Unsupervised Word Sense Disambiguation Using The WWW

Ioannis P. KLAPAFTIS & Suresh MANANDHAR
Department of Computer Science
University of York
York, UK, YO10 5DD
{giannis,suresh}@cs.york.ac.uk

Abstract.
This paper presents a novel unsupervised methodology for automatic disambiguation of nouns found in unrestricted corpora. The proposed method is based on extending the context of a target word by querying the web, and then measuring the overlap of the extended context with the *topic signatures* of the different senses by using Bayes rule. The algorithm is evaluated on Semcor 2.0. The evaluation showed that the web-based extension of the target word's local context increases the amount of contextual information to perform semantic interpretation, in effect producing a disambiguation methodology, which achieves a result comparable to the performance of the best system in SENSEVAL 3.

Keywords. Unsupervised Learning, Word Sense Disambiguation, Natural Language Processing

1. Introduction

Word Sense Disambiguation is the task of associating a given word in a text or discourse with a definition or meaning (sense) which is distinguishable from other meanings potentially attributable to that word [7]. WSD is a long-standing problem in NLP community.

The outcome of the last SENSEVAL-3 workshop [12] clearly shows that supervised systems [6,16,10] are able to achieve up to 72.9% precision and recall[1] [6] outperforming unsupervised ones. However, supervised systems need to be trained on large quantities of high quality annotated data in order to achieve reliable results, in effect suffering from the knowledge acquisition bottleneck [18].

Recent unsupervised systems [14,17,4] use semantic information (glosses) encoded in WordNet [13] to perform WSD. However, descriptive glosses of WordNet are very sparse and contain very few contextual clues for sense disambiguation [14]. This problem, as well as WordNet's well-known deficiencies, i.e. the lack of explicit links among semantic variant concepts with different part of speech, and the lack of explicit relations between topically related concepts, were tackled by *topic signatures* (TS) [2].

[1] This result refers to the precision and recall achieved by the best supervised system with fine-grained scoring in the English sample task.

A TS of i_{th} sense of word w is a list of words that co-occur with i_{th} sense of w. Each word in a TS has a weight that measures its importance. Essentially, TS contain topically related words for nominal senses of WordNet [13]. TS are an important element of the proposed methodology[2] and thus, our assumption is that TS contain enough contextual clues for WSD.

Another problem of WSD approaches (e.g. like approaches of Lesk [11]) is the criteria based on which we can determine the window size around the target word to create its local context. A large window size increases noise, while a small one decreases contextual information. The common method is to define the window size based on empirical and subjective evaluations, assuming that this size can capture enough amount of related words to perform WSD.

To our knowledge, the idea of extending the local context of a target word has not been studied yet. The proposed method suggests the novelty of extending the local context of the target word by querying the web to obtain more topically related words. Then, the overlap of the extended context with the topic signatures of the different senses is measured using Bayes rule.

The rest of the paper is structured as follows: section 2 provides the background work related to the proposed one, section 3 presents and discusses the disambiguation algorithm, section 4 contains the evaluation of the proposed methodology , section 5 identifies limitations and suggests improvements for future work and finally section 6 summarizes the paper.

2. Background

A topic signature (TS) of i_{th} sense of a word w is a list of the words that co-occur with i_{th} sense of w, together with their respective weights. It is a tool that has been applied to word-sense disambiguation with promising results [1].

Let s_i be the WordNet synset for i_{th} sense of a word w. Agirre's [2] method for the construction of WordNet's TS is the following.

1. **Query generation**
   In this stage, a query is generated, which contains all the monosemous relatives to s_i words as positive keywords, and the words in other synsets as negative keywords. Monosemous relatives can be hypernyms, hyponyms and synonyms.
2. **Web documents download**
   In the second stage, the generated query is submitted to an Internet search engine, and the n first documents are downloaded.
3. **Frequency calculation**
   In the third stage, frequencies of words in documents are calculated and stored in a vector vf_i excluding common closed-class words(determiners, pronouns e.t.c). The vector contains pairs of $(word_j, freq_{i,j})$, where j is the j_{th} word in the vector and i is the i_{th} sense of w.
4. **TS weighting**
   Finally, vf_i is replaced with a vector vx_i that contains pairs of $(word_j, w_{i,j})$, where $w_{i,j}$ is the weight of word j for TS of i_{th} sense of word w. Weighting

[2] In the next section we will provide the exact automatic process for the construction of TS

measures applied so far to TS are the tf/idf measure [15], x^2, t-score and mutual information.

TS were applied to a WSD task, which showed that they are able to overcome WordNet's deficiencies [1]. According to this approach, given an occurrence of the target word in a text, a local context was formed using a window size of 100 words[3]. Then, for each word sense the weights for the context words appearing in the corresponding topic signature were retrieved and summed. The highest sum determined the predicted sense.

A similar WSD method was proposed by Yarowsky [19]. His method performed WSD in unrestricted text using Roget's Thesaurus and Grolier's Encyclopedia and involved 3 stages as summarised below.

1. **Creation of context discrimination lists.**
   Representative contexts are collected to create context discriminators for each one of the 1041 Roget's categories (sense categories). Let $RCat$ be a Roget category. For each occurrence of a word w in a category $RCat$, concordances of 100 surrounding words in the encyclopedia are collected. At the end of this stage, each Roget category is represented by a list of topically related words. From this the conditional probability of each w given $RCat$, $P(w|RCat)$, is calculated.
2. **Salient Word weighting**
   In the second stage, salient words of each list are identified and weighted. Salient words are detected according to their probabilities in a Roget category and in the encyclopedia. The following formula calculates the probability of a word appearing in the context of a Roget category, divided by the overall probability of the word in the encyclopedia corpus.

   $$\frac{P(w|RCat)}{P(w)}$$

   This formula, along with topical frequency sums are multiplied to produce a score for each salient word, and the n highly ranked salient words are selected. Each selected salient word is then assigned the following weight:

   $$\log(P(w|RCat)P(w))$$

   At the end of this stage, each Roget category is represented by a list of topically related words, along with their respective weights. As it is possible to observe, there is a conceptual similarity between TS and Yarowsky's [19] context discriminators.
3. **Disambiguation of a target word.**
   The process to disambiguate a target word was identical to the TS based WSD. A local context was formed, and then for each Roget sense category, the weights for the context words appearing in the corresponding sense discriminator were retrieved and summed. The highest sum determined the predicted Roget sense.

Both of these approaches attempt to disambiguate a target word w by constructing sense discrimination lists, and then measuring the overlap between the local context and each sense discrimination list of w using Bayes rule. Both of these approaches calculate empirically the window size of the local context. In the proposed method, we take the same view of using sense discriminators, but we attempt to extend the local context, in order to provide more information, aiming for a more reliable and accurate WSD.

[3] This window size was chosen by Agirre et. al [1] as the most appropriate one, after several experiments

3. Disambiguation Method

The proposed method consists of three steps which can be followed to perform disambiguation of a word w.

1. **Collect external web corpus WC.**
   In this stage, sentence s containing target word w is sent to Google and the first r documents are downloaded. Part-of-speech (POS) tagging is applied to retrieved documents to identify nouns within a window of $+/-$ n number of words around w. A final list of nouns is produced as a result, which is taken to represent the external web context WC of w.
2. **Retrieve topic signatures TS_i for each nominal sense i of w.**
   In this stage, TS_i for each sense i of w is retrieved or constructed. In our approach, we have used TS weighted by the tf/idf measure [15].
3. **Use WC and TS_i to predict the appropriate sense.**
   When any of the words contained in TS_i appears in WC, there is evidence that i_{th} sense of w might be the appropriate one. For this reason, we sum the weights of words appearing both in TS_i and in WC for each sense i, and then we use Bayes's rule to determine the sense for which the sum is maximum.

   $$\arg\max_{TS_i} \sum_{w \in WC} \left(\frac{P(w|TS_i) * P(TS_i)}{P(w)}\right)$$

   In case of lack of evidence[4], the proposed methodology makes a random choice of the predicted sense.

Essentially, the weight of a word w in a TS_i is $P(w|TS_i)$, the probability of a word w appearing in topic signature TS_i. Consequently, the sum of weights of a word among all senses of a word is equal to 1. The probability $P(TS_i)$ is the *a priori* probability of TS_i to hold, which is essentially the *a priori* probability of sense i of w to hold. Currently, we assume $P(TS_i)$ to be uniformly distributed. Note that $P(w)$ may be omitted, since it does not change the results of the maximization.

3.1. Method Analysis

The first step of the proposed method attempts to extend the local context of the target word by sending queries to Google, to obtain more topically related words. In most WSD approaches, the local context of a target word is formed by an empirically calculated window size, which can introduce noise and hence reduce the WSD accuracy. This is a result of the fact that the appropriate size depends on many factors, such as the writing style of the author, the domain of discourse, the vocabulary used etc.

Consequently, it is possibly impracticable to calculate the appropriate window size for every document, author etc. Even if this is empirically regarded or randomly chosen as the best possible, there is no guarantee that the target word's local context contains enough information to perform accurate WSD.

In our approach, we attempt to overcome this problem by downloading several web documents and choosing a small window size around the target word. Our initial pro-

[4] We were unable to download any documents from the web

jections suggest that through this tactic, we will increase the amount of contextual clues around the target word, in effect increasing the accuracy of WSD.

The proposed method is heavily based on TS and its performance depends on its quality. It has already been mentioned, that TS are conceptually similar to the Roget's sense discrimination (RSD) lists [19]. But Yarowsky's method [19] of constructing RSD lists suffers from noise. We believe that our proposed method is less likely to suffer from noise for the reasons following.

Firstly, TS are built by generating queries containing only monosemous relatives. In contrast, RSD lists were constructed by taking into account each word w appearing in a Roget category. However, many of these words used to collect examples from Grolier's encyclopedia were polysemous, in effect introducing noise.

Secondly, noise reduction is enhanced by the fact that TS are constructed by issuing queries that have other senses's keywords as negative keywords, in effect being able to exclude documents that contain words that are semantically related with senses of the target word, other than the one we are interested in.

Finally, noise reduction is enhanced by the fact that in *tf/idf* [15] based TS, words occurring frequently with one sense, but not with the other senses of the target word, are assigned high weights for the associated word sense, and low values for the rest of word senses. Furthermore, words occurring evenly among all word senses are also assigned low weights for all the word senses [3].

On the contrary, Yarowsky's measure [19] only takes into account the probability of a word appearing in a Roget's category representative context, divided by the overall probability of the word in corpus. Figure 1 shows the conceptual architecture of the afore mentioned WSD methods and of the proposed one.

4. Evaluation

4.1. Preparation

At the first step of the disambiguation process (section 3), we mentioned two parameters that affect the outcome of the proposed disambiguation method. The first one was the number of documents to download and the second was the window size around the target word within the retrieved documents. Let n be the window size and r the number of downloaded documents.

Our purpose at this stage, was to perform WSD on a large part of SemCor 2.0 [9] with different values on r and n, and then choose these values, for which our WSD algorithm was performing better. These values would be used to perform evaluation on the whole SemCor.

Two measures are used for this experiment, $Prank1$ and $Prank2$. $Prank1$ denotes the percentage of cases where the highest scoring sense is the correct sense, and is equal to recall and precision. Note that our recall measure is the same as the precision measure, because every word was assigned a sense tag [5]. $Prank2$ denotes the percentage of cases when one of the first two highest scoring senses is the correct sense.

A part of our experiments is shown in Table 1. We obtained the best results for $r = 4$ and $n = 100$. It seems that when r is above 4, the system retrieves inconsistent

[5] In the seldom case of having lack of evidence to output a sense, the predicted sense was randomly chosen

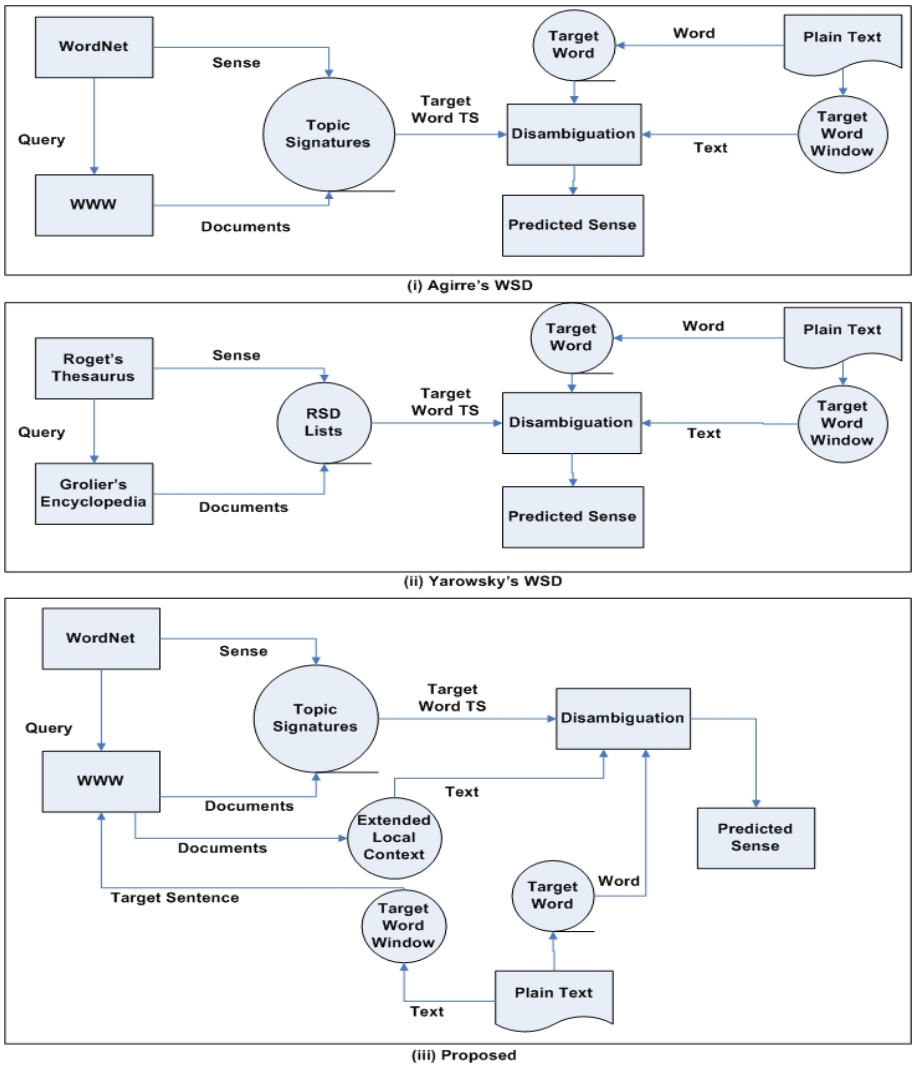

Figure 1. WSD Methods Conceptual Architectures

web documents, which increase noise. In contrast, when r is below 4, the amount of contextual clues decreases. Additionally, for all values of r, when n is above 100, the system receives noisy information, while when n is below 100 (not shown), performance is lowered due to the small window size.

4.2. Results on SemCor

Table 2 shows the results of our evaluation, when n is set to 100 and r is to 4 for the first 10 SemCor files.

Table 2 shows that our system achieved 67.4% on a large part of SemCor 2.0. In 81.2% of cases one of the first two highest scoring senses was the correct one. Results

n	r	Prank1	Prank2
100	3	0.685	0.800
150	3	0.675	0.801
200	3	0.670	0.811
100	4	0.731	0.846
150	4	0.681	0.830
200	4	0.677	0.823
100	5	0.695	0.800
150	5	0.675	0.790
200	5	0.660	0.793

Table 1. Experiments on the br-a01 file of SemCor 2.0

File	Nouns	Prank1	Prank2
br-a01	573	0.731	0.846
br-a02	611	0.729	0.865
br-a11	582	0.692	0.812
br-a12	570	0.685	0.808
br-a13	575	0.600	0.760
br-a14	542	0.706	0.854
br-a15	535	0.691	0.818
br-b13	505	0.570	0.774
br-b20	458	0.641	0.779
br-c01	512	0.671	0.791
Total	5463	0.674	0.812

Table 2. Results from the first 10 files of Brown 1 Corpus

on the whole SemCor did not change significantly. In particular, our approach achieved 69.4% Prank1 and 82.5% Prank2.

Our first baseline method was the performance of TS (without querying the web) using a window of 100 words as Agirre et. al. [1] did. The comparison between TS WSD and the proposed method would allow us to see, if the web-based extension of local context is useful for WSD.

Our second baseline method was a method similar to that of Lesk [11], which is based on extending the content of the target word by querying the web (as in our approach), and then measuring the overlap of the extended context with WordNet-based lists of the different senses as in [8]. Each sense list is constructed by taking into account all the hypernyms, hyponyms, meronyms, holonyms and synonyms of the particular sense. This method is used to show that TS overcome the WordNet deficiencies (section 1) and are useful for WSD. Table 3 shows the comparison between the proposed and the baseline methods.

Table 4 shows a comparison of our method's performance with other recent WSD approaches on same evaluation data set[6]. We compare our method with a method similar

[6]The compared systems performance is mentioned in their literature

Methods	**Prank1** (%)	**Prank2** (%)
Proposed	69.4	82.5
TS WSD	60.79	76.34
Web based Lesk-like WSD	59.90	69.91

Table 3. Comparison between the proposed and the baseline methods

to that of Lesk [11], which creates sense lists for a target word w using WordNet hypernym glosses [5]. Each word in a sense list is assigned a weight inversely proportional to its depth in the WordNet hierarchy. At the end, they measure the overlap of each sense list with the local context[7]. This system is referred to in Table 4 as *WordNet-based Lesk-like*.

The second system is the best performing system in the last SENSEVAL 3 workshop [14] and is similar to the previous one, differing on the weighting of words in the sense lists. In particular, Ramakrishnan et al. [14] use a variation of the tf/idf [15] measure, which they call tf/igf [14]. The inverse gloss frequency (igf) of a token is the inverse of the number of glosses, which contain that token and it captures the commonness of that particular token. This system is referred to in Table 4 as *gloss-centered*.

Methods	**Prank1**(%)	**Prank2**(%)
Gloss-centered	71.4	83.9
Proposed	69.4	82.5
WordNet-based Lesk-like	49.5	62.4

Table 4. Comparison of modern approaches to WSD.

We intend to compare our approach with SENSEVAL workshop approaches in the near future. At this stage, this was infeasible due to different WordNet versions (SENSEVAL uses 1.7.1 while ours is 2.0). As it is possible to observe, our method achieves better results than the WordNet-based Lesk-like method [5], and a comparable performance to the gloss-centered approach.

5. Identified Limitations & Further Work

The current work is an ongoing research and hence, we have identified two significant limitations by manual identification of incorrect predictions.

The first identified limitation is the retrieval of inconsistent (noisy) web documents. This shortcoming is a result of the query which is sent to Google. The proposed methodology generates a string query, which is essentially the sentence containing the target word. If this sentence is not large enough, then Google will return irrelevant documents that will negatively affect the performance of the system.

We propose two solutions to this problem. The first one is to send a n number of adjacent sentences to Google. That way, the particular search engine will be able to return more consistent web documents, possibly increasing the accuracy of WSD.

The second solution is to use NP chunking and enclosing in quotes. This technique will allow Google to search for the exact sequence of words enclosed in quotes. As a

[7] Local context is equal to the sentence containing the target word

result, returned web documents will be less noisy and the accuracy of WSD will possibly increase.

The second identified limitation is the noise included in topic signatures. There were cases in the evaluation, in which the retrieved web documents were relevant, but we were unable to predict the correct sense, even when we were using all the possible combinations of the number of downloaded documents and the window size around the target word within the documents.

This limitation arose from the fact that word senses were similar, but still different. TS were unable to discriminate between these senses, which means that TS of the corresponding senses have high similarity. Experiments on calculating semantic distance between word senses using TS and comparison with other distance metrics have shown that topic signatures based on mutual information (MI) and t-score perform better than tf/idf-based TS [3].

This means that our WSD process would possibly achieve a higher performance using the MI or t-score based TS. Unfortunately, this was infeasible to test at this stage, since these TS were not available to the public. Future work involves experimentation with other kinds of TS and exploration of the parameters of their construction methodology targeted at more accurate TS.

Finally, *verbs and pre-nominal modifiers* are not considered in the particular approach. We intend to extend topic signatures by developing appropriate ones for verbs and pre-nominal modifiers. Thus, their disambiguation will also be feasible.

6. Conclusions

We have presented an unsupervised methodology for automatic disambiguation of noun terms found in unrestricted corpora. Our method attempts to extend the local context of a target word by issuing queries to the web, and then measuring the overlap with topic signatures of the different senses using Bayes rule.

Our method outperformed the TS-based WSD, indicating that the extension of local context increases the amount of useful knowledge to perform WSD. Our method achieved promising results, which are comparable to the result of the best performing system participating in SENSEVAL 3 competition.

Finally, we have identified two main limitations, which we intend to overcome in the future in order to provide a more reliable WSD.

Acknowledgments

The first author is grateful to the *General Michael Arnaoutis* charitable foundation for its financial support. We are more than grateful to our colleague, George Despotou, and to Georgia Papadopoulou for proof reading this paper.

References

[1] E. Agirre, O. Ansa, E. Hovy, and D. Martinez, 'Enriching very large ontologies using the www', in *ECAI Workshop on Ontology Learning. Berlin, Germany*, (2000).

[2] E. Agirre, O. Ansa, E. Hovy, and D. Martinez, 'Enriching wordnet concepts with topic signatures', *ArXiv Computer Science e-prints*, (2001).
[3] Eneko Agirre, Enrique Alfonseca, and Oier Lopez de Lacalle, 'Approximating hierarchy-based similarity for wordnet nominal synsets using topic signatures', in *Sojka et al. [SPS+03]*, pp. 15–22, (2004).
[4] Timothy Chklovski, Rada Mihalcea, Ted Pedersen, and Amruta Purandare, 'The senseval-3 multilingual english-hindi lexical sample task', in *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, eds., Rada Mihalcea and Phil Edmonds, pp. 5–8, Barcelona, Spain, (July 2004). Association for Computational Linguistics.
[5] K. Fragos, Y. Maistros, and C. Skourlas., 'Word sense disambiguation using wordnet relations', *First Balkan Conference in Informatics, Thessaloniki*, (2003).
[6] Cristian Grozea, 'Finding optimal parameter settings for high performance word sense disambiguation', in *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, eds., Rada Mihalcea and Phil Edmonds, pp. 125–128, Barcelona, Spain, (July 2004). Association for Computational Linguistics.
[7] N. Ide and J. Veronis, 'Introduction to the special issue on word sense disambiguation: The state of the art.', *Computational Linguistics 24(1)*, 1Ű40, (1998).
[8] Ioannis P. Klapaftis and Suresh Manandhar, 'Google & WordNet based Word Sense Disambiguation', in *Proceedings of the International Conference on Machine Learning (ICML-05) Workshop on Learning and Extending Ontologies by using Machine Learning Methods, Bonn, Germany*, (August 2005).
[9] S. Lande, C. Leacock, and R. Tengi, 'Wordnet, an electronic lexical database', in *MIT Press, Cambridge MA 199-216*, (1998).
[10] Yoong Keok Lee, Hwee Tou Ng, and Tee Kiah Chia, 'Supervised word sense disambiguation with support vector machines and multiple knowledge sources', in *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, eds., Rada Mihalcea and Phil Edmonds, pp. 137–140, Barcelona, Spain, (July 2004). Association for Computational Linguistics.
[11] Michael Lesk, 'Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone', in *Proceedings of the AAAI Fall Symposium Series*, pp. 98–107, (1986).
[12] Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff, 'The senseval-3 english lexical sample task', in *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, eds., Rada Mihalcea and Phil Edmonds, pp. 25–28, Barcelona, Spain, (July 2004). Association for Computational Linguistics.
[13] G. Miller, 'Wordnet: A lexical database for english', *Communications of the ACM*, **38(11)**, 39Ű–41, (1995).
[14] Ganesh Ramakrishnan, B. Prithviraj, and Pushpak Bhattacharya, 'A gloss-centered algorithm for disambiguation', in *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, eds., Rada Mihalcea and Phil Edmonds, pp. 217–221, Barcelona, Spain, (July 2004). Association for Computational Linguistics.
[15] G. Salton and C. Buckley, 'Term weighting approaches in automatic text retrieval', *Information Processing and Management*, **24**(5), 513–523, (1988).
[16] Carlo Strapparava, Alfio Gliozzo, and Claudiu Giuliano, 'Pattern abstraction and term similarity for word sense disambiguation: Irst at senseval-3', in *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, eds., Rada Mihalcea and Phil Edmonds, pp. 229–234, Barcelona, Spain, (July 2004). Association for Computational Linguistics.
[17] Sonia Vázquez, Rafael Romero, Armando Suárez, Andrés Montoyo, Iulia Nica, and Antonia Martí, 'The university of alicante systems at senseval-3', in *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, eds., Rada Mihalcea and Phil Edmonds, pp. 243–247, Barcelona, Spain, (July 2004). Association for Computational Linguistics.
[18] Xinglong Wang and John Carroll, 'Word sense disambiguation using sense examples automatically acquired from a second language', in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 547–554, Vancouver, British Columbia, Canada, (October 2005). Association for Computational Linguistics.
[19] David Yarowsky, 'Word-sense disambiguation using statistical models of roget's categories trained on large corpora', in *Proceedings COLING-92 Nantes, France*, (1992).