MINISTRY OF HIGHER EDUCATION AND SCIENRIFIC RESEARCH

Yahia Fares University of Medea

Faculty of Letters and Languages

Department of Foreign Languages-English Section

LINGUISTIC STUDY OF ARABIC LEXICAL DISAMBIGUATION WITH THE AIM OF IMPROVING AUTOMATIC TRANSLATION

Thesis Submitted in Partial Fulfilment

Of the Requirements for Master of Arts Degree in English (Linguistics)

Submitted by

- **-Haouari Boumediene Hammidouche**
- **-Yacine Hamadache**

Supervised by

Mohamed Douifi

Academic Year: 2018 – 2019

ABSTRACT

The work presented in this thesis explores the question of the automatic acquisition of senses for lexical disambiguation in a translation framework. the issue of the identification of the senses is situated in a linguistic framework and the treatment is oriented towards translation. Thus, We propose a linguistic method of meaning acquisition and making it possible to establish semantic correspondences through creating a linguistic annotated corpora by taking advantage of the large amount of annotated English corpus Princeton WordNet sense, and using a machine translation systems. This method is applied within the Arabic language and evaluated on the Arabic corpus alone in a translation-linguistic framework.

The results of this method are exploited by a lexical disambiguation method, which gives meaning to new instances of ambiguous words in context, and by a lexical selection method, which proposes their most adequate translation. Finally, a weighted assessment of the results of disambiguation and lexical selection is proposed, based on the inventory constructed by the word-sense disambiguation method.

Key words: Lexical disambiguation, **Automatic translation**, **Corpus enrichement**.

ملخص

يستكشف العمل المقدم في هذه الأطروحة مسألة إيجاد المعاني الصحيحة بشكل اوتوماتيكي من اجل ازالة الغموض عن الكلمات التي لها أكثر من معنى واحد و هذا في إطار الترجمة. مسألة تحديد المعاني الصحيحة للكلمات هي مسألة لغوية بحتة. **كما ان** عملية معالجة اللغة في هذه الاطروحة هي عملية موجهة لتطوير الترجمة الآلية. لهذا فإننا نقترح استعمال بعض المناهج اللغوية من اجل تحديد المعاني الصحيحة للكلمات و كذا تحقيق تطابق على مستوى الدلالات اللفضية بين اللغات. يتم هذا عن طريق تصحيح بعض المجامع اللغوية, حيث نعتمد على المجمع الانجليزي الكبير برينستون وورد نت للمعاني **و كذا لستعمل نظام ترجمة آلي**. سيتم تطبيق هذا المنهج على اللغة العربية كما سيتم تقييمه على مجامع للغة العربية أيضا في اطار لغوي و ترجمي.

هذا و سيتم استغلال النتيجة المحصل عليها من هته المنهجية في عملية ازالة الغموض اللغوي , **التي من** شأنه إعطاء معاني جديدة و صحيحة للكلمات التي لها أكثر من معنى عن طريق منهجية اختيار **الدلالات**, والتي تقترح أنسب ترجمة لها .

في الأخير , سيتم تقديريا تقييم نتائج ازالة الغموض الدلالي و اختيار الكلمات المحصل عليها إستنادا الى الاختبارات التي تم إنشاءها من أجل منهجية إزالة الغموض عن معاني الكلمات.

الكلمات الدلالية : إزالة الغموض عن الكلمات, الترجمة الآلية, إثراء المجامع.

General introduction

This general introduction sets the stage for this thesis. It is devoted for describing the background of the study, identification of the problem, research questions, objectives of the study, significance of the study, and clarification of the used terms.

1. Background of the Study

Language is a system of communication, written or spoken, which consists of the use of words in a conventional way. It is very important because it is one of the main ways that keeps individuals in contact with each other. The English language is an example for the importance of a language which has become a global phenomenon in the 21st century, and its learning has become an important educational issue. Consequently, people from different countries and cultures who speak different languages can communicate with each other easily as long as they acquire the ability to use English.

As said previously, due to globalization, people who come from non-English countries, just like Algeria, are positively forced to learn this language in order to be able to communicate with others throughout the world. Actually, a lot of them depend on translation as a mean of communication. “In this respect we can speak about an ethos of translation, whose aim is that of repeating the act of linguistic hospitality in the cultural and spiritual field.” Alexandria(2001, p42). In light of the above, developing the accuracy and efficiency of translation comes to be regarded as the role of not only translators, but linguists as well. Nowadays, many people from different cultures who speak different languages in order to communicate are using automatic translation (AT) softwares, applications, and websites. Unfortunately, many of these tools provide users with mistranslations, what gives rise to communication problems. This was the fundamental reason why natural language processing (NLP) engineers are seeking to work on automatic translation in order to enhance the quality of translation. One of the thorniest problems of NLP is the issue of word sense disambiguation (WSD). The latter was defined by Singh as follows, **“**Word Sense Disambiguation is the determination of sense of a particular word used in a particular context. Since words can have multiple senses in dictionary, it is very much important for the machine

to determine the correct meaning of word in the context. For instance consider the following**”** Singh (2015). Thus, Text clarification is a central task innatural language processing, which can improve many applications such as the extraction of multilingual information, the automatic summary or the automatic translation.

Either manually or automatically, It is a must to reveal a certain number of ambiguities, such as the lexical ambiguity which represents one of the huge difficulties of languages automatic translation, since some words may have several meanings. Lexical disambiguation is one of the important problems which should be resolved directly or indirectly in order to process natural language automatically. This thesis sheds light upon the lexical disambiguation and Focuses mainly on Arabic language which is regarded as a poorly endowed language for this task.

2. Problem of the Study

We have a corpus where the occurrences of many words are lexically disambiguated. We want to use linguistic methods in the process of translation from English into Arabic and vice versa within a lexical disambiguation tool that can automatically disambiguate new instances of these words in new sentences. This type of problem falls into the category of problems that supervised learning techniques address.

Early works on learning take us back probably to Chomsky’s studies. A simple way to learn is to observe situations for which the adequate response is known. Perhaps, a system can learn if it takes advantages of this experience to build a knowledge base that allows him to make a decision about a situation in which the answer is unknown. In other words, a language processing system can disambiguate an ambiguous word by analyzing the form of it. What is meant by observation is the base phrase analysis. Therefore, words determination procedure may cause a problem to WSD in the process of defining if, for example, the word ‘well’ can be a noun, verb or adverb, especially if this operation is applied to the Arabic language which has a different properties such as diacritical marks which can completely change the meaning of an a word.

3. Research Questions

Based on the discussion above, this study addresses the following questions:

1. In which manner linguistics can contribute to the field of WSD?

2. To what extent diacritical marks theory can enhance the process of WSD for Arabic?
3. Can Chomsky's syntactic parsing theory improve in lexical disambiguation of English morphemes?

4. Objectives of the Study

We suggest two main objectives for the thesis:

1. Lexical disambiguation of Arabic: firstly, we present the linguistic method used for the automatic creation of the corpus annotated in meaning for Arabic language, taking advantage of the large amount of English corpus annotated in WordNet sense, and using a machine translation system. This method is used for the Arabic language. We will exploit these annotated corpora to improve lexical disambiguation for Arabic.

2. Improving automatic translation with lexical disambiguation:
   In a second step we use lexical disambiguation for automatic translation. More precisely, we annotate the parallel corpora used during the process of automatic translation systems using Chomsky's syntactic parsing method.

5. Significance of the Study

Many techniques are used by many teams of research to try to solve the problem of lexical disambiguation. Thus, we chose to use these techniques, in addition to our linguistic background, to design our method to be used in a automatic lexical disambiguation tool. This tool must go through a learning phase and linguistic analyses. At the end of this phase, this tool should be able to automatically determine the correct meaning or interpretation of each ambiguous word in a new corpora. To evaluate its performance, we use the part of the corpus set aside during the learning phase. The design of such a lexical disambiguation tool requires:

- Linguistics and computational linguistics skills for developing criteria.
- Good background knowledge of the linguistics properties of the language (which are English and Arabic in our thesis).

6. Clarification of Key Terms.

Naturally, many terms are used in this thesis. Perhaps, some terms are not widely used. Thus, they are defined in order to clarify their meaning according to the context and to avoid any unnecessary misunderstanding. First of all, **disambiguate**, it express the act of removing uncertainty of meaning from (an ambiguous sentence, phrase, or other linguistic unit). (Oxford Advanced Learner's Dictionary, 2010,). Secondly**, lexical disambiguation** stand for "disambiguation of the sense of a polysemantic word". (TheFreeDictionary.com n.d). Thirdly and finally, what we mean by **corpus** is a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject.(Oxford Advanced Learner's Dictionary, 2010,).

7. Outline of the Thesis

This manuscript is organized in two main parts: First of all, we present the state of the art of the Arabic language in NLP(natural language processing) , and lexical disambiguation in general, and that of Arabic in particular. We show that the lack of resources in Arabic causes several problems in the process of improving, evaluation of the Arabic systems of lexical disambiguation and the process of comparing them.

Secondly, we develop through our contributions in this Thesis the field of WSD, the implementation of the proposed a linguistic method for the construction of a lexical disambiguation for Arabic, the development of resources for both the creation of a lexical disambiguation theory and for the evaluation of itself, moreover, the exploitation of the produced resources to build the intended system, and finally, the study of the contribution of the task of Arab lexical disambiguation for the automatic translation.

Thus, we present in the first chapter, the Arabic language by focusing on the main features of this language. Also, we present some existing Arabic NLP resources and tools. As well as, we present the task of lexical disambiguation, its implementation process, useful generic resources, existing approaches in general and Chomsky's syntactic parsing particularly. Then we present more particularly the lexical disambiguation of Arabic by illustrating the state of the Arabic language for this task. The second chapter is associated with the methodology followed in conducting this research, The last chapter aims to exploit

the resources produced in order to create an Arabic lexical disambiguation systems, and then to study the impact of the lexical disambiguation task on automatic translation.

CHAPTER ONE

LITERATURE REVIEW

1.1Introduction

Word-sense disambiguation is the process of identifying the meaning of a word in a given context (Mitkov, Jan 13, 2005). It was first presented as new separate computational task during the early days of machine translation in the late 1940s, making it one of the oldest problems in computer linguistic(Agirre,2007). (Weaver and Warren1955) acknowledged that context is crucial, and they recognized that "statistical semantic studies should be undertaken, as a necessary primary step. Lexical disambiguation is one of the main characteristics of a language; it is commonplace that a word have multiple meanings. For example in GPS application if user asks that he wants to go the bank. So which location should GPS give? The noun bank may has several meanings, it could be a financial institution to deposit a check or it could be a supply or a stock held in reserve for future use. It needs to understand how words are linked together in a specific context in order to know the right sense. Humans may simply understand a context of word simply by looking at the surrounding words and comparing them in a different usage. The same mechanism might be used by machines to determine the correct sense.

(Zipf, George Kingsley1972) published his "Law of Meaning" that accounts for skewed distribution of words by number of senses, which is more frequent words have more senses than less frequent words in a power law relationship. Our work will be divided into two sections; In the first section we present the state of the of the Arabic language in NLP(natural language processing), and lexical disambiguation in general, and that of Arabic in particular. We show that the lack of resources in Arabic causes several problems in the process of improving, evaluation of the Arabic systems of lexical disambiguation and the process of comparing them. However in the second section we show how applying linguistics in natural language processing (NLP) helps systems such as machine translation (MT) to be improved.

1.2 Theoretical framework

1.2.1 Computational linguistics (CL)

Computational linguistics is an interdisciplinary field that concerned with processing of language by computers (Mitkov, Jan 13, 2005). At the beginning computational linguistics (CL) around 1949 ,it kept into existence shyly when Warren and Weaven wrote his famous memorandum suggesting that translation machine might be possible .The first conference one Machine translation(MT) took place at MIT in 1952 and the first journal, Mechanical translation, began in 1954.However,the term 'Computational linguistics' started to appear only in the mid of 1960's.The journal changed its name to Mechanical translation and computational linguistics in 1965 but, the word 'computational linguistics' appeared in a small type.

1.2.2Machine translation (MT)

"The term Machine translation(MT) refers to computerized system responsible for the production of translations with or without human assistance" (Hutchin, 1995). A distinction is commonly mad between human-aided MT (HAMT) and machine-aided human translation (MAHT).The latter comprises computer-based translation tools which support translators by providing access to on-line dictionaries and reception of texts, stores previously translated texts (translation memories) and integrated resources referred to as translator workstation. Although the aim of machine translation system is to produce high-quality translation without the intervention of human in any stage. But, this still not possible (Hutchin, 1995). However researchers always try their best in order to improve this system.

1.2.3Lexical disambiguation (LD)

The task of lexical disambiguation (Word Sense Disambiguation) is to find for each word of a text the most appropriate meaning among a pre-defined inventory of meaning (Singh, 2015, p.32) . For example, in the sentence "I see the mountain through my window". « The algorithm should choose the meaning of 'window' which corresponds to the carpentry rather than the one that corresponds to the the graphical interface. In this task, only English can be considered a truly endowed language. Indeed, two types of resources that require particularly important human work can be used to create a system of lexical disambiguation:

knowledge-based resources (dictionaries, lexical bases ...) and corpora annotated in meaning. Poorly endowed languages rarely have these resources.

1.3 Lexical disambiguation of Arabic language

1.3.1 State of the Arabic language for lexical disambiguation

As mentioned before, the lexical disambiguation approach requires the existence of large lexical resources. Moreover, missing diacritics in Arabic texts is the most difficult feature for lexical disambiguation because it increases the number of possible meanings of a word and makes the task of disambiguation more difficult. Many publications on the lexical disambiguation of Arabic language often consider that the task is complicated because of certain characteristics of the Arabic language.

1.3.2 Contribution of lexical disambiguation in MachineTranslation

The lexical disambiguation task can improve various applications such as machine translation, information retrieval, word processing, and so on.

Machine translation consists of translating a written text from the source language into the target language while going through a stage of understanding the source language before proceeding with its translation. As a result, a system of lexical disambiguation can play an important role in overcoming the ambiguity of words (having multiple meanings), whether those of the source language or those of the target. For example, the English word "Want" can have various meanings (close words) such as "wish", "desire", etc. More precisely, if an automatic translation system knows the correct meaning of the words of the source language, then it could more effectively determine the corresponding appropriate words in the target language.Therefore, annotating a text through lexical disambiguation can help a machine translation system preselect the most relevant meaning of an ambiguous word drawn from its training data.

In our work, we are interested in studying the contribution of lexical disambiguation in machine translation, which requires working on data where the meaning of words is present. To achieve these goals, we need parallel English-Arabic corpora annotated in sense.Thus, a question arises: Does the task of lexical disambiguation actually improve the quality of automatic translation, especially for a morphologically richness and complex language such

as Arabic?. This is in vivo evaluation of our Arabic lexical disambiguation system based on its contribution to the performance of machine translation.

Among the few works that have been done to improve the task of machine translation using lexical disambiguation we cite the work of (Carpuat and Wu, 2007), which showed that the integration of predictions of a system in a segment-based statistical machine translation system improves the quality of translation by evaluating it on the three different Chinese-English test sets. Similarly, they have achieved improvements on NIST's major Chinese and English machine translation task. They evaluated the performance of the machine translation system in terms of the BLUE score (BLEU, or the Bilingual Evaluation Understudy, is a score for comparing a candidate translation of text to one or more reference translations.) and other automatic assessment measures.

In addition, (Nguyen et al. 2018) presented a case study using lexical disambiguation of Korean as well as morphological information to improve the performance of neural translation systems. First, they built a Korean lexico-semantic network (LSN) as a lexico-semantic knowledge base on a large scale. Then, based on the Korean LSN, they created a Korean lexical disambiguation system that can annotate the correct meaning of the Korean words in the training corpus. Finally, they conducted translation experiments using Korean-English, Korean-French, Korean-Spanish and Korean-Japanese language pairs. Experimental results show that the use of the Korean lexical disambiguation system as well as morphological information improved the translation quality of neuronal machine translation systems (for all language combinations).

1.4 Arabic in natural language processing

Many recent research are interested in Arabic natural language processing (NLP).This language has two important characteristics that make natural processing more complex: agglutination and missing diacritics of words. The performances of applications and classical tasks in this field such as automatic speech recognition, automatic summarization, machine translation or lexical disambiguation difficulties.

In this section, we are going to present the main characteristics of the Arabic language, showing the morphological richness of this language, as well as some useful resources and tools for its automatic processing. First, we begin by describing Arabic language in a succinct way, while clarifying the difference between modern standard Arabic and dialectal Arabic.

Then .we discuss some basic concepts of Arabic morphology. Finally, we will introduce its different grammatical categories, before exposing, at the end of this chapter, some linguistic resources and existing tools will be provided

1.4.1 Characteristics of Arabic language

1.4.1.1 Arabic language

In the Arabic language, we find that modern standard Arabic (MSA), that appears in formal and official situations such as (teaching, religious speeches, radio broadcasts and television, administrative documents etc.) is then most commonly used and understood form instead of the classical one. Furthermore the dialectal Arabic (DA) which is the daily life language and which varies according to the region and the country (Maghrebian Arabic, Egyptian Arabic, Iraqi Arabic, Golf Arab etc.). However, as it is described in Figure 1.1, these two varieties of the Arabic language are totally distinct

In our work, we are interested in standard Arabic, which is an international and intercontinental language, classified among the Semitic languages, and spoken by more than 263 million people. As illustrated in Figure 1, it is the official language in 25 countries, and ranks third after English and French according to Wikipedia.

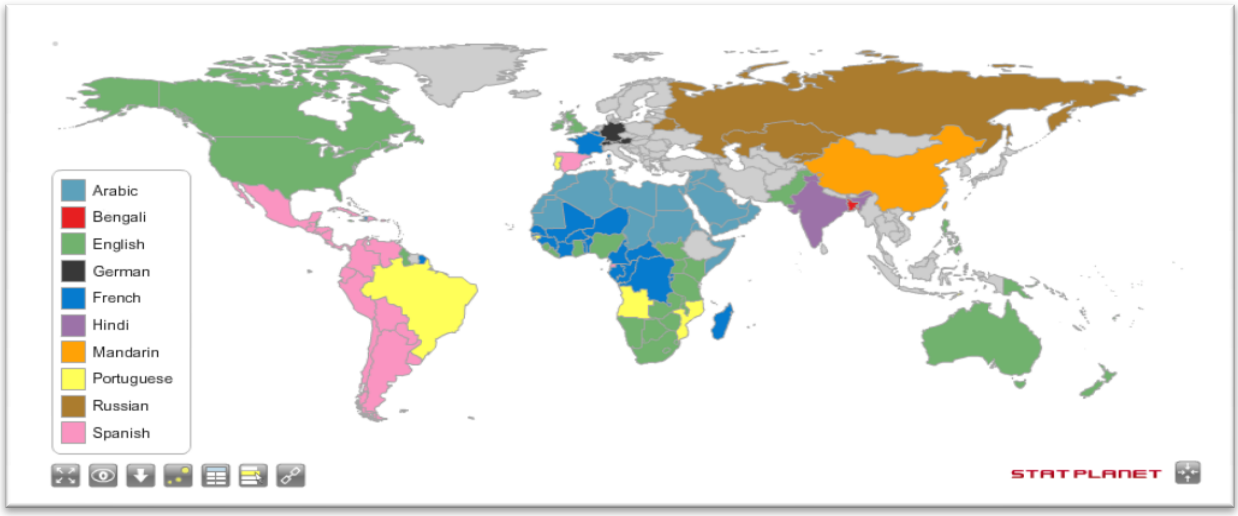


FIGURE 1: Situation of the Arabic language in the (worldRFI 2009)

It is a language written from right to left, using cursive writing, with 28 letters in its alphabet.These letters are consonants and long vowels(ا « Alif »,"و« Waw » and' ي

« Ya »).However, we find the "Hamza" (ء) which can take different forms(ا,إ,أ,آ),short vowels1.3 above or below the letters allowing their articulation, and which include: the "fatha" (the sound A), the "kasra" (the sound I), and the " damma " (the sound Ou) .

Soukoun [..]	بْ [b]	تْ [t]	ثْ [s]	جْ [j]	حْ [h]	نْ [n]	لْ [l]
Fatha [a]	بَ [ba]	تَ [ta]	ثَ [sa]	جَ [ja]	حَ [ha]	نَ [na]	لَ [la]
Damma [ou]	بُ [bou]	تُ [tou]	ثُ [sou]	جُ [jou]	حُ [hou]	نُ [nou]	لُ [lou]
Kasra [i]	بِ [bi]	تِ [ti]	ثِ [si]	جِ [ji]	حِ [hi]	نِ [ni]	لِ [li]

FIGURE 2: Short Vowels in Arabic [Verschaere, 2016]

"soukoun" corresponds to the absence of short vowel.

In addition to that, there is the "Tanwine" which is a second type of vowels used to vocalize consonants, it marks the repetition of two identical vowels at the end of a word (name), indicating whether it is a definite or indefinite name. The "chadda" is a double vowel that indicates the doubling of the consonant or semi-consonant concerned. This symbol is placed only above the letter, and is written with the "fatha", the "kasra", the "damma" and the "Tanouine", as it can be placed above the letter without the presence of a short vowel.

Unlike languages like English or French, in Arabic there is no difference between lower- or uppercase letters. Moreover, the writing pattern of some Arabic letters may vary depending on their location in the word (at the beginning, middle, or at the end) and the letters surrounding them.

Modern standard Arabic (MSA) uses similar punctuation to those used in European languages (.,;,:,!, "", (), Etc.), but it differs slightly with respect to some punctuation marks, because of writing style from right to left. Thus, the inverted question mark (؟) and the inverted comma (،) are inverted. Tat-weel (-) is a particularly unique Arabic symbol used to lengthen characters without changing the meaning of the word (egنـلـ, and نل)

Arabic is considered as a poorly endowed language with few applications and difficult to process in the field of natural language processing .It is a very rich language in terms of several meanings (polysemous terms), which can cause comprehension problems or even cause a real confusion especially when we treat unguarded Arab corpora. Actually, vowels play a very important role not only to remove the ambiguity of the word but, also to identify its grammatical category of its position in dependently in a given sentence. Words with the same representation, as it shown in Table 1.(Debili et al. 2002) confirmed that Arabic texts with Diacritic marks contain 43% ambiguous words, while texts without diacritics may contain 72%.An Arabic word may have a complete new meaning after changing a simple vowel. Moreover, the morphological richness of the Arabic language can cause serious problems of ambiguity in the lexical segmentation of a word, which influences the determination of its exact meaning. Taking the example of the word "ورد" [wrd] which has two possible interpretations: a first that considers it as a single word "ورد" [warada] while the second that considers that the letter "و" [w] is a prefix of the word "رَدّ"[radda]

Word	interpretation 1	interpretation 2	interpretation 3
شِعْرٌ	شِعْرٌ (poetry)	شَعْرٌ (hair)	شَعَرَ (felt)
كتبكَتَبَ	(wrote)	كُتُبٌ (books)	كُتِبَ (waswritten)
بيَّنَ	(declared)	بَيِّنٌ (obvious)	بَيْنَ (among)
وردوَرَدَ	(mentioned)	وَرْدٌ (roses)	وُرِدَ (wasmentioned)

TABLE 1: Examples of ambiguity of Arabic words due to lack of Diacritic marks

1.4.2 Basic Concepts of the of Arabic Morphology

1.4.2.1 The root

Each Arabic word is based on three letters (consonants) or sometimes four and very rarely two (a few special words).These letters constitute the root **(جذر)** or the heart of the word, to which we can add other letters to form different words (lexemes, several derivations) belonging to the same semantic field. We find about 6,000 roots in Arabic .Indeed, the root is the most important part in defining the main meaning of a word..For instance the triliteral root

[ك،ت،ب] [K + t + b] which means (signify) the general idea (the notion) of writing, from which several words can be derived as:

Writer, the one who writes.,كاتب / [katib]

-Book., كتاب [kitab]

-You write.,تكتبو [taktubuna]

-We write. "نكتب, [naktubu]

-Library.مكتبة [maktaba]

-Writing.كتابة [kitaba]

-What is written.مكتوب[maktub]

-Office. مكتب [maktab]

1.4.2.2 The scheme

An Arabic word comes from a root and is integrated into a schema (وزن) that represents the form of the word. The Arabic schemata of the root [ف،ع،ل], that allows to determine the structure of words. By associating a root with one or more schemas, we get new vocabulary. In Arabic, we find few hundred schema. So, we brought some examples of schemes applied on the root [ك،ت،ب] [K + t + b]:

-The wordكتب [kataba] corresponds to schema فعل [faEala]

-The wordكاتب / [katib] corresponds to scheme فاعل [faEiL]

-The wordمكتوب [maktub] corresponds to the schema مفعول [mafEul]

-The wordكتاب [kitab] corresponds to the scheme فِعَال[fiEal]

1.4.2.3 Affixes

Affixes (prefixes, suffixes) are letters that can be added at the beginning and / or end of the base of a word, used to identify grammatical features. They mark the appearance, the mode of the verb, the person, the genre as well as the number of the subject, etc.

1.4.2.4 Prefixes

Prefixes are verbal morphemes placed at the beginning of a word to form another one or change its meaning.. In table 2 we present some examples of words that derived from the root [ك+ت+ب] [K + t + b] with prefixes in Arabic:

Prefix	signification	word
ت	You write	تَكْتُبُ
ي	He writes	يَكْتُبُ
ن	We write	نَكْتُبُ

TABLE 2: Examples of words with prefixes in Arabic

1.4.2.5 Suffixes

Suffixes are morphemes (verbal or nominal) which are placed at the end of the base (verbal or nominal) of a word. In Table 3 we give some examples of words with suffixes in Arabic.

Suffix	signification	word
تُ	I wrote	كَتَبْتُ
تَ	You wrote	كَتَبْتَ
نا	We have written	كَتَبْنَا
وا	They wrote	كَتَبُوا
ه	His book	كِتَابُهُ
ها	Her book	كِتَابُهَا

TABLE 3: Examples of words with suffixes in Arabic

1.4.2.6 Clitics

Clitics (proclitic and enclitic) in Arabic are linguistic units that are pronounced and written as affixes but they are grammatically independent. In fact, an Arabic graphic word can be divided into five elements: proclitic (s), prefix, root, suffix (s) and enclitic (s).

1.4.2.6.1 Proclitics

In Arabic, the proclitics generally depend on the word that precedes it to form a new word. Therefore, we find simple proclitics (conjunctions, prepositions, coordinators, etc.) with

a single letter and other compounds (a combination of the former) with more than one letter. We present, below, some examples of proclitics in Arabic:

-The definite article: ل

-The interrogation tool:أ

-Coordinators: e,,وف

-The prepositions:لب

-The letter that indicates the future:س

Proclitic (s)	signification	word
ل	The office	المكتب
أ	Have youwritten	أكتبت
ف	So hewrote	فكتب
و	And hewrote	وكتب
ل	For writing	ليكتب
ب	Withher book	بكتابها
س	He willwrite	سيكتب

TABLE 4: Examples of words in Arabic with proclitics

1.4.2.6.2 Enclitics:

Enclitics in Arabic represent the letters that can be attached to the end of a word (verb, noun or preposition). Like proclitics, enclitics can be combined with each other to produce another compound word. Table5 shows some examples of existing Arab enclitics:

Enclitic(s)	signification	word
ي	They have taught me	علّموني
ك	They told you	أخبروك
كم	they wrote to you	كتبولكم

TABLE 5: Examples of words in Arabic with enclitics

In Table 5 we present an example of a word agglutinated in Arabic consisting of a pre-basic (prefix and proclitic), a basic, and a post-basic (suffix and enclitic).

1.4.2.6.3 The radical

In the Arabic language there is a difference between the radical (çz, ¥,) and the root. Indeed, the radical or stem is the smallest lexical unit derived from a root in a given form. This can be obtained from a word (agglutinated form)

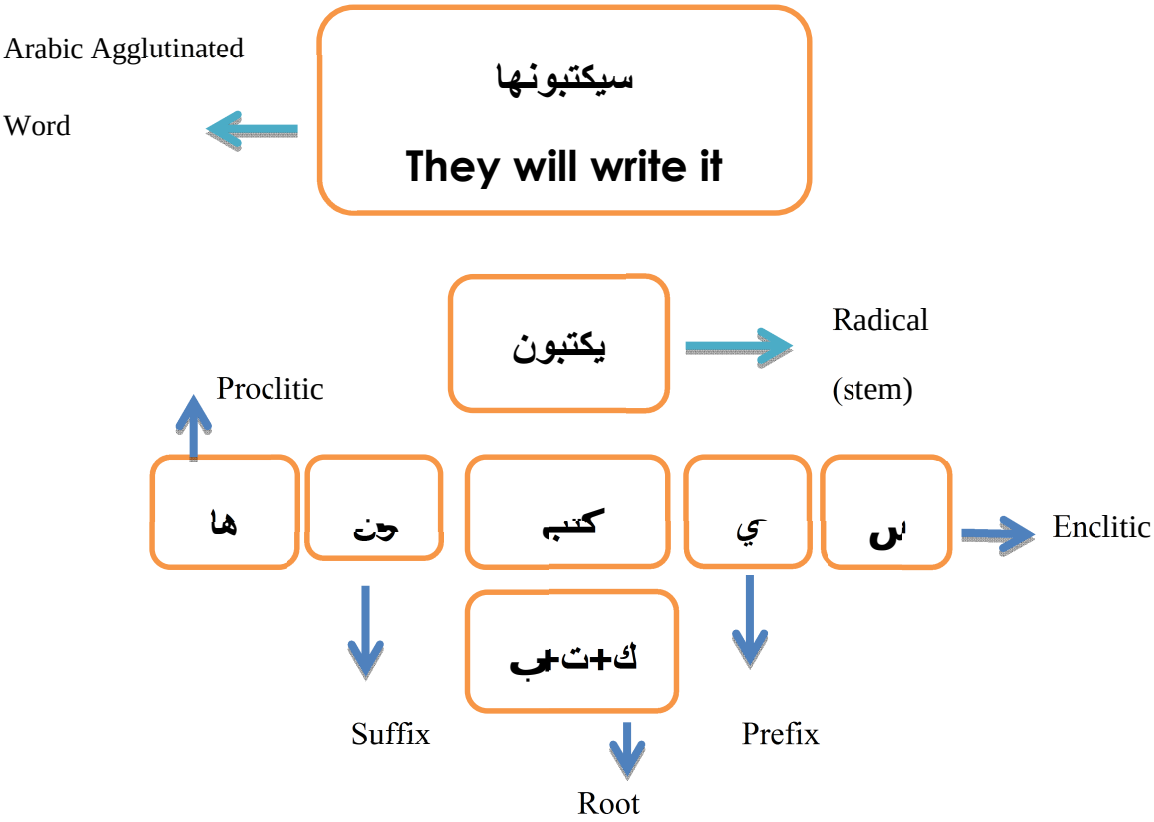


FIGURE 3: Structure of an agglutinated Arabic word

After removing the affixes (prefixes and or affixes) that surround it. Knowing that, in Arabic, the radical of the triliterate verbs is formed of three letters according to the schema فَعَلَ More precisely, take the previous example of the root [ك+ت+ب] [K + t + b], from which we can generate different radicals such as:: the verbكتب [kataba] (he wrote), the verb.,"يكتب"[yaktubu] (he writes), the name كاتب / [katib] (writer), the name كتاب [kitab] (book) etc.

1.4.2.6 The lemma

The lemma is the input form of the dictionary. It is the canonical (basic) form of a word, obtained after removal of affixes (prefixes, infixes, and suffixes) and application of rules allowed by the grammatical class of the word:

-Third person singular of the accomplished active verbs.

-Singular nominative for variable nouns (masculine for qualifying nouns) and classical voweled form for invariable nouns (singular masculine for pronouns).

Moreover, a radical is itself a flexed form and is derived from a lemma. Indeed, an inflected form in Arabic can be composed as follows:

-Lemme + Conjunction

-Lemme + Preposition

-Lemma + definedArticle

-Lemma + possessive article

Here we quote some examples of lemmas produced from radicals of the root [ك+ت+ب] [K + t +b]:

-The word يكتبون [yaktubuna] (theywrite) wasderivedfrom the lemma كَتَبَ[kataba] (hewrote)

-The word كُتُبٌ[kutub] (books) wasderivedfrom the lemma كِتأبٌ[kitab] (book).

-The wordمكاتب [makateb] (offices) wasderivedfrom the lemma مَكْتَبٌ[maktab] (Office).

-The wordمكتوبةٌ [maktuba] (written) wasderivedfrom the lemma مَكْتُبْ[maktub] (written).

1.4.2 The parts of the speech in Arabic

In the Arabic lexicon, words are divided into three types of grammatical categories:

-The Name: an element which indicates a place, a person or a thing and which can be of masculine or feminine and it van be of singular, dual or plural number. It can be defined or undefined. In Arabic, the adjective is a subclass of the noun which is placed after the one that it qualifies, it corresponds to the kind and the number of the name which it follows and undergoes the same rules of formation.

-Verb: A word to which the different terms of a sentence relate directly or indirectly is considered a verb in Arabic. It refers to an action or state of being.

-Particle: An entity used with the noun and the verb to identify a certain role in the sentence, which allows a coherent reading of it.

1.5 Applying linguistics on NLP

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken (Mihalcea, et al, 2006). NLP is a component of artificial intelligence. Whereas, applied linguistics is the branch of linguistics concerned with practical applications of languages studies (Wikipedia, 2018). This section sheds lights upon the issue

of the collaboration between applied linguistics and natural language processing (NLP), generally, and between applied linguistics and automatic translation particularly. The fundamental interest is about applications of linguistics to NLP. First of all, this section discusses the reasons why linguistic theories should be applied within NLP and what results of neglecting it are. Secondly, it will be followed by a discussion on selecting the most appropriate method that linguistics can contribute to the Arabic language particularly, and automatic translation in general. Thirdly, it analyses some linguistic studies that can be shared with NLP in order to improve the lexical disambiguation process.

1.5.1 The importance of applying linguistics to NLP

As it is well known, linguistics is the scientific study of language, it is assumed to cover everything about the organization of a piece of writing. According to an educational view, one of the essential aims of linguistics is to study the cognitive mechanism of language.

The impossibility of this process, led linguists to another process which is the attempt to match, or at least model, the language comptence of native speakers. This can be done by supervising the indirect outcome off his/her speech output, in addition to detecting the linguistic rules governing this output. Linguists attempt to detect and presuppose the basal units and rules of the targeted language in order to improve the translation work in general, and the process of WSD.

Previous studies and experiences in NLP have shown that the description of morphology, syntax, and semantics of the targeted language; which is Arabic in our thesis, in a natural language processing framework is necessary. The direct use of grammatical descriptions and monolingual dictionaries comes to be regarded as wrong step due to the imperfection and inconvenient of these sources which have a huge lack of quality in term of coverage, methodology and consistency. Moreover, in several researches, disregarding linguistics and linguists contributions has clearly shown negative results in terms of fineness and performance. Actually, the main link between applied linguistics and NLP can be clearly seen at the level of WSD; one of NLP huge pillars.

1.5.1.1 Morpheme

Morpheme is the smallest meaningful unit in a language, and the lowest level of language structure that is linked directly to NLP. The latter is concerned with the meaning of the treated text. Morphology and morphology are the two dimensions managing the

morpheme. Morphology realizes that although there are a few morphemes have various spellings (and articulations), they still identical. Morphology contains information and guidelines on the different exemptions from apparently clear guidelines. Thus, whereas adding (s) pluralize a huge number of nouns in English, quite few are not. On the other hand, the connotation of the same morpheme in a particular language can exceed to several interpretations and meanings, it depends on the position and function of the morpheme within a sentence. For instance, the same English suffix (s) can be interpreted as a verb, third person singular, plural, or non-possessive. The process of WSD considers the apostrophe as a huge obstacle due to the several possible interpretations. Thus, a full understanding of each morpheme's function in the sentence is essential. As there is a huge difference between sentence structure of standard Arabic and sentence structure of English, any error of interpretation leads to a wrong lexical disambiguation in general, and wrong translation particularly.

There are parts of speech such as Noun, Verb, Adjective, etc,. As it is known in linguistics, each part of them has an independent paradigm of word forms. Thus, NLP has shown a great interest in morphemes, especially when the same morpheme in English can be interpreted in a different part of speech, for instance:

- Noun: The well was empty.
- Verb: Tears well in her eyes.
- Adjective: she is feeling very well.

In order to determine whether the morpheme in the sentence is a noun, a verb, or an adjective, a lexical analysis of each sentence is required.

1.5.1.2 Homonyms & Homographs

In a WSD framework, homonyms and homographs are considered as the regular source of lexical ambiguity. The latter is connected with syntax and semantics in many ways, both of the syntactic structure of the sentence and the philosophical meaning of the sentence are major factors in the process of lexical disambiguation. Thus, the deep hole meaning of 'well' is excluded from consideration as a verb or adjective, simply because it is a noun. However the syntactic structure of 'Tears well in her eyes', determines 'well' as a verb.

Previewing the previous examples and other similar pieces of lexical disambiguation and linguistics are promptly available to a linguist. In a way or another, these studies are

linked and important for NLP, one significant issue for NLP is that the majority of these pieces of linguistic knowledge are not obtainable in any source and definitely not in any suitable shape. Thus, the one and only way to earn these linguistic abilities when they are necessary, is by having a permanent linguist in any project done within NLP frameworks. Presently, the reason the linguistics knowledge and studies are not available for NLP engineers and scientists is not because linguists monopolize it, but rather because of the sensitivity of it and the damages that can occur in language through any misuse of linguistic theories on NLP projects. Therefore, in any NLP project that takes into consideration efficiency and accuracy there must be a very qualified linguist.

1.5.2 Sublanguage:

A linguist can easily figure out the significant difference between linguistics and natural language processing, Linguistics is concerned with the nature of language and communication in general. However, NLP shows its interest in a restricted section of it. Sometimes, the one needs to zoom in to see some details in language, and that is exactly what the linguist does, and sometimes, he/she needs to zoom out to see the whole picture, and that is precisely the work of a NLP-er.

According to Victor Raskin : “Typically, an NLP project deals with a limited sublanguage of natural language, such as the language of a narrow area of science or technology. By doing that, NLP puts linguistics even further on the spot because to be useful, it would have to shed its most important, though for the most part unconscious idealization, namely that one native speaker's competence is identical to any other's. It is true that there are areas in linguistics, such as dialectology, sociolinguistics, and - recently and most unsurely of itself - linguistic pragmatics, which do not subscribe to the idealization.’’ (Raskin, 1985). Linguistics neglects the truth that, in many countries all around the globe the Arabic, English, Chinese, Spanish, etc., are languages that do not exist in one particular form. Instead, there are more than 20 Arabic dialect, and 150 or so Mandarin Chinese dialects, and several English and Spanish idiolects etc. What pursues is that it is possible that the guidelines formulated for a language will not match many of its dialects and idiolects. Moreover, for instance the English lexicon may be not used in its totality by any of English speakers. In addition to that, the syntax of English is used only slightly in any dialect.

1.5.3 Linguistics and MachineTranslation

Translation is considered as an independent field of study. However, since it involves the use of one or more languages, linguistics knowledge about languages cannot be ignored. Thus, it can contribute to the field of translation via an applied linguistic theory. Raskin stated that:“linguistics can be presented as the ability to determine whether some two texts, each in a different language, are translations of each other. In order to be translations of each other, the texts should probably satisfy two texts in different languages are translations of each other if they have the same meaning, illocutionary force and perlocutionary effect and inference’’(Raskin, 1985). In light of the above, the researchers decided to adopt and apply two linguistic theories; on the one hand, the use of Arabic diacritical marks. On the other hand, Chomsky’s syntactic structures.

1.5.4 Arabic lexicon composition

The traditional Arabic grammar knows only three subsets: nouns, verbs and particles. This ranking shows, quickly, its limits when it comes to a natural language processing. Actually, the class of particles has been extended to include grammatical morphemes which, in reality, belong to other classes such as demonstrative or relative pronouns which constituted particular nominal entries.This extension leads to a reorganization into four subsets of the Arabic lexicon: verbs, nouns, pronouns and word tools [Kouloughli, 1991] and [Khoja et al., 2001].

1.5.4.1 Verbs

a verb is an entity expressing a time-dependent meaning. the majority of the Arabic verbs are formed on radicals of three consonants as is the case of the verb and possibly 4 consonants as is the case of the verb“ كَتَبَ (*kataba* – write). And possibly 4 consonants as is the case of the verb “ دَحرَجَ (*daḥrağa*- strundle) .These roots can give rise to several schema or patterns as a result of one or more morphological transformations. Several linguistic studies have been conducted on the verbal system in Arabic.

1.5.4.1.1 Verbs with developed root

Several schema were added to the simple verbs. These schemas are formed from simple roots by a set of morphological operations to give a particular meaning to the verbs results.

Verbal schemata	Associatedmorphologicaloperation
<<فَعَلَ>>(fa‘‘ala)	repitition of the second radical consonant.
<<فَاعَلَ>>(faā‘ala)	lengthening of the first radical consonant by adding a " (ā – álif).
<<أَفْعَلَ>>(áfa‘ala)	Adding "أ"at the beginning of the root.
<<تَفَعَّل>>(tafa‘‘ala)	Adding "ت"at the beginningog the root in addition to the repitition of the second radical consonant.
<<تَفَاعَلَ>>(tafaā‘ala)	Adding "ت"at the beginningog the root in addition to the lengthening of the first radical consonant by adding a "أ".
<<إِفْتَعَلَ>>(ìfta‘ala)	Adding "إ"at the beginning of the root in addition to the insertation of the new morpheme "ت".
<<إِنْفَعَلَ>>(ìnfa‘ala)	Adding a bi-consonant morpheme " إِنْ at the beginning of the root.
<<إِسْتَفْعَلَ>>(ìstaf‘ala)	Adding a tri-consonant morpheme " إِسْتَ at the beginning of the root.

Table 6 : verbs with developed root

1.5.4.2 Nouns

The morphological system of Arabic nouns distinguishes three subcategories:

- **1.5.4.2.1 Primitive nouns**: they are nouns that cannot be attached to a verbal root. They constitute the fundamental glossary of the concrete language. Example : يَدْ (hand), قِطٌ(cat), etc,.
- **1.5.4.2.2 Derived Or deverbal nouns:** these are the nouns that can be derived from a verbal root. The number and the nature of these forms vary according to the status of the verb to which they are attached. As names, they can receive the marks of the case, gender, and indeterminacy.

- **1.5.4.2.3 Numbers:** this category of nouns consists of simple numerals representing the units: from "صِفْرٌ (Zero, 0) to "تِسْعَة (Nine, 9), ; dozens : "عَشَرَة (Ten, 10), "عِشْرُونَ(Twenty, 20), etc,. In addition to hundreds and composed number such as: " أَحَدَ عَشَرْ (Eleven, 11).

  In their decomposition, Arabic grammarians have equated adjectives with nouns since they take almost all morphological forms; they can, for example, be defined or indefinite and bend according to the case, the number and the kind.

1.5.4.3 Pronouns

The class of pronouns has been introduced as an extension of the traditional decomposition system of the lexicon of the Arabic language. This class groups together some forms that were considered as particular names. This set escapes any rule of derivation. Pronouns form a closed list of words. In this list, we distinguish :

- **1.5.4.3.1 Demonstrative pronouns**"أسماء الإشارة" :they represent a subcategory of pronoun expressing a demonstration idea. They makeit possible to indicate that the represented object is found, either in the text, or in the space or the time. Thus, There are two subsets: the demonstratives of proximity, for example :
  "هَذَا"(This), "هَؤُلاَ ءْ"(these).
- **1.5.4.3.2 Relative pronouns**"الاسماء الموصولة" : they refer to the name or personal pronoun which proceedes them and which we designate by antecedent. Relatives agree with the antecedent and can only be adapted to the duel (as demonstratives). Among the relative pronouns, for instance : "الّذي" (That), "اللَّتَيْنَ" (Those, used for female dual) etc,.
- **1.5.4.3.3Personal pronouns**"الضمائر المفصلة" :they serve to designate the three grammatical types of people :
  - ✓ the first person, thatis, the enunciator or speaker : "أَنا"(i), or "نَحنُ (we).
  - ✓ the second person, thatis, the recipient or interlocutor : "أَنتَ(you, male), "أَنتِ" (you, female), "أَنتُم"(you, male, plural), "أَنتُنّ (you, femlae, plural).

1.5.5 Arabic diacritical marks problem

In order to describe the problem of the Arabic diacritical marks , we depend on the definition of Joseph Dichy: "The current Arabic writing does not note the short vowels, the case marks composed of a brief vowel followed, for the names and the indefinite adjectives, a consonant "ن" (n- tanwin), etc. Thus, we discuss the issue of 'non-diacritized' writing. These signs of diacritical marks, which are realized, when noted, in the form of diacritic signs placed above or below the letters, appear in certain religious texts (holy Quran or hadith) or literary (classical poetry, in particular): we will say that they are edited in vowel. In addition, there are two practices, that of fully diacritized writing and that which is partially diacriticalized. The partial singulation answers, in the neat editions, the lifting of certain ambiguities of first reading [...]. It should be noted that the partial diacriticalisation does not adopt a codification based on a tradition: it does not therefore have a systematic character."(Dichy, 1997). The

automatic analysis of Arabic faces the problem of the absence of diacritics in texts. This problem leads to several cases of lexical ambiguity in Arabic given the polysemous nature of non-diacritirized words. The problem of multiple diacriticalization, is comparable to that of the multiple Russian language letters. For instance, the non-diacritized Arabic word"علم" can have several diacritical marks which lead to different interprétations:

Diacrx²itical marks	Transliteration	Translation	Grammatical category
عَلِمَ	'alima	Knew	Verb, active voice, 3rd person, masculine, singular
عُلِمَ	'olima	Wasknown	Verb, passivevoice, 3rd person, masculine, singular
عَلَّمَ	'allama	Taught	Verb, active voice, 3rd person, masculine, singular
عَلَمٌ	'alamun	A flag	Noun ,Masculin, singular
عِلْمٌ	'ilmun	A science	Masculin, singular

Table 7:potential of the word "علم"

This example shows that the unvoweled simple word (علم) has a severalpotentialdiacriticalmarks ,many possible grammatical categories. A statistical study was done by F. Debili shows that only 19% of the words in a Arabic corpus are unambiguous and the average of six possible diacriticals per word [Debili, 2001]. Thus, given the large proportions of the ambiguous words, we can conclude that, for Arabic, this phenomenon is very frequent and requires serious studies.

1.6 Chomsky's generative grammar

Chomsky has had to weigh the terms carefully since he seeks to avoid a continuous misunderstanding. He did not say that the sentence is engendered by competence (or

grammar, which is the same thing for him); he did not say that the grammar assigns a certain structure to the sentence: it assigns a description to it. We are far from traditional grammar, which serves not only to recognize the structure of a heard sentence, but also to give a certain structure to the uttered sentence. Chomsky resumed from traditional grammar the indication of the syntactic functions (subject, predicate, object, etc.) and the grammatical characteristics (common noun, transitive verb, etc.). He concluded " it seems to me that the information presented is, without question, substantially correct and is essential to any account of how language is used or acquired. The main topic I should like to consider is how information of this sort can be formally presented in a structural description, and how such structural descriptions can be generated by a system of explicit rules". (Chomsky,1957)

As we see, Chomsky takes over the traditional description of the syntactic structure of this sentence; he simply wants to present it formally; and he does it the following way:

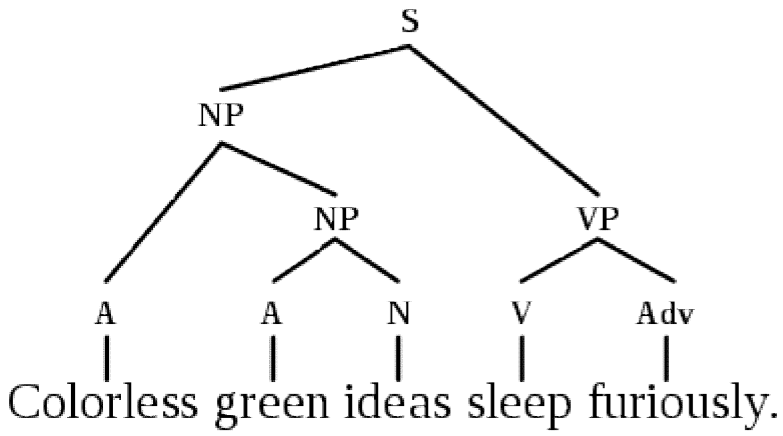


Figure 4: a tree diagram of the sentence "colorless green ideas sleep furiously"

The issue of lexical disambiguation will be discussed later; it is clearly here that the Figure 1 faithfully represents the traditional analysis. It is also good to return to a detail: grammar does not generate the sentence, but the description of a given sentence: Chomsky first poses ‘S’, that is to say the known sentence, and not at the recreates. This is what the professor does. Moreover, the man who already masters the rules of grammar: he explains to other linguists how to apply the rules to a spoken or written sentence.

In light of the above, and depending on the researchers’ background knowledge, Chomsky’s syntactic structure theory comes to be regarded as an excellent and prominent theory to be applied within the process of lexical disambiguation. By giving the grammatical

features of a single word, it would be easy for the translation engine to identify the appropriate word in the second language; which is Arabic in our thesis.

1.7 Conclusion

In this chapter we have defined the task of lexical disambiguation in general by presenting the process of implementation, the useful resources, as well as the approaches for implementation. We identified two ways to evaluate a lexical disambiguation system. In addition to that, we presented the methods of lexical disambiguation applied to the Arabic language and its contribution to statistical machine translation.Moreover, as we have mentioned, many researchers have noted the absence of reference corpus in the Arabic language and thus justify the constitution of their own corpus, but they have not yet produced a corpus of free access reference built on a common resource base, also we have mentioned the importance of applying linguistics into the field of natural language processing in both statistical and automatic translation, and how the tree diagram model that was introduced by Chomsky helps to identify the word category in order to ease the process of disambiguation.

CHAPTER TWO

RESEARCH METHODOLOGY

2.1 Introduction

We can indicate that lexical disambiguation is the task of choosing the right meaning of a polysemic word in a given context. Lexical disambiguation or word-sense disambiguation is a field of research that has attracted a lot of interest since the 1950s. Indeed, the accomplishment of this task is necessary and useful in most work in natural language processing studies and projects. The focus in this chapter is on the methods, tools and the procedures followed in this research and it is presented as follow:

The first part deals with the research design that is applied in the research. The next part covers the description of the research setting, this part is concerned with the several tools used in the study, it is related to the choice of the English and Arabic analyzers and translation engines. The following part examines the data collection instruments employed in this research. The fourth part, and the last, is devoted for explaining how the researchers are going to analyze the data that is obtained from the instruments.

2.2 Research design

When conducting any research, researchers use a variety of approaches and methods for gaining deep understandings of the phenomenon being studied. Hence, in this research, In order to find out the most suitable approaches and methods that are effective and can improve the process of word sense disambiguation, and test the efficiency and accuracy of the translation, the researchers decided to use both of qualitative and quantitative (mixed method) research design, A single data set that is either collected qualitatively and analyzed quantitatively or vice versa.

The investigators used this approach for the following reasons. First of all, mixed methods research design’s studies are the most appropriate methods that are concerned with the field of computational linguistics in general, and word sense disambiguation particularly. Thus, the researchers are able to study and analyze linguistic characteristics of Arabic and English, translation issues and problems that could occur inside A setting which will shed new light on an established theory that might result in new and further explorations. Moreover,

when conducting a mixed method case study, researchers will use a variety of instruments and tools in the process of data collection, the most relevant tool in mixed method case study is the theoretical drive. The latter leads the researchers to focus on the importance of the accuracy and the efficiency of the several translation tests beside other instruments that are to be mentioned and identified in the section of data collection. Furthermore, data are collected qualitatively and analyzed quantitatively by developing themes that emerged from the findings that are found through data collection, to be finally presented in a form of detailed reports. In addition to that, the choice of case study's type was based on that it is the researchers' intention to explore the way in which the process of improving the lexical disambiguation of the Arabic language within the automatic translation.

2.3 Research setting

This study has been done during the second semester of the academic year 2018/ 2019, it lasted four months, and specifically it was from March until June. As long as word sense disambiguation studies are done in a computational linguistic framework, the research setting of this study is boils down to the different tools and software used in the process of disambiguating ambiguous words, such as resources, lexical bases of Arabic and other tools.

2.3.1 Lexical bases

2.3.1.1 WordNet

The Arabic WordNet (AWN) (Black and ElKateb , 2004 , Christiane et al. , 2005 , El-kateb et al. , 2006) is a lexical database available online, which follows the design and content of Princeton WordNet (PWN) (Miller , 1995) and which contains 23,481 Arabic words (including 21,833 translated words) and 11,269 Arabic synsets (of which 10,127 synsets translated by various contributors). The sets of synonyms are linked between them by linguistic relations in general, and specifically by semantic relations (hyperonymy, antonymy, etc.). In Fact, the Arabic WordNet is aligned with Princeton WordNet as well as with all WordNets that are aligned with the PWN, either directly or indirectly through an interlingual index (ILI) and the SUMO ontology (Suggested Upper Merged Ontology) (Black et al. , 2006)

2.3.1.2 BabelNet

BabelNet is a large online multilingual linguistic network, providing glosses (lexicalized entries) in several languages, including the Arabic language. It is built by automatically integrating WordNet, Wikipedia, Wiktionary, Wikidata as well as OmegaWiki. The Arabic part of BabelNet version 4.0, includes 2 942 886 synsets and 4,267,782 meanings of words among which 102,704 meanings translated from WordNet and 144,952 senses translated from Wikipedia.

2.3.1.3 Arabic PropBank

The Propbank is a semantically annotated corpus that existed at first in glais and Chinese. Since 2009, the state of the art of Arabic TAL has been enriched by Propbank Arab (APB) (Palmer et al. 2010). This corpus includes annotated sentences with their predicate-argument structures as well as their labels of linguistic roles. The Arabic Propbank consists of two parts: frames files containing the different predicates with their arguments, and an annotated syntactic corpu with the help of Penn Treebank Arabic (Maamouri et al. , 2004) to annotate in semantics roles the predicates.

2.3.1.4 Wikipedia in Arabic

Wikipedia is a free online encyclopedia, universal, multilingual. However the Wikipedia edition in Arabic is still poor and limited (with 589 917 articles) in comparison to some languages such as English (5 721 476 articles), French (2,042,399 items), etc.

2.3.2 Tools

2.3.2.1 MADAMIRA

The morphological analyzer MADAMIRA (Pasha et al. , 2014): is a system of morphological analysis and Arabic disambiguation that is freely available, it exploits some the aspects of the major two existing systems and widely used for the automatic processing of the Arabic language are: MADA (Habash , 2005) and AMIRA (Diab , 2009). In fact, MADA-MIRA allows tokenization, lemmatisation, rooting, morphosyntactic, morphological disambiguation, diacritisation, recognition of named entities, etc. As described in Table 3.1, MADAMIRA proposes the two token schemes. following:

ATB: consists of segmenting all clitics except
defined items, and to standardize the characters ALIF and YA using the character 'PLUS' as a marker of clitics.

D3: consists in tokenizing the different kind of proclitics as well as all articles and enclitics . In addition, it normalizes the characters Alif and Ya after the devoyelization process of Arabic characters.

Segmentation	Sequance of words
Brut	ربراب سيخلى سبيله
D3 (Tokenized form)	ربرابس+يخلى سبيل+ه
ATB (Diacritized form)	رَبْرَابْ سْ+ يُخْلَى سَبِيلُ+هْ

Table 3.1 : Example of D3 and ATB

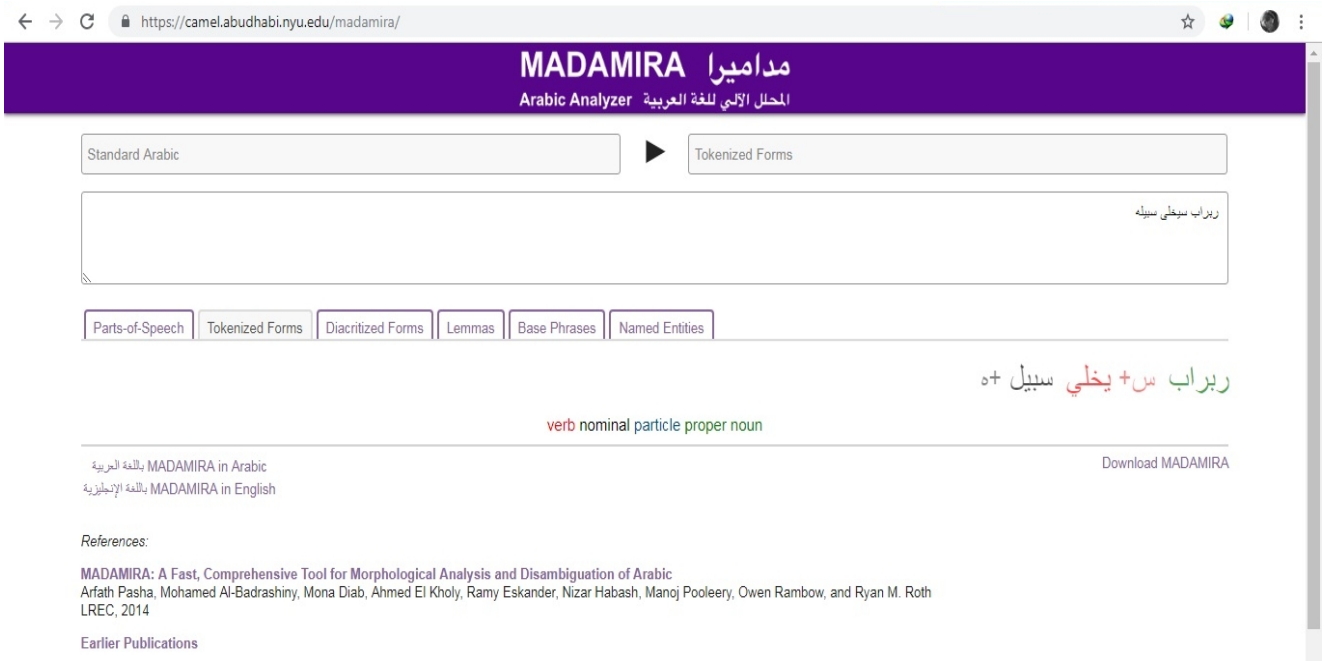


FIGURE 3.1 : Example of tokenized form of MADAMIRA

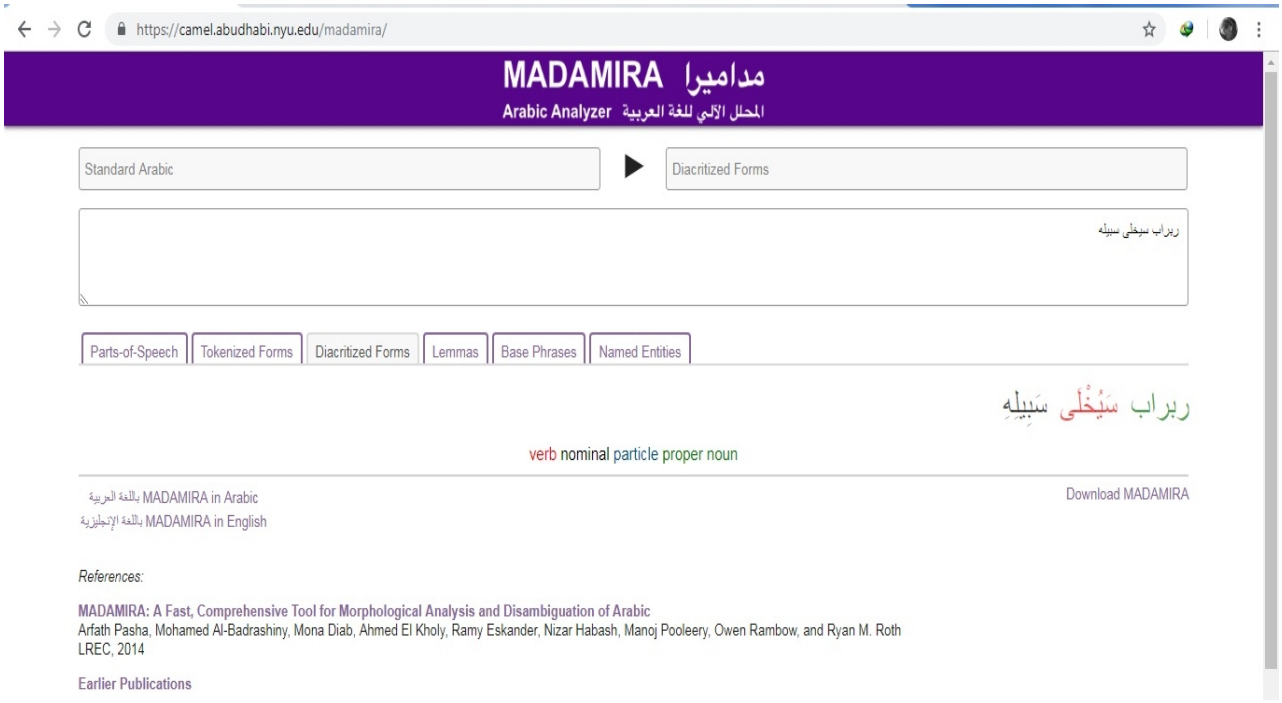


FIGURE 3.2 : Example of diacritized form

2.3.2.4 The morphological Stemmer of Khoja

In their Arabic morphological stemmer available online, Khoja (2001 , 2003) combining statistical techniques and linguistic rules to obtain is a magnificent method in order to achieve high accuracy rate. They used 131 different labels that are ranked in five categories: noun, verb, particle, residual and punctuation. the adjective, the pronoun, the name the common name and the numeral belong to the category name. For train the system, they exploited a corpus of 50,000 words from the Saudi newspaper Al-Jazeera, and got 90% in accuracy.

2.4 Data Collection

This section presents the instruments that are employed to gather the necessary data for this study. The present research seeks to investigate whether analyzing and applying linguistics methods to automatic translation would enhance the level of Arabic-English and English-Arabic translations. For this reason, we are going to explore the case through involving several instruments in the process of collecting data. These instruments include: automatic translation based on: applying linguistic approaches to translation and a lexical analysis of the translated words, especially ambiguous words. In addition to the use of translation engines for more accurate and sharp translation results' analysis in order to

examine the process of translation not only at the level of lexical disambiguation, but also semantically. Thus, varying sources of data helps make this research more comprehensive and valid.

2.5 Description of the experiment:

"researchers from different linguistics areas employ methods originating in experimental linguistics and psycholinguistics to test various linguistic theories, models and concrete research hypotheses. The choice of a method or a methodological approach always means commitment to a particular experimental design" Mertins, B. (2016). In order to collect different answers about the linguistic performance of diacritical marks and Chomsky's syntactic parsing theories, many experiment were done in pursuance of evaluating the impact of the theories mentioned above on lexical disambiguation. Thus, we intend to implement a two-phase experiment in order to linguistically compare translation results of pre-implementation and post-implementation. Moreover, we propose four sentences; two sentences in Arabic and two sentences in English. The latter share the same English morpheme "stone" which is going to be the center of ambiguity. However, the former share the same Arabic word "علم".

2.6 Researchers needs

Researchers in linguistics, or even more generally in NLP, often need powerful softwares so that they run them to work on morpho-syntacticaly disambiguated corpora. These softwares must be able to generate the learning data needed for implementation process in order to achieve a fine automatic lexical disambiguation. Since our main focus is on the linguistic approaches used in WSD, their needs should be in the same theoretical framework. Thus, a powerful search tool of linguistic phenomena allowing editing of the possible interpretations and meaning of a single word comes to be regarded as the most important need due to its huge importance.

2.7 Implementation process

Three steps are needed in order to implement the automatic lexical disambiguation (Schwab , 2017).

1. The use of a generic resource: several resources not dedicated to the lexical disambiguation are possible: dictionaries, encyclopedias, non-annotated corpus, annotated corpora, lexical bases ...etc,.This optional step is often conducted after a linguistic edition process.

2. Establishment of a resource dedicated to lexical disambiguation: use one or more raw resources to give an informative representation for every single meaning of a bunch of possible interpretations of a word. In addition to a linguistic segmentation process which would help researchers define the linguistic features of every single word of the sentence. This linguistic segmentation is based on Chomsky's syntactic structure studies. Thus, it is a question of constituting a resource dedicated to the task. These senses are defined by human expertise or induced from contexts of use in texts.

3. Applying linguistic theories to disambiguate texts:
it is about the disambiguation algorithm itself. Several factors can be taken into consideration. Some are common to every algorithm like the size of the context considered for the word to be disambiguated (for example a few words before or after it, the sentence that contains it, even the type of text). In addition to the syntactic analysis tree diagram which provides researchers with grammatical structure of the sentence, and the Arabic diacritical marks.

2.8 Evaluation of the lexical disambiguation

In order to evaluate a system of lexical disambiguation, Two approaches are possible:

- **2.8.1 In vivo**: the systems are evaluated according to their contrubution to the performance of a particular application (automatic translation, search of information, etc.). In this thesis, the translation of an ambiguous word is going to be evaluated within the sentence. In other words, in vivo evaluation is another way to describe the semantic evaluation.

- **2.8.2 In vitro**: systems are evaluated using reference annotated corpora, in particular those of the evaluation campaign Ontonotes 5.0 that have existed for twenty years. In this thesis, the translation of an ambiguous word is going to be evaluated separately. Thus, in vitro evaluation comes to be regarded as a syntactic evaluation.

2.9 Evaluation tool :

For the evaluation, we use the Arabic part that we have updated from OntoNotes Release 5.0 and the classical evaluation measures of lexical disambiguation as used in this software is defined as the classical measures which are the precision P, the recall R and the F1 score which corresponds to the harmonic mean of P and R. The precision is defined as :

$$P = \frac{correctly\ annotated}{total\ annotated}$$

$$R = \frac{correctly\ annotated}{total.to\ annotate}$$

$$F1 = \frac{2 * P * R}{P + R}$$

2.10 Trustworthiness of the research

In order to obtain the trustworthiness for the conducted research, and to test the relevancy and accuracy of the provided evidences, the researchers used several techniques to validate their collected data by establishing:

2.10.1 Credibility is defined as the confidence that can be placed in the truth of the research findings (Holloway & Wheeler, 2002; Macnee& McCabe, 2008). Hence, Credibility establishes whether the research findings represent plausible information drawn from the softwares' original data. Therefore, the researchers adopted triangulation as a first tool to validate their data; triangulation "involves the use of multiple and different methods, investigators, sources and theories to obtain corroborating evidence" (Onwuegbuzie & Leech, 2007, p. 239). As it was mentioned above, this research strived to use four instruments in the process of data collection, namely theoretical linguistic approaches, translation softwares, disambiguation tests and their results. Another tool that is used to authenticate the collected data is the selectivity of different corpora used in this research. Mixed method research data collection requires the researcher in the field of NLP to put high standards in selecting corpora and tools. This helps the researcher to gain an insight into the context of the study, which minimizes the distortions of information that might arise due to the presence of the researcher in the field. The researcher's extended time in the field improves the trust of the respondents and provides a greater understanding of participants' culture and context (Onwuegbuzie & Leech, 2007). In This thesis a great understanding of the linguistic features was provided.

2.10.2 Transferability: Transferability refers to the degree to which the results of mixed method research can be transferred to other contexts with other respondents – it is the interpretive equivalent of generalizability (Bitsch, 2005; Tobin & Begley, 2004). According to Bitsch (2005), the "researcher facilitates the transferability judgment by a potential user

through ‘thick description” (p. 85). This means that an extensive set of details of research processes, from data collection, context of the study to production of the final report must be mentioned in the research for enabling future researchers to replicate the study with similar conditions in other settings. Simultaneously, all these information have been mentioned within this chapter.

2.10.3 Dependability: dependability refers to “the stability of findings over time”. (Bitsch, 2005, P.86). Consequently, in order to address the dependability issue more directly, the processes that took place within the study have been reported in detail, thereby enabling future researchers to repeat the work, if not necessarily to gain the same results.

2.10.4 Conformability: Conformability is “concerned with establishing that data and interpretations of the findings are notfigments of the inquirer’s imagination, but are clearly derived from the data” (Tobin & Begley, 2004, p.392). In order to achieve this criterion, researchers used two tools. The first one is audit trial, which is a transparent result of the research finding report and experiments. Second, the use of reflexivity was necessary; reflexivity is “an assessment of the influence of the investigator’s own background, presuppositions and interests on the mixed method research process that includes the researcher’s personal history.” (Krefting, 1991, p. 218). It helps in reducing the effect of investigator bias; as a result, the researchers major which is applied linguistics, in addition to the huge interests of NLP to help ensure as far as possible that the work’s findings are the result of the experiences and ideas of the informants, rather than the researchers’ subjectivity.

2.11 Conclusion

This chapter is devoted for describing the methods and the procedures followed in this research. Researchers adopted a mixed method Case Study for the reasons that have been already mentioned; it is conducted in a computational linguistic framework due to its match with the characteristics of the research. Many tools and softwares were used in this research in order to facilitate the implementation process of the linguistic theories and approaches. The mother tongue of researchers is Arabic. Therefore, Arabic was selected to be the target language in the experiment of the lexical disambiguation. In order to have a deep understanding to the case being studied, the researchers employed several instruments for collecting data. The way these instruments are analyzed is what has been mentioned before the end of this chapter. The last point that has been tackled in this chapter is related to the establishment of this research's trustworthiness, the researchers adopted four tools to validate their collected data. These tools are credibility, transferability, dependability and conformability. Each one of these tools has sub-tools that are used to achieve the desired aims. The results obtained in this study will be presented and findings will be discussed in the coming chapter

CHAPTER THREE

IMPLEMENTATION AND RESULTS

3.1 Introduction

The present study aims to enhance the efficiency and the accuracy of the lexical disambiguation process in translation in general, and Arabic in particular, through implementing linguistic theories. This chapter provides a detailed description and analysis of the implementation process and the data obtained from the observations and test results. The analysis takes the form of thematic analysis which is a method of identifying, analyzing and reporting patterns within data. As a result, the purpose of this chapter is to present and discuss the emergent themes which consist of the main findings of this study, where we are going to interpret the findings in order to answer the research questions set in the general introduction of this thesis. After stating our linguistic theories, we proceed to the implementation of lexical disambiguation systems for Arabic by exploiting our resources. First, on the one hand, we present a theory of lexical disambiguation for Arabic: a system based on the morphological analysis which provides diacritical marks for Arabic words. It will be followed by an explanation on how this morphological analysis enhance the process of lexical disambiguation. This theory is evaluated on the corpus OntoNotes 5.0 and Yandex translation engine. On the other hand, we present another theory of lexical disambiguation for English: a system based on Chomsky's syntactic structures. It will be followed by an explanation on how the syntactic analysis improves lexical disambiguation. This theory is evaluated on the corpus OntoNotes 5.0 and Yandex translation engine. Moreover, in order to prove the effectiveness of the our methods of translating, a comparison between translations before and after applying our theories is going to be done.

3.2 Pre-implementation

In This phase, we decided to translate several examples before applying any linguistic theory. Thus, any linguistic or morphological difference between pre-implementation translation and the post-implementation translation is considered as the contribution of researchers. Example 1 and Example 2 refer to the pre-implementation translation test of diacritical marks theory. However, Example 3 and Example 4 refer to pre implementation test of Chomsky's syntactic structures.

Example 1 :



FIGURE 3.1: Pre-implementation of the sentence "علم ياسين بالأمر"

Example 2 :



FIGURE 3.2: Pre-implementation of the sentence "هواري رفع علم بلاده"

Example 3 :

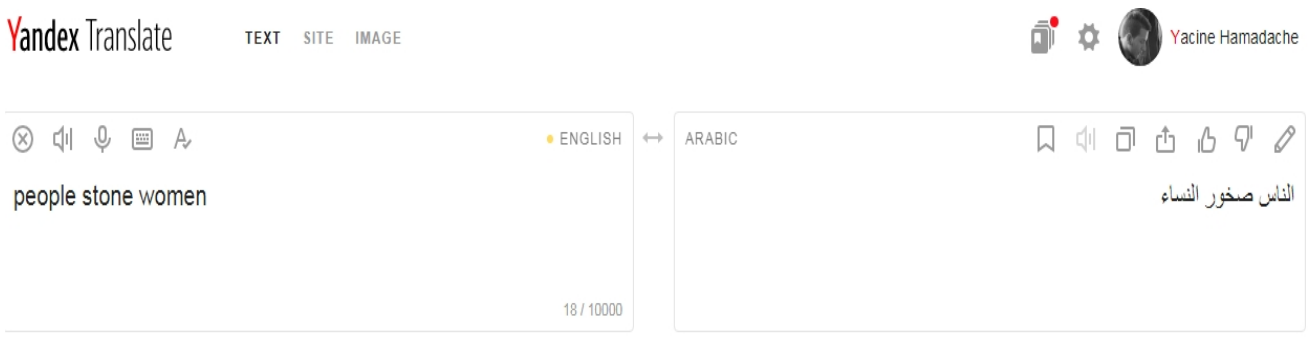


FIGURE 3.3: Pre-implementation of the sentence “people stone women”

Example 4 :



FIGURE 3.4: Pre-implementation of the sentence “this is a stone wall”

3.3 Implementation

In this section, the we decided to implement two theories; diacritical marks theory and Chomsky’s syntactic structures.

3.3.1 Diacritical marks theory

In order to implement diacritical marks theory, we chose two Arabic sentences:

- علم ياسين بالأمر
- هواري رفع علم بلاده

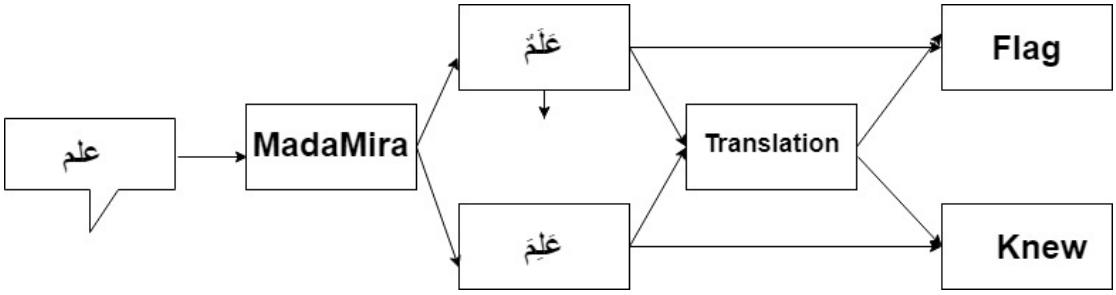


FIGURE 3.5: Process if providing diacritical marks to the Arabic word "علم"

We applied the following process to the chosen sentences:

1. Segmentation of the sentence; using the tokenized form schemata of Madamira
2. Diacritization of the sentence; using the diacritized form schemata of Madamira
3. Exploiting Madamira's morphological analysis and linking it to Ontonotes 5.0 and Yandex translation engine
4. Translating two Arabic sentences that include the same word 'علم'.

3.3.1.1 Segmentation

''The top-scoring analysis of each word can then be Passed the Tokenization Component to Generate Customized Tokenization (or Several) for the Word, According to the schemes requested by the user''. Pasha et al,2014

In Arabic we find several clitics that stick to the word, leading to Morphological and orthographic ambiguities. Thus, a segmentation process is necessary.

Example1:



FIGURE 3.6: Segementation of the sentence "علم ياسين **بالأمر**

Example 02:



FIGURE 3.7: Segmentation of the sentence "هواري رفع علم بلاده"

3.3.1.2 Diacritization

‘‘Arabic has a high degree of ambiguity resulting from its diacritic-optional writing system and common deviation from spelling standards’’. Pasha et al,2014

in light of the above, a diacritized form of the two sentences was provided:

Example 1:

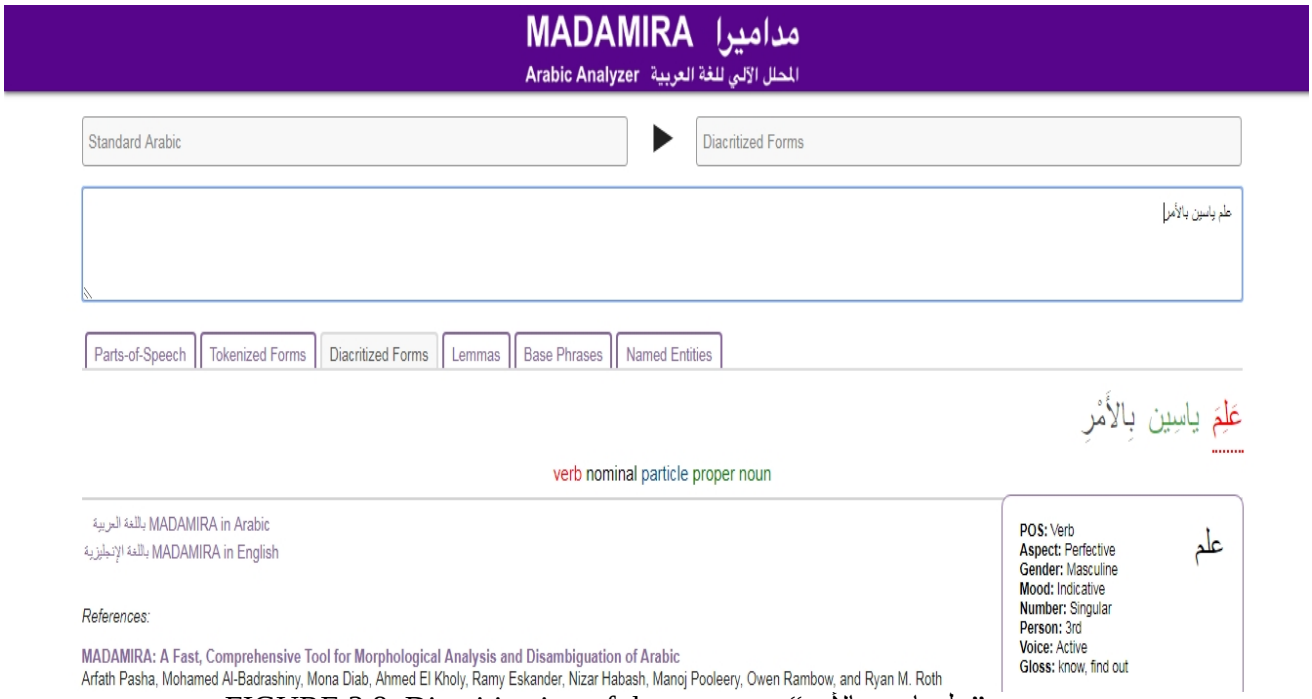


FIGURE 3.8: Diacritization of the sentence "علم ياسين بالأمر"

Example 2:



FIGURE 3.9: Diacritization of the sentence "هواري رفع علم بلاده"

3.3.1.3 exploiting Madamira

(Pado and lapata, 2009) used annotation transfer to automatically obtain annotations for new languages, using available resources for English and exploiting parallel corpora. Their

objective was to transfer semantic roles from English to languages less rich in resources. They carried out an experimental evaluation of a parallel English German corpus that proves the feasibility of their approach. In light of the above, we decided to adopt Pado and Lapata's method and apply it to their thesis. Thus, an exploiting procedure was done aims to exploit the morphological analyzer Madamira within OntoNotes 5.0 and Yandex translation engine. The focus was on the different interpretations of the Arabic word 'علم'. It will be argued that the diacritical marks provided by Madamira are the major disambiguation items in the process of the lexical disambiguation of the Arabic word.

3.3.1.4 Translation

The translation test is done on the platform of Yandex translation engine and OntoNotes 5.0. The OntoNotes Project [Weischedel et al., 2015] is the result of collaborative work between BBN Technologies, the University of Colorado Th latter comes to be regarded as an open source software used by many NLP researchers. Thus, we have chosen these tools mentioned above, to be our implementation platforms.

3.3.2 Chomsky's syntax tree

In order to implement Chomsky's syntax structure parsing, we chose two english sentences:

1. 'People stone women'
2. 'this is a stone wall'

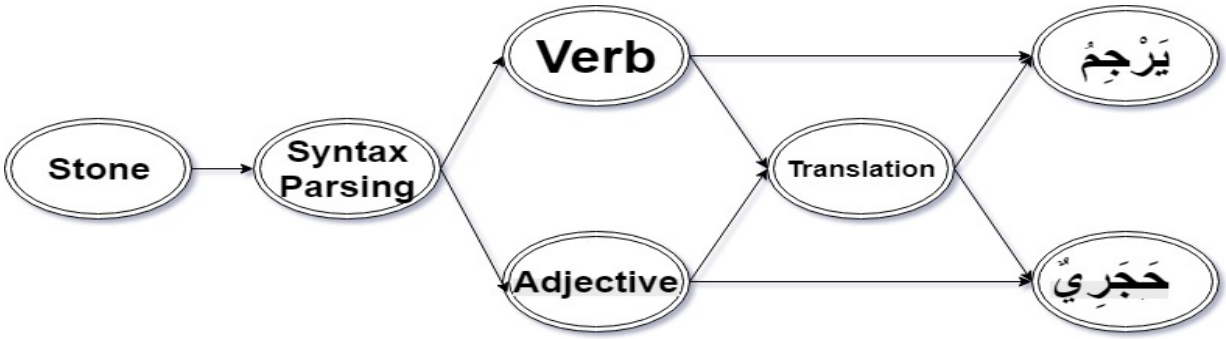


FIGURE 3.10: Syntactic Parsing process of the the morpheme "stone"

We applied the following process to the chosen sentences:

1. Syntax parsing
2. Word classification

3. translation

3.3.2.1 Syntax parsing

Syntactic parsing is the task of recognizing a sentence and assigning a syntactic structure to it. (Jurafsky, 2018). Thus, a syntactic parsing was done according to Chomsky's theory

Example 3:

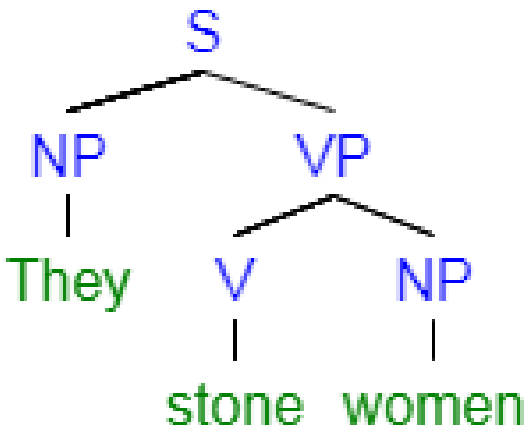


FIGURE 3.10: Syntactic Parsing of the sentence "they stone women"

Example 4 :

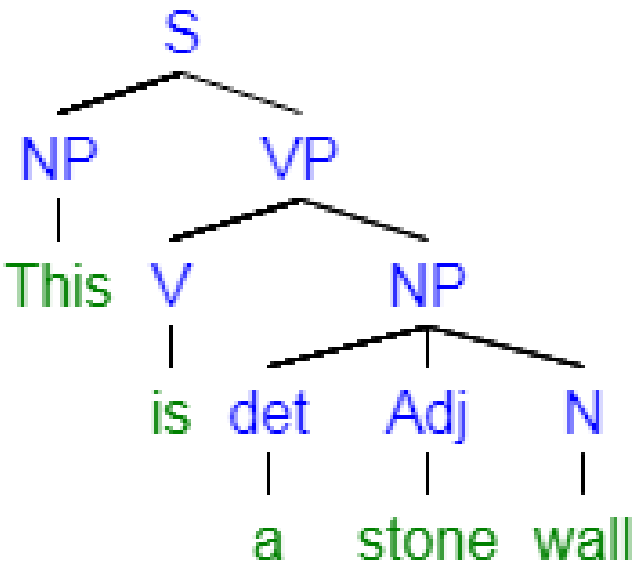


FIGURE 3.11: Syntactic Parsing of the sentence “This is a stone wall”

3.3.2.2 word classification

as we can see in the two example (see 3.2.2.1) the same English morpheme ‘stone’ can signify a different part of speech. Thus, a word classification process is needed in order to classify it to a verb, in the first example, and as an adjective, in the second example.

3.3.2.3 Translation

we provided two different translations of the English word ‘stone’ into arabic, the first translation; ‘يَرْجِمُ’ is provided for the verb form. However, the second translation; ‘حَجَرِيٌ’ is provided for the adjective form of the word ‘stone’. we integrated these different translations in Yandex translation engine’s bilingual corpus.

3.4 Reporting Results and Discussing Findings

In this section, we are interested in evaluating the impact of the lexical disambiguation on machine translation systems which allow us to evaluate our system of lexical disambiguation by automatic translation with an in vivo evaluation method (see section 2.9.1). The present study aims to investigate whether the use of diacritical marks and Chomsky’s syntactic structures theories are effective in enhancing the process of lexical disambiguation. Therefore, this section provides a detailed description and analysis of the data

obtained from the several test results. As illustrated in both Figures 3.5 and 3.10, we are going to evaluate the effect of lexical disambiguation through AR-EN translation results. First of all, we applied on each sentence, the schemata of tokenization D3 of MADAMIRA (see 3.6 and 3.7), which consists in tokenizing the proclitics and the clitiques, as well as all articles and enclitics. Furthermore, the vocabulary is thus reduced and the words are treated uniformly without prefixes or suffixes. They were followed by applying the schemata of diacritized forms which provides a diacriticalization to the Arabic words. Thus, the difference between the first and the second diacriticalizations of the word '**علم**' is the fundamental aspect that led to the different translations of the same word. As we can see in 3.12 The Arabic word '**علم**' has been translated into 'knew' and in 3.13 it has been translated into 'flag'.

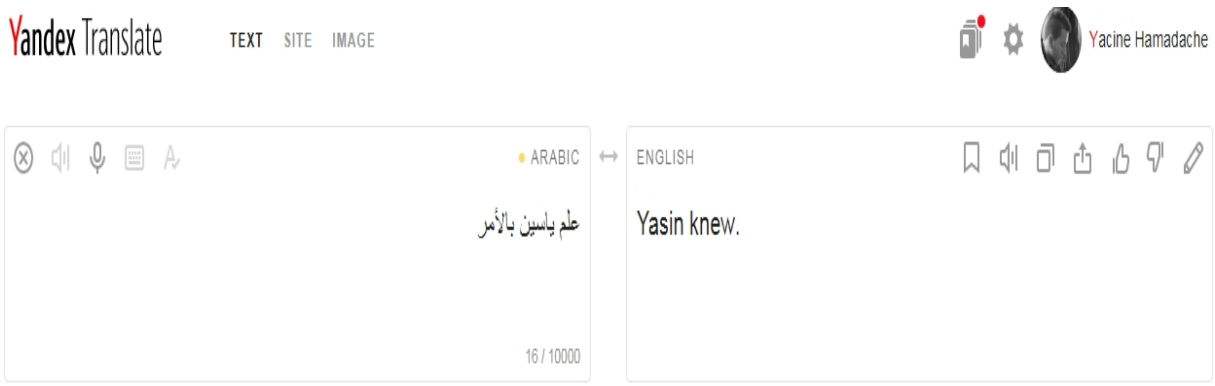


FIGURE 3.12: Implementation result of the sentence "علم ياسين بالأمر"

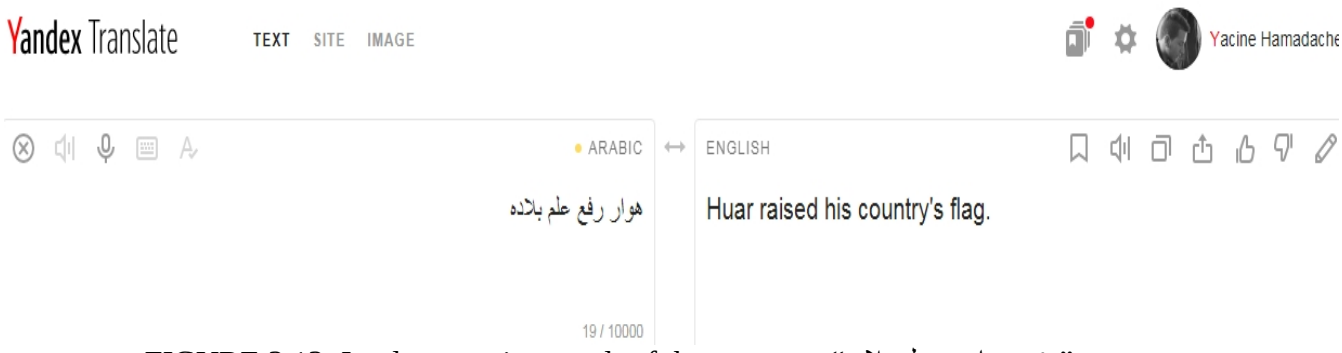


FIGURE 3.13: Implementation result of the sentence "هواري **رفع** علم بلاده"

Figures 3.14 and 3.15 present the performances obtained by integrating the Arabic diacritized marks in translation systems .On AR-EN systems, by exploiting Arabic

morphological information with the words (Word+Sense+Lemma+POS+Sense). We significantly improve our translation by +8.71% and +23.85% in 3.14 and 3.15 respectively compared to the factorised system in terms of BLUE.

	Données
Source: Mot	<علم ياسين بالأمر.>
Source: Mot+Sens	علم+|علم+ ياسين|+|+ياسين ب-|ب- الأمر+|الامر+ knew+|علم+><yasin+|ياسين><|command+|امر+|
Source: Mot+Sens+Lemme+POS	knew|علم<f1\> </body></html> ياسين|?????. </body></html> ب|in<F9>. </body></html> thing|امر<wsl>. </body></html>
Référence Yacine knew about it.	

Système	Bleu	Hypothèse
		Système de base
Mot→Mot	37.44	flag yasin command.
		Systèmes factorisés
Mot+Sens→Mot	66.18	flag yasin the case.
Mot+Sens +Lemme+POS →Mot	74.89	yasin knew the case.

FIGURE 3.14: Ontonotes 5.0 Implementation result of the sentence "علم ياسين بالأمر"

The factorized system of Ontonotes has translation the word "**علم**" into "flag" which is a sementically wrong interpretation . However, the researchers factorised system has translated it into "knew". The latter comes to be regarded as the correct form of translation for the word "**علم**"in "علم ياسين بالامر".

Données	
Source: Mot	<هواري رفع علم بلاده.>
Source: Mot+Sens	|هواري+|هواري+|رفع+|رفع+|علم+|علم|بلاد-|بلاد+|ه-|ه| |رفع|+><?????|+><هواري|+upload|+taught|+علم|+بلاد|-><country+|ه-|><it-
Source: Mot+Sens+Lemme+POS	هواري|????(f9) </body></html> raise|رفع </body></html> علم|><flag><c1><taught><. </body></html> بلاد|country><7x><. o|his</body></html></body></html>
Référence	Haouari raised the flag of his country.

Système	Bleu	Hypothèse
		Système de base
Mot→Mot	41.23	hawari upload taught country his.
		Systèmes factorisés
Mot+Sens→Mot	57.47	hawari raises country taught.
Mot+Sens +Lemme+POS →Mot	81.32	hawari raises country flag.

FIGURE 3.15: OntoNotes Implementation result of the sentence "رفع هواري علم بلاده"

We conclude that lexical and morphological disambiguation information is more relevant when Arabic is a source language. As well as, the use of linguistic annotated data for a morphologically rich language (like Arabic) is relevant for the automatic translation task to reduce the ambiguity of the data.

Figures 3.16 and 3.17 present the result obtained by integrating Diacritical marks in EN-AR translation systems.

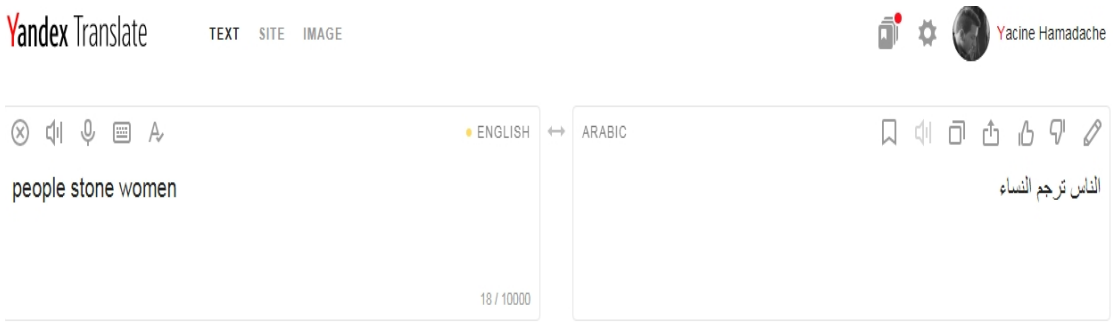


FIGURE 3.16: Implementation result of the sentence "people stone women"

Through exploiting the morphological analyzer of ontonotes 5.0, we have obtained two different parts of speech; verb and adjective. Thus, the Arabic translation of the verb form of "stone", is totally different from the translation of the adjective form of the same word. The latter was translated into the Arabic word "حجري". However, the former was translated into "ترجم ".

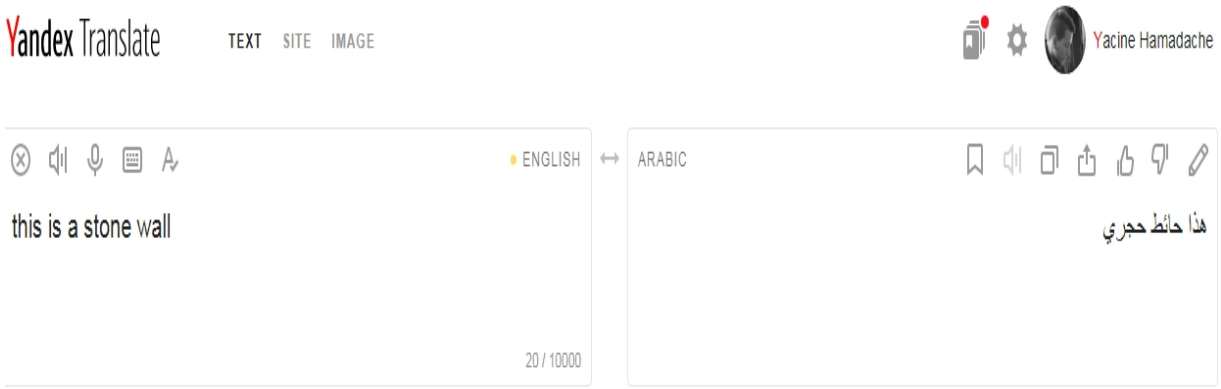


FIGURE 3.17: Implementation result of the sentence "this is a stone wall"

3.5 Conclusion

This chapter presented the descriptive thematic analysis and discussed the findings that were obtained from the research instruments. The analysis and discussion of these results were aimed at achieving the main objective of the thesis which is improving the process of Arabic lexical disambiguation, the several instruments did help in developing a deeper understanding of the case being studied.

Data obtained through experiments were either collected qualitatively and analyzed quantitatively or vice versa (collected quantitatively and analyzed quantitatively) in order to facilitate our analysis. Thus, on the one hand, we performed an in vivo evaluation to our theory of Arabic lexical disambiguation by comparing it with previous translations of the same sentences. The result of the translation of (see Figure 3.12) in comparison to its pre-implementation result (see Figure 3.1) shows huge syntactic and semantic difference. Moreover, applying diacritical marks theory to the Arabic sentence "علم ياسين بالأمر", has changed completely the meaning of the sentence and improved it by providing a correct translation to the Arabic word "عَلِمَ' which is "**knew**", instead of "**Copy that,**". As well as, through applying the same theory to the second sentence "هواري رفع علم بلاده", a purely lexical disambiguation of the same Arabic word "**علم**"was achieved by substituting "**science**" (see

Figure 3.2) with "**Flag**" (see Figure 3.13). On the other hand, we perform an in vitro evaluation to our theory of Arabic lexical disambiguation within the framework of OntoNotes 5.0 Corpus, We evaluate our in vitro approach (in terms of precision P, recall R and F1 score) based on a system of lexical disambiguation, while exploiting the Arabic part of the OntoNotes 5.0 corpus that we have enriched with the diacritical marks theory as reference corpus. On AR-EN systems, by exploiting the morphological information (diacritical marks) within the word "علم", wa could raise translation quality percentage by +8.71% and +23.85% in Figure 3.14 and Figure 3.15 respectively compared to the factorised system in terms of BLUE. As we can see, the factorized system translation of "علم" was "**flag**" in Figure 3.14 and "**taught**" in Figure 3.15, these translations come to be regarded as syntactically correct. However, The most important issue of translation is semantics. Thus, the translation provided by our theory serves translation on both levels; semantic and syntactic. Moreover, "**knew**" and "**flag**", in Figure 3.14 and 3.15 respectively, present a high quality of the lexical disambiguation process in general and Arabic translation particularly, especially that they have shown some linguistic development and differences between the two translations; pre-implementation and post-implementation.

Concerning EN-AR translation, comparing (see Figure 3.3) to (see Figure 3.16) makes us understand how to overcome errors made by the system. By exploiting the morphological information; Chomsky's syntactic parsing, we improved the translation system as it can be seen in the example of "they stone women". Furthermore, through Chomsky's syntactic parsing, two different parts of speech for the same word "stone" were obtained; verb form and adjective one. In the pre-implementation translation, the word "stone" was translated into "صخور", (see Figure 3.3) Perhaps, the translation engine could not identify the grammatical feature of the word, which led to a completely erroneous translation that is neither semantically nor syntactically correct. However, for the post implementation translation test (see Figure 3.16), The morphological alanyzer that was linked to Yandex translation engine could identify "stone" as a verb. Thus the translation was "تَرْجُمُ" which is semantically and syntactically correct.

The second example of EN-AR translation did not know any changes as Figure 3.17 shows. The English morpheme "stone" was translated into "حجري" knowing that it is considered as an adjective in both languages; English and Arabic.

All in all, and in light of the above, The findings of this research obtained from the analysis of implementations and their results are highly significant as they help in answering the main research questions addressed in this study regarding the effectiveness of implementing the diacritical marks and Chomsky syntactic parsing theories in the process of WSD. the following chapter will propose some implications and practical suggestions, which may be of a great help to educators, translators and NLP engineers.

Implications

There are many applications of computational linguistics (CL) techniques to computer-assisted language learning (CALL).which is a program to help people learn foreign languages. Machine translation has witnessed a tremendous improvement due to some new developing technologies. Nowadays machines can take something written or spoken in one language and give the same meaning in another language. This led people from around the world to use this technology for day-to-day activities. For instance, Google company in moving toward a new model of translation that provides a better result for learners. That system called neutral machine translation (NMT),This system uses large amount of computer information that enable to produce a translation which appear just like human language.

Other companies and organizations started to develop this model of translation, this goes along with machine learning method that is best known with Deep learning (DL).It takes a large collected data passing by a certain process (algorithm) to organize and recognize the different word and sentences, Now people can have a fast translations to different website, texts and social media messages in seconds .Which ease the way for researchers and students to learn in a second language, that's may help preserving a diversity of languages.

Another thing that using different machine translation models such as Google translate helps Learners to avoid misspelling words and sentences, especially from native language (Arabic) into the target language (English) .Actually this consider as a supporting tools to help teachers to teach foreign languages .Also teachers sometimes ask learners to check out their dictionaries to identify the word classes (part of speech).Automatic-speech recognition took a central part in speech -to-speech translation that witness a huge development in the last decade in order to help people to communicate in different languages ,also it translate the transcribes text into target language

Limitations

Arabic language comparing to other languages such as English and French must have a real organized researches to reach a level of maturity in machine translation. People may face struggles in gathering resources duo to the limited resources accept for some institution. Having a tools and machine being smart to have a conversation with or order something still be seen as a hard task to accomplish

GENERAL CONCLUSION

The work presented in our thesis focuses on the lexical disambiguation of the Arabic language, which is considered as a poorly endowed language. On the one hand, our goal has been to implement a diacritical method to build a disambiguation systemsupervised lexical for a language L using a good system of automatic translation of a language rich in annotated corpora, such as English, to a less well-endowed language, such as Arabic, and bearing annotations of the language rich towards the less well-endowed language.

On the other hand, we were interested in the evaluation of our system of lexical disambiguation created by following two approaches: an in vitro evaluation where our system of disambiguation is evaluated on an annotated corpus of reference which we enriched semi-automatically, and an in vivo assessment where our disambiguation system is evaluated for its contribution to the performance of machine translation systems. To our knowledge, this type of evaluation has not yet been applied in the evaluation campaigns.

Lexical disambiguation is a task that aims to associate the most appropriate meaning to each word of a text among the senses proposed by a predefined meaning inventory. It is necessary to have a large amount of data annotated in sense. One or more linguistic experts can be used to manually annotate the training data with the appropriate meanings. Such as a task is very expensive in time and manpower. To evaluate a system of lexical disambiguation, it is necessary to have an annotated corpus of reference. In order for different systems and approaches to be evaluated comparatively, it is necessary that reference corpus is available for the community and free. The availability of training data and a reference evaluation corpus are much more relevant questions for languages that have virtually no textual resources annotated in the royalty-free sense. The Arabic language belongs to this category of languages with very few resources available for this task.

This thesis work is based on several axes. As a first step, we described the difficulties related to languages with less than English for the construction of lexical disambiguation system. The difficulty is mainly due to the lack of corpus annotated in sense and how missing diacritical marks may create vagueness into the word meaning. We thus provide the community with corpora translated automatically into Arabic and annotated with Princeton

WorldNet. These corpora can be used during the learning step of a lexical disambiguation system.

In a second step, for in vitro evaluation of the lexical disambiguation of Arabic, we have updated the Arabic part of the OntoNotes Release 5.0 corpus, and created our own English-Arabic translation system, to create a disambiguation system. Arabic by having annotations from English.

The results on Arabic prove the effectiveness of our method and are very encouraging for the future of the lexical disambiguation of languages with a small amount of annotated data. Therefore, it is now possible to produce good lexical disambiguation systems for any language provided that we have a good system of automatic translation from English to this one. We also propose, how applying linguistics into natural language processing (NLP) can improve both translation and WSD.

In addition, we performed an in vivo evaluation of our system of lexical disambiguation of Arabic by comparing it to an English lexical disambiguation system, for the annotation of large corpora. Next, we used these large English and Arabic corpora in order to have a better result of translation

Finally we saw how Chomsky's syntactic structure theory helps researchers to be taken as a model applied in word sense disambiguation process by giving the grammatical category of each word in a given sentence

Perspective

From the contributions presented in this work, various perspectives can be envisaged. First, we plan to increase the size of the English-Arabic parallel data annotated in sense. The latest improvements in both conceptual (meaning vocabulary reduction) and implementation performed on the system of lexical disambiguation have greatly accelerated treatment. The annotation of the 9 Million missing sentences thus becomes possible in reasonable time (of the order of the month against 4 years ago). We are considering

We also plan to use the aligned corpus and their translingual information to reinject them into factored statistical translation systems, in order to compare the output hypotheses of the statistical and neuronal TA systems. Another important perspective would be to exploit the

outputs of the English-Arabic and Arabic-English factored neural translation systems for the creation of new lexical disambiguation systems supervised for Arabic and English in order to improve the task of lexical disambiguation. . The idea behind it would be to try to see how far it is possible to obtain a virtuous circle where automatic translation improves thanks to lexical disambiguation, which in return improves thanks to machine translation.

In addition, we propose to use new factors such as synonymy to improve lexical disambiguation and automatic translation tasks. Since the output of our factored English-Arabic translation system is annotated with meanings from Princeton WordNet, we plan to use human annotators to quickly annotate these Arabic corpora by validating or replacing the annotations hypotheses produced to confirm that our Factorized translation system can simultaneously play the role of a machine translation system and a system of lexical disambiguation of Arabic.

This work could also allow a contrastive study of the attributes relevant to the disambiguation of a language (Arabic, French, etc.), a study that could not be conducted before the existence of corpora for as many languages.

Finally, this work is starting to be exploited in the team for other applications than machine translation. Our methods have been adapted to English to create a system for generating pictograms from oral statements or writing in which the choice of pictograms is done by removing lexical ambiguities. This type of application is useful to allow to establish an alternative communication for example for people in a situation of multiple disabilities or for people not mastering the language.

APPENDIX A

arabic transileration Table : The Arabic letters Names (Reference sheet)

The Arabic Letters Names (Reference Sheet)

'alif	=	ا	ṭā'	=	ط
bā'	=	ب	ẓā'	=	ظ
tā'	=	ت	ayn	=	ع
thā'	=	ث	ghayn	=	غ
jīm	=	ج	fā'	=	ف
ḥā'	=	ح	qāf	=	ق
khā'	=	خ	kāf	=	ك
dāl	=	د	lām	=	ل
dhāl	=	ذ	mīm	=	م
rā'	=	ر	nūn	=	ن
zāy	=	ز	hā'	=	هـ
sīn	=	س	wāw	=	و
shīn	=	ش	lām 'alif	=	لا
ṣād	=	ص	hamza	=	ء
ḍād	=	ض	yā'	=	ي

APPENDIX B

Ontonotes experiments

	Données
Source: Mot	<علم ياسين بالأمر.>
Source: Mot+Sens	علم+|علم+ ياسين|+|+ياسين ب-|ب- الأمر+|الامر+ |+امر|+command|<ياسين|+yasin><+علم|+knew
Source: Mot+Sens+Lemme+POS	knew|علم<f1\> </body></html> ياسين|?????. </body></html> ب|in<F9>. </body></html> thing|امر<wsl>. </body></html>
Référence Yacine knew about it.	

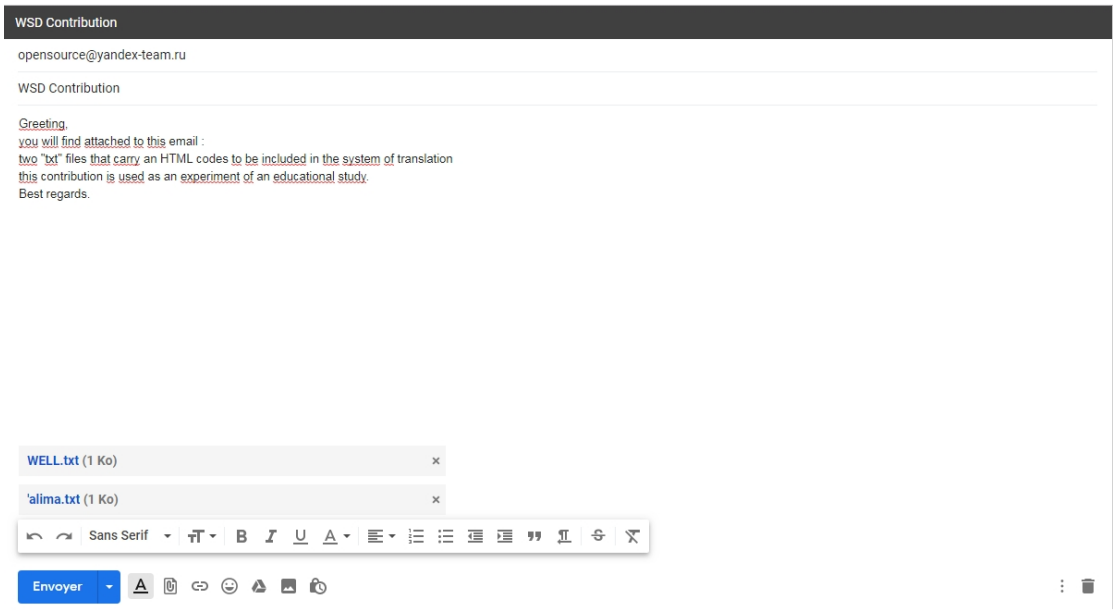
Système	Bleu	Hypothèse
		Système de base
Mot→Mot	37.44	flag yasin command.
		Systèmes factorisés
Mot+Sens→Mot	66.18	flag yasin the case.
Mot+Sens +Lemme+POS →Mot	74.89	yasin knew the case.

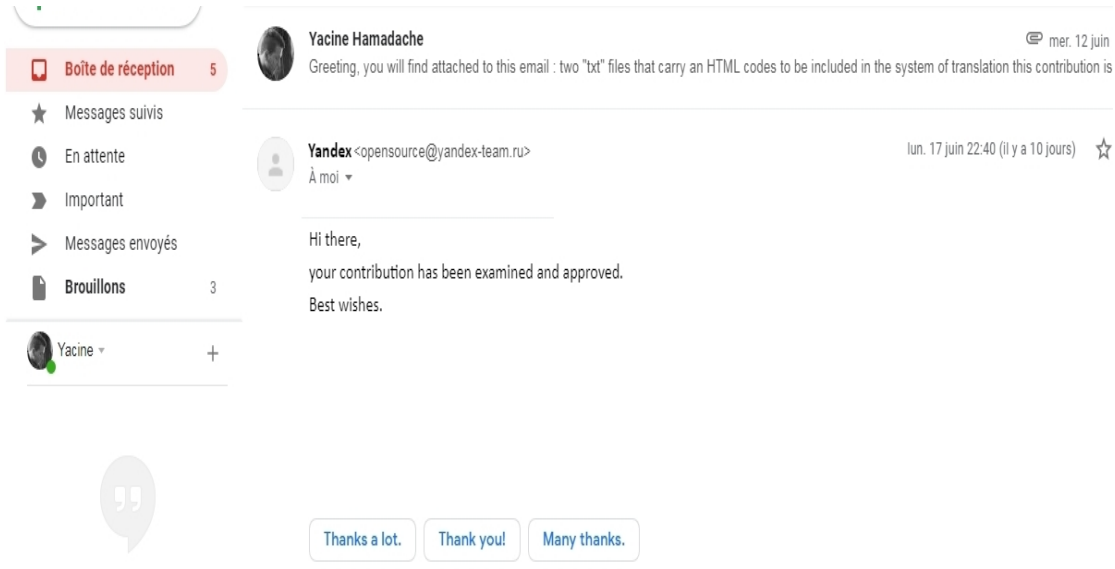
	Données
Source: Mot	<هواري رفع علم بلاده.>
Source: Mot+Sens	|هواري+|هواري+|رفع+|رفع+|علم+|علم|بلاد-|بلاد+|ه|-ه|-it-|><country+|ه-|><taught+|بلاد-|><هواري+|upload+|علم+|taught+|رفع|+><?????|+><هواري+|upload+|علم+|taught+|بلاد-|><country+|ه-|><it-
Source: Mot+Sens+Lemme+POS	هواري|????(f9) </body></html> raise|رفع </body></html> علم|><flag><c1><taught><. </body></html> بلاد|country><7x><. ه|his</body></html></body></html>
Référence Haouari raised the flag of his country.	

Système	Bleu	Hypothèse
		Système de base
Mot→Mot	41.23	hawari upload taught country his.
		Systèmes factorisés
Mot+Sens→Mot	57.47	hawari raises country taught.
Mot+Sens +Lemme+POS →Mot	81.32	hawari raises country flag.

APPENDIX C

E-mail conversation with Yandex





APPENDIX D

An example of HTML code for the sentence :

"علم ياسين بالأمر"

```
# Yandex
# Dictionary

# Input Language is: ar

# Alphabetical order is not required.

# Use inflectional & derivational paradigms' and properties' definition files (.nof or .def), e.g.: # Special Command: #
# Special Command: #use paradigms.nof

# Special Features: +NW (non-word) +FXC (frozen expression component) +UNAMB (unambiguous lexical entry)
#       +FLX= (inflectional paradigm) +DRV= (derivational paradigm) # Special Characters: '\' "" '+' ',' '#' ' '
#use _Exemple.flx #use _Properties.def

,PREP
,ADV

# nams

,N+Conc+FLX=Flag+DRV=thing:FlexionPL
•?      ,N+Conc+FLX=jazIra+DRV= thing:FlexionPL
%o,N+Conc+Hum+FLX=thing+DRV=rijAl:FlexionPL
,N+Abst+FLX= thing+DRV=rijAl:FlexionPL # verbs
# Les deux verbes "     " ('alima - knew) et "? " ('alima - knew, taught) (V_'alima)

# three verbs "('alima - knew), "? " (taught) and "y copy that -
knew) <'alima><knew> <AR>,<EN> (D_faiIla)

,V+Tr+FLX=V_'alima +DRV=D_    faiIla:FlxDRV
,V+Tr+FLX= V_'âllama +DRV=D_ faâala:FlxDRV
```

Bibliography

1. Alexandria, VA Event: February, 2001 | The International Research Foundation for English Language Education. (2019). Retrieved from https://www.tirfonline.org/2001/02/alexandria-va-event-february-2001/

2. Singh, H., & Gupta, V. (2015). An Insight into Word Sense Disambiguation Techniques. *International Journal of Computer Applications*, *118*(23).

3. Turnbull, J., Lea, D., Parkinson, D., Phillips, P., Francis, B., Webb, S., ... & Ashby, M. (2010). Oxford advanced learner's dictionary. *International Student's Edition*.

4. Dictionary, Encyclopedia and Thesaurus - The Free Dictionary. (2019). Retrieved from http://www.thefreedictionary.com/.

5. Weaver, W. (1995). Translation. Machine translation of languages, 14, 15-23.

6. Hutchins, J. (2003). Machine translation: general overview. In *The Oxford handbook of computational linguistics*.

7. Carpuat, M., & Wu, D. (2007). Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

8. Nguyen, Q. P., Vo, A. D., Shin, J. C., & Ock, C. Y. (2018). Effect of word sense disambiguation on neural machine translation: A case study in Korean. IEEE Access, 6, 38512-38523.

9. Padó, S., & Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. Journal of Artificial Intelligence Research, 36, 307-340.

10. Verschaere, A. (2016, September 23). Apprendre l'alphabet arabe et ses 28 lettres. Retrieved June 26, 2019, from https://jeretiens.net/lalphabet-arabe-les-28-lettres/

11. Debili, F., Achour, H., & Souissi, E. (2002). La langue arabe et l'ordinateur: de l'étiquetage grammatical à la voyellation automatique. *Correspondances*, *71*, 10-28.

12. Raskin, V. (1987). Linguistics and natural language processing. Machine translation: Theoretical and methodological issues, 42-58.

13. Kouloughli, D. E. (1991). Lexique fondamental de l'arabe standard moderne. Editions L'Harmattan.

14. Khoja, S. (2001, June). APT: Arabic part-of-speech tagger. In Proceedings of the Student Workshop at NAACL (pp. 20-25).

15. Dichy, J. (2019). The Intriguing Issue of Dictionary Arrangement in Medieval Arabic Lexicography. The Foundations of Arabic Linguistics IV. doi:10.1163/9789004389694_008

16. Mihalcea, R., Liu, H., & Lieberman, H. (2006). NLP (Natural Language Processing) for NLP (Natural Language Programming). Computational Linguistics and Intelligent Text Processing, 319-330. doi:10.1007/11671299_34

17. Applied linguistics. (2019, June 28). Retrieved June 11, 2019, from https://en.wikipedia.org/wiki/Applied_linguistics

18. Durrheim, K. (2004). *Research design*. In M. Terre Blanche, & K. Durrheim (Eds.)

19. *Research in practice: Applied methods for the social sciences* (29-53). Cape Town:University of Cape Town

20. Black, W. J., & ElKateb, S. (2004). A prototype English-Arabic dictionary based on WordNet. In Proceedings of 2nd Global WordNet Conference, GWC2004, Czech Republic (pp. 67-74).

21. Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41

22. Palmer, M., Babko-Malaya, O., Bies, A., Diab, M. T., Maamouri, M., Mansouri, A., & Zaghouani, W. (2008). A Pilot Arabic Propbank. In LREC.

23. Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004, September). The penn arabic treebank: Building a large-scale annotated arabic corpus. In NEMLAR conference on Arabic language resources and tools (Vol. 27, pp. 466-467).

24. Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., ... & Roth, R. (2014, May). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In LREC (Vol. 14, pp. 1094-1101).

25. Diab, M. (2009). Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In 2nd International Conference on Arabic Language Resources and Tools (Vol. 110, p. 198).

26. Habash, N., & Sadat, F. (2006, June). Arabic preprocessing schemes for statistical machine translation. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers (pp. 49-52). Association for Computational Linguistics.

27. Mertins, B. (2016). The use of experimental methods in linguistic research: advantages, problems and possible pitfalls. Slavic Languages in Psycholinguistics, 15-33.

28. Schwab, D., Goulian, J., & Tchechmedjiev, A. (2013). Worst-case complexity and empirical evaluation of artificial intelligence methods for unsupervised word sense disambiguation. Int. J. Web Eng. Technol., 8(2), 124-153.

29. Holloway, I., & Wheeler, S. (2002). *Qualitative research in nursing* (2 Ed.). Malden, MA: BlacLemetyinen,H.(2012).*Language acquisition.* Retrieved fromhttps://www.simplypsychology.org/language.html.

30. . Onwuegbuzie, A. J., & Leech, N. L. (2007). *Validity and qualitative research: Anoxymoron?. Quality & Quantity,* 41(2), 233-249.

31. Bitsch, V. (2005). Qualitative research: *A grounded theory example and evaluation criteria. Journal of agribusiness*, 23(345-2016-15096), 75.

32. Tobin, G. A., & Begley, C. M. (2004). *Methodological rigour within a qualitative.*

33. *framework. Journal of advanced nursing*, 48(4), 388-396.

34. Krefting, L. (1991). *Rigor in qualitative research: The assessment of trustworthiness. TheAmerican journal of occupational therapy,* 45(3), 214-222

35. Lees, R. B., & Chomsky, N. (1957). Syntactic Structures. Language, 33(3), 375. doi:10.2307/411160.

36. Padó, S., & Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. Journal of Artificial Intelligence Research, 36, 307-340.

37. Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L& El-Bachouti, M. (2013). Ontonotes release 5.0 ldc2013t19. Linguistic Data Consortium, Philadelphia, PA.

38. Jurafsky, D. (2018). Syntactic parsing is the task of recognizing a sentence and assigning a syntactic structure, Stanford University.

39. Mitkov, R. (Ed.). (2004). The Oxford handbook of computational linguistics. Oxford University Press.

40. Agirre, E., & Edmonds, P. (2007). Introduction. Word Sense Disambiguation, 27-28. doi:10.1007/1-4020-4809-2_1

41. Hutchins, W. J. (1995). Machine translation: A brief history. In Concise history of the language sciences (pp. 431-445). Pergamon.

42. Singh, H., & Gupta, V. (2015). An Insight into Word Sense Disambiguation Techniques. *International Journal of Computer Applications*, 118(23), 32-39. doi:10.5120/20888-3666.