ARABIC WORD SENSE DISAMBIGUATION

Laroussi Merhbene, Anis Zouaghi
UTIC , Higher School of Techniques Sciences of Tunis, Tunis, Tunisia

Mounir Zrigui
Department of Computing, Faculy of sciences of Monastir, Monastir, Sousse, Tunisia

Keywords: Arabic ambiguous words, LSA, Harman, Okapi, Croft, Lesk algorithm, Signatures and stemmer.

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 76%.

1 INTRODUCTION

This work is part of the understanding of the Arabic speech (Zouaghi and al., 2008). In this paper we are interested in determining the meaning of Arabic ambiguous words that we can meet 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 (Ide and Verronis, 1998).

To assign the correct meaning, our method starts with the application of several pre-treatment (rooting (Al-Shalabi and al., 2003) and the tf × idf (Salton and Buckley, 1988)) on words belonging to the context of the ambiguous word, subsequently we have applied the measures of similarities (Latent Semantic Analysis (Drewster, 1990), Harman (Harman, 1986), Croft (Croft, 1983) and Okapi (Robertson and al., 1994)) 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 (Lesk, 1986) to distinguish the exact sense of the different senses given by this measures of similarity.

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

2 PROPOSED METHOD

2.1 Principe of the Method

We use and test a non-supervised method. The Principe of our method is as follows: First, we started by collecting, from the web, various Arabic texts to build a corpus (Section 3, Table1).

We have applied several pre-treatments (paragraph 2.2) to the words 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 (Black and al., 1990) (lexical database of electronic Arabic words), we extract the synonyms of each word considered ambiguous. We collect the contexts of use of these synonyms, this step enhances the number of contexts of use of each ambiguous word.

Using the algorithm of Al-shalabi (Al-Shalabi and al., 2003) we extract the root of each word, after

Merhbene L., Zouaghi A. and Zrigui M. (2010).
ARABIC WORD SENSE DISAMBIGUATION.
In *Proceedings of the 2nd International Conference on Agents and Artificial Intelligence - Artificial Intelligence*, pages 652-655
DOI: 10.5220/0002762106520655

that, we used the tf × idf measure (Salton and Buckley, 1988) to extract the different signatures, (the words that affect the meaning of each ambiguous word).

Following these pre-treatment steps, we have implemented and tested several methods used in information retrieval: the latent semantic analysis (Drewster, 1990), Harman (Harman, 1986), Croft (Croft, 1983) and Okapi (Robertson and al., 1994), 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 then has the great similarity score with the current context is the most likely sense of the ambiguous word.

In the following sub-paragraphs, we detail the different steps of the proposed disambiguation method.

2.2 Pre-processing

2.2.1 Word Rooting

(Sawalha and al., 2008) compared three stemming algorithms, the experimental results show that the Al-Shalabi, Kanaan, and Al-Serhan algorithm was in the second place, his advantage is that he does not use any resource.

This algorithm extracts word roots by assigning weights to word's letters (The weights are real numbers between 0 and 5) multiplied by the rank which depends of the letters position.
The three letters with the lowest weights are selected. This algorithm achieves accuracy in the average of 90%.

2.2.2 Extraction of the Signatures

Several methods have been proposed to find for each given word the other words that appear generally next to him. In this experience, we have used the tf × idf measure (Salton and Buckley, 1988), it allow to assess the importance of a word in relation to a document, which varies depending on the frequency of the word in the corpus. This encoding allows us to eliminate the few informative words such as:

كان، له، فوق، حتّى، من، قد، بها ،...

(he was, to him, on, to, from, then, with, …)

These signatures represent the most basic part of our model because they represent the words that affect the meaning of each ambiguous word. 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.

2.3 Estimation of the Most Relevant Sense using LSA, Okapi, Harman and Croft

Let CC = m_1, m_2, …, m_k the context where the ambiguous word m appear. 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 results of each comparison are a score indicating the degree of semantic similarity 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, and are respectively the sums of weights of all words belonging at the same time, the current context CC and the context of use CU. $FE(m_i)$, correspond to the first member of $E(m_i)$, or E (m_i) can be replaced by one of the information retrieval methods : Croft, Harman or Okapi, whose equations are respectively:

2.3.1 Harman Measure (Harman, 1986)

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

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.

2.3.2 Croft Measure C(m)(Croft, 1983)

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

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

2.3.3 Okapi Measure (Robertson and al., 1994)

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

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

2.3.4 Latent Semantic Analysis (Drewster, 1990)

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.

2.4 Applying the Lesk Algorithm

We adapted lesk algorithm simplified (Vasilescu, 2003) that adapt the lesk algorithm (Lesk, 1986), to calculate the number of words that appear in the current context of ambiguous word and the different contexts of use, which was considered as semantically closer to the results of methods used previously. 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:

```
Begin
   Score ← 0
   Sens ← 1 // Choose the sense
   C ← context(t)//Contexte 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
        Sens ← 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: score (D, w) = Score (description (s), w).

The application of this algorithm allowed us to obtain a rate of disambiguation up to 76%.

3 EXPERIMENTAL RESULTS

The table 1 below describes the size of the corpus collected representing all contexts of use (texts) of ambiguous words considered in our experiments.

Table 1: 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
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

Table 2 below shows the rates of disambiguation obtained corresponding to ten Arabic ambiguous words. We note that we used the following metric to measure the rate of disambiguation:

$$\text{Exact rate} = (\text{Number of senses obtained correctly} / \text{Number of senses assigned}) \times 100 \quad (5)$$

Table 2: Rate of disambiguation of Arabic ambiguous words after pre-treatment.

Methods applied	The rate of disambiguation
LSA	72.3%
Harman	64.2%
Croft	64.5%
Okapi	59.4%
Lesk	76%

From our experiments we 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. 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 (Lesk, 1986) used a list of words appearing in the definition of each sense of the ambiguous word achieved 50% - 70% correct disambiguation; our system achieved 76% correct disambiguation. Karov and Edelman (Karov and al., 1998) (in this issue), propose an extension to similarity-based methods, which gives 92% accurate results on four test words.

4 CONCLUSIONS

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 76% 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 .

We propose that in the future works we can use the syntactic level to disambiguate words.

REFERENCES

Al-Shalabi, R., Kanaan, G., and Al-Serhan, H., 2003. *New approach for extractingArabic roots*. Paper presented at the International Arab Conference on Information Technology (ACIT'2003), Egypt.

Black, W. J. and Elkateb, S., 2004. *A Prototype English-Arabic Dictionary Based on WordNet*, Proceedings of 2nd Global WordNet Conference, GWC2004, Czech Republic: 67-74

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

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.

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

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

Ide, N. and Verronis, J., 1998. *Word Sense Disambiguation: The State Of the Art. Computational Linguistics*, pp. 2424:1, 1-40.

Karov, Y. and Shimon, E., 1998. *Similarity-based word sense disambiguation*. In this issue.

Lesk, M., 1986. *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; pp. 24 – 26. ISBN 0897912241.

Robertson, S., Walker, M., Hancock-Beaulieu and Gatford, M., 1994. *Okapi at TREC-3* ; Third Text Retrieval Conference (TREC-3), NIST special publication 500-225; pp. 109-126; Gaithersburg, Maryland, USA.

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

Sawalha and al., 2008. *Comparative Evaluation of Arabic Language Morphological Analysers and Stemmers*. Coling 2008: Companion volume – Posters and Demonstrations, pages 107–110, Manchester, August 2008.

Vasilescu, F., 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.

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