Word Sense Disambiguation for Arabic Text Categorization

Meryeme Hadni¹, Said El Alaoui¹, and Abdelmonaime Lachkar²
¹Department of Computer Science, Sidi Mohamed Ben Abdellah University, Morocco
²Department of Electrical and Computer Engineering, Sidi Mohamed Ben Abdellah University, Morocco

Abstract: In this paper, we present two contributions for Arabic Word Sense Disambiguation. In the first one, we propose to use both two external resources Arabic WordNet (AWN) and WN based on term to term Machine Translation System (MTS). The second contribution consists of choosing the nearest concept for the ambiguous terms, based on more relationships with different concepts in the same local context. To evaluate the accuracy of our proposed method, several experiments have been conducted using Feature Selection methods; Chi-Square and CHIR, two machine learning techniques; the Naïve Bayesian (NB) and Support Vector Machine (SVM). The obtained results illustrate that using the proposed method increases greatly the performance of our Arabic Text Categorization System.

Keywords: WSD, arabic text categorization system, AWN, MTS.

Received September 1, 2015; accepted October 18, 2015; Publisher on line January 28, 2016

1. Introduction

Word Sense Disambiguation (WSD) is the problem of identifying the sense (meaning) of a word within a specific context. In Natural Language Processing (NLP), WSD is the task of automatically determining the meaning of a word by considering the associated context. It is a complicated but crucial task in many areas such as topic Detection and Indexing [12, 20], Information Retrieval [2, 6, 26], Information Extraction [2, 10], Machine Translation [5, 6], Semantic Annotation [23], Cross-Document Co-Referencing [3, 22] and Web People Search [2, 18, 30]. Given the current explosive growth of online information and content, an efficient and high-quality disambiguation method with high scalability is of vital importance.

All approaches to WSD [1, 7, 21, 28] make use of words in a sentence to mutually disambiguate each other. The distinction between various approaches lies in the source and type of knowledge made by the lexical units in a sentence. Thus, all these approaches can be classified into corpus-based approaches and knowledge-based ones. Corpus-based methods use machine-learning techniques to induce models of word usages from large collections of text examples. In [8, 11], the authors extract statistical information from corpora that may be monolingual or bilingual, and raw or sense-tagged. Knowledge-based methods use external knowledge resources which define explicit sense distinctions for assigning the correct sense of a word in context. In [24, 29] the authors have utilized Machine-Readable Dictionaries MRD, thesauri, and computational lexicons, such as WordNet (WN). Since most MRD and thesauri were created for human use and display inconsistencies, these methods have clear limitations. Like WN extends knowledge resource for the English language, Arabic WordNet (AWN) has been developed for the Arabic language, but it is an incomplete project. To overcome the above cited problem, we propose to use the WN resource to search the terms not existing in AWN.

In this work, we present an efficient method for Arabic WSD based knowledge external resource AWN. For the terms don’t exist in AWN, we traduce the terms from Arabic into English using Machine Translation System (MTS) and search the corresponding concepts in WN resource. After extracting the concepts, or the list of concepts, we choose the nearest concept based on the semantic similarity measure. Then, these concepts will be translated into Arabic language using MTS and the text document is represented as a vector of concepts.

The rest of this paper is structured as follows: Section 2 summarizes the related work. Section 3 introduces the different strategies of mapping and disambiguation. Section 4 describes the architecture of our proposed methods. In section 5, we evaluate the results of the experiments. Finally, in the last section, we present the conclusion and future work.

2. Related Works

WSD is the process of automatically determining the meanings of ambiguous words based on their context, which is one of problematic issues in NLP. Various works on WSD can be found in English and other European languages that solve the problem of the
terms that have several meanings. The authors in [7] have proposed a WSD strategy based on dependency parsing tree matching. In this strategy: Firstly, a large scale dependency knowledge base is built. Secondly, with the knowledge base, the matching degree between the parsing trees of each sense gloss and the sentence are computed. The sense with the maximum matching degree would be selected as the right sense. McCarthy et al. [19] have proposed a method to disambiguate the ambiguous words based on distributional similarity and semantic relatedness. Firstly, they select feature words based on direct dependency relationships. They parse a corpus with the dependency parser to get a great deal of dependency triples. Based on the dependency triples, distributional similarities among words are computed and top-N similar words are chosen as feature words [17]. Secondly, the relatedness between each sense of ambiguous words and feature words is computed. The sense with the maximum weighted sum of relatedness is selected as the right sense. Agirre et al. [1] have presented the method for WSD with a personalized PageRank [19]. They collect feature words with direct dependency like relationships. Knowledge from Wikipedia is injected into WSD system by means of a mapping to WN. Previous efforts aimed at automatically linking Wikipedia to WN include; full use of the first WN sense heuristic [27], a graph-based mapping of Wikipedia categories to WN synsets [21], a model based on vector spaces [25] and a supervised approach using keyword extraction [23].

Unlike European languages, there are few works and contributions that deal with Arabic WSD.

Yarowsky [29] proposed a new approach for text categorization, based on incorporating semantic resource (WN) into text representation, using the Chi-Square, which consists of extracting the k better features best characterizing the category, compared to others representations. The main difficulty in this approach is that it is not capable of determining the correct senses. For a word that has multiple synonyms, they choose the first concept to determine the nearest concept. The work in [14] is a comparative study with the other usual modes of representation; Bag of Word (BoW), Bag of Concepts (BoC) and N-Gram, and uses the first concepts after mapping on WN to determine the correct sense for an ambiguous term. Zouaghi et al. [31] proposed a new approach for determining the correct sense of Arabic words. They proposed an algorithm based on Information Retrieval measures to identify the context of use that is the closest to the sentence containing the word to be disambiguated. The contexts of use represent a set of sentences that indicate a particular sense of the ambiguous word. These contexts are generated using the words that define the meanings of the ambiguous words, the exact String-Matching algorithm, and the corpus. They used the measures employed in the domain of Information Retrieval, Harman, Croft, and Okapi combined with the Lesk algorithm, to assign the correct sense of those words proposed. In the Lesk algorithm [15], when a word to disambiguate is given, the dictionary definition or gloss of each of its senses is compared to the glosses of every other word in the phrase. A word is assigned the meaning which gloss shares the largest number of words in common with the glosses of the other words. The algorithm begins new for each word and does not utilize the senses it previously assigned.

These works show some weakness [15, 31] uses the dictionaries gloss for each concept. For example, the term “عين” has two glosses in the Al-Wasit dictionary: Gloss 1 “eyes”; “العين, the visual organ of humans and of animals” and gloss 2 “source”: “يتنوع الماء ببعض من الأرض يجري”; “the source of water that comes from the earth”, which gives an ambiguity in the gloss of concepts. In [9, 14] the authors present the systems that use BoC and choose the first concepts after mapping on AWN for determining the correct concepts, and the first concept is random and therefore not always the best choice.

Table1. Difference between AWN and WN.

<table>
<thead>
<tr>
<th></th>
<th>WN</th>
<th>AWN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Concepts</td>
<td>117,659</td>
<td>10,165</td>
</tr>
<tr>
<td>Number of Nominal</td>
<td>117,798</td>
<td>6,252</td>
</tr>
<tr>
<td>Number of Verbal</td>
<td>11,529</td>
<td>2,260</td>
</tr>
<tr>
<td>Number of Adjectival</td>
<td>21,479</td>
<td>606</td>
</tr>
<tr>
<td>Number of Adverbial</td>
<td>4,481</td>
<td>106</td>
</tr>
</tbody>
</table>

However, one major problem when dealing with AWN is the lack of many concepts because AWN is an incomplete project (e.g., Table 1). Therefore, for the terms that do not exist in AWN we search for the corresponding concepts on WN based on MTS.

Therefore, for the terms that do not exist in AWN, we search the corresponding concepts on WN based on MTS. In this paper, for each term that has a different meaning, we propose a new method for Arabic WSD based on relationships with different concepts in the same local context.

3. Mapping and Disambiguation Strategies

In Natural Language, the assignment of terms to concepts is ambiguous. Mapping the terms into concepts is achieved by choosing a strategy of matching and disambiguation for an initial enrichment of the representation vector. In this section, we will describe the different strategies of mapping and disambiguation.

3.1. Mapping Strategies

The words are mapped into their corresponding concepts. From this point, three strategies for adding or replacing terms by concepts can be distinguished [9].

3.1.1. Add Concepts

This strategy extends each term vector $\bar{t}_d$ by new entries for WN concepts $C$ appearing in the texts set.
Thus, the vector $\vec{t}_d$ will be replaced by the concatenation of $\vec{t}_d$ and $\vec{c}_d$ where $\vec{c}_d = (cf(d,c_1),...,cf(d,c_l))$. The concept vector with $l = |C|$ and $cf(d,c)$ denotes the frequency that a concept $c \in C$ appears in a text $d$.

The terms, which appear in WN as a concept [14] will be accounted at least twice in the new vector representation; once in the old term vector $\vec{t}_d$ and at least once in the concept vector $\vec{c}_d$.

### 3.1.2. Replace Terms by Concepts

This strategy is similar to the first strategy; the only difference lies in the fact that it avoids the duplication of the terms in the new representation; i.e. the terms which appear in WN will be taken into account only in the concept vector. The vector of the concepts will thus contain only the terms which do not appear in WN.

### 3.1.3. Only Concept

This strategy differs from the second strategy in that it excludes all the terms from the new representation including the terms which do not appear in WN; $\vec{c}_d$ is used to represent the category.

### 3.2. Strategies for Disambiguation

When mapping terms into concepts, the assignment of terms to concepts is ambiguous since we deal with natural language [9]. One word may have several meanings and thus one word may be mapped into several concepts. In this case, we need to determine which meaning is being used, which is the problem of sense disambiguation. Two simple disambiguation strategies exist:

### 3.2.1. All Concepts Strategy

This strategy [9] considers all proposed concepts as the most appropriate one for augmenting the text representation. This strategy is based on the assumption that texts contain central themes that in all cases will be indicated by certain concepts having height weights. In this case, the concept frequencies are calculated as follows:

$$Cf(d,c) = tf\{d, \{t \in T \setminus c \in ref_c(t)\}\}$$  \hspace{1cm} (1)

When $cf(d,c)$ denotes the frequency that a concept $c \in C$ appears in a text $d$. $ref_c(t)$ mapping the term into concept.

### 3.2.2. First Concept Strategy

This strategy considers only the most often used sense of the word as the most appropriate concept. This strategy is based on the assumption that the used ontology returns an ordered list of concepts in which more common meanings are listed before less common ones in hierarchical order [9].

$$Cf(d,c) = tf\{d, \{t \in T \text{first}(\text{ref}_c(t)) = c\}\}$$  \hspace{1cm} (2)

### 4. Proposed Method for Arabic WSD

In this section, we present a new method for Arabic WSD using External Knowledge Resources like AWN and WN. Our proposed method utilizes the AWN resource to Map terms into concepts. However, AWN is an incomplete project as previously shown in Table 1, and contains less concepts, less nominal and less verbal phrases than the English version of WN. Hence, when mapping terms into AWN, it may be any concept corresponding to the original term in the text. To overcome this problem, in this paper we suggest two potential solutions: The first stage relates to the mapping strategy. For a concept that does not exist in AWN (e.g., الزراعة), we use MTS from Arabic to English (e.g., agriculture) to find the corresponding concept using Knowledge External Resources like WN (e.g., department of agriculture, agriculture department). Finally, we use the MTS from English to Arabic to yield the corresponding translated concept in the Arabic language (e.g., قسم الزراعة). The second stage relates to the disambiguation strategy. It consists of choosing the nearest concept to the ambiguous term based on more relationships with different concepts in the same local context.

#### 4.1. Mapping Terms into Concepts

In this strategy, after omitting the stop words, for example: {“واب”, “same”, “بعض”, “some”, “من”, “from”, “إلى”, “to”}, the text is analyzed sentence by sentence. The sentence defines the local context of each term that appears.

The local context is the bi-gram on the left and on the right of term (± 2). Then, for process mapping the term into concepts, we extract the concepts of all terms of the documents using AWN.

For example the term “استكمال” has some synset corresponding to: “Achievement”, “الإنجاز”, “كامل”, “أكمل”, “Completed”, “Continue”, “Integrate”, “ første”, “دمج”. For the term “زراعة” we search the translation “agriculture” in WN. The synset corresponding are: “department of agriculture, agriculture department” which are equivalent to the concepts “قسم الزراعة”.

In our approach, we adopt the only concept strategy for vector representation and for the term that has several meanings (concepts) we present a new method to choose the nearest concept, based on more relationships with different concepts to the same local context. More details of our proposed method are described in the next section.
4.2. Strategy for Word Sense Disambiguation

WSD allows us to find the most appropriate sense of the ambiguous word. One word may have several meanings and thus one word may be mapped into several concepts, therefore we need to determine the correct concept. The main idea behind this work is to propose a new and efficient method for Arabic WSD based on the Knowledge approach. In this, to determine the most appropriate concept for an ambiguous term in a sentence, we select the concepts that have a more semantic relationship with other concepts in the same local context. The nearest concept

\[
d_{\text{nearest}} = \max_{c} \max_{c'} S_{c,c'}
\]

where \(n\): number of concepts in the local context of the ambiguous term, \(m\): number of the concept of the ambiguous term, and \(w_{j}\): The concept in the local context.

Figure 1 below describes the proposed method for Arabic WSD. We then describe the similarity measures in more detail. Algorithm 1 presents the algorithm for Arabic WSD.

4.2.1. Semantic Similarity Measures

Measures of text similarity have been used for a long time in NLP applications and related areas.

4.2.2. Feature Selection

Feature Selection [14, 16] studies how to select the list of variables that are used to construct models describing data. Its purposes include reducing dimensionality, removing irrelevant and redundant
features, reducing the amount of data needed for learning and improving accuracy. In this work, we used the Chi-Square statistics for feature selection.

- **Chi-Square**: The Chi-Square statistics can be used to measure the degree of association between a term and a category [14]. Its application is based on the assumption that a term whose frequency strongly depends on the category in which it occurs will be more useful for discriminating it among other categories. For the purpose of dimensionality reduction, terms with small Chi-Square values are discarded. The Chi-Square multivariate is a supervised method allowing the selection of terms by taking into account not only their frequencies in each category but also the interaction of the terms between them and the interactions between the terms and the categories. The principal consists in extracting k better features characterizing best the category compared to the others, this for each category. An arithmetically simpler way of computing chi-square is the following:

\[
X^2_{w,c} = n \frac{(p(w,c) * p(w,c) - p(w,c) * p(w,c))^2}{p(w) * p(w) * p(c) * p(c)}
\]  

(5)

Where \( p(w, c) \) represents the probability that the documents in the category c contain the term w, \( p(w) \) represents the probability that the documents in the corpus contain the term w, and \( w(c) \) represents the probability that the documents in the corpus are in the category c, and so on. These probabilities are estimated by counting the occurrences of terms and categories in the corpus.

The feature selection method chi-square could be described as follows. For a corpus with m classes, the term-goodness of a term w is usually defined as either one of:

\[
X^2_{max}(w) = \max \left\{ X^2_{w,c} \right\}
\]

(6)

\[
X^2_{avg}(w) = \sum_{j=1}^{m} p(c_j) * X^2_{w,c_j}
\]

(7)

Where \( p(c_j) \): The probability of the documents to be in the category \( c_j \) then, the terms whose term-goodness measure is lower than a certain threshold would be removed from the feature space. In other words, chi-square selects terms having strong dependency on categories.

### 4.2.3. Weighting Concepts

The weight \( W(C'_d) \) of a concept \( C'_i \) in a document \( d \) is defined as the combined measure of its local centrality and its global centrality, formally:

\[
W(C'_d) = cc(C'_i,d) * idc(C'_i)
\]

(8)

The local centrality of a concept \( C'_i \) in a document \( d \), noted \( cc(C'_i,d) \) based on its pertinence in the document, and its occurrence frequency. Formally:

\[
cc(C'_i,d) = \alpha * tf(C'_i,d) + (1-\alpha) \sum_{i=1}^{n} Sim(C'_i,C'_i)
\]

(9)

Where \( \alpha \) is a weighting factor that balances the frequency in relation with the pertinence (this factor is determined by experimentation), \( Sim(C'_i,C'_i) \) measures the semantic similarity between concepts \( C'_i \) and \( C'_i \), \( tf(C'_i,C'_i) \) is the occurrence frequency of the concepts \( C'_i \) in the document \( d \).

The global centrality of a concept is its discrimination in the collection. A concept which is central in too many documents is not discriminating. Considering that a concept \( C'_i \) is central in a document \( d \), if their centrality is superior to a fixed threshold \( s \), the document centrality of the concept is defined as follows:

\[
dc(C'_i) = \frac{n}{N}
\]

(10)

### 5. Evaluation and Discussion

In the following section, we describe the corpus utilized in our experiment and the preprocessing algorithm of the input of text. We present a brief description of AWN Ontology. And finally, we outline the results and discussion.

#### 5.1. Corpus Description and Preprocessing

In this work, we use the data provided by Arabic natural language resource: Essex Arabic Summaries Corpus (EASC). It contains 153 Arabic articles and 765 human-generated extractive summaries of those articles. These summaries were generated using http://www.mturk.com/. Among the major features of EASC are: Names and extensions are formatted to be compatible with current evaluation systems. The data are available in two encoding formats UTF-8 and ISO-8859-6 (Arabic).

This corpus is classified into 10 categories as shown in Table 2. In this Arabic dataset, each document was saved in a separate file within the corresponding category's directory.

The dataset was divided into two parts: training and testing. The training data consist of 60% of the documents in each category. The testing data, on the other hand consist of 40% of the documents in each category. Figure 2 presents a sample text of the finance category.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Number of Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art and Music</td>
<td>10</td>
</tr>
<tr>
<td>Education</td>
<td>97</td>
</tr>
<tr>
<td>Environment</td>
<td>34</td>
</tr>
<tr>
<td>Finance</td>
<td>17</td>
</tr>
<tr>
<td>Health</td>
<td>17</td>
</tr>
<tr>
<td>Politics</td>
<td>21</td>
</tr>
<tr>
<td>Religion</td>
<td>08</td>
</tr>
<tr>
<td>Science and Technology</td>
<td>16</td>
</tr>
<tr>
<td>Sports</td>
<td>10</td>
</tr>
<tr>
<td>Tourism</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 2. EASC’s Arabic text corpus.
The preprocessing of the texts is an important phase in NLP. It is necessary to clean the texts by:

- Removing punctuation, numbers, words written in other languages, and any Arabic word containing special characters.
- Removing the diacritics of the words, if it exists.
- Normalizing the documents by doing the following: replacing the letter (١٠) with (٠), and replacing the letter (٣٢٤) with (٣٢٤).

5.2. AWN Ontology

AWN is a lexical resource for standard modern Arabic based on Princeton WN and is built according to methods developed for Euro WN. AWN can be related to other WN of other languages, allowing for translation from and into tens of languages. The connection of WN to SUMO ontology (Suggested Upper Merged Ontology) is also an asset.

AWN contains 9,228 concepts or synsets (6,252 nominal; 2,260 verbal; 606 adjectival; and 106 adverbial), 18,957 expressions and 1,155 named concepts. The files bases ANW under XML format contain the four tags:

- **Item Tag**: Contains (Synset) concepts, classes and instances of the ontology.
- **Word Tag**: Contains words.
- **Form Tag**: Contains the roots of Arabic words.
- **Link Tag**: Contains the relationships between concepts.

In our work, similar words (synonyms) are represented by one concept.

5.3. Results and Discussion

Our method is measured in terms of precision and recall. Precision and recall are defined as:

\[
Recall = \frac{a}{a+c} \quad a+c > 0 \quad \text{and} \quad Precision = \frac{a}{a+b} \quad a+b > 0
\]  

Where \(a\) counts the assigned and correct cases, \(b\) counts the assigned and incorrect cases, \(c\) counts the not assigned but incorrect cases and \(d\) counts the not assigned and correct cases.

The values of precision and recall often depend on parameter tuning; there is a trade-off between them. This is why we also use another measure that combines both the precision and recall: the F1-measure which is defined as follows:

\[
F1 - measure = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}, a+c > 0
\]  

To evaluate the methods proposed, we explore the semantic similarity measure to choose the nearest concept, and we propose to use the Chi-Square method to reduce dimensionality.

A result of our proposed method with two classifiers: SVM and NB, and to using CHI method to feature selection, is presented in Figure 3. Overall, the proposed method achieved the best performance. Specifically, the best accuracy 73.2% (Table 3) was achieved with the proposed method with Wu and Palmer's measure using the CHI to features selection and the SVM classifier.

6. Conclusions and Future Work

WSD plays a vital role in many Text Mining applications. WSD problem has been widely investigated and solved in English and other European languages. Unfortunately, for Arabic language this problem remains a very difficult task. Yet no complete WSD method for this language is available.

In this paper, to overcome this problem, we proposed an efficient method based Knowledge approach. In fact, two contributions have been proposed and evaluated. In the first one, we suggested to use both two external resources AWN and WN based on term to term MTS. The second contribution relates to the disambiguation strategies, it consists of choosing the nearest concept for the ambiguous terms, based on more relationships with different concepts in the same local context.
To illustrate the accuracy of our proposed method, this later has been integrated and evaluated using our Arabic TC System [13]. The obtained results illustrate clearly that the proposed method for Arabic WSD outperforms greatly the other ones. In the future work, we propose a generalized method exploring the use of Wikipedia as the lexical resource for disambiguation.

References


Meeting of the Association for Computational Linguistics, Stroudsburg, USA, pp. 1522-1531, 2010.


Meryeme Hadni is a phd student in laboratory of Computer and Modelization, Faculty of Sciences, University Sidi Mohamed Ben Abdellah (USMBA), Fez, Morocco. She has also presented different papers at different National and International conferences.

Abdelmonaime Lachkar received his PhD degree from the USMBA, Morocco in 2004 in computer science. He is working as a professor and head of Computer Science and Engineering (E.N.S.A), in University Sidi Mohamed Ben Abdellah (USMBA), Fez, Morocco. His current research interests include: Arabic text mining applications: Arabic web document clustering and categorization, Arabic information and retrieval systems, Arabic text summarization, image indexing and retrieval, 3D shape indexing and retrieval in large 3D objects data bases, color image segmentation, unsupervised clustering, cluster validity index, on-line and off-line Arabic and Latin handwritten recognition, and medical image applications.

Said Ouatk is working as a professor in Department of Computer Science, Faculty of Science Dhar EL Mahraz (FSDM), Fez, Morocco. His research interests include high-dimensional indexing and content-based retrieval, Arabic document categorization, 2D/3D shapes indexing and retrieval in large 3D objects database.