Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online)
An Open Access, Online International Journal Available at www.cibtech.org/sp.ed/jls/2015/01/jls.htm
2015 Vol.5 (S1), pp. 859-864/Omidzadeh

Research Article

WORD SENSE DISAMBIGUATION USING ADABOOST AND J48 COMBINATION

***Razieh Omidzadeh**
Department of Education, Khorramabad, Lorestan, Iran
**Author for Correspondence*

ABSTRACT

Nowadays, increasing in use of digital contents causes to increase in usage of machine translation systems. Unfortunately there is a big problem in front of these systems which called word sense disambiguation. This problem comes from the words with different meanings in the source language. In this paper a novel algorithm has been proposed to overcome this problem which works based on both Adaboost and j48 algorithms. To evaluate this method, we have used a standard benchmark. The result of this evaluation confirms that this method could be useful in word sense disambiguation.

Keywords: *Word Sense Disambiguation, Adaboost, J48, Machine Translation*

INTRODUCTION

Nowadays there are a lot of data which has to be translated from a language to another language (Saad *et al.*, 2013). In this regards, researchers try to use machine for translating. In this way translation could be done so cheap and rapidly. On the Other hands using this kind of systems decrease the quality of translation. However during the times the ability of machine translations systems has been increased, but there is a problem available in front of these systems. This problem which called word sense disambiguation caused from the different meanings of a special word in a language. For example, bank in English can either mean a financial institution, or a sloping raised land (Bhattacharyya *et al.*, 2013). The task of Word Sense Disambiguation (WSD) is to assign the correct sense to such ambiguous words based on the surrounding context. This is an important problem which has many applications in natural language processing.

In this paper, WSD is cast as a problem in supervised learning because different meanings of a word are a class of the problem. In the following at first some related works will be introduced in section 2. Next in section 3 the proposed method will be explained. The details of evaluation of this method are given in section 4. At last the conclusion and some hints for future works are expressed in the section 5.

Related Works

In the last three decades, a large body of work has been presented that concerns the development of automatic methods for the enrichment of existing resources such as WordNet (Fellbaum and Christiane, 2010). These include proposals to extract semantic information from dictionaries (e.g. Chodorow *et al.*, (1985) (Leacock *et al.*, 1998) and Rigau *et al.*, (1998) (Agirre *et al.*, 1996), approaches using lexicon syntactic patterns (Hearst, 1992; Cimiano *et al.*, 2005; Cimiano *et al.*, 2004; Girju *et al.*, 2003; Girju *et al.*, 2006); Harabagiu *et al.*, 1999), heuristic methods based on lexical and semantic regularities (Harabagiu *et al.*, 1999; Pantel *et al.*, 2002) taxonomy based ontologization (Pen-nacchiotti and Pantel (Snow *et al.*, 2008, 2006). Other approaches include the extraction of semantic preferences from sense-annotated (Agirre and Martinez, 2001);

Agirre *et al.*, 2000) and raw corpora (McCarthy and Carroll, 2003; McCarthy *et al.*, 2004), as well as the disambiguation of dictionary glosses based on cyclic graph patterns (Navigli, 2009). Other works rely on the dis-ambiguation of collocations, either obtained from specialized learner’s dictionaries (Navigli and Ve-lardi, 2005; Navigli *et al.*, 2005) or extracted by means of statistical techniques (Cuadros and Rigau, 2008; Cuadros *et al.*, 2008), e.g. based on the method proposed by Agirre and de Lacalle, 2004; Agirre *et al.*, 2004). But while most of these methods represent state-of-the-art proposals for enriching lexical and taxonomic resources, none concentrates on augmenting WordNet with associative semantic relations for many domains on a very large scale.

MATERIALS AND METHODS

Methodology

In this section the proposed method was introduced. The architecture of this method is illustrated in the figure 1.

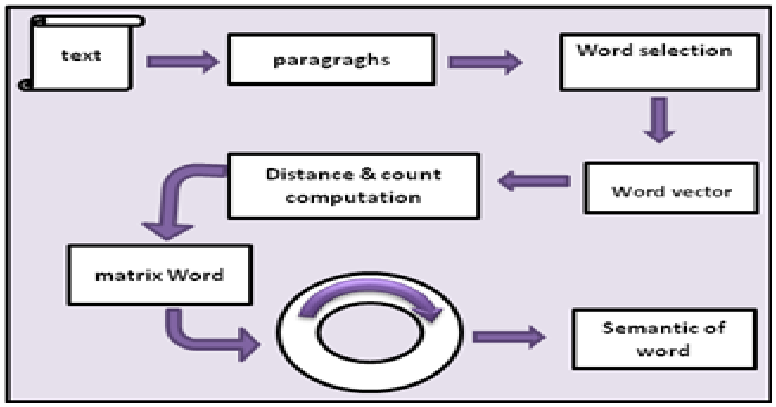

Figure 1: The architecture of proposed system

As shown in this figure at first the texts are desecrated in the separated paragraphs. This work should be done because in most of languages the more relevant context is near the given word. So to find the meaning of the word the words in the other paragraphs have negligible effects. Therefore in this paper to find the meaning of the ambiguous word just the words which in its paragraph are considered too.
As it seems obvious, not all the words in the paragraph have the value to be processed while many of the words are just used to connect the other words. Therefore to decrease the computation and increasing the performance the words should be selected. This process is shown in figure 2. The words with low priority like pronouns and prepositions are omitted using a stop list. Then using stemming techniques, words with same roots are replaced with their roots. For example, the word "running" would convert to "run". Important words should be found in the next step. To do that, TF-IDF technique was used which is based on word frequency in the text. More information about TF-IDF is available at (Li *et al.*, 2007). In abstract in TF-IDF process a word has more value when the frequency of it in that paragraph is high and its frequency in the other paragraphs is low. To decrease the effect of length of paragraph the TF-IDF values are normalized based on the paragraph length. Then the words are ready to be selected as selected words. This process could be done based on two methods; threshold or top N. In threshold method if normalized TF-IDF value is higher than a threshold, the given word is selected. On the other hand in top N method words sorted based on their TF-IDF values and the N top words are selected. Because the count of selected words in threshold method could be different, in this paper top N method has been used.

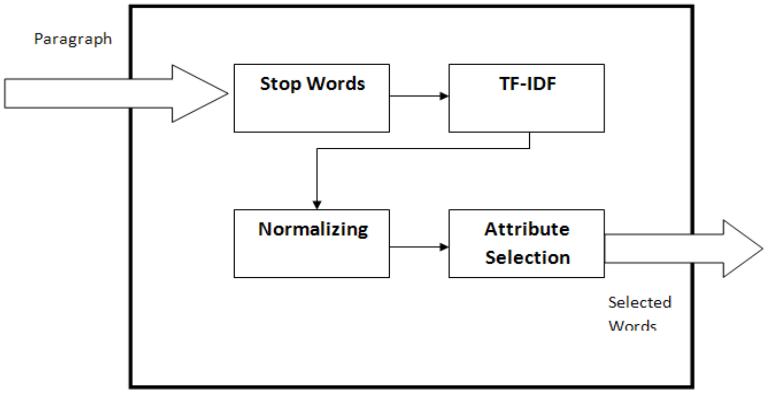

Figure 2: Selecting the words

We considered in this paper that if the selected word is more close to ambiguity word it has more effect to find the meaning. Therefore the distance of selected words with the word with ambulation is extracted. Then computing the distances and the count of selected words a vector is created for each paragraph which for per selected word there is to column in it, one for its count and the other for its distance.

Now we have an ordinary dataset with some columns for features and one column for its label or class which here is the meaning of the word. In figure 3 a sample of this dataset is illustrated. Now we could use an appropriate supervised machine learning algorithms on it. In this paper Adaboost algorithm is used which using j48 algorithm in its adaboost core. Adaboost is short for Adaptive Boosting, which is a machine learning algorithm, formulated by Yoav Freund and Robert Schapire (Freund *et al.*, 1995). It could be say that this algorithm is a meta-algorithm, and can be used in conjunction with many other learning algorithms to improve their performance. AdaBoost is adaptive in the sense that subsequent classifiers built are tweaked in favor of those instances misclassified by previous classifiers. For more information about j48 and adaboost see (Blackmore *et al.*, 2002) and (Rätsch *et al.*, 2001).

In the next section the details of implementation and evaluation of this method are explained.

Evaluation

In this section at first the benchmark that used in the evaluation is introduced then the details of implementation and results are explained.

Benchmark

In this paper to evaluate proposed method a standard benchmark was used. This benchmark that named TWA could be downloaded from: http://lit.csci.unt.edu/~rada/downloads/TWA/TWA.tar.gz. Table 1 shows some basic information of this bench mark.

Table 1: Basic information of the TWA benchmark

Language of the Benchmark	**English**
The number of words with ambiguity	6
The unit of each benchmark	Paragraph
The limitation of words' number in the paragraph	NO
Number of distinct words with ambiguity in the paragraph	1
Ambiguate word could be repeated in the paragraph	Yes
Basic structure of benchmark	Nonstandard XML
Different meanings of each ambiguate word	2
Different meanings of ambiguate word in the paragraph	1

In table 2 the statistical information of this benchmark is explained.

Table 2: statistical information of TWA benchmark

Ambiguity Word	**First meaning**	**Second meaning**	**Number of paragraphs**	**Number of paragraphs with first meaning**	**Number of paragraphs with first meaning**
Bass	Fish	Music	107	10	97
Crane	Bird	machine	95	23	72
Motion	Legal	movement	201	59	142
Palm	Tree	hand	201	58	143
Plant	Living	Factory	188	86	102
Tank	container	vehicle	201	126	75

Implementation

To implement the proposed method Weka which is an open source machine learning tool was used. Also for counting the distance and frequencies of selected words an application eas developed with .net framework.

Results

The evaluation results of proposed method based on different measurements are illustrated in table 3.

Table 3: The evaluation results for using Adaboost with j48 algorithm

Word	Correct classified	Precision	Recall	TP	F-Measure
Bass	100	1	1	1	1
Crane	93.7	0.94	0.93	0.94	0.93
Motion	97	0.97	0.97	0.97	0.97
Palm	98	0.97	0.98	0.97	0.98
Plant	81.4	0.86	0.81	0.81	0.82
Tank	97	0.97	0.98	0.97	0.97

As the results show, using the Adaboost with j48 algorithm cause to very high performance for WSD. To compare this result with results of other methods in table 4 results of different methods are gathered. For better illustration in this table just correct classification is considered.

Table 4: Compression of different methods based on correct classification

Method	Tank	Plant	Palm	Motion	Crane	Bass	Ave
Adaboost with j48	97	81.4	98	97	93.7	100	94.52
SVM	83.7	62.8	91.4	90.3	79.1	94.6	83.65
Naïve Bayes	74.1	71.8	89.9	83.6	81	82.6	80.50
IBL 1	63.4	62.8	66.6	76.6	54.7	93.5	69.60
IBL2	70.1	63.3	68.2	71.1	40	94.4	67.85
IBL3	69.7	59.6	64.1	73.13	41.2d	91.6	71.63
Rep Tree	72.1	69.7	71.1	80.6	76.8	95.3	77.60
CART	69.6	76.6	71.7	82.1	75.8	95.3	78.52
J 48	72.6	68.6	73.6	80.1	73.7	96.3	77.48
ID3	72.1	75.5	76.6	82.3	74.7	95.3	79.42
RBF	64.1	51	71.1	70.15	75.8	96.7	71.48
AVG	73.50	67.55	76.57	80.63	72.53	94.15	94.52

The results confirm that the proposed method could increase the performance of WSD. It should be mentioned that these 10 methods were developed with Weka.

Conclusion And Future Works

In this paper a novel method was proposed for WSD problem in machine translation. This method after extracting the selected words from the paragraph using TF-IDF and preprocesses, makes a vector from each paragraph based on the counts of these words and distances of them with ambiguate words. Then using combination of adaboost and j48 a model which constructed that could determine the meaning of disambiguate word from the paragraph. To evaluate this method, a TWA benchmark was used. The results show that proposed method has better performance than 10 other methods.

To continue this research, it is possible to apply this method on the multi languages benchmark or using POS tagging to increase the performance of proposed method. Also combination of this method with dictionary based method could increase the ability of proposed method in WSD on different repositories.

ACKNOWLEDGEMENT

I am grateful of department of lorestan education, for their useful collaboration.

REFERENCES

Saad Farag and Andreas Nürnberger (2013). Translation Ambiguity Resolution Using Interactive Contextual Information. *Computational Linguistics* (Springer Berlin Heidelberg) 219-240.

Bhattacharyya Pushpak and Mitesh Khapra (2013). Word Sense Disambiguation. *Emerging Applications of Natural Language Processing: Concepts and New Research* 22.

Fellbaum Christiane (2010). "WordNet: An electronic lexical database (1998). WordNet, Available: http://www.cogsci.princeton.edu/wn.

Leacock Claudia and Martin Chodorow (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database* **49**(2) 265-283.

Agirre Eneko and German Rigau (1996). Word sense disambiguation using conceptual density. *Proceedings of the 16th Conference on Computational Linguistics* **1**, Association for Computational Linguistics.

Schütze Hinrich and Jan O Pedersen (1995). Information retrieval based on word senses.

Cimiano Philipp and Johanna Völker (2005). Text2Onto. *Natural Language Processing and Information Systems* (Springer Berlin Heidelberg) 227-238.

Girju Roxana, Adriana Badulescu and Dan Moldovan (2003). Learning semantic constraints for the automatic discovery of part-whole relations. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Association for Computational Linguistics **1**.

Harabagiu Sanda, George Miller and Dan Moldovan (1999). Wordnet 2-a morphologically and semantically enhanced resource. *Proceedings of SIGLEX* 99.

Pantel Patrick and Dekang Lin (2002). Discovering word senses from text. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM.

Snow Rion *et al.*, (2008). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. *Proceedings of the conference on empirical methods in natural language processing*, Association for Computational Linguistics.

Agirre Eneko and David Martinez (2000). Exploring automatic word sense disambiguation with decision lists and the Web. *Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content,* Association for Computational Linguistics.

McCarthy Diana *et al.*, (2004). Finding predominant word senses in untagged text. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics.*

Navigli Roberto (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* **41**(2) 10.

Navigli Roberto and Paola Velardi (2005). Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. Pattern Analysis and Machine Intelligence, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(7) 1075-1086.

Cuadros Montse and German Rigau (2008). Knownet: Building a large net of knowledge from the web. *Proceedings of the 22nd International Conference on Computational Linguistics*, Association for Computational Linguistics **1**.

Agirre Eneko and Oier Lopez De Lacalle (2004). Publicly Available Topic Signatures for all WordNet Nominal Senses. LREC.

Li Juanzi and Kuo Zhang (2007). Keyword extraction based on tf/idf for Chinese news document. *Wuhan University Journal of Natural Sciences* **12**(5) 917-921.

Freund Yoav and Robert E Schapire (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. *Computational Learning Theory* (Springer Berlin Heidelberg).

Blackmore K and Bossomaier TRJ (2002). Comparison of See5 and J48. PART Algorithms for Missing Persons Profiling. *First International Conference On Information Technology & Applications (ICITA 2002)*.

Rätsch Gunnar, Takashi Onoda and Müller KR (2001). Soft margins for AdaBoost. *Machine Learning* **42**(3) 287-320.