# Word Sense Disambiguation for Arabic Text Categorization

Said OUATIK EL ALAOUI[1,3], Meryeme HADNI[1], Abdelmonaime LACHKAR[2], Driss ABOUTAJDINE[3]
[1]LIM, Department of Computer Science, FSDM, USMBA, Morocco
[2] LISA, Department of Electrical & Computer Engineering, E.N.S.A, USMBA, Morocco
[3]LRIT- CNRST URAC29, FSR, Mohammed V-Agdal 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 AWN and WN based on Term to Term Machine Translation System (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 evaluate the accuracy of our proposed method, several experiments have been conducted using Feature Selection methods; Chi-Square and CHIR, and 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:** Word Sense Disambiguation, Arabic Text Categorization System, Arabic WordNet, Machine Translation System.*

## 1. Introduction

Word Sense Disambiguation is the problem of identifying the sense (meaning) of a word within a specific context. In Natural Language Processing (NLP), Word Sense Disambiguation 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 [1, 2], Information Retrieval [3], Information Extraction [4], Machine Translation [5,6], Semantic Annotation [7], Cross-Document Co-Referencing [8, 9], Web People Search [10- 12]. 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 Word Sense Disambiguation [13-15] 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 [16, 17], 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 [18, 19] the authors have utilized Machine-Readable Dictionaries MRD, thesauri, and computational lexicons, such as WordNet. Since most MRD and thesauri were created for human use and display inconsistencies, these methods have clear limitations. Like WordNet extends Knowledge Resource for the English language, Arabic WordNet has been developed for the Arabic language, but it is an incomplete project. To overcome the above cited problem, we propose in this work an efficient method for Arabic WSD based Knowledge External resource (Arabic WordNet). For the terms don't exist in Arabic WordNet, we traduce the terms from Arabic into English using MTS and search the corresponding concepts in WordNet 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 the 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 work

Word Sense Disambiguation 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 [13] 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. In [21], the authors 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 [22]. 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. In [14], the authors have presented the method for WSD with a personalized PageRank. [21], they collect feature words with direct dependency like relationships. Knowledge from Wikipedia is injected into a WSD system by means of a mapping to WordNet. Previous efforts aimed at automatically linking Wikipedia to WordNet include; full use of the first WordNet sense heuristic [23], a graph-based mapping of Wikipedia categories to WordNet synsets [15], a model based on vector spaces [24] and a supervised approach using keyword extraction [25].

Unlike European languages, there are few works and contributions that deal with Arabic WSD. In [26], the authors propose a new approach for text categorization, based on incorporating semantic resource (WordNet) 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. Another work, [20] 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 WordNet to determine the correct sense for an ambiguous term. The authors in [27] 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 [28], 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, [27, 28] uses the dictionaries gloss for each concept. For example, the term"عين"has two glosses in the Al-Wasit dictionary[1] : 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 [20,26] the authors present the systems that use Bag-of-Concept and choose the first concepts after mapping on Arabic WordNet for determining the correct concepts, and the first concept is random and therefore not always the best choice.

Table1. Difference between Arabic WN and WN

| | WordNet | Arabic WordNet |
|---|---|---|
| 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 Arabic WordNet is the lack of many concepts because Arabic WordNet is an incomplete project (e.g. Table1). Therefore, for the terms that do not exist in AWN we search for the corresponding concepts on WordNet based on Machine Translation System (MTS).

Therefore, for the terms that do not exist in AWN we search for the corresponding concepts on WordNet based on Machine Translation System (MTS). In this paper, for any term that has a different meaning, we propose a new method for Arabic Word Sense Disambiguation (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

[1] http://www.al3arabiya.org/2010/01/arabic-arabic-dictionary.html

or replacing terms by concepts can be distinguished as proposed by [26]:

### 3.1.1. Add Concepts

This strategy extends each term vector $\vec{t}_d$ by new entries for WordNet 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 $\vec{c}_d = \left(\mathrm{cf}(d, c_1), \ldots, \mathrm{cf}(d, c_l)\right)$. The concept vector with $l = |C|$and $\mathrm{cf}(d, c)$ denotes the frequency that a concept c ϵ C appears in a text d.

The terms, which appear in WordNet as a concept [20] will be accounted for 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$\vec{c}_d$. The example illustrates the vector representation

### 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 WordNet will be taken into account only in the concept vector. The vector of the terms will thus contain only the terms which do not appear in WordNet.

### 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 WordNet; $\vec{c}_d$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 [26]. 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 [26] 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:

$$Cf(d,c) = tf\{d, \{t \in T \backslash c \in ref_c(t)\}\} \quad (1)$$

When: $Cf(d,c)$ denotes the frequency that a concept c ϵ 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 [26].

$$Cf(d,c) = tf\left\{d, \{t \in T\ first\left(ref_c(t)\right) = c\}\right\} \quad (2)$$

## 4. Proposed Method for Arabic WSD

In this section, we present a new method for Arabic WSD using External Knowledge Resources like Arabic WordNet and WordNet. Our proposed method utilizes the Arabic WordNet resource to Map terms into concepts. However, Arabic WordNet 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 WordNet. 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 the Machine Translation System from Arabic to English (e.g. agriculture) to find the corresponding concept using Knowledge External Resources like WordNet (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. For terms that have different synonyms, we suggest a new method for Word Sense Disambiguation 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 Arabic WordNet.

For example the term "استكمال" has some synset corresponding to: "Achievement إنجاز،", "To complete إكمال "," Complete أكمل"," Completed أنجز"," Continue

ناضج ", " Integrate دمج" . For each term that does not exist in AWN, we propose to translate the term by using Machine Translation System from Arabic to English, in order to restart the search of the meanings in WordNet. For example: the term "agriculture الزراعة" does not exist in AWN, so we search the translation "agriculture" in WordNet. 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

Word Sense Disambiguation 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.

Where Sim will be detailed in section4.3.

The nearest concept is calculated as follows:

$$C_{nearst} = \max \quad S_c \qquad (3)$$

Where: n is the number of concepts proposed and c is the concept.

Figure 1 below describes the proposed method for Arabic WSD. We then describe the similarity measures in more detail. Figure 2 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.

In this section, we present the similarity measure [29] 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 Arabic WordNet from synset to synset (measured in edges or nodes) is denoted by $\mathrm{len}(c_1, c_2)$ .

Depth: The depth of a node is the length of the path to it from the global root, i.e.,

$\mathrm{depth}(c_1, c_2) = \mathrm{len}(c_1, c_2)$ .

Lso: We write $\mathrm{lso}(c_1, c_2)$ for the lowest super-ordinate of $c_1$ and $c_2$.

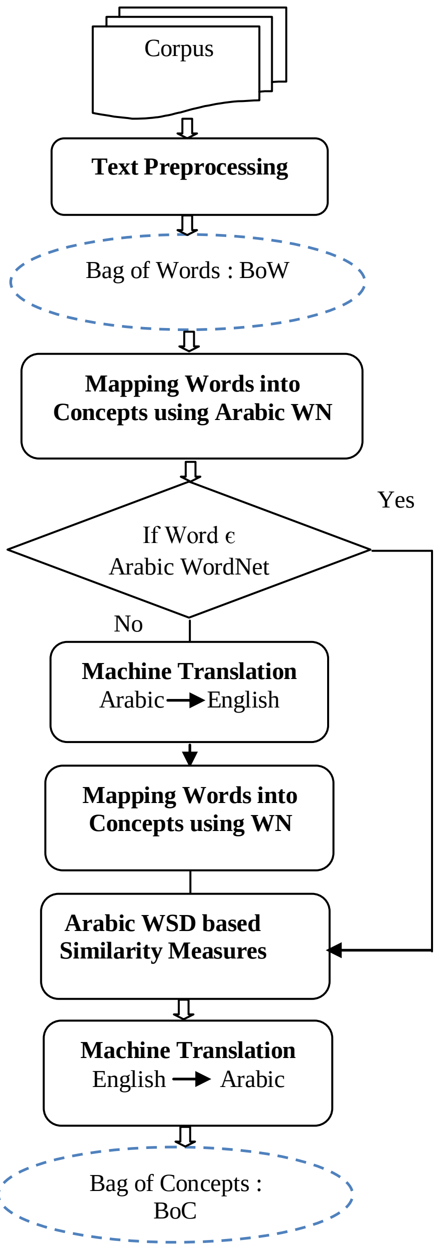


Figure 1. Flowchart of the Proposed Method for Arabic WSD

**Wu and Palmer's Similarity:** Wu [29] introduce a scaled metric for what they call conceptual similarity between a pair of concepts in a hierarchy such as:

$$\mathrm{sim}_{wp}(c_1, c_2) = \frac{2 * \mathrm{depth}(\mathrm{lso}(c_1, c_2))}{\mathrm{len}(c_1, \mathrm{lso}(c_1, c_2)) + \mathrm{len}(c_2, \mathrm{lso}(c_1, c_2)) + 2 * \mathrm{depth}(\mathrm{lso}(c_1, c_2))} \quad \textbf{(4)}$$

W: Ambiguous term.
S: Sentence containing w.
N: Number of the concepts of term w.
$LC = \{c_1, c_2, \ldots, c_N\}$: List of the concepts of W.
K: Number of concepts in the local context of W.
$LW = \{c_1, c_2, \ldots, c_k\}$ : List of the concepts of Local Context (± 2 terms).
MTS: Machine Translation System
WN: WordNet Ontology
AWN: Arabic WordNet Ontology
$Sim(c_i, c_j)$ : The similarity measure between two concepts $c_i$ and $c_j$ .

```
For each term W ϵ S do{
  -  Map W into concepts using AWN.
  -  If W ϵ AWN then LC = {c1, c2, ..., cN}
  -  Else
     -  Use MTS (Arabic to English) for term W.
        W'←MTS(W)
     -  Map W' into concepts using WN
          o  If W'∉WN then omit the term
          o  Else LC = {c1, c2, ..., cN}
          o  End If
  -  End If
/* Calculate the score with the other concepts in the local context*/
S(C)← 0
  For each concept ci ϵ LC
  {
     For each concept wj ϵ LW
        S(ci) ← S(ci) + Sim (ci, wj)
  }
/* Select the nearest concept*/
     Cp(W) = cp / max_{i=1..N} S(ci) = S(cp)
}
```

Figure 2. The Algorithm of the proposed method for Arabic WSD

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 [20, 30] 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 [20]. 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:

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

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 $p(c)$ 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:

$$X^2_{max}(w) = \max_j \left\{X^2_{w,c_j}\right\} \quad (6)$$

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

Where $p(c_j)$ is 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:

$$W(C^i_d) = cc(C^i, d) * idc(C^i) \quad (8)$$

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:

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

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

$$dc(C^i) = \frac{n}{N} \quad (10)$$

## 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 Arabic WordNet 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: the EASC (Essex Arabic Summaries Corpus). 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).

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 |

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

تعهدت مؤسسات مستثمرة أمريكية وبريطانية تدير أصولا تتجاوز قيمتها 3 تريليونات دولار باستثمار مليار دولار في شركات الطاقة النظيفة في محاولة للحد من المخاطر التي تسببها التغيرات المناخية
.
وقال رئيس مكتب خدمات الإدارة المالية في كاليفورنيا ستيف وستلي إن الأموال ستستثمر في أي مشروعات سواء لتوليد الطاقة الكهربائية أو استخدام توربينات اكثر كفاءة في محطات الكهرباء أو شركات صناعة السيارات مثل تويوتا التي تنتج سيارات تعمل

Figure 3. A sample of an Arabic Text

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 3 presents a sample text of the finance category.

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.Arabic WordNet Ontology

Arabic WordNet[2] (AWN) is a lexical resource for standard modern Arabic based on Princeton WordNet and is built according to methods developed for Euro WordNet. AWN can be related to other WordNet[3] (WN) of other languages, allowing for translation from and into tens of languages. The connection of WordNet to SUMO ontology (Suggested Upper Merged Ontology) is also an asset.

Arabic WordNet 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:

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

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 and recall: the F1-measure which is defined as follows:

[2] http://globalwordnet.org/arabic-wordnet/

[3] http://wordnet.princeton.edu/

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

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.

A result of our proposed method with two classifiers: SVM and NB, and to using CHI method to feature selection, is presented in Figure 4.

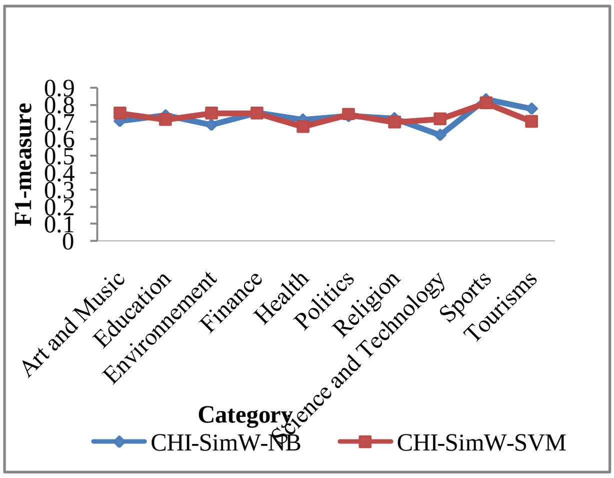


Figure 4. The results (F1-measure) obtained with Chi-square reduction techniques using SVM and NB classifiers.

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. Conclusion and Future Work

Word Sense Disambiguation 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 a 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 Machine Translation System 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 [31]. 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.

**Reference**

[1]M. Grineva, M. Grinev, and D. Lizorkin. Extracting key terms from noisy and multitheme documents. In WWW, pages 661-670, 2009.

[2]O. Medelyan, I. H. Witten, and D. Milne. Topic indexing with wikipedia. In AAAI Workshop on Wikipedia and Arti¯cial Intelligence, 2008.

[3]M. Sanderson .Word sense disambiguation and information retrieval. In SIGIR, Proceeding of the 17th annual international ACM SIGIR. Pages 142-151.

[4] J. Ellman, I. Klincke and J. TAIT. Word Sense Disambiguation by Information Filtering and Extraction. Computers and the Humanities, 2000. Volume 34, Issue 1-2, pp 127-134.

[5]Y. S. Chan, H.T.Ng, and D.Chiang. 2007. Word Sense Disambiguisation improves statistical machine translation. In proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pages 33-40.

[6]M. Carpuat and D. Wu. 2007. Improving statistical machine translation using word sense désambiguisation. In Proceedings of the 2007 Joint conference on Empirical Methods in Natural Language Processing, and Computational Natural Language Learning, pages 61-72.

[7]V. Lonneke and M. Apidianaki. Cross-Lingual Word Sense Disambiguation for Predicate Labelling of French. TALN, Marseille 2014.

[8] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In Int'l Conf. on Computational Linguistics, pages 79-85,1998.

[9]Y. Ravin and Z. Kazi. Is hillary rodham clinton the president?: disambiguating names across documents. In Workshop on Coreference and its Applications (CorefApp), pages 9-16, 1999.

[10] J. Artiles, J. Gonzalo, and S. Sekine. Weps 2 evaluation campaign: overview of the web people search clustering task. In Web People Search Evaluation Workshop (WePS), WWW Conference, 2009.

[11] G. S. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In HLT-NAACL, pages 33-40, 2003.

[12] M. Yoshida, M. Ikeda, S. Ono, I. Sato, and H. Nakagawa. Person name disambiguation by bootstrapping. In ACM SIGIR, pages 10-17, 2010.

[13]Chen P., Ding W., Bowes C. and Brown D., 2009. 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 ACL. ACL, 28-36.

[14] Agire E., Lacalle O. L. d. and Soroa A., 2009. Knowledge-based WSD and specific domains: performing over supervised WSD. In Proceedings of the International Joint Conference on Artificial Intelligence 2009. AAAI Press, 1501-1506.

[15] Ponzetto, S. P. and Navigli, R. (2010), Knowledge-rich Word Sense Disambiguation rivaling supervised system. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11-16 July 2010, pp. 1522-1531.

[16] Dagan, 1994 Dagan, I., and Itai, A. 1994. Word sense disambiguation using a second language monolingual corpus. Computational Linguistics 20(4):563–596.

[17] Gale et al., 1992 Gale, W., K. Church, and D. Yarowsky. A Method for Disambiguating Word Senses in a Large Corpus. Computers and the Humanities. 26, pp. 415-439, 1992.

[18] Resnik et al., 1997 Resnik, P. and D. Yarowsky. A Perspective on Word Sense Disambiguation Methods and Their Evaluation. In Proceedings of SIGLEX '97, Washington, DC, pp. 79-86, 1997.

[19]Yarowsky, 1992 Yarowsky, D. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora.' In Proceedings, COLING-92. Nantes, pp. 454-460, 1992.

[20] Elberrichi Z. Abidi K., 2012.Arabic Text Categorization: A Comparative Study of Different Representation Modes, in IAJIT, Vol. 9 Issue 5.

[21] McCarthy, Koeling R., Julie Weeds and Carroll J, 2007. Unsupervised Acquisition of Predominant Word Senses. Computational Linguistics, 553-590.

[22] Lin D. 1998. Automatic retrieval and clustering of similar words. In Proceedings of COLING-ACL 98. ACL,pp. 768-774.

[23] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. Yago: A large ontology from Wikipedia andWordNet. Journal of Web Semantics, 6(3):203–217.

[24] Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In Advances in Web Intelligence, volume 3528 of Lecture Notes in Computer Science. Springer Verlag.

[25] Nils Reiter, Matthias Hartung, and Anette Frank. 2008. A resource-poor approach for linking ontology classes to Wikipedia articles. In Johan Bos and Rodolfo Delmonte, editors, Semantics in Text Processing, volume 1 of Research in Computational Semantics, pages 381–387. College Publications, London, England.

[26] Elberrichi Z. and Rahmoun A., 2008. Using WordNet for Text Categorization, in IAJIT, Vol.5, No.1.

[27] Merhbene L., Zouaghi A. and Zrigui M. 2012. Arabic Word Sense Disambiguation. In Proceeding of International Conference on Agents and Artificial Intelligence, Volume 1, Valencia, Spain, 22-24 January, p.p:652-655.

[28] Lesk M., 1986, 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 (SIGDOC'86), pp. 24–26.

[29] Budanitsky A., Hirst G.,2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness, Association for Computational Linguistics.

[30] Yanjun Li ., Congnan Luo. , and Soon Chung M., 2008. Text Clustering with Feature Selection by Using Statistical Data, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.

[31]M.Hadni, S. E. Ouatik, A. Lachkar. Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text. International Journal on Natural Language Computing (IJNLC) Vol. 2, No.6, December 2013.