation time. In addition to this loader class, there are sub packages to extract information from each used resource. For example, the "safar.resource.awn" contains three main packages: (i) the interface package that proposes interfaces for the information that can be extracted from AWN such as the methods "GetSynetsOfWord" or "GetWordsOfSynset", etc.; (ii) the implementation packages that provides concrete implementation of the AWN interfaces. For instance, the concrete implementation "AwnStandardImpl" class implements the above methods to extract synsets and words from the standard AWN, while the class "AwnEnrichedImpl" extracts this information from the enriched AWN. This will give us more flexibility at evaluation time to switch from an implementation to another and, therefore, to have the ability to measure the impact of the

enriched AWN for example; (iii) the model package contains the needed model classes to store the information extracted by the implementation methods.

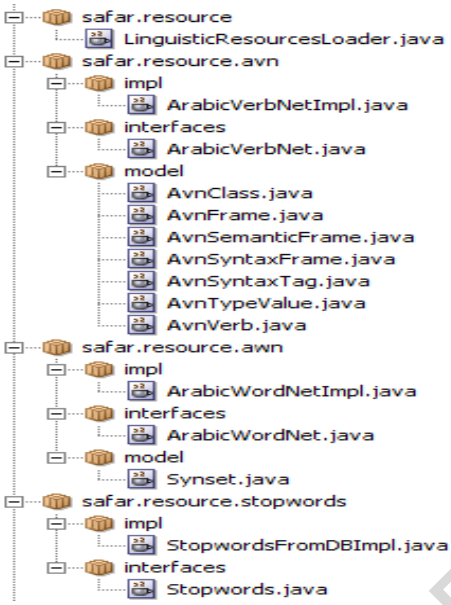


Figure 42. SAFAR-RSL resources used in IDRAAQ

Similarly, the "safar.resource" package provides the three sub packages related to each used resource (Arabic VerbNet, stopwords, Stanford Grammar for Arabic used by the Stanford parser, etc.).

6.4.2.3 Tools Layer

In the SAFAR-TL, we have integrated statistical and language independent tools such as the JIRS system, the amine platform that we adopted for semantic processing, the Yahoo API used as a baseline system compared to our approach, the different tools for accessing various information database and document collections (see Figure 43).

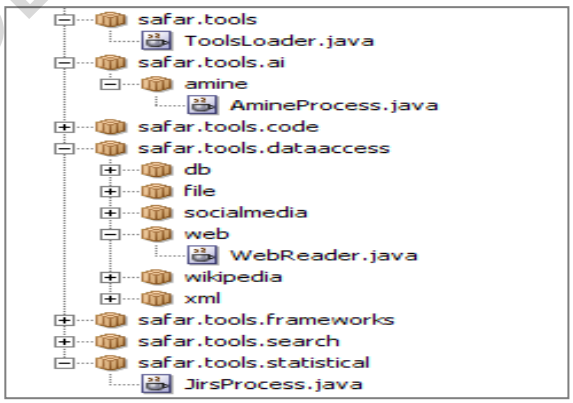


Figure 43. SAFAR-TL tools used in IDRAAQ

One of the processes that can be consolidated in this layer is the goldstandard of the evaluation campaigns, so developers are not requested to re-develop this part to test a further version of IDRAAQ or another Arabic QA system.

Regarding the semantic level, we have integrated the AWN-AVN ontology constructed and described in Chapter 5 in the SAFAR-RSL and we have performed CG operations (such as MaximalJoint, Generalization, Projection, etc.) by loading this ontology using the Amine Platform (Kabbaj 2006) java library added to to SAFAR-TL layer. Let us briefly recall that Amine is a Java Open Source Platform for the development of intelligent systems and multi-Agent systems. Previous works showed the compliancy of this platform with Arabic NLP (Bouzoubaa and Kabbaj 2007).

Amine Platform has also been chosen due to its use of CGs as a knowledge representation formalism. Moreover, the platform is a modular environment which provides: (i) an Ontology layer: we use this layer for manipulating the AWN ontology; (ii) an Algebraic layer: in addition to the elementary and the structured data types, this layer provides also various matching-based operations (like match, equal, unify, subsume, compare, maximalJoin, generalize, analogy, etc.); (iii) dynamic and basic ontology processes and (iv) Knowledge Base (KB) support.

6.4.2.4 Basic Services Layer

The SAFAR-BSL was the most important layer that were used by the IDRAAQ system. Indeed, this layer allowed us to directly integrate the morphology and syntactic processing using existing analysers and parsers (Alkhalil and BAMA analysers, Stanford parser). The integration of IDRAAQ in SAFAR is important since this provide interfaces for the morphology analysis task without being dependent on a specific analyser. For instance, we can evaluate the impact of using Alkhalil versus using BAMA without changing the code of IDRAAQ. The only prerequisite is to mention the implementation to be used (see Figure 44).

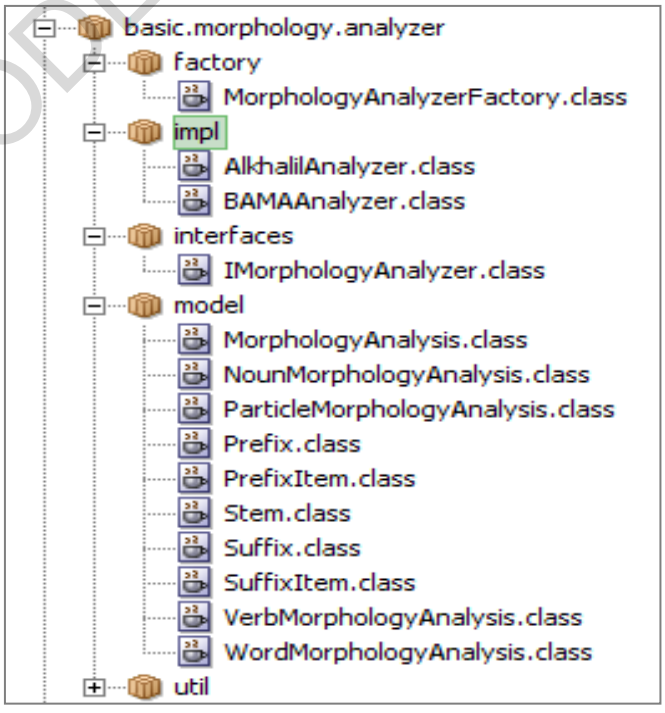


Figure 44. SAFAR-BSL used by IDRAAQ

Let us take the following example of java code that uses the implementation of the BAMA analyzer:

```
IMorphologyAnalyzer analyzer = MorphologyAnalyzerFactory
        .getImplementation(Analyzer.BAMA);
```

To switch to another morphological analyzer such as Alkhalil Analyzer, the only requirement is to use its implementation in SAFAR instead of the BAMA one as described in the changed code:

```
IMorphologyAnalyzer analyzer = MorphologyAnalyzerFactory
        .getImplementation(Analyzer. ALKHALIL);
```

The same development can be adopted for the syntactic parsing. So far, SAFAR integrates the Stanford parser, but if further parsers such as Bickel's parser[7] are integrated, the process of changing the parser is easier with SAFAR.

6.5 Chapter summary

In this chapter, we have introduced the IDRAAQ system initiated on the basis of the previously described three-level approach. This system is designed around the integrated NLP platform SAFAR which facilitates the development of core modules of Arabic QA as well as their evaluation.

The development of Arabic QA systems is a challenging and complicated task since it involves various tasks and resources of NLP. The existing platforms such as GATE, OpenNLP and NooJ present some limitations to their usage in the context of QA in general and Arabic QA in particular.

We have shown how IDRAAQ is integrated in each SAFAR layer such as the BSL, the TL, and RSL layers. This integration has the objective to consolidate what have been done in the framework of this project in terms of adoption of different tools and resources as well as to help further developments related to IDRAAQ or another Arabic QA using SAFAR.

Finally, we have seen that the evaluation process is highly improved by this integration, since much developers effort and time is saved by making available a number of gold standards and CLEF and TREC java models (classes for reading and writing in the formats used by these campaigns) that can be used within SAFAR.

[7] Daniel M. Bickel has developed a parser at the University of Pennsylvania, http://www.cis.upenn.edu/˜dbikel/software.html

Chapter 7

General Conclusions

In this thesis, we have presented our achievements in the field of Arabic Question Answering (QA) systems. The main objective behind these systems is saving the users‘ efforts and time considering the growth of the underexploited Arabic content on the Web. Such systems help in reaching this goal by providing precise and direct answers as a response to user questions instead of displaying unmanagearable lists of document links needing manual filtering. In comparison to other IR tools such as Search Engines (SEs), the accuracy of QA systems depends on the understanding of the given question and not only on its keyowrds occurrence in the text.

Generally, a question is analyzed following a pipline of three modules: (i) Question Anslysis and Classification, (ii) Passage Retrieval and (iii) Answer Extraction and Validation. The existing work on QA for English and other targeted languages shows the raise of different approaches and techniques in order to tackle the challenges of each module. The reported experiments highlight the importance of passage retrieval as a key module of a QA system. The performance of this module has a great impact on the accuracy of the whole system. To improve this performance in terms of accuracy, MRR and c@1, the most used measures in the field, researchers explored two famillies of approaches: (i) *surface-based approaches* relying on language-independent and statistical tools having the focus on keyword and structure similarity between question and passages, and (ii) *deeper approaches* having the aim of understanding the meaning of the question and the passages in order to compare their semantic similarity later. The decision of which approach using depends on the availability of components and resources related to other NLP tasks (syntactic parsers, semantic analyzers, ontologies, etc.).

Although Arabic QA field was investigated in early 1990s, the matury has not been reached with respect to the few attempts and research that were proposed so far. Various reasons were behind this state, especially:

- In comparison to other languages such as English and Spanish, Arabic combines many levels of complexity, for NLP in general and QA in particular, due to the nature of its script, morphology and syntax, to its high ambiguity level, to the low maturity of some QA-related NLP tasks such as NER and syntactic parsing, etc.

- The lack of advanced resources that cover not only the lexical part of the Modern Standard Arabic but also the syntactic and semantic parts.

Many limitations were highlighted in the existing Arabic QA attempts since the conducted experiments only used few types of questions (in most cases factoid questions are highlighted), small sized collections without complexity significance (Web-based systems were not studied), surface-based approaches without exploration of semantic ones, etc.

7.1 Findings and Research Directions

The main subject of this thesis was to propose a hybrid approach for Arabic QA systems with a special focus on the most important module, i.e., passage retrieval. The proposed approach combines the advantages of the surface-based and those of the deeper approaches to answer different types of questions as well as to overcome the specific challenges of the Arabic language. This hybrid approach was evaluated on the basis of experiments presented in Chapters 3, 4 and 5. Therefore, three sets of experiments were conducted as follows:

- *Experiments using the surface-based approach based on keywords and structure:* using the 2,264 TREC and CLEF questions, it was shown that after the implementation of the surface-based side of our approach relying on the AWN semantic relations-based QE and the Distance Density N-gram Model, a significant improvement of performance in terms of accuracy (increase from 9.66% to 20.20%), MRR (3.41 to 9.66) and number of answered questions (20.27% to 26.47%) was registered with respect to a baseline system (e.g. the Yahoo! API). These results encouraged us to asset the effectiveness of the proposed keyword-based and structure-based levels as well as the *usability* of the considered linguistic resources and the statistical tools for Arabic PR. They are promising even more considering the facts that: (i) this improvement was achieved by targeting a challenging Web collection of snippets, (ii) in most cases, the returned snippets contain just a few lines of content not enough to display together the question terms and the expected answer, (iii) the answer of a question may not be found by the system if the available Arabic Web content does not cover its topic; this happens considering the set of 2,264 CLEF and TREC are just translations into Arabic of questions originated from the European and the American cultures respectively, and (iv) the major part of this set is composed of factoid questions where the important keywords that are mainly NEs could have a translation which is different from the expected answers (that are also NEs).
- *Arabic WordNet coverage and usability experiments:* the AWN lexical resource and its semantic relations showed the ability to support the surface-based approach giving rise to the improvement of performance in comparison to the baseline system. Nevertheless, this resource has many *coverage* shortcomings that we emphasized through the theoretical and experience-based perspectives. These shortcomings impact the *usability* of this resource

and have been the reasons behind its limited use in Arabic NLP applications. To tackle this problem, we proposed an enrichement of AWN by targeting three types of content needed by Arabic QA as observed in the experience-based analysis:

- *Instances or NEs enrichment*: since our aim is to answer questions from the Web, we were interested in linking AWN to the YAGO ontology after the automatic translation of its entities into Arabic and their validation. This kind of dynamic information is widely used in questions. The added links between AWN synsets and the translated YAGO entities improved the effectiveness of the structure-based level with the injection of the AWN synset in the Distance Density N-gram Model (DDN) model;
- *Verbs and nouns enrichment*: the coverage of these main Common Linguistic Categories is poor in AWN with respect to the Arabic lexicon and the *coverage* registered in experiments for TREC and CLEF nouns and verbs. The proposed enrichment consists in: (i) extending the list of verb senses in AWN using the translation of both English VerbNet and Unified Verb Index by means of three heuristic rules already used in the EuroWordNet project and (ii) refining the hyponymy relation among AWN noun synsets using a technique based on pattern discovery and Maximal Frequent Sequences over Web snippets and starting from a list of seed AWN synsets. Both enrichements allowed to improve the Query Expansion recall by generating a higher number of related terms required by the keyword-based level. The loss in terms of precision (as usually the case in similar IR approaches) is avoided by applying the DDN model on top of the keyword-based level.
- *Broken plurals enrichment*: BP is among the forms of plural that are widely and specifically used in Arabic. The analyzed questions showed that the enrichment of AWN forms in terms of BP is important to apply the QE process for a higher number of questions in real-world applications, especially QA.

The above proposed enrichment of AWN improved the performance with a significant gain in terms of accuracy, MRR and number of answered questions. This improvement was achieved not only on the set of 2,264 TREC and CLEF questions that served as a basis to analyze AWN *coverage* shortcomings, but also on the set of 160 questions provided by the 2012 edition of the Question Answering for Machine Reading workshop. The participation of our system in the 2012 edition allowed us to compare the performance with other systems for different languages using the c@1 measure. The obtained performance is about 0.21 which is higher than the baseline and gives an acceptable ranking for our system among the participating ones. Moreover, the use of the enriched AWN resource allowed us to obtain the best c@1 score (0.36) regarding Topic #1 (i.e., AIDS) which is higher than the mean score

(0.32) over all best runs registered in this topic by all the participating systems for different languages including English.

- *Experiments using the deeper approach based on semantic representation and comparison:* to enhance the performance of Arabic QA, we investigated the effectiveness of the semantic-based level for the processing of complex questions beyond the factoid ones. Indeed, the overall performance obtained with the surface-based approach was penalized by the other types of questions. This is the case for example in the 2012 QA4MRE test-set where the factoid questions only represent 22%. The semantic level combines two well-known approaches and proposes two steps:
    - *Step 1*: The Question and its candidate passages are respresented into Conceptual Graphs. This step makes use of both the syntactic parsing using the Standford parser and the ontology we built from the Arabic VerbNet and AWN resources. To construct the CGs, we designed 11 rules that test the typed dependencies in each syntactic analysis to decide the CG pattern to be assigned.
    - *Step 2*: The semantic similarity score is measured between the CG of the question and the CG of a given candidate passage. The passages are ranked according to this score.

  To show the effectiveness of this level, experiments were conducted on the 2013 edition of the QA4MRE test-set as well as the CLEF-TREC 1999-2008 test-set. The former test-set contains 284 questions, most of them require semantic processing rather than surface-based approaches that are more suitable to factoid questions; these are more represented in the latter test-set. We used the 2012 QA4MRE as a training set for the rules of Step 1 and the 2013 edition as a test set in these experiments. The main investigations in this new experiments are: (i) the importance of the AWN enrichment, especially in terms of verbs to support the semantic-based level, (ii) the impact of the different used components (the Standford parser, Alkhalil analyzer and the AWN-AVN ontology) for the final performance of the semantic-based level, and (iii) the gain in terms of performance regarding the processing of non factoid questions. This gain was registered in both test-set and considering the number of answered questions as well as the c@1 measure.

7.2 Thesis contributions

This thesis confirmed the fact that leveraging the current advances registered in different Arabic NLP tasks such as morphology analysis (e.g. Alkhalil Analyzer), syntactic parsing (e.g. Stanford parser for Arabic) and semantic resources (e.g. AWN and AVN), it is possible to build a Question Answering system for Arabic with the ability to tackle various challenges.

We consider that the major contributions of this thesis can be summarized under the main challenges of the Arabic QA task:

1. *Language challenge*: We used different resources and tools for the Arabic language (AWN, AVN, Stanford parser, Distance Density N-gram model, etc.). Also, the existing methods for automatic enrichment of WordNets and for the semantic representation of text in conceptual graphs were adapted to tackle the particularities of Arabic in terms of complex syntax and morphology, non diacritized text ambiguity, challenges related to Named Entities, etc.
2. *Web challenge*: We conducted experiments using the Web as a targeted collection. This allowed the measure of the real *usability* of the proposed method and also of the considered resources and tools. In addition, a comparison of performance with a baseline search engine illustrated the significance of the work.
3. *Question and Answer challenge*: We addressed various types of questions starting by the classical factoid questions and ending by the questions requiring deeper understanding of the meaning such as the case of the QA4MRE task. This has a contribution to make clearer the *usability* of the considered resources and tools to deal with different levels of question complexity.
4. *Evaluation challenge*: We contributed to the evaluation challenge by conducting a set of experiments with different sets of questions with the aim to show the significance of the proposed approach and to compare the Arabic QA systems with other languages. Indeed, our participation in the QA4MRE task on behalf of CLEF 2012 allowed a benchmarking of our system with other evaluated systems and approaches. This also highlighted the gap and the remaining work for Arabic QA.

This thesis also presents a number of contributions for the QA community.

- First, we proposed and evaluated a new hybrid approach that combines surface-based and deep approaches.
- Second, we introduced a new factor (i.e., the TDRC factor) in the formula of Montes-y-Gómez (2001) related to the semantic similarity score; this factor can also be used in other languages' QA systems.
- Third, we analyzed the TREC and CLEF questions provided between 1999-2008 for the enrichment of AWN; this method can be followed to extend other WordNets.
- Finally, we translated a number of resources (TREC and CLEF questions, YAGO, etc.) into the Arabic language and make them available for the community for monolingual or cross-language tasks.

7.3 Further challenges

The construction of usable systems for Question Answering is a long term project. In the current thesis, we showed the different challenges that researchers may face, especially for the Arabic language. According to the contributions described above, we believe that further challenges might be tackled to achieve even better performance. The main directions that we propose are:

1. *Coming up with significant resources for Arabic QA*: the iterative method that we used for the enrichment of AWN, i.e., conducting experiments, analyzing the shortcomings, extending the resource and reconducting experiments can be followed in further works. A special focus might be devoted to the formal semantic information (conceptual graphs representation) in this kind of resources.

2. *Exploring other semantic methods*: the participation in the QA4MRE task highlighted the impact of new semantic methods such deduction and text entailment for answering questions that are more complicated.

3. *Integration of work*: the number of involved resources and tools in a QA system give rise the necessity of carrying on the work in the framework of integrated NLP platforms such as SAFAR. The architecture of the system plays a key role in the time response, especially when experiments are conducted on large test-sets of questions and documents.

Appendix A

Publications produced in the context of the thesis

Reference	*Index et/ou Editeur*	*Status*
L. Abouenour, M. Nasri, K. Bouzoubaa, A. Kabbaj, P. Rosso. Construction of an ontology for intelligent Arabic QA systems leveraging the Conceptual Graphs representation. Journal of Intelligent and Fuzzy Systems. DOI: 10.3233/IFS-141248. IOS Press 2014.	▪ **Thomson Reuters ISI Science Citation Index** ▪ **IOS Press** ▪ **Impact Factor : *0.788***	Published in 2014
L. Abouenour, K. Bouzoubaa, P. Rosso. On the Evaluation and Improvement of Arabic WordNet Coverage and Usability. In: Languages Resources and Evaluation, vol. 47, issue 3, pp. 891-917. DOI: 10.1007/s10579-013-9237-0	▪ **Thomson Reuters ISI Science Citation Index** ▪ **Springer** ▪ **Impact Factor : *0.659*** http://dx.doi.org/10.1007/s10579-013-9237-0	Published in 2013
Y. Benajiba, P. Rosso, **L. Abouenour**, O. Trigui, K. Bouzoubaa, L. H. Belguith. Question Answering . Chapter 11 in: Natural Language Processing of Semitic Languages. Zitouni I. (Ed.). Series: Theory and Applications of Natural Language Processing. Springer http://www.springer.com/computer/ai/book/978-3-642-45357-1	▪ **Springer** ▪ **DOI:** 10.1007/978-3-642-45358-8_11	Published in 2014
L. Abouenour. On the Improvement of Passage Retrieval in Arabic Question/Answering (Q/A) Systems. Lecture Notes in Computer Science, 2011, Volume 6716/2011, 336-341, DOI: 10.1007/978-3-642-22327-3_50. R. Muñoz et al. (Eds.), NLDB'11. Springer-Verlag, Berlin-Heidelberg.	▪ **Scopus** ▪ **ACM Digital Library** ▪ **DOI:** 10.1007/978-3-642-22327-3_50 http://www.springerlink.com/content/0302-9743/	Published in 2011
L. Abouenour, K. Bouzoubaa, P. Rosso. On the extension of Arabic Wordnet Named Entities and its Impact on Question/Answering. In Proceedings of the 2nd International Conference on Knowledge Engineering and Ontology Development. Valencia, Spain, October 2010.	▪ **Scopus** http://dblp.uni-trier.de/db/conf/ic3k/keod2010.html	Published in 2010
L. Abouenour, K. Bouzoubaa, P. Rosso. An evaluated semantic QE and structure-based approach for enhancing Arabic Q/A. In the Special Issue on "Advances in Arabic Language Processing" for the IEEE International Journal on Information and Communication Technologies (IJICT), ISSN: 0973-5836, Serial Publications, June 2010.	▪ **IEEE** http://www.ieee.ma/IJICT/IJICT-SI-Bouzoubaa-3.3/B-F_SI.htm	Published in 2010

Reference	*Index et/ou Editeur*	*Status*
L. Abouenour, K. Bouzoubaa, P. Rosso. IDRAAQ: New Arabic Question Answering System Based on Query Expansion and Passage Retrieval. CLEF 2012 (Online Working Notes/Labs/Workshop).	http://www.informatik.uni-trier.de/~ley/db/conf/clef/clef2012w.html#AbouenourBR12	Published in 2012
L. Abouenour, K. Bouzoubaa, P. Rosso. Using the Yago ontology as a resource for the enrichment of Named Entities in Arabic WordNet. Workshop on Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages Status, Updates, and Prospects, *LREC'10* Conference, Malta, May 2010.		Published in 2010
L. Abouenour, K. Bouzoubaa. Using an Arabic Ontology to Improve the Q/A Task. In Proc. of the 11th *International Business Information Management Association Conference* IBIMA 2009, Cairo, January, 2009.	▪ **Thomson Reuters ISI Index to Scientific and Technical Proceedings** http://www.ibima.org/past.html http://www.ibima.org/Cairo2009/papers.html	Published in 2009
L. Abouenour, K. Bouzoubaa, P. Rosso. Structure-based evaluation of an Arabic semantic Query Expansion using the JIRS Passage Retrieval system. In Proc. Workshop on Computational Approaches to Semitic Languages, E-ACL-2009, pp. 62-68. Athens (Greece), April 2009.	▪ **ACL - Association for Computational Linguistics** http://aclweb.org/anthology-new/semitic.html	Published in 2009
L. Abouenour, K. Bouzoubaa, P. Rosso. "Three-level approach for Passage Retrieval in Arabic Question /Answering Systems", Proc. of the 3rd *International Conference on Arabic Language Processing* CITALA2009, Rabat, Morocco, May, 2009.		Published in 2009

Appendix B

Typed dependencies rules for Conceptual Graph construction from Arabic text

This appendix provides the list of the rules that we designed to construct Conceptual Graphs from the typed dependencies provided by the Stanford Arabic parser. The tags used in these rules refer to the tag set adopted in the Stanford Arabic parser (see Appendix C).

Rule 1: "GTag=JJ and DTag=NN"

If the Governor Tag (GTag) is "JJ" and the Dependent Tag (DTag) is a noun, then there are two cases:

- The dependent tag is neither "NNP" nor "NNPS": in this case the conceptual graph of the dendency is constructed following the pattern:

  CG-dep = [cg : [Conc(G)]<-attributeOf-[Conc(D)]]

  Where Conc(G) and Conc(D) are the corresponding ontology concepts of the gouvernor and dependent respectively.

- The dependent tag is "NNP" or "NNPS": in this case the dependent is tagged by the Stanford parser as a singular or plural proper noun respectively, therefore, we follow the pattern:

  CG-dep = [cg : [Conc(G) : Conc(D)]]

Rule 2: "GTag ={NN,NNS} and DTag ={NNP,NNPS}"

If the GTag is NN (or plural noun NNS) and the DTag is NNP (or plural proper noun NNPS), then the pattern is:

CG-dep = [cg : [Conc(G) : Conc(D)]]

Rule 3: "GTag ={NN} and DTag ={DTNN,DTNNS}"

If the GTag is NN and the DTag is DTNN or DTNNS, then the pattern is:

CG-dep = [cg : [Conc(G)]<-attributeOf-[Conc(D)]]

Rule 4: "GTag ={V* } and DTag ={NN} and dependency-type=dobj"

If the GTag is a tag of a verb (such as VBP) and the DTag is NN and the dependency type returned by the Stanford parser is dobj (Direct object), then two cases occur:

- The dependent tag is neither “NNP” nor “NNPS”: in this case the conceptual graph of the dendency is constructed following the pattern:

  CG-dep = [cg : [Conc(G)]<-objOf-[Conc(D)]]

- The dependent tag is “NNP” or “NNPS”: in this case the dependent is tagged by the Stanford parser as a singular or plural proper noun respectively, therefore, we follow the pattern:

  CG-dep = [cg : [SupConc(D) : D]<-objOf-[Conc(G)]]

  Where SupConc(D)is the super concept of the NE corresponding to D in the ontology . Let us recall that almost all NEs were extracted from YAGO (see Chapter 4).

Rule 5: “GTag ={V* } and dependency-type={iobj, nsubj, dep, xcomp}”

If the GTag is a tag of a verb (such as VBP) and the dependency type returned by the Stanford parser is iobj (Indirect object), nsubj (Nominal subject), dep (General dependent) or xcomp (clausal complement with external subject) then two cases occur:

- The dependent tag is neither “NNP” nor “NNPS”: in this case the conceptual graph of the dendency is constructed following the pattern:

  CG-dep = [cg : [Conc(D)]<-agentOf-[Conc(G)]]

- The dependent tag is “NNP” or “NNPS”: in this case the dependent is tagged by the Stanford parser as a singular or plural proper noun respectively, therefore, we follow the pattern:

  CG-dep = [cg : [SupConc(D) : D]<-agentOf-[Conc(G)]]

Rule 6: “GTag ={NN } and DTag ={NN}”

If the GTag is NN and the DTag is also NN, then the pattern is:

CG-dep = [cg : [Conc(G)]<-is-[Conc(D)]]

Rule 7: “GTag ={CD }”

If the GTag is CD then the pattern is:

CG-dep = [cg : [Number = D]<-attributeOf-[Conc(G)]]

Rule 8: “DTag ={CD }”

If the DTag is CD then the pattern is:

CG-dep = [cg : [Conc(G) = D]]

Rule 9: "DTag ={JJ } and dependency-type={amod}"

If the DTag is JJ and the dependency type returned by the Stanford parser is amod (adjectival modifier), then we have two cases:

- The GTag is a tag of a verb (such as VBP): in this case no CG pattern is applied;
- The GTag is not a tag of a verb: in this case we follow the pattern:

  CG-dep = [cg : [Conc(D)]<-propertyOf-[Conc(G)]]

Rule 10: "dependency-type={prep}"

If the dependency type returned by the Stanford parser is a prepositional modifier (such as in, for, etc.), then the applied CG pattern is:

CG-dep = [cg : [prep : *p_i "D"]]

Where i is the rank of the preposition D in the list of the prepositions existing in the processed text.

Rule 11: "dependency-type={rcmod} and DTag={V*}"

If the dependency type returned by the Stanford parser is rcmod (Relative clause modifier), then the applied CG pattern is:

CG-dep = [cg : [Conc(G)]-attributeOf->[cg : Conc(D)]]

Appendix C

Stanford parser tags used in the designed rules

This appendix provides the description of the tags[1] and the dependencies types[2] used in the 11 rules designed to construct Conceptual Graphs from typed dependencies.

Tags:

CD	numeral, cardinal
DT	Determiner
IN	preposition or conjunction, subordinating
JJ	adjective or numeral, ordinal
NN	noun, common, singular or mass
NNP	noun, proper, singular
NNPS	noun, proper, plural
VBP	verb, present tense, not 3rd person singular

Dependencies types:

amod	adjectival modifier An adjectival modifier of an NP is any adjectival phrase that serves to modify the meaning of the NP. *Example:* "Sam eats red meat" amod(meat, red)
dep	dependent A dependency is labeled as dep when the system is unable to determine a more precise dependency relation between two words. This may be because of a weird grammatical construction,a limitation in the Stanford Dependency conversion software, a parser error, or because of an unresolved long distance dependency. *Example:* "Then, as if to show that he could, . . . " dep(show, if)
dobj	direct object The direct object of a VP is the noun phrase which is the (accusative) object of the verb. *Example:* "She gave me a raise" dobj (gave, raise) "They win the lottery" dobj (win, lottery)
iobj	indirect object The indirect object of a VP is the noun phrase which is the (dative) object of the

[1] *Source*: The University of Pennsylvania (Penn) Treebank Tag-set available at https://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html

[2] *Source*: Marie-Catherine de Marneffe and Christopher D. Manning. Stanford typed dependencies manual. September 2008. Revised for Stanford Parser v. 1.6.8 in June 2011.

	verb. *Example*: “She gave me a raise” iobj (gave, me)
nsubj	<u>nominal subject</u> A nominal subject is a noun phrase which is the syntactic subject of a clause. The governor of this relation might not always be a verb: when the verb is a copular verb, the root of the clause is the complement of the copular verb, which can be an adjective or noun. *Example:* “Clinton defeated Dole” nsubj (defeated, Clinton) “The baby is cute” nsubj (cute, baby)
prep	<u>prepositional modifier</u> A prepositional modifier of a verb, adjective, or noun is any prepositional phrase that serves to modify the meaning of the verb, adjective, noun, or even another prepositon. In the collapsed representation, this is used only for prepositions with NP complements. *Example*: “I saw a cat in a hat” prep(cat, in) “I saw a cat with a telescope” prep(saw, with) “He is responsible for meals” prep(responsible, for)
xcomp	<u>open clausal complement</u> An open clausal complement (xcomp) of a VP or an ADJP is a clausal complement without its own subject, whose reference is determined by an external subject. These complements are always non-finite. The name xcomp is borrowed from Lexical-Functional Grammar. *Example*: “He says that you like to swim” xcomp(like, swim) “I am ready to leave” xcomp(ready, leave)

Bibliography

Abbès, R., Dichy, J., & Hassoun, M. (2004). The architecture of a standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source program. In *Workshop on Computational Approaches to Arabic Script-based Languages*, Coling 2004. Geneva, Switzerland.

Abdelbaki, H., Shaheen, M., & Badawy, O. (2011). ARQA High-Performance Arabic Question Answering System. In: *Proceedings of Arabic Language Technology International Conference (ALTIC)*.

Abuleil, S., & Evens, M. (1998). Discovering Lexical Information by Tagging Arabic Newspaper Text, Workshop on Semantic Language Processing. COLING-ACL '98, University of Montreal, Montreal, PQ, Canada, Aug. 16 1998, pp. 1-7.

Abu-Salem, H., Al-Omari, M., and Evens, M. 1999. Stemming Methodologies Over Individual Query Words for an Arabic Information Retrieval System. In *JASIS*, 50(6), May, 1999.

Adda G., Lecomte J., Mariani J., Paroubek P., Rajman M., The GRACE French Part-of-Speech Tagging Evaluation Task, *in Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, vol. 1, ELDA, Granada, p. 433-441, May, 1998.

Agirre E., Màrquez L., Wicentowski R. (eds), *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Association for Computational Linguistics, Prague, Czech Republic, June, 2007.

Agosti M., Nunzio G. M. D., Ferro N., Harman D., Peters C., *Proceedings of the 11th Conference on Research and Advanced Technology for Digital Libraries*, vol. 4675 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin - Heidelberg, chapter The Future of Large-Scale Evaluation Campaigns for Information Retrieval in Europe, p. 509-512, 2007. ISBN 3540748504, 9783540748502.

Ahonen-Myka, H. (2002). Discovery of frequent word sequences in text. In *Proceedings of the ESF exploratory workshop on pattern detection and discovery* (pp. 180-189), London, UK: Springer-Verlag.

Al Khalifa, M., & Rodríguez, H. (2009). Automatically extending NE coverage of Arabic WordNet using Wikipedia. In *Proceedings of the 3rd international conference on Arabic language processing CITALA'09*, May, Rabat, Morocco.

Aljlayl, M. A., & Frieder, O. (2002). On Arabic search: improving the retrieval effectiveness via a light stemming approach. In Proceedings of the International Conference on Information and Knowledge Management, pages 340{347, McLean, Virginia. ACM Press. ISBN 1-58113-492-4.

Alotaiby, F., Alkharashi, I., & Foda, S. (2009). Processing large Arabic text corpora: Preliminary analysis and results. In *Proceedings of the second international conference on Arabic language resources and tools* (pp. 78-82), Cairo, Egypt.

Anwarus Salam, K. M., Khan, M., & Nishino, T. Example based English-Bengali machine translation using WordNet. In: *Proceedings of Triangle Symposium on Advanced ICT (TriSAI)*, Tokyo.

Attia, M., M. Rashwan, M. Al-Badrashiny, Fassieh (R), A semi-automatic visual interactive tool for morphological, pos-tags, phonetic, and semantic annotation of arabic text corpora. In: *IEEE Trans Audio Speech Lang Process* 17 (2009), 916–925.

Attia, Mohammed A. 2006a. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In *Challenge of Arabic for NLP/MT Conference*, The British Computer Society, London, UK, pp. 48-67.

Babko-Malaya, O., M. Palmer, N. Xue, A. Joshi, S. Kulick, Proposition Bank II: Delving deeper, frontiers in corpus annotation. In: *Workshop in conjunction with HLT/NAACL* (2004), Boston, MA, May 6.

Baker, C. F., Fillmore, C. J., & Cronin, B. (2003). The structure of the FrameNet database. *International Journal of Lexicography*, 16(3), 281-296.

Baldwin, T., Pool, P., & Colowick, S. M. (2010). PanLex and LEXTRACT: Translating all words of all languages of the world. In Proceedings of *Coling 2010*, Demonstration Volume, (pp. 37–40), Beijing.

Bekhti, S., & Al-Harbi, A. (2011). AQuASys: An Arabic Question-Answering Systembased on Extensive Question Analysis and Answer relevance Scoring, *International Journal of Academic Research*.

Benajiba, Y., & Rosso, P. (2008). Arabic Named Entity Recognition using Conditional Random Fields. In: Proc. Workshop on HLT & NLP within the Arabic world. Arabic Language and local languages processing: Status Updates and Prospects, 6th Int. Conf. on Language Resources and Evaluation, LREC-2008, Marrakech, Morocco, May 26-31.

Benajiba, Y., Diab M., & Rosso, P. (2009a). Arabic Named Entity Recognition: A Feature-Driven Study. In: IEEE Transactions on Audio, Speech and Language Processing, vol. 15, num. 5. *Special Issue on Processing Morphologically Rich Languages*, pp. 926-934. 2009. DOI: 10.1109/TASL.2009.2019927.

Benajiba, Y., Diab, M., & Rosso, P. (2009b). Using language independent and language specific features to enhance Arabic named entity recognition. In *IEEE Transactions on Audio, Speech and Language Processing. Special Issue on Processing Morphologically Rich Languages*, Vol. 17, No. 5, July, 2009.

Benajiba, Y., Rosso, P., & Gomez, J. M. (2007a). Adapting the JIRS Passage Retrieval System to the Arabic Language. In: *Proceedings of CICLing 2007 Conference*, volume 4394 of Lecture Notes in Computer Science, pages 530–541. Springer-Verlag.

Benajiba, Y., Rosso, P., & Lyhyaoui, A. (2007). Implementation of the ArabiQA question answering system's components. In *Proceedings of workshop on Arabic natural language processing, 2nd Information Communication Technologies Int. Symposium, ICTIS-2007*, April 3-5, Fez, Morocco.

Benamara, F. (2004). Cooperative Question Answering in Restricted Domains: the WEBCOOP Experiment. In: *Proceedings of the ACL Workshop on Question Answering in Restricted Domains*.

Benamara, F., & Saint-Dizier, P. (2004). Advanced Relaxation for Cooperative Question Answering. In: *New Directions in Question Answering*. MIT Press.

Bilotti, M. W. (2004). *Query expansion techniques for question answering*. Master's thesis, Massachusetts Institute of Technology.

Bilotti, M. W., Katz, B., & Lin, J. (2004). What Works Better for Question Answering: Stemming or Morphological Query Expansion?. In: *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*.

Black, W., Elkateb, S., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Introducing the Arabic WordNet project. In *Proceedings of the third international WordNet conference*. Sojka, Choi: Fellbaum & Vossen (eds).

Boudelaa, S., & Gaskell, M. G. (2002). A reexamination of the default system for Arabic plurals. Language and Cognitive Processes, 17, 321-343.

Boytcheva, S., Dobrev, P., & Angelova, G. (2001). CGExtract : towards extraction of conceptual graphs from controlled English. Proceedings 9th Intl. Conference on Conceptual Structures (ICCS'01), Stanford, Calif., USA, July.

Breck, E., Burger, J., House, D., Light, M., & Mani, I. (1999). Question Answering from Large Document Collections Our Approach : Mixing IR and KR via NLP. *AAAI Fall Symposium on Question Answering Systems*. To appear : 1999,

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual (Web) search engine. In: *Proceedings of the 7th International World Wide Web conference (WWW7)/Computer Networks*, 30(1-7),p.p. 107—117.

Brini, W., Ellouze & M., Hadrich, B. L. (2009a). QASAL : Un système de question-réponse dédié pour les questions factuelles en langue Arabe. In *9th Journées Scientifiques des Jeunes Chercheurs en Génie Electrique et Informatique*, Tunisia.

Brini, W., Trigui, O., Ellouze, M., Mesfar, S., Hadrich, L., & Rosso, P. (2009b). Factoid and definitional Arabic question answering system. In *Post-proceedings of NOOJ-2009*, June 8-10, Tozeur, Tunisia.

Buckwalter, T. (2004). *Arabic Morphological Analyzer 2.0*. Linguistics Data Consortium (LDC).

Buscaldi, D., Rosso, P., Gómez, J. M., & Sanchis, E. (2010). Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems*, 34(2), 113-134.

Callan, J.P. (1994). Passage-level evidence in document retrieval. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and Development in Information Retrieval. pp 302-310.

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton, The Hague.

Chu-Carroll, J., Fan, J., Boguraev, B. K., Carmel, D., Sheinwald, D., & Welty, C. (2012). Finding needles in the haystack: Search and candidate generation. IBM J. Res. & Dev., vol. 56, no. 3/4.

Chung,,H., Han, K., Rim, H., Kim, S., Lee, J., Song, Y., & Yoon, D. A. (2004). Practical QA System in Restricted Domains. In: *Proceedings of the ACL Workshop on Question Answering in Restricted Domains*.

Clark, P., & Fellbaum, C., & Hobbs, J. (2008). Using and extending WordNet to support question-answering. In: *Proceedings of the Fourth Global WordNet Conference*, University of Szeged, Hungary, pp. 111–119.

COLING, pages 42.488.

Costa, R.P., & Seco, N. (2008). Hyponymy extraction and Web search behavior analysis based on query reformulation. In *Proceedings of the 11th Ibero-American Conference* on AI: Advances *in Artificial Intelligence*, (pp. 1–10).

Croft, W. B., Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35, pp. 285-295.

Cui, H., Li, K., Sun, R., Chua, T. S., & Kan, M. Y. (2004). National University of Singapore at the TREC-13 Question Answering Main Task. In: *Proceedings of TREC-13*.

Cui, H., Sun, R., Li, K., Kan, M.-Y., & Chua, T.-S. (2005). Question answering passage retrieval using dependency relations. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '05*, 400. doi:10.1145/1076034.1076103.

Dang, C., & Luo, X. (2008). WordNet-based document summarization, in Proceedings of Seventh WSEAS Int. Conf. on Applied Computer & Applied Computational Science (ACACOS '08), Hangzhou, China, April 6-8.

Darwish, K., & Oard, D. W. (2003). CLIR experiments at Maryland for TREC 2002: Evidence combination for Arabic-English retrieval. Technical Report LAMP-TR-101,CS-TR-4456,UMIACS-TR-2003-26, University of Maryland, College Park, February.

Denicia-carral, C., Montes-y-Gõmez, M., Villaseñor-pineda, L., & Hernandez, R. G. (2006). A text mining approach for definition question answering. In *Proceedings of the 5th international conference on natural language processing, FinTal'2006*, Turku, Finland.

Diab, M. Second Generation AMIRA Tools for Arabic Processing: Fast and Robust Tokenization, POS tagging,

and Base Phrase Chunking. In: *Proceedings of the 2nd Intl. Conference on Arabic Language Resources and Tools* (2009), Cairo, Egypt.

Diab, M. T. (2004). Feasibility of bootstrapping an Arabic Wordnet leveraging parallel corpora and an English Wordnet. In *Proceedings of the Arabic language technologies and resources, NEMLAR, Cairo, Egypt*.

Diekema, A. R, Yilmazel, O., & Liddy E. D. (2004). Evaluation of Restricted Domain Question-Answering Systems. In: *Proceedings of the ACL2004 Workshop on Question Answering in Restricted Domain*, pp. 2-7.

Diekema, A. R., Yilmazel, O., & Liddy, E. D. (1999). Evaluation of Restricted Domain Question-Answering Systems. Center for Natural Language Processing.

Dridan, R., & Baldwin, T. (2007). What to classify and how: Experiments in question classification for japanese. In: *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 333–341,Melbourne, Australia.

Echihabi, A., & Marcu, D. (2000). A Noisy-Channel Approach to Question Answering.

Edmonds P., Kilgarriff A., Special issue based on Senseval-2, *Journal of Natural Language Engineering*, January, 2003.

El Amine, M. A. (2009). Vers une interface pour l'enrichissement des requêtes en arabe dans un système de recherche d'information. In *Proceedings of the 2nd conférence internationale sur l'informatique et ses applications (CIIA'09)*, May 3-4, Saida, Algeria.

Elberrichi, Z., Rahmoun, A., & Bentaalah, M.A. (2008). Using WordNet for text categorization, in The International Arab Journal of Information Technology, Vol. 5, No. 1, January.

Elghamry, K. (2008). Using the Web in building a corpus-based hypernymy-hyponymy lexicon with hierarchical structure for Arabic. *Faculty of computers and information* (pp. 157-165).

Elkateb, S., Black, W., Vossen, P., Farwell, D., Rodríguez, H., Pease, A., & Al khalifa, M. (2006). *Arabic WordNet and the challenges of Arabic*. In *Proceedings of Arabic NLP/MT conference*, London, U.K.

Elkateb, S., Black, W., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., and Fellbaum, C., (2006). Building a WordNet for Arabic. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.

Etzioni, O., M. J. Cafarella, D. Downey, S. Kok, A. M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, Web-scale information extraction in KnowItAll. In: *WWW* (2004).

Ezzeldin, A. M., & Shaheen, M. (2012). A survey of Arabic Question Answering: Challenges, Tasks, Approaches, Tools, and Future Trends. In Proceedings of the 13th International Arab Conference on Information Technology ACIT ' 2012 Dec . 10-13

Farghaly, A., & Shaalan, K. (2009). Arabic Natural Language Processing : Challenges and Solutions, *8*(4), 1–22. doi:10.1145/1644879.1644881.

Fellbaum, C. (ed.) (1998). WordNet: *An electronic lexical database*. Massachusetts: MIT Press.

Ferrucci, D., Brown, E., Chu-carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., & Prager, J. (2010). Building Watson : An Overview of the DeepQA Project, 59–79.

France, F. D., Yvon, F., & Collin, O. (2003). Learning Paraphrases to Improve a Question-Answering System. In: *Proceedings of the 10th Conference of EACL Workshop Natural Language Processing for Question-Answering*.

Gaizauskas, R., & Humphreys, K. (1998). A Combined IR/NLP Approach to Question Answering Against LargeText Collections. University of Sheffield UK .

García-Blasco, S., Danger, R., & Rosso, P. (2010). Drug-Drug interaction detection: A new approach based on maximal frequent sequences. *Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN*, 45, 263-266.

García-Hernández, R. A. (2007). Algoritmos para el descubrimiento de patrones secuenciales maximales. Ph.D. thesis, *INAOE*. September, Mexico.

García-Hernández, R. A., Martínez Trinidad, J. F., & Carrasco-ochoa, J. A. (2010). Finding maximal sequential patterns in text document collections and single documents. *Informatica*, 34(1), 93-101.

Gelbukh, A. (Ed.): CICLing 2006, LNCS 3878, pp. 319–330, 2006. c_Springer-Verlag Berlin Heidelberg 2006.

Gerard D. M., Suchanek F. M., Pease A. Integrating YAGO into the Suggested Upper Merged Ontology. 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008). Dayton, Ohio, USA (2008).

Gildea, D. & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Gómez, J. M., Buscaldi, D., Rosso, P., & Sanchis, E. (2007a). JIRS Language-independent Passage Retrieval system: A comparative study. In: *Proceedings of the 5th Int. Conf. on Natural Language Processing*, ICON-2007, Hyderabad, India, January 4-6.

Gómez, J. M., Rosso, P., & Sanchis, E. (2007b). Re-ranking of Yahoo snippets with the JIRS Passage Retrieval system. In: *Proceedings of the Workshop on Cross Lingual Information Access*, CLIA-2007, 20th Int. Joint Conf. on Artificial Intelligence, IJCAI-07, Hyderabad, India, January 6-12.

Gong, Z., Wa Cheang, C. and Leong Hou, U. (2005). Web Query Expansion by WordNet. DEXA 2005, LNCS 3588, pp. 166 – 175.

Goweder, A., & De Roeck, A. (2001). Assessment of a significant Arabic corpus. In *Proceedings of the Arabic NLP Workshop at ACL/EACL*, (pp. 73–79), Toulouse, France.
Graesser, A. C., Lang, K. L., & Roberts, R. M. (1991). Question answering in the context of stories. *Journal of Experimental Psychology: General*, 120(3).
Graff, D. (2007). Arabic Gigaword third edition. Linguistic Data Consortium. Philadelphia, USA.
Graff, D., Kong, J., Chen, K., & Maeda, K. (2007). English Gigaword third edition. Linguistic Data Consortium. Philadelphia, USA.
Green, S., & Manning, C. D. (2010). Better arabic parsing: baselines, evaluations, and analysis. In 23rd Conference on Computational Linguistics, pages 394–402, Beijing, China.
Green, W., Chomsky, C., & Laugherty, K. (1961). BASEBALL: An automatic question answerer. In: *Proceedings of the Western Joint Computer Conference*, p.p. 219-224.
Grefenstette, G. (1992). Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of the Annual International Conference on Research and Development in Information Retrieval (SIGIR)*.
Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.
Grimm, S., A. Abecker, J. Völker and S. Rudi, Ontologies and the Semantic Web. In *J. Domingue, D. Fensel, & J. A. Hendler, Handbook of Semantic Web Technologies*, 2011, S. 508-537, Berlin, Heidelberg: Springer.
Gruber, T., A translation approach to portable ontology specifications. In: *Knowledge Acquisition* (1993), 5, 2, 199–220.
Habash, N., O. Rambow, R. Roth, MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In: *Proceedings of the 2nd Intl. Conference on Arabic Language Resources and Tools* (2009), Cairo, Egypt.
Hacioglu, K. (2005). Semantic Role Labeling Using Dependency Trees. In: COLING. (2004).
Hammou, B., Abu-salem, H., Lytinen, S., & Evens, M. (2002). QARAB: A question answering system to support the Arabic language. In *Proceedings of the workshop on computational approaches to Semitic languages*, ACL, (pp. 55-65), Philadelphia.
Harabagiu, S., Moldovan, D., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Gîrji, R., Rus, V., & Morarescu, P. (2001). FALCON: Boosting knowledge for answer engines. Proceedings 9th Text Retrieval Conference (TREC-9). NIST Special Publication 500-249.
Harman D., The DARPA TIPSTER project, *ACM SIGIR Forum*, vol. 26, n° 2, p. 26-28, 1992. ISSN:0163-5840.
Hauser R., Results of the 1. Morpholympics, *LDV-FORUM*, June, 1994. ISSN 0172-9926.
Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics, COLING '92, Vol. 2 (pp. 539-545).
Hensman, S., & Dunnion, J. (2004). Automatically building conceptual graphs using VerbNet and WordNet In Proceedings of the 3rd International Symposium on Information and Communication Technologies (ISICT), Las Vegas, June 16-18, pp.115-120.
Hensman, S., & Dunnion, J. (2005). Constructing conceptual graphs using linguistic resources. In: *Proceedings of the 4th WSEAS International Conference on Telecommunications and Informatics*, Prague.
Hermjakob,.U. (2001). Parsing and Question Classification for Question Answering. In: *Proceedings of the ACL Workshop on Open-Domain Question Answering*.
Hovy, E., Gerber, L., Hermjakob, U., Lin, C.Y., & Ravichandran, D. (2001). Towards semantics-based answer pinpointing. In: Proc. First Internat. Conf. Human Language Technology Research. Association for Computational Linguistics, Morristown, USA, pp. 1–7.
Hsu, M. H., Tsai, M. F., & Chen, H. H. (2006). Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison. In: *Ng, H.T., Leong, M.-K., Kan, M.-Y., Ji, D. (eds.) AIRS 2006*. LNCS, vol. 4182, pp. 1–13. Springer, Heidelberg.
Hsu, M. H., Tsai, M. F., & Chen, H. H. (2008). Combining wordnet and conceptnet for automatic query expansion: a learning approach. In Asia Information Retrieval Symposium, LNCS 4993, pp. 213–224 Springer.
James, F. R., Dowdall, J., Kaljur, K., Hess, M., & Mollá, D. (2003). Exploiting Paraphrases in a Question Answering System. In: *Proceedings of the Workshop on Paraphrasing at ACL2003*.
Jarrar, M. (2011). Building a Formal Arabic Ontolog y(Invited Paper). In: *Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks*. Alecso, Arab League. Tunis, April 26-28.
João Pinto, F. (2008). Automatic query expansion and word sense disambiguation with long and short queries using WordNet under vector model. *Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos*, Vol. 2, No. 2.
Kabbaj, A. An Overview of Amine. In: *P. Hitzler and H. Scharfe (eds.), Conceptual Structures in Practice*, CRC Press, Taylor & Francis Group, 2009, pp. 321-347.
Kabbaj, A. Development of Intelligent Systems and Multi-Agents Systems with Amine Platform. In: *Proceeding of the 15th Int. Conference on Conceptual Structures, ICCS* (2006), Springer-Verlag.
Kaisser, M. Web Question Answering by Exploiting Wide-Coverage Lexical Resources. In: *Proceedings of the 11th ESSLLI Student Session* (2006). J. Huitink & S. Katrenko (eds.).

Kanaan, G., Hammouri, A., Al-Shalabi, R., & Swalha, M. (2009). A new question answering system for the Arabic language. *American Journal of Applied Sciences* 6(4), 797-805.

Kaplan, R. M. 2005. A Method for Tokenizing Text. In Festschrift in Honor of Kimmo Koskenniemi's 60th anniversary. CSLI Publications.

Katz, B. & Lin.J. (2002). START and Beyond. In: *Proceedings of the 6th World Multi conference Systemics, Cybernetics and Informatics.*

Katz, B., & Lin, J. (2003). Selectively Using Relations to Improve Precision in Question Answering. *Proceedings of the EACL-2003 Workshop on Natural Language Processing for Question Answering*, April.

Katz,.B. (1997). Annotating the World Wide Web using Natural Language. In: *Proceedings of the 5th Conference on Computer Assisted Information Searching on the Internet.*

Khalid, M. A. (2008). Passage Retrieval for Question Answering using Sliding Windows. In: *Proceedings of COLING 2008*, (August), 26–33 Workshop IR4QA.

Khoja, S. (1999). Stemming Arabic text. Computer Science Department, Lancaster University, Lancaster, UK.

Kim, H., Chen, S., & Veale, T. (2006). Analogical reasoning with a synergy of HowNet and WordNet. *In Proceedings of GWC'2006, the 3rd global WordNet conference*, January, Cheju, Korea.

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2006). Extensive Classifications of English verbs. *Proceedings of the 12th EURALEX International Congress.* Turin, Italy. September.

Kipper-Schuler, K. (2006). VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. Thesis.

Kipper-Schuler, K. VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, PA, 2005.

Kupiec, J. (1993). Murax: a robust linguistic approach for question answering using an on-line encyclopedia. In: *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)*, pp. 181–190, Pittsburgh, PA. ACM Press.

Kurata, G., Okazki, N., & Ishizuka, M. (2004). GDQA: Graph driven question answering system - NTCIR-4 QAC2 Experiments. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages pp 338–344. Tokyo, Japan.

Kwok, C., Etzioni, O., Weld, D. (2001). Scaling Question Answering to the Web. In: *Proceedings of WWW10*, Hong Kong.

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2007). *Light Stemming for Arabic Information Retrieval*, volume 38 of *Text, Speech and Language Technology*, pages 221-243. Springer Netherlands, 2007. ISBN 978-1-4020-6045-8.

Larkey, L.S., Ballesteros, L., & Connell, M.E. (2002). Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. In: *Proceedings of ACM SIGIR*, pp. 269-274.

Levin, B. (1993). *English Verb Classes And Alternations: A Preliminary Investigation*. The University of Chicago Press.

Li, X., & Croft, W. (2001). *Incorporating Syntactic Information in Question Answering* (Tech. Rep. No. CIIR-239). Amherst, Massachusetts, USA: University of Massachusetts.

Liddy, E. D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc

Liddy, L., Eduard Hovy, Jimmy Lin, John Prager, Dragomir Radev, Lucy Vanderwende, Ralph Weischedel. Natural language processing. In: In: Encyclopedia of library and information science. New York: Marcel Dekker; 2003;

Liebeskind, C., Ido Dagan, Jonathan Schler . Semi-automatic Construction of Cross-period Thesaurus. Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 29–35, Sofia, Bulgaria, August 8 2013. Association for Computational Linguistics

Lin, D. (1994). PRINCIPAR an efficient, broad-coverage, principle-based parser. In: *Proceedings of*

Lin, J. (2007). An exploration of the principles underlying redundancy-based factoid question answering. ACM Trans. Inf. Syst., 25, 1-15.

Liu, H., & Singh, P. (2004). ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal.*

Llopis, F., & Vicedo, J.L. (2002). IR-n: A Passage Retrieval System at Clef-2001. *Workshop of The Cross-Language Evaluation Forum (Clef 2001). Lecture Notes in Computer Science 2406:244-252.* Springer-Verlag.

Llopis, F., Vicedo, J. L., & Ferrandez, A. (2002). Passage Selection to Improve Question Answering. In: *Proceedings of the COLING* 2002 *Workshop on Multilingual Summarization and Question Answering*.

Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K., & Sutcliffe, R. (2005). *Overview of the CLEF 2004 Multilingual Question Answering Track.* Multilingual Information Access for Text, Speech and Images, CLEF 2004, Revised Selected Papers, Lecture Notes in Computer Science 3491, Springer-Verlag.

Manning, C. D., & Schutze, H. (1999). *Foundations of Statistical natural Language Processing*. The MIT Press, ISBN 0-262-13360-1.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Manning, C., & Jurafsky, D. (2012). StanfordNLP Group Official Website, http://nlp.stanford.edu/software/index.shtml, checked July 14th.

Matuszek, J. Cabral, M. Witbrock, J. Deoliveira, An introduction to the syntax and content of cyc. In: *Proceedings of the AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and its Applications to Knowledge Representation and Question Answering* (2006).

Milward, D., & Thomas, J. (2000). From information retrieval to information extraction. Proceedings Asocition for Computational Linguistics (ACL) Workshop on Recent Advances in Natural Language Processing and Information Retrieval.

Mitkov R. (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford Handbooks in Linguistics, Oxford University Press, Januray, 2005. ISBN-13: 978-0-19-927634-9.

Moens, M., and M. Steedman, Temporal ontology and temporal reference. In: *Computational Linguistics* (1988) 14, 15–28.

Mohammed, F.A., Nasser, K., Harb, H. M. (1993). A knowledge-based Arabic question answering system (AQAS). In *ACM SIGART Bulletin* (pp. 21-33).

Moldovan, D., Harabagiu, S., Pasca, M., & Girgu, R. (2000). The Structure and Performance of an Open-domain Question Answering System. In*: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp 563-570. Hon Kong.

Moldovan, D., Pasca, M., Harabagiu, S., Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system. Proc. of the 40th Annual Meeting Association of Computational Linguistics, New York.

Moll Aliod, D., Zaanen, M. Van, & Smith, D. (2006). Named Entity Recognition for Question Answering, 51–58.

Molla Aliod, D., Berri, J., & Hess, M. (1998). A real world implementation of answer extraction. Proceedings 9th International Conference on Database and Expert Systems Applications Workshop Natural Language and Information Systems (NLIS'98), pp. 143-148.

Mollá, Aliod, D., & Van Zaanen, M. (2005). Learning of Graph Rules for Question Answering. In: *Proceedings of ALTW05*, Sydney, December.

Montes-y-Gómez, M., Gelbukh, A., López-López, A., & Baeza-Yates, R. (2001). Flexible Comparison of Conceptual Graphs . 12th International Conference on Database and Expert Systems Applications DEXA 2001, Munich, Germany, September 2001. Lecture Notes in Computer Science, vol 2113, Springer-Verlag.

Monz, C. (2003). *From Document Retrieval to Question Answering*. Ph.D. dissertation, Institute for Logic, Language, and Computation, University of Amsterdam.

Mousser, J. A Large Coverage Verb Lexicon For Arabic. In: *Proceedings of the 7th conference on International Language Resources and Evaluation* (LREC) (2010), Valetta, Malta.

Mousser, J. Classifying Arabic Verbs Using Sibling Classes. In: *Proceeding of the International Conference on Computational Semantics (IWCS)* (2011), Oxford, UK.

Murdock, J. W., & Tesauro, G. (2012). *Statistical Approaches to Question Answering in Watson*. I. B. M. (n.d.).

Nadeau, D., & Sekine, S. A. (2007). Survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1).

Nanba, H. (2007). Query Expansion using an Automatically Constructed Thesaurus. In Proceedings of NTCIR-6 Workshop Meeting, May 15-18, 2007, Tokyo, Japan.

Navigli, R. (2009). Word sense disambiguation: a survey, ACM Computing Surveys, Vol. 41, No. 2, pp. 1–69.

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of FOIS-2* (pp. 2–9), Ogunquit, Maine.

Niles, I., & Pease, A. (2003). Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 2003 international conference on information and knowledge engineering*, Las Vegas, Nevada.

Noguera, E., Toral, A., Llopis, F., & Munoz, R. (2005). Reducing question answering input data using named entity recognition. In Proceedings of the 8th International Conference on Text, Speech & Dialogue, pages 428–434.

Nwesri, A. (2008). *Effective retrieval techniques for Arabic text*, PhD Thesis, School of Computer Science and Information Technology, RMIT University.

Ortega-Mendoza, R. M., Villaseñor-pineda, L., & Montes-y-Gõmez, M. (2007). Using lexical patterns to extract hyponyms from the Web. In *Proceedings of the Mexican international conference on artificial intelligence MICAI 2007*. November, Aguascalientes, Mexico. *Lecture Notes in Artificial Intelligence* 4827, Springer.

Pallett, D. (2003). A look at NIST'S benchmark ASR tests: past, present, and future, *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, IEE, Virgin Islands, USA, p. 483-488, November. ISBN:0-7803-7980-2 / DOI:10.1109/ASRU.2003.1318488.

Palmer, M., P. Kingsbury & D. Gildea. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 21. USA: MIT Press.

Pantel, P., & Pennacchiotti, M. (2006). Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of conference on Computational Linguistics Association for computational linguistics*, (pp. 113-120), Sydney, Australia.

Paroubek, P., Chaudiron, S., & Hirschman, L. (2007). Principles of Evaluation in Natural Language Processing. TAL, 48(1), 7–31.

Pasca, M., & Harabagiu, S. (2001). The informative role of wordnet in open-domain question answering. In: *Proceedings of the NAACL 2001Workshop on WordNet and Other Lexical Resources*.

Pazienza, M. T., M. Pennacchiotti and F. M. Zanzotto, Mixing WordNet, VerbNet and Propbank for studying verb relations. In: *Proceedings of the 5th Language Resource and Evaluation Conference LREC* (2006), Genoa, Italy.

Peetz, M., & Lopatka, M. (2008). Query Expansion with Wikipedia. In: *Proceedings of IIR*, University of Amsterdam.

Peñas, A., Hovy, E. H., Forner, P., Rodrigo, A., Sutcliffe, R. F. E., Sporleder, C., Forascu, C., Benajiba, Y., & Osenova, P. (2012). Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation. *CLEF (Online Working Notes/Labs/Workshop)*.

Peñas, A., Rodrigo, A., Sama, V., & Verdejo, F. (2008). Special Issue on Natural Language and Knowledge Representation, *Journal of Logic and Computation*. To appear (draft version).

Peng, F., Ralph Weischedel , Ana Licuanan , Jinxi Xu, Combining deep linguistics analysis and surface pattern learning: a hybrid approach to Chinese definitional question answering, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, p.307-314, October 06-08, 2005, Vancouver, British Columbia, Canada. DOI:10.3115/1220575.1220614.

Pound J., Ihab F. I., and Weddell. G. 2009. QUICK: Queries Using Inferred Concepts fromKeywords Technical Report CS-2009-18. Waterlo, Canada.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J., Jurafsky, D.: Semantic Role Labeling Using Different Syntactic Views. In: ACL'2005.

Prager, J. (2001) One search engine or two for question answering. Proceedings 9th Text Retrieval Conference (TREC-9). NIST Special Publication 500-249.

Radev, D. R., Fan, W., Qi, H., Wu, H., & Grewal, A. (2002). Probabilistic Question Answering from the Web. In: *Proceedings of the 11th World Wide Web Conference*, Hawaii.

Rashwan, M., M. Al-Badrashiny, M. Attia, S. Abdou, and A. Rafea, A stochastic Arabic diacritizer based on a hybrid of factorized and unfactorized textual features. In: *IEEE Transactions on Audio, Speech and Language Processing* (2011), vol. 19, no. 1, pp. 166-175.

Rassinoux, M., Baud, R. H., & Scherrer, J. R. (1994). A Multilingual Analyser of Medical Texts. In: W. Tepfenhart, J. Dick, J. Sowa (eds.), Conceptual Structures: Current Practices. Proc. ICCS94, LNAI 835, pp. 84-96.

Reddy, R. R. N., & Bandyopadhyay, S. (2004). Dialogue based Question Answering System in Telugu. Dept. of Comp. Sc. & Engg, Jadavpur University, Kolkata.

Rila, M., Tokunaga, T., & Tanaka, H. (1998). The use of WordNet in information retrieval, in Proceedings of Workshop on Usage of WordNet in Natural Language Processing Systems.

Riloff, E., & Thelen, M. (2003). A Rule Based Question Answering System for Reading Comprehension Tests.

Rocchio, J. J. (1971). Relevance feedback. In The SMART Retrieval System. Prentice-Hall, Englewoad Cliffs, N.J., 313-323.

Rodriguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., & Martí, A. (2008a). Arabic WordNet: Semi-automatic extensions using Bayesian Inference. In *Proceedings of the the 6th Conference on Language Resources and Evaluation LREC2008*, May, Marrakech, Morocco.

Rodriguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Mart., M., Black, W., Elkateb, S., Kirk, J., Pease, A., Vossen, P., & Fellbaum, C. (2008b). Arabic WordNet: Current state and future extensions. In *Proceedings of the fourth global WordNet conference*, January 22-25, Szeged, Hungary.

Sagot, B., & Fišer, D. (2008). Building a free French WordNet from multilingual resources. Workshop on Ontolex 2008, LREC'08, June, Marrakech, Morocco.

Salloum, W. (2009). A Question Answering System Based on Conceptual Graph Formalism. In: *Second International Symposium on Knowledge Acquisition and Modeling*, 383–386. doi:10.1109/KAM.2009.38.

Sarmento, L. (2008). Experiments with query expansion in the RAPOSA (FOX) question answering system. *Communications of the ACM*, 8, 792–798.

Schlaefer, N., Gieselmann, P., Schaaf, T., & Waibe, A. (2006). A pattern learning approach to question answering within the ephyra framework. In Text, speech and dialogue (p. 687-694). Springer Berlin, Heidelberg.

Schroeder, M. (1992). Knowledge Based Analysis of Radiology Reports using Conceptual Graphs. In: H. Pfeiffer, T. Nagle (eds.), Conceptual Structures: Theory and Implementation, Proc. 7th Annual Workshop, July, LNAI 754.

Scott, S., & Gaizauskas, R. (2001) University of She_eld TREC-9 Q & A System. Proceedings 9th Text Retrieval Conference (TREC-9). NIST Special Publication 500-249.

Shaheen, M., & Ezzeldin, A. M. Arabic Question Answering: Systems, Resources, Tools, and Future Trends. In: *Arabian Journal for Science and Engineering* (2014), pp 1-24, Springer Berlin Heidelberg.

Shannon, C. (1948) A Mathematical Theory of Communication, *Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656.

Sharaf, A. M. (2009). The Qur'an annotation for text mining. First year transfer report. School of Computing, Leeds University. December.

Shen, D., & Lapata, M. (2007). Using semantic roles to improve question answering. In: *Proceedings of the 12th Conference on Empirical Methods in Natural Language Processing*, pp. 12–21, Prague, Czech Republic.

Shen, D., Klakow, D. (2006). Exploring correlation of dependency relation paths for answer extraction. In: *Proceedings of the COLING/ACL*, 889–896.

Shen, D., Leidner, J. L., Merkel, A., & Klakow, D. (2006). The Alyssa System at TREC'2006 : A Statistically-Inspired Question Answering System.

Siddiqui, T. J. (2006). Intelligent Techniques for Effective Information Retrieval (A Conceptual Graph Based Approach). ACM SIGIR Forum Vol. 40 No. 2 December.

Sidrine, S., Y. Souteh, K. Bouzoubaa and T. Loukili, SAFAR: vers une Plateforme Ouverte pour le Traitement Automatique de la Langue Arabe. In: *Proceeding of the 6th Conference of Intelligent Systems: Theory and Applications SITA* (2010), May, Rabat, Morocco.

Simmons, R. F. (1965). Answering English questions by computer: A survey. Communications Association for Computing Machinery (ACM), 8(1): 53{70.

Snow, R., Jurafsky, D., & Andrew, Y. N. (2005). Learning syntactic patterns for automatic hypernym discovery. In Lawrence K. In Saul et al. (ed.), *Advances in Neural Information Processing Systems*, 17. Cambridge, MA: MIT Press.

Souteh, Y., & Bouzoubaa, K. SAFAR platform and its morphological layer, In Proceeding of the Eleventh Conference on Language Engineering ESOLEC'2011, Cairo, Egypt, 14/ 12/ 2011.

Sowa J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley Company.

Sowa, F. *Conceptual Structures: Information Processing in Mind and Machine*, 1984, Addison-Wesley Company.

Srihari, R., & Li, W. Cymfony A Question Answering System Supported By Information Extraction.

Stephane, N. (2003). *Sesei : un filtre sémantique pour les moteurs de recherche conventionnels par comparaison de structures de connaissance extraites depuis des textes en langage naturel*. Master report (M.Sc.), Faculté des études supérieures de l'Université, September.

Steven Abney. 1989. Parsing by chunks. *The MIT Parsing Volume*.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In *Proceedings. of 16th international World Wide Web conference WWW'2007*, (pp. 697-706), May, Banff, Alberta, Canada: ACM Press.

Suchanek, F. M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In Proc. of the 16th WWW, pp. 697-706 (2007).

Sun, R., Ong, C.-H., & Chua, T.-S. (2006). Mining dependency relations for query expansion in passage retrieval. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '06*, 382. doi:10.1145/1148170.1148237.

Sutcliffe, R., A. Peñas, E. Hovy, P. Forner, A. Rodrigo and C. Forascu, Overview of QA4MRE Main Task, *CLEF* (2013).

Swier, R., & Stevenson, S. (2005). Exploiting a Verb Lexicon in Automatic Semantic Role Labelling. In: HLT/EMNLP.

Tellex, S., Katz, B., Lin, J., Fernandes, A., & Marton, G. (2003). *Quantitative evaluation of passage retrieval algorithms for question answering*. In Proc. 26th Ann. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, New York, 2003, pp. 41–47.

Tjong Kim Sang, E., & Hofmann, K. (2007). Automatic extraction of Dutch hypernym-hyponym pairs. In *Proceedings of CLIN-2006*, Leuven, Belgium.

Toral, A., Munoz, R., & Monachini, M. (2008). Named entity WordNet. In *Proceedings of the Sixth international conference on language resources and evaluation (LREC'08)*, Marrakech, Morocco.

Trigui, O., Belguith, H. L., & Rosso, P. (2010). DefArabicQA: Arabic Definition Question Answering System. In: *Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC*, Valletta, Malta (pp. 40-45).

van Zaanen, M. (2002). *Bootstrapping Structure into Language: Alignment-Based Learning*. PhD thesis, University of Leeds, Leeds, UK, January 2002.

Vanyo G. Peychev, Jimmy C. Dubuisson, Vladimir T. Dimitrov, Zhechka A. Toteva Structured Documents Represented by Conceptual Graphs - A Simple Formalism for Presenting Structured Documents. ICETE (1), 2004, pp:257-262.

Vargas-Vera, M., & Motta, E. (2004). AQUA – Ontology-based Question Answering System. In: *Proceedings of the Third Mexican International Conference on Artificial Intelligence*, pp 468–477, Mexico City, Mexico, April.

Velardi, P., Pazienza, M., DeGiovanetti, M. (1998). Conceptual Graphs for the Analysis and Generation of Sentences. In: IBM J. Res. and Develop. Vol 32 (2), March 1988, pp. 251-267.

Voorhees E. M., Harman D. K. (eds), *TREC: Experiment and Evaluation in Information Retrieval*, Digital libraries and electronic publishing series, william y. arms edn, The MIT Press, Cambridge, MA, 2005. ISBN 0-262-22073-3.

Voorhees, E.M. (1993). Query expansion using lexical semantic relations. In: *Proceedings of the Annual International Conference on Research and Development in Information Retrieval (SIGIR)*.

Vossen P. (ed). EuroWordNet, *A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, 1999, The Netherlands.

Vossen, P. (ed). (1998). EuroWordNet, a multilingual database with lexical semantic networks. *Kluwer Academic Publishers*, The Netherlands.

Wagner, A. (2005). Learning thematic role relations for lexical semantic nets. PhD. thesis, University of Tübingen, 2005.

Woods, W. A, Kaplan, R. M, & Webber, B. N. (1972). The Lunar Sciences Natural Language Information System. Final Report. BBN Report 2378, Bolt Beranek and Newman Inc., Cambridge, Massachusetts.

Wu, M., Duan, M. Y., Shaikh, S., Small, S., & Strzalkowski, T. (2005). University at albany's ilqua in trec 2005. In *Proceedings of the TREC*, 77–83.

X Li, & D Roth. (2006). Learning question classifiers: the role of semantic information. Natural Language Engineering 12 (3), 229-249.

Yang, H., & Chua, T. S. (2003). Qualifier: question answering by lexical fabric and external resources. In: *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*.

Zaafrani, R. (1998). Al-Mu'allim 2 Software : An Arabic Computer Learning System Using Conceptual Sentence Generation. Proceedings of 6th ICEMCO, International Conference and Exhibition on Multi-lingual Computing, Cambridge, England. April 16-19.

Zaghouani, W. (2012). RENAR: A Rule-Based Arabic Named Entity Recognition System, ACM Transactions on Asian Language Information Processing (TALIP), v.11 n.1, p.1-13, March. DOI: 10.1145/2090176.2090178.

Zhang, D., & Lee, W. S. (2002). A Web-based Question Answering System, October 31.

Zhang, K., & Zhao, J. (2010). A Chinese question-answering system with question classification and answer clustering. In: *Proceedings of the Fuzzy Systems and Knowledge Discovery (FSKD), 2010 7th Int. Conf.*, Yantai, Shandong.

Zheng, Z. (2002a). AnswerBus Question Answering System. In: *Proceeding of HLT Human Language Technology Conference*.San Diego,CA. HLT March,.pp.24 –27.

Zheng, Z. (2002b). Developing a Web-based Question Answering System. In: *Proceedings of the 11th International Conference on World Wide Web*.

Zweigenbaum, P. (2003). Question Answering in Biomedicine. In: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*.