Word Sense Disambiguation for Arabic Text Categorization

Meryeme Hadni[1], Said El Alaoui[1], and Abdelmonaime Lachkar[2]
[1]Department of Computer Science, Sidi Mohamed Ben Abdellah University, Morocco
[2]Department 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”: “عضو الإبصار للإنسان و غيره من الحيوان, the visual organ of humans and of animals” and gloss 2 “source”: “ينبوع الماء ينبع من الأرض و 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.

	WN	AWN
Number of Concepts	117.659	10.165
Number of Nominal	117.798	6.252
Number of Verbal	11.529	2.260
Number of Adjectival	21.479	606
Number of Adverbials	4.481	106

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 $\vec{t}_d$ 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 $\vec{c}_d = \left(cf\left(d,c_1\right),\ldots,cf\left(d,c_1\right)\right)$. The concept vector with $1=|C|$ and $Cf\left(d,c\right)$ denotes the frequency that a concept *c* ™*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:

$$Cf\left(d,c\right) = tf\left\{d,\left\{t \in T \setminus c \in ref_c\left(t\right)\right\}\right\} \tag{1}$$

When *cf*(*d*, *c*) denotes the frequency that a concept *cϵC* 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].

$$Cf\left(d,c\right) = tf\left\{d,\left\{t \in Tfirst\left(ref_c\left(t\right)\right) = c\right\}\right\} \tag{2}$$

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: {"سواء, same", "بعض, some", "من, from", "الى, 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 انجاز", "To complete إكمال", "Complete أكمل", "Completed أنجز", "Continue ناضج", "Integrate دمج" . For the term "الزراعة" 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 meaning 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 is calculated as follows:

$$C_{nearst} = max\ max\ S_c$$
$$S_c = \sum_{j=1..n}^{c=1..m} Sim\left(c, w_j\right) \tag{3}$$

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.

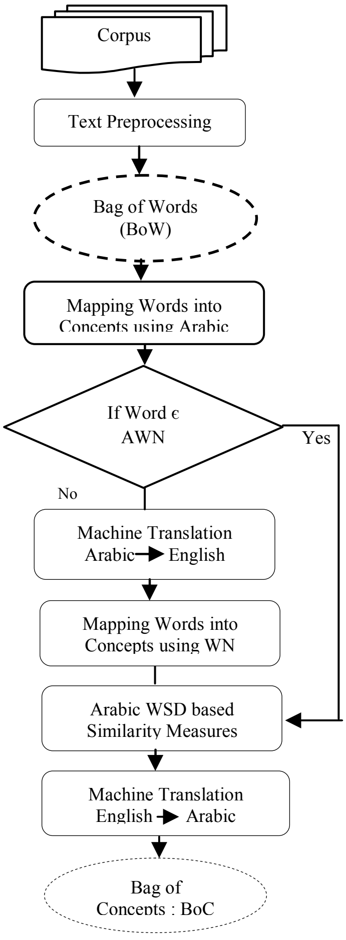
Corpus
Text Preprocessing
Bag of Words (BoW)
Mapping Words into Concepts using Arabic
If Word є AWN
Yes
No
Machine Translation Arabic ➔ English
Mapping Words into Concepts using WN
Arabic WSD based Similarity Measures
Machine Translation English ➔ Arabic
Bag of Concepts : BoC

Figure 1. Flowchart of the proposed method 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.

In this section, we present the similarity measure which can be applied to find the concept that corresponds to the correct sense of the ambiguous words. We use the following definitions and notations: Len: The length of the shortest path in AWN from synset to synset (measured in edges or nodes) is denoted by $len(c_1, c_2)$, depth: The depth of a node is the length of the path to it from the global root, i.e., $depth(c_1, c_2)=len(c_1,c_2)$, lso: We write $lso(c_1, c_2)$ for the lowest super-ordinate of c_1 and c_2.

- Wu and Palmer's Similarity: Budanitsky and Hirst [4] introduce a scaled metric for what they call conceptual similarity between a pair of concepts in a hierarchy such as:

$$sim_{wp}(c_1,c_2) = \frac{2*depth\left(lso\left(c_1,c_2\right)\right)}{len\left(c_1,lso\left(c_1,c_2\right)\right)+len\left(c_2,lso\left(c_1,c_2\right)\right)+2*depth\left(lso\left(c_1,c_2\right)\right)} \tag{4}$$

Algorithm 1: The proposed method for Arabic WSD.

```
W: Ambiguous term.
S: Sentence containing w.
N: Number of the concepts of term w.
LC={$c_1$, $c_2$, ……, $c_N$} List of the concepts of W.
K: Number of concepts in the local context of W.
LW={$c_1$, $c_2$, ……, $c_K$}List of the concepts of Local Context (± 2 terms).
MTS: Machine Translation System
WN: WN Ontology
AWN: AWN Ontology
$Sim(c_i, c_j)$: The similarity measure between two concepts $c_i$ and $c_j$.
    For each term W є S do{
        Map W into concepts using AWN.
        If W є AWN then LC={$c_1$, $c_2$, ……, $c_N$}
        Else
            Use MTS (Arabic to English) for term W.
            W' ←MTS(W)
            Map W' into concepts using WN
                If W' ∉WN then omit the term
                Else LC={$c_1$, $c_2$, ……, $c_N$}
                End If
            End If
/* Calculate the score with the other concepts in the local context*/
S(C) ← 0
    For each concept$c_i$ є LC
    {
    For each concept$w_j$є LW
        S($c_i$) ← S($c_i$) + Sim ($c_i$, $w_j$)
    }
/* Select the nearest concept*/
    $C_p(W)=C_p/max_{i=1....N}S(ci)=S(C_p)$
```

In the next section, we describe the Feature Selection methods applied to reduce dimensionality and remove irrelevant features.

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:

$$X^2_{w,c} = \frac{\mathrm{n} * \left(p(w,c) * p(\overline{w,c}) - p\left(w,\bar{c}\right) * p\left(\bar{w},\mathrm{c}\right)\right)^2}{p(w) * p(\bar{w}) * p\left(\bar{c}\right) * p(c)} \tag{5}$$

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:

$$X^2_{max}(w) = \overset{max}{\underset{j}{}}\left\{X^2_{w,c_j}\right\} \tag{6}$$

$$X^2_{avg}(w) = \sum_{j=1}^{m} p\left(c_j\right) * X^2_{w,c_j} \tag{7}$$

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 $W(C^i_d)$ of a concept C^i, in a document d is defined as the combined measure of its local centrality and its global centrality, formally:

$$W\left(C^i_d\right) = cc\left(C^i, d\right) * idc\left(C^i\right) \tag{8}$$

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:

$$cc\left(C^i, d\right) = \alpha * tf\left(C^i, d\right) + (1-\alpha) \sum_{i \neq 1} Sim\left(C^i, C^1\right) \tag{9}$$

Where α is a weighting factor that balances the frequency in relation with the pertinence (this factor is determined by experimentation), $Sim(C^i, C^1)$ measures the semantic similarity between concepts C^i and C^1, $tf(C^i, C^1)$ 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:

$$dc\left(C^i\right) = \frac{n}{N} \tag{10}$$

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 2. EASC's Arabic text corpus.

Categories	Number of Documents
Art and Music	10
Education	07
Environment	34
Finance	17
Health	17
Politics	21
Religion	08
Science and Technology	16
Sports	10
Tourism	14

تعهدت مؤسسات مستثمرة أمريكية وبريطانية تديرأصولا تتجاوز قيمتها 3
تريليونات دولارباستثمارملياردولارفي شركات الطاقة النظيفة في محاولة
للحد من المخاطرالتي تسببها التغيرات المناخية .
وقال رئيس مكتب خدماتا لإدارة المالية في كاليفورنيا ستيف وستلي إن
الأموال ستستثمرفي أي مشروعات سواء لتوليد الطاقة الكهربائية أواستخدام
توربينات اكثر كفاءة في محطات الكهرباء أو شركات صناعة السيارات مثل
تويوتا التيتنتج سيارات تعمل بالوقود والكهرباء. وقال وستلي "مهمتنا تشجيع
الشركات على التفكيرفي البيئة ."
جاء ذلك خلال اجتماع قمة المؤسسات المستثمرة لللتصدي للمخاطرالمناخية
الذي يعقد في مقرالأمم المتحدة بحضور مسؤولين ماليين من الولايات
الأمريكية وممولين ومستثمرين كبارلبحث كيفية التعامل مع المخاطرالمالية
للتغيرات المناخية.

Figure 2. A sample of an Arabic text.

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 (“إ آ أ”) with (“ا”), and replacing the letter (“ؤ ء”) with (“ا”).

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:

$$Recall = \frac{a}{a+c}, a+c > 0 \text{ and } Precision = \frac{a}{a+b}, a+b > 0 \quad (11)$$

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 andrecall: the F1-measure which is defined as follows:

$$F1 - measure = \frac{2*(Precision * Recall)}{Precision + Recall}, a+c > 0 \quad (12)$$

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.

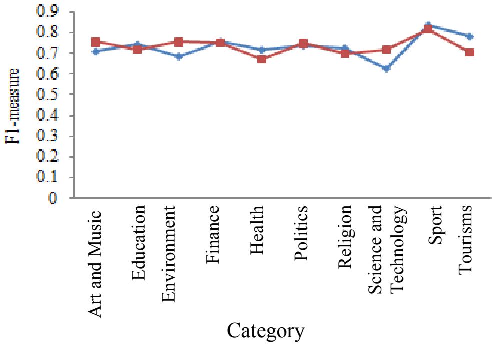

Figure 3. The results (F1-measure) obtained with chi-square reduction techniques using SVM and NB classifiers.

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.

Table 3. The comparison of performance on EASC’s corpus.

	Rappel	Precision	F1-Mesure
SVM	0,746	0,718	0,732
Naive Bayesien	0,747	0,71	0,782

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

[1] Agirre E., Lacalle O., and Soroa A., "Knowledge-Based WSD on Specific Domains: Performing Better than Generic Supervised WSD," *in Proceedings of the 21st International Joint Conference on Artificial Intelligence*, San Francisco, USA, pp. 1501-1506, 2009.

[2] Artiles J., Gonzalo J., and Sekine S., "WePS 2 Evaluation Campaign: Overview of the Web People Search Clustering Task," *in Proceedings of In Web People Search Evaluation Workshop* (*WePS*), 2009.

[3] Bagga A. and Baldwin B., "Entity-Based Cross-Document Coreferencing using the Vector Space Model," *in Proceedings of the 17th international conference on Computational linguistics*, Stroudsburg, USA, pp. 79-85, 1998.

[4] Budanitsky A. and Hirst G., "Evaluating WN-based Measures of Lexical Semantic Relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13-47, 2006.

[5] Carpuat M. and Wu D., "Improving Statistical Machine Translation using Word Sense Disambiguation," *in Proceedings of the Proceedings of 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, pp. 61-72, 2007.

[6] Chan Y., Ng H., and Chiang D., "Word Sense Disambiguation Improves Statistical Machine Translation," *in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 33-40, 2007.

[7] Chen P., Ding W., Bowes C., and Brown D., "A fully Unsupervised Word Sense Disambiguation Method using Dependency Knowledge," *in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, USA, pp. 28-36, 2009.

[8] Dagan I. and Itai A., "Word Sense Disambiguation using a Second Language Monolingual Corpus," *Computational Linguistics*, vol. 20, no. 4, pp. 563-596, 1994.

[9] Elberrichi Z., Rahmoun A. and Bentaalah M., "Using WN for Text Categorization," *the International Arab Journal of Information Technology*, vol. 5, no. 1, pp. 16-24, 2008.

[10] Ellman J., Klincke I. and Tait J., "Word Sense Disambiguation by Information Filtering and Extraction," *Computers and the Humanities,* vol. 34, no. 1-2, pp. 127-134, 2000.

[11] Gale W., Church K., and Yarowsky D., "A Method for Disambiguating Word Senses in a Large Corpus," *Computers and the Humanities*, vol. 26, no. 5, pp. 415-439, 1992.

[12] Grineva M., Grinev M., and Lizorkin D., "Extracting Key Terms from Noisy and Multitheme Documents," *in Proceedings of the 18th International Conference on World Wide Web*, New York, USA, pp. 661-670, 2009.

[13] Hadni M., Ouatik S., and Lachkar A., "Hybrid Part-of-Speech Tagger for Non-Vocalized Arabic Text," *International Journal on Natural Language Computing*, vol. 2, no. 6, pp. 1-15, 2013.

[14] Karima A., Zakaria E. and Yamina T., "Arabic Text Categorization: A Comparative Study of Different Representation Modes," *Journal of Theoretical and Applied Information Technology*, vol. 38, no. 1, pp. 1-5, 2012.

[15] Lesk M. "Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," *in Proceedings of the 5th Annual International Conference on Systems Documentation*, New York, USA, pp. 24-26, 1986.

[16] Li Y., Luo C., and Chung S., "Text Clustering with Feature Selection by Using Statistical Data," *IEEE Transactions on Knowledge and Data Engineering,* vol. 20, no. 5, pp. 641- 652, 2008.

[17] Lin D., "Automatic Retrieval and Clustering of Similar Words," *in Proceeding of the17th International Conference on Computational Linguistics*, Stroudsburg, USA, pp. 768-774, 1998.

[18] Mann G. and Yarowsky D., "Unsupervised Personal Name Disambiguation," *in Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL*, Stroudsburg, USA, pp. 33-40, 2003.

[19] McCarthy D., Koeling R., Weeds J., and Carroll J., "Unsupervised Acquisition of Predominant Word Senses," *Acquisition of Predominant Word Senses*, vol. 33, no. 4, pp. 553-590, 2007.

[20] Medelyan O., Witten I and Milne D., "Topic Indexing with Wikipedia," *in Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence*, Chicago, USA, pp. 19-24, 2008.

[21] Ponzetto S. and Navigli R., "Knowledge-Rich Word Sense Disambiguation Rivaling Supervised Systems," *in Proceedings of the 48th Annual*

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

[22] Ravin Y. and Kazi Z., "Is Hillary Rodham Clinton the President? Disambiguating Names across Documents," *in Proceedings of the ACL'99 Workshop on Coreference and its Applications*, pp. 9-16, 1999.

[23] Reiter N., Hartung M., and Frank A., "A Resource-Poor Approach for Linking Ontology Classes to Wikipedia Articles," *in Proceeding of Johan Bos and Rodolfo Delmonte, editors, Semantics in Text Processing STEP Conference Research in Computational Semantics*, pp. 381-387, 2008.

[24] Resnik P., Yarowsky D., "A Perspective on Word Sense Disambiguation Methods and Their Evaluation," *in Proceedings of SIGLEX'97*, Washington DC, USA, pp. 79-86, 1997.

[25] Ruiz-Casado M., Alfonseca E., and Castells P., "Automatic Assignment of Wikipedia Encyclopedic Entries to WN Synsets," *Springer-Verlag Berlin Heidelberg*, vol. 3528, pp. 380–386, 2005.

[26] Sanderson M., "Word Sense Disambiguation in Information Retrieval," *in Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, USA, pp. 142-151, 1994.

[27] Suchanek F., Kasneci G., and Weikum G., "Yago: A Large Ontology from Wikipedia and WN," Journal *of Web Semantics*, vol. 6, no. 3, pp. 203-217, 2008.

[28] Van L. and Apidianaki M., "Cross-Lingual Word Sense Disambiguation for Predicate Labelling of French," *in Proceedings of TALN*, Marseille, France, pp. 46-55, 2014.

[29] Yarowsky D., "Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora," *in Proceeding of 14th Conference on Computational Linguistic*, Stroudsburg, USA, pp. 454-460, 1992.

[30] Yoshida M., Ikeda M., Ono S., Sato I., and Nakagawa H., "Person Name Disambiguation by Bootstrapping," *in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, USA, pp. 10-17, 2010.

[31] Zouaghi A., Merhbene L., and Zrigui M. "A Word Sense Disambiguation for Arabic Language using the Variants of the Lesk Algorithm," *in Proceeding of International Conference on Agents and Artificial Intelligence*, Valencia, Spain, pp. 22-24, 2012.

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 Ouatik 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.