Ambiguous Arabic Words Disambiguation: The results

Laroussi Merhben
UTIC(Monastir unit) higher school of techniques sciences of Tunis.
Aroussi_Merhben@hotmail.com

Anis Zouaghi
UTIC (Monastir Unit) superior Institute of informatics of Medenine
Anis.zouaghi@gmail.com

Mounir Zrigui
UTIC (Monastir unit) Faculty of sciences of Monastir
Mounir.Zrigui@fsm.rnu.tn

Abstract

In this paper we propose an hybrid system of Arabic words disambiguation. To achieve this goal we use the methods employed in the domain of information retrieval: Latent semantic analysis, Harman, Croft, Okapi, combined to the lesk algorithm. These methods are used to estimate the most relevant sense of the ambiguous word. This estimation is based on the calculation of the proximity between the current context (Context of the ambiguous word), and the different contexts of use of each meaning of the word. The Lesk algorithm is used to assign the correct sense of those proposed by the LSA, Harman, Croft and Okapi. The results found by the proposed system are satisfactory, we obtained a rate of disambiguation equal to 73%.

Keywords

Arabic ambiguous words, LSA, Harman, Okapi, Croft, Lesk algorithm, signatures and syntactic tagger.

1. Introduction

This work is part of the understanding of the Arabic speech [15]. In this paper we are interested in determining the meaning of Arabic ambiguous words that we can encounter in the messages transcribed by the module of speech recognition.

The word sense disambiguation (WSD) involves the association of a given word in a text or discourse with a definition or meaning (sense) which is distinguishable from other meanings potentially attributable to that word [12].

To assign the correct meaning, our method starts with the application of several pre-processing (tf × idf [14], normalization and syntactic tagging [2]) on words belonging to the context of the ambiguous word, subsequently we have applied the measures of similarities (Latent Semantic Analysis [5], Harman [8], Croft[3] and Okapi [13]) which will allow the system to choose the context of using the most closer to the current context of the ambiguous word, and we have applied Lesk algorithm [10] to distinguish the exact sense of the different senses given by this measures of similarity.

This paper is structured as follows, in section 2 we present the ambiguity of the Arabic language, after that in section 3 we describe the proposed method for disambiguation of ambiguous Arabic words later in section 4, we present the results of tests of our model.

2. Disambiguation of Arabic

The Arabic language is considered a difficult language to be automatically processed [10]. Among the characteristics that make this language processing ambiguous, we quote:

• The non vocalization of the Arabic language: a non vocalized Arabic word has several possible meanings. However, in modern editions, the texts in Arabic languages are not vocalized. We recall that vocalization in Arabic language is the addition of signs to the consonant to precise the pronunciation. Here is an example of a non-vocalized word: كتب (Kataba), this word might mean by way of his vocalization: كَتَبَ (he wrote), كُتُبُ (books), كُتِبَ (it was written). This phenomenal makes the problem of disambiguation more difficult;

• The structure of an Arabic word has a big problem for the automatic disambiguation. Indeed, an Arabic word can mean any expression in English or french. Here are some examples: the word وتتذكروننا (watatathakarounana) expresses the sentence in french " and you remember us ", the same word (وبقوله) (wabikawlihi) which means in English" and by his word". Thus the automatic understanding of such words requires a prior segmentation, a task that is not obvious;

• Another source of problems is the lack of language resources such as dictionaries, previously tagged corpus, and so on. This lack of resources with the characteristics of this language makes automatic processing more difficult;

In what follows, we describe the proposed method for disambiguation of the meaning of ambiguous Arabic words.

3. Proposed System

3.1 Method

Because of the lack of linguistic resources necessary for the automatic processing of the Arabic language, we preferred to use and test a non-supervised method. We note that Unsupervised methodology identifies patterns in a large sample of data, without the benefit of any manually labeled examples or external knowledge sources, on the other hand the supervised methodology Create a sample of training data where a given target word is manually annotated with a sense from a predetermined set of possibilities.

The Principe of our method is as follows: First, we started by collecting, from the web, various Arabic texts to

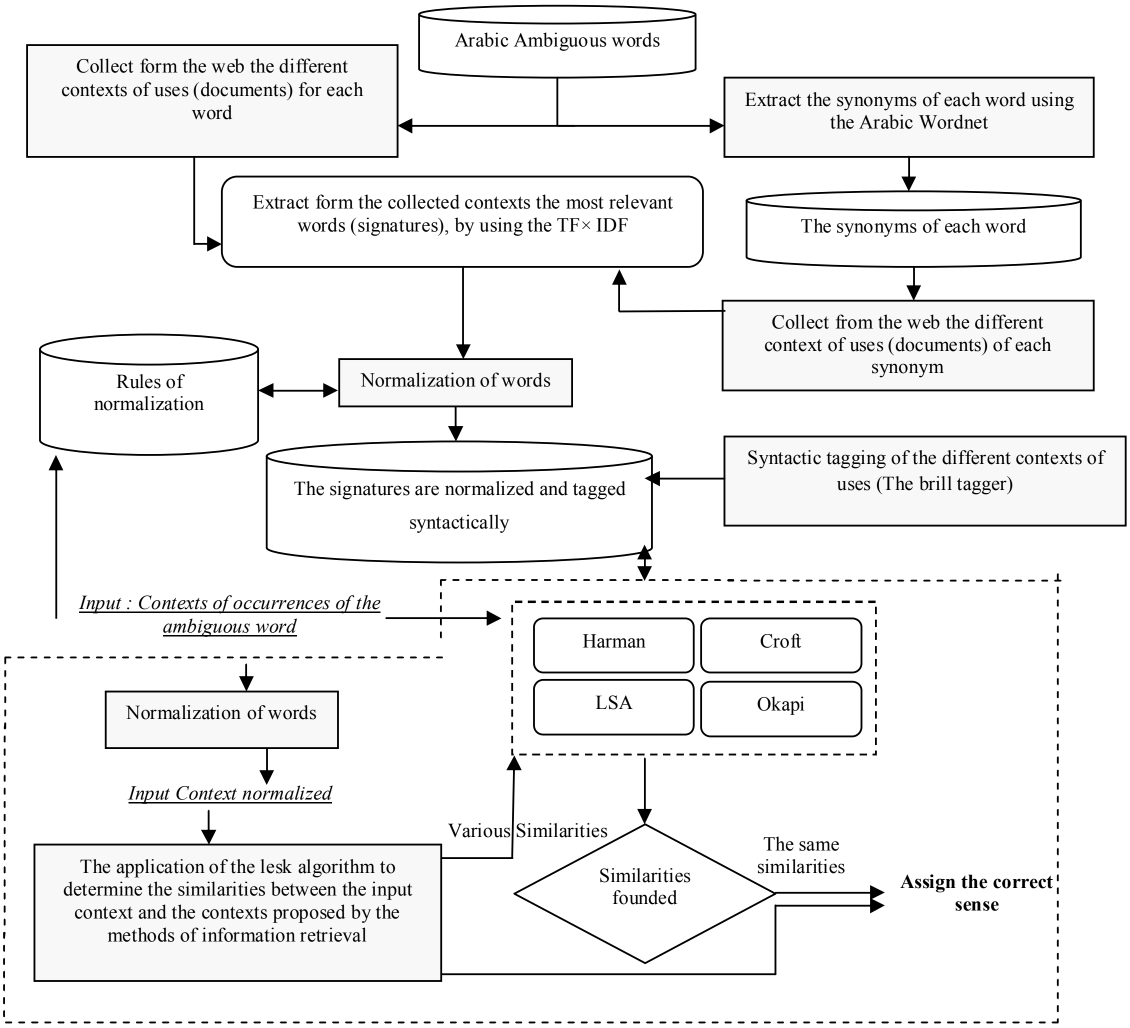
Arabic Ambiguous words
Collect form the web the different contexts of uses (documents) for each word
Extract the synonyms of each word using the Arabic Wordnet
Extract form the collected contexts the most relevant words (signatures), by using the TF× IDF
The synonyms of each word
Collect from the web the different context of uses (documents) of each synonym
Rules of normalization
Normalization of words
The signatures are normalized and tagged syntactically
Syntactic tagging of the different contexts of uses (The brill tagger)
Input : Contexts of occurrences of the ambiguous word
Harman
Croft
LSA
Okapi
Normalization of words
Input Context normalized
Various Similarities
The application of the lesk algorithm to determine the similarities between the input context and the contexts proposed by the methods of information retrieval
Similarities founded
The same similarities
Assign the correct sense

Figure 1. Method proposed to disambiguate the ambiguous Arabic word senses

build a corpus (see Section 4, Table 2) for several areas (i.e. sport, politics, religion, science, etc.).

From the corpus collected with the help of a linguist, we extracted the ambiguous words (words with several possible meanings out of context).

We note that we have applied several pre-processing steps (see Section 3.3) to the words that belonging to different contexts of use of the ambiguous word to improve the performance of the proposed system. We mean by context of use of an ambiguous word all sentences or texts in which the word has the same meaning.

From the Arabic WordNet [1] (lexical database of electronic Arabic words), we extract the synonyms of each word considered ambiguous. Then we collected the different contexts of use of these synonyms. This step enhances the number of contexts of use of each ambiguous word.

From the collection of possible contexts of use of each ambiguous word, and using the tf × idf measure [14] we were able to extract the different signatures, (the words that affect the meaning of each ambiguous word). Thus, each collection of signatures extracted from a context of use of an ambiguous word describes a unique sense of it. We also tested the contribution of syntactic knowledge on the outcome of disambiguation; we measured the similarity between the current context of an ambiguous word and its various contexts of use after tagging using the Brill tagger [2].

Following these pre-processing steps, we have implemented and tested several methods used in information retrieval: the latent semantic analysis [5], Harman [8], Croft [3] and Okapi [13], to measure the similarity between the current context of occurrence of the ambiguous word and the different possible contexts of use (possible meaning) of the word to disambiguate. The context which has a high similarity score with the current context is the most likely sense of the ambiguous word.

We note that we have tested these methods for measuring the similarity between the current context and all the possible contexts of use of each ambiguous word (Contexts represented in the form of texts and sentences) (see experimental results in section 4) in the first experiment we give the results obtained after pre-processing (Contexts represented by their signatures and tagged syntactically) Noticing that these methods do not always give the same result, we have tested the algorithm of Lesk to judge what is the most likely senses among those proposed by the methods listed above. In the following sub-paragraphs, we detail the different steps of the proposed disambiguation method. Figure 1 below describes our method.

3.2 Constitution of our Corpus

As mentioned previously, we have collected the various contexts of uses of each ambiguous word from the web and we do the same work for their synonyms that are obtained from a predefined lexical resource such as Arabic Wordnet [1]. With the help of linguistics we have given for each context the corresponding sense. Contexts (texts) are extracts of newspaper articles, which were recorded without restriction as to their nature and volume (see paragraph 1 in section 4).

All collected texts are non-vocalized. We used dictionary Al Wassit [6] to determine the definition of ambiguous words used for the test.

3.3 Pre-processing

3.3.1 Extraction of the signatures

The Several methods have been proposed to find for each given word the other words that appear generally next to him. In this experiment we have used the tf × idf measure (Term Frequency × Inverse Document Frequency) [14] it allows to assess the importance of a word in relation to a document, which varies depending on the frequency of the word in the corpus. For each context we take only the 20 words that have the maximum score (tf × idf), this encoding allows us to eliminate the stop words and the non-content words such as:

كان، له، فوق، حتّى، من، قد، بها، في،...

(he was, to him, on, to, from, then, with, whereas, ...)

These signatures represent the most basic part of our model because they represent the words that affect the meaning of each ambiguous word; these words have a higher likelihood of appearing together. If we don't find these signatures in the current context, in this case we extract from this context all the words that affect the meaning of ambiguous word and we add them to our database, this will ameliorate the performance of our system. Table 1 below shows some examples of signatures.

Table 1. Example of signatures describing the possible meanings of the word "عين" (ayn)

Different Senses	Number of signatures	Signatures
العين المبصرة (eye)	50	ترى ,الجسم ,الجمجمة ,تابعت ,النّور …, See, the body, the crane, follow, the light, ...
عين الماء الجارية (source of water)	46	الماء، الجبل، الضّيعة، الجارية، تسقي،... Water, mountains, the companion, which flows, baste,
الشيء عينه أي نفس الشيء (the same thing)	39	شيء, مسألة , حكم, منهج, الأصح ,… Thing, Problem, trial program, rather ,…
عينا على الأعداء أي جاسوسا (spy)	49	يتجسّس، يراقب، المفتّش، خائن،... Spy, monitor, inspector, traitor,
حرف العين (the word ayn)	56	حرف، اللّغة، كلمة، الجملة، تتركّب،... The letter, language, word, sentence, consists ...

3.3.2 Word Normalisation

In this step we group all the words that are derived from the same root in one cluster. The words that we have considered in this step are the signatures obtained previously. Our goal is to create a partitioning of set of data (signatures) into a set of relevant subclasses called clusters represented by a root. For example we take the words (ذهب (thahaba) - ذاهب (thaaheb) - مذهب (mathhab) - ذهبنا (thahabna)) (go, someone that will go, way, we went), all these words are derived from the same root ذهب. So, all the words are represented and replaced by the root ذهب. This grouping was done manually using linguists and dictionaries. We have thus constructed bags containing the words derived from each root. This treatment also contributed to the improved performance of the proposed system. Indeed disambiguation rate changed from 52% to 56% (see experimental results, Section 4).

3.3.2 Syntactic Tagging of contexts

To test the influence of syntactic knowledge on semantic disambiguation task, we tagged syntactically the different contexts of use of ambiguous words, using the transformation based learning in the Brill tagger [2]. Syntactic tags used in our experiment are three, they can indicate if the word in question is a particle, a verb or a noun. Syntactic tagging of the corpus allowed our system to study the contribution of syntactic information on the result of determining the correct orientation for each ambiguous term. To achieve this goal, we measured the similarity (see next paragraph) between the current context and contexts of use while taking into account the syntactic tags assigned to different words. The syntactic tagging system give a success rate of 78 %. This study has enabled us to obtain a gain of performance in terms of accuracy. Indeed disambiguation rate changed from 52% to 64.3%. (see experimental results, Section 4)

3.4 Estimation of the most relevant sense using LSA, Okapi, Harman and Croft

Let CC = m_1 m_2 … m m_{-1} … the context where the ambiguous word m appears. Suppose that S_1, S_2, .., S_k are the possible senses of m out of context. And CU_1, $CU_{2,}$… CU_K are the possible contexts of use of m for which the meanings of m are respectively: S_1, S_2, …S_K.

To determine the appropriate sense of m in the current context CC we have used the information retrieval methods (LSA, Okapi, Harman and Croft) which allow the system to calculate the proximity between the current context (Context of the ambiguous word), and the different use contexts of each possible sense of this word.

The result of each comparison is a score indicating the degree of semantic similarity (see equation 1) between the CC and CU given. This allows our system to infer the exact meaning of the ambiguous word. The following equation (1) describes the method used to calculate the score of similarity between two contexts:

$$S_t(CC, CU) = (\Sigma_{i\in RC} E(m_i) + \Sigma_{i\in LC} E(m_i)) / (\Sigma_{i\in RC} FE(m_i) + \Sigma_{i\in LC} FE(m_i)) \quad (1)$$

Where, $\Sigma_{i\in RC} E(m_i)$ et $\Sigma_{i\in LC} E(m_i)$ are respectively the sums of weights of all words belonging at the same time to the current context CC and to the context of use CU.

FE(mi), correspond to the first member of E(mi), where E (mi) can be replaced by one of the information retrieval methods : Croft, Harman or Okapi, whose equations are respectively:

- Harman measure [8]:

$$H(m) = W_H(m, CU(t)) = - \log (n(m) / N) \times [\log(n_{CU}(m) + 1) / \log(T(CU))] \quad (2)$$

Where, WH(m, CU(t)) is the weight attributed to m in the use contexts CU of the ambiguous word t by the Harman measure ; n(m) is the number of the use contexts of t containing the word m ; N is the total number of the use contexts of t ; $n_{CU}(m)$ is the occurrence number of m in the use context CU ; and T(CU) is the total number of words belonging to CU.

- Croft measure C(m)[3]:

$$C (m) = W_C(m, CU (t)) = - \log (n(m) / N) \times [k + (1\text{-}k) \times (n_{CU}(m) / \mathrm{Max}_{x\in CU}\, n_{CU}(x))] \quad (3)$$

Where, W_C(m, CU (t)) is the weight attributed to m in the user context CU of t by the Croft measure; k is a constant that determines the importance of the second member of C(m) (here, k = 0,5) ; and and $\mathrm{Max}_{x\in c}$ $n_{CU}(x)$ is the maximal number of occurrences of word m in CU.

- Okapi Measure [13]:

$$O(m) = W_O(m, CU(t)) =$$

$$\log [(N - n(m) + 0,5) / n(m) + 0.5] \times [n_c(m) / (n_{CU}(m) + (T(CU) / T_m(B)))] \quad (4)$$

Where, W_O(m, CU(t)) is the weight attributed to m in CU of t by the Okapi measure ; and Tm(B) is the average of the collected use contexts lengths.

- Latent Semantic Analysis [5]:

After the construction of the matrix A (term × documents) LSA find an approximation of the lowest rank of this matrix, by using the singular value decomposition which reduce obtains N singular values, where N = min (number of terms, number of docs). After that, the K highest singular values are selected and produces an approximation of k-dimension to the original matrix (It's the semantic space) In our experiments we used the Cosine to compare the similarities in the semantic space and k = 8.

3.5 Applying the Lesk algorithm to assign the correct sense

We adapted the Lesk algorithm [11] to calculate the proximity between the words that appear in the different definitions given by the methods used previously and the current context. The input of the algorithm is the word t and S = (s_1, ..., s_N), are the candidates senses corresponding to the different contexts of use achieved by applying methods of information retrieval. The output is the index of s in the sense candidates.

The lesk algorithm simplified [7] :

```
Begin
   Score ← 0
   Sense ← 1 // Choose the sense
   C ← Context (t) // Context of the word t
   For all I ∈ [1, N]
      D ← description (si)
   Sup ← 0
   For all w ∈ C do
      w ← description (w)
      sup ← sup + score (D, w)
   if sup > score then
    Score ← sup
    Sense ← i
End.
```

The choice of the description and context varies for each word tested by this algorithm.

The function Context (t) is obtained by the application of the input context. The function description (s_i) finds all the candidate senses obtained by the information retrieval methods. The function score return the index of the candidate sense to take: score (D, w) = Score (description (s), w).

The application of this algorithm allowed us to obtain a rate of disambiguation up to 73% (see paragraph 3 in section 4).

4. Experimental results

4.1 Characteristics of our corpus

The table 2 below describes the size of the corpus collected representing all contexts of use (texts) of ambiguous words considered in our experiments. We note that we intend to increase the size of the corpus in our next experiments.

Table 2. Characteristics of the collected Corpus

Total size of the corpus	1900 texts
Number of ambiguous words	10 words
Average number of synonyms of each ambiguous word	4
Average number of the possible senses	5
Total number of contexts of uses	300 texts
Average size of each context of use	560 words, 40 sentences

All the methods that were applied by our system consider all this characteristics of corpus in the different tests. We note that the corpus is manually created and evaluated.In our experiments we have used 10 ambiguous words to test our model. The Table 3 below describes an example of some contexts of use of the ambiguous word "الظّلمات" (atholoumat) for each sense.

Table 3. Example of context of use for each sense of the ambiguous word "الظلمات" (atholoumat)

Sense	Example of a contexts of use used in the test
انعدام الضّوء darkness	....سميت تفاعلات **الظلام** بعملية التمثيل الضوئي بسقوط الضوء على مجموعة من الخلايا ... The reactions of darkness called photosynthesis of the light because of the concentration of light on a set of cells
الجهل ignorance	... مشاعل هذه الحضارة الفتية تبدد **ظلمات** الجهل من خلال التمدن الإسلامي... The occupation of youth culture dissipates the darkness of ignorance to the Islamic civilization
العمى Blind	... يمكن ان يؤدي إلى ضعف البصر الدائم أو إلى **ظلمات** العمى May cause permanent visual impairment or a darkness of blindness

In table 4 below, we will give an example of data tested by our model and the sense given by every method. The data tests are randomly created.

Table 4. Results given by disambiguating the word "عين" in an example of test data

Example of test data	**Sense affiliated**				
	LSA	**Harman**	**Croft**	**Okapi**	**Lesk**
تبدو عين الإنسان كرويّة الشّكل	العين المبصرة	العين المبصرة	العين المبصرة	حرف العين	العين المبصرة
The eye of the human appear like a spherical form	eye	eye	eye	The word ayn	eye

4.2 Comparison of results obtained by the methods of information retrieval:

The figure 2 below presents the results obtained by using the methods ASL, Okapi, Croft and Harman. We note that we used the following metric to measure the rate of disambiguation:

Exact rate = (Number of senses obtained correctly / Number of senses assigned) × 100

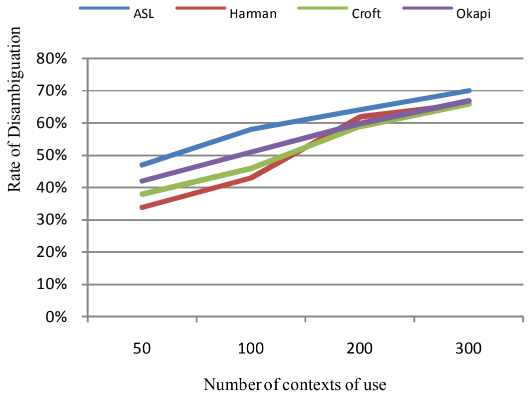

Figure 2. Comparison of results obtained by the disambiguation methods: LSA, Harman, Okapi and Croft

The results presented in Figure 2 above show that a bad disambiguation is obtained whenever there is a lack of contexts of use. Indeed, we note that the rates of disambiguation become satisfactory when the number of contexts of use is equal to 300. The worst results are obtained when the number is less than 100.

We can therefore conclude that the lowest rate of disambiguation is mainly due to the insufficient number of contexts of use, which result in the failure to meet all possible events. We also note that LSA provides the best results.

4.3 Experiment 1: Results obtained by the application of the LSA, Okapi, Harman, Croft and Lesk algorithm

The Table 5 below shows the rates of disambiguation obtained corresponding to ten Arabic ambiguous words. We validate results with 25 randomly selected samples. We note that the proposed hybrid system successfully disambiguate 76% of ambiguous words.

Table 5. Rate of disambiguation of arabic ambiguous words after pre-processing (extraction of signatures, word normalization and syntactic tagging)

Ambiguous words	Rate of sense affiliated correctly (%)				
	LSA	**Harman**	**Croft**	**Okapi**	**Lesk**
عين (ayn)	74	65	67	62	68
حسب (hasaba)	64	57	59	58	69
هان (hana)	62	54	49	52	74
الظلمات (atholoumat)	83	75	73	71	81
النّور (annour)	78	72	70	71	78
شعر (cheer)	69	65	63	61	70
فجر (Fajara)	57	51	51	53	62
نبع (nabaa)	71	62	60	60	73
دجم (dajama)	68	65	66	66,5	72
عقل (aakl)	75	65	61	64	84
the rate of disambiguation (%)	70.1	63.1	61.9	55.2	73.1

From Table 5, we note that the rate of disambiguation of the word "عين" (ayn) is lower than the other words, since it has more senses and more signatures, which makes the disambiguation of the term complex than the other words. Figure 3 below shows the influence of the number of signatures on the rate of disambiguation obtained.

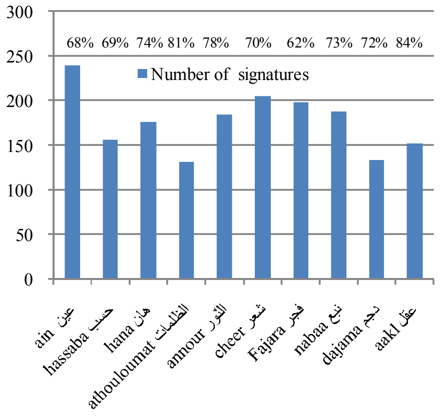

Figure 3. Influence of the number of signatures of the ambiguous words on the rates of disambiguation

We also note that the results obtained by the methods used in information retrieval: Harman, Croft and Okapi are very close.

The average rate of disambiguation is equal to 60%.

While disambiguation results obtained from the latent semantic analysis are often different from those found by Harman, Croft and Okapi. The average of disambiguation obtained by LSA is equal to 70.1%. We can then infer that the LSA gives better results. After some tests it was noted that these measures do not have in all cases the same meaning to be assigned (see Table 6 below). This makes the system unable to make a decision on the correct orientation. This explains why we decided to use the algorithm of Lesk. This algorithm allows our system to improve the results (we have achieved an average of disambiguation equal to 73%), it allow the system to choose the adequate sense.

Table 6. Example of the results of test of the word "هان" (hana) in the sentence:

"أكبر **مهانة** يتعرض لها الفرد عندما يعبد حجرا أو شجرا أو حيوانا أو يخضع لبشر
حي أو ميت"

(The greatest **humiliation** that a person may encounter when making a prayer to a stone or a tree or an animal or it becomes the subject of a human being living or dead.)

Ambiguous word	LSA	Harman	Croft	Okapi	Lesk
هان	ذلّ	سهل	رخص	سهل	ذلّ
hana	humiliate	simplify	Lowering	simplify	humiliate

4.4 Experiment 2: Results obtained before pre-processing

In this experiment we have tested the influence of the use of signatures on the results of disambiguation of the meaning of a word. Table 7 below shows that the rates of disambiguation of Arabic words obtained using the contexts of use without going through the signatures are increased from 52% to 73%.

Table 7. The rate of disambiguation of ambiguous words before pre-treatement

Ambiguous words	Rate of sense affiliated correctly (%)				
	LSA	Har man	Croft	Okapi	Lesk
عين(ayn)	42	36	37	36	48
حسب (hasaba)	50	43	45	41	54
هان (hana)	42	39	37	36	45
الظلمات (atholoumat)	62	51	54	53	64
النّور (annour)	51	42	41	41	54
شعر (cheer)	46	34	34	33	49
فجر (Fajara)	41	38	40	40	46
نبع (nabaa)	53	45	46	47	49
دجم (dajama)	45	36	34	34	48
عقل (aakl)	59	56	57	56	62
the rate of disambiguation (%)	49.1	42	42.5	41.7	52

4.5 Experiment 3: Studying the influence of the syntactic knowledge

We also tested the contribution of syntactic knowledge on the obtained results. For that, we have used the Brill tagger (see section 3.4). The table 8 shows that the rates of disambiguation of Arabic words obtained before syntactic tagging of contexts of use of each ambiguous word decreased compared to Experiment 1 (using syntactic tags), the rate decreased from 73% to 64.3%, while it was increased compared to the previous experiment 2 (use of contexts as they are, without the use of signatures) from 52% to 64.3%.

Table 8. The rate of disambiguation using syntactic tags

Ambiguous words	Rate of sense affiliated correctly (%)				
	LSA	Har man	Croft	Okapi	Lesk
عين(ayn)	42	36	37	36	48
حسب (hasaba)	63	56	58	60	59
هان (hana)	57	53	51	50	60
الظلمات (atholoumat)	50	49	46	47	57
النّور (annour)	72	64	61	64	74
شعر (cheer)	67	54	59	55	71
فجر (Fajara)	61	56	54	54	67
نبع (nabaa)	51	48	47	49	58
دجم (dajama)	65	62	62	60	68
عقل (aakl)	56	49	48	46	61
the rate of disambiguation (%)	65	62	63	63	70

4.6 Comparison of the proposed hybrid system with other systems disambiguation:

In this part we do a comparison of the results founded by our system with other system of disambiguation, comparing these results with the various works is a difficult task, because we do not work on the same corpus, or the same language, or with the same methods:

The method created by Lesk [11] used a list of words appearing in the definition of each sense of the ambiguous word achieved 50% - 70% correct disambiguation, Our system achieved 73% correct disambiguation

Karov and Edelman [9] (in this issue) propose an extension to similarity-based methods which gives 92% accurate results on four test words.

5. Conclusion

We have proposed a system for disambiguation of words in Arabic. This system is based simultaneously on the methods of information retrieval and the algorithm of Lesk used to calculate the proximity between the current context (i.e. the occurrence of ambiguous word) and the different contexts of use of the possible meanings of the word. While Lesk algorithm is used to help the system to choose the most appropriate sense proposed by previous methods. The results founded are satisfactory. For a small sample of 10 ambiguous words, the proposed system allows to determine correctly 73% of ambiguous words. We have tried to establish a sufficiently robust system based on methods that have improved their success in many system of word disambiguation. On the other hand, during the pre-processing we tried to make the ambiguous Arabic words known by the system we proposed a database containing the possible contexts of use for each sense of an ambiguous word, synonyms, signatures identifying the meaning of each one and syntactic tags.

We propose that in the future works we can use a multi-agent system that takes the more appropriate result given by all the methods applied by our system.

6. References

[1] Black, W. J. & Elkateb, S. (2004) A Prototype English-Arabic Dictionary Based on WordNet, Proceedings of 2nd Global WordNet Conference, GWC2004, Czech Republic: 67-74.

[2] BRILL E. (1993), A Corpus-Based Approach to Language Learning, Thesis non-published, University of Pennsylvania, Department of Computer and Information Science.

[3] Croft.W, 1983. Experiments with representation in a document retrieval system; Research and development, 2(1) ; pp. 1-21 ; 1983.

[4] De Loupy, 2000. Assessing the contribution of linguistic knowledge in semantic disambiguation and information retrieval. THESIS presented in the University of Avignon and the country of Vaucluse.

[5] Derwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshmann, R. 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Informartion Science, 41 : 391-407.

[6] Dictionary Al Wasit, 4th Edition, 2003, Academy of the Arabic language, Sunrise International Library.

[7] Florentina Vasilescu , 2003. Monolingual corpus disambiguation by the approaches of Lesk : University of Montreal, Faculty of Arts and Sciences; Paper presented at the Faculty of Graduate Studies to obtain the rank of Master of Science (MSc) in computer science.

[8] Harman D., 1986. An experimental study of factors important in document ranking ; Actes de ACM Conference on Research and Development in Information Retrieval ; Pise, Italy ; 1986.

[9] Karov, Yael and Edelman, Shimon (1998). “Similarity-based word sense disambiguation”. In this issue.

[10] Larkey L. S., Ballesteros L. and Connell M., « Improving Stemming for Arabic Information Retrieval : Light Stemming and Cooccurrence Analysis », In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, August 2002, p. 275-282

[11] Michael Lesk, Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone , ACM Special Interest Group for Design of Communication Proceedings of the 5th annual international conference on Systems documentation, p. 24 - 26, 1986. ISBN 0897912241

[12] Nancy Ide, Jean Verronis. 1998.Word Sense Disambiguation: The State Of The Art. Computational Linguistics, 2424:1, 1-40.

[13] Robertson et al., 1994.: S. Robertson, S. Walker, M. Hancock-Beaulieu, M. Gatford ; Okapi at TREC-3 ; Acts de Third Text Retrieval Conference (TREC-3), NIST special publication 500-225 ; pp. 109-126 ; Gaithersburg, Maryland, USA ; 1994.

[14] Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513-523.

[15] Zouaghi A., Zrigui M., Antoniadis G. 2008. Understanding of the Arabic spontaneous speech: A numeric modelisation, Revue TAL VARIA.