A Mapping Between Classifiers and Training Conditions for WSD

Aarón Pancardo-Rodríguez[1], Manuel Montes-y-Gómez[1,2], Luis Villaseñor-Pineda[1], and Paolo Rosso [2]

[1] National Institute of Astrophysics, Optics and Electronics, Mexico
{aaron_cyberman, mmontesg, villasen}@inaoep.mx
[2] Polytechnic University of Valencia, Spain
{mmontes, prosso}@dsic.upv.es

Abstract. This paper studies performance of various classifiers for Word Sense Disambiguation considering different training conditions. Our preliminary results indicate that the number and distribution of training examples has a great impact on the resulting precision. The Naïve Bayes method emerged as the most adequate classifier for disambiguating words having few examples.

1 Introduction

The objetive of Word Sense Disambiguation (WSD) is to distinguish between the different senses of a word, that is, to identify the correct sense of a word in a context. The state of the art of WSD [1] shows that the supervised paradigm is the most efficient. Under this approach, the disambiguation process is carried out using information that is estimated from data. Several statistical and machine learning techniques have been applied to learn classifiers from disambiguated corpora. For instance, statistical classifiers, decision trees, decision lists, memory-based learners, and kernel methods such as Support Vector Machines (SVM).

The comparison among the different approaches to WSD is difficult. The last edition of the Senseval competition showed that the SVM is emerging as one of the most powerful supervised techniques for WSD [3]. Although important, this comparison focuses on the entire systems as black boxes, and does not consider the details about the individual classifiers and the fine tunning of their paramethers.

Some researchers have attempted to compare the performace of classifiers under equal training conditions. For instance, Paliouras et al [2] disambiguated all content words from Semcor using various classifiers (e.g., J48, Naïve Bayes, PART, k-nn and a decision table). Their results indicated that the decision tree induction outperforms other algorithms. Zavrel et al [4] investigated the performance of some classifiers (neuronal networks, memory-based leraning, rule induction, decision trees, maximum entropy, winnow perceptrons, Naïve-Bayes, and SVM) and some ensembles on a diverse set of natural language processing tasks. Their results showed that the SVM algorithm is the most prommising for WSD.

In the study of the global execution of some classifiers, we focus our attention on providing information about the behaviour of the classifiers under different training

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 246 – 249, 2005.

conditions. Basically, in this paper we analyze the influence of the number of training examples and context words over the output precision for each classifier.

2 Analysis of the Semantically Tagged Corpora

The supervised methods for WSD requiere a semantically tagged corpus in order to learn the disambiguation rules. Traditionally, the Semcor[1] corpus has been used for this purpose. It is a subset of the English Brown corpus containing almost 700,000 running words tagged by POS, and more than 200,000 content words lemmatized and sense-tagged according to Wordnet.

The Senseval[2] corpora are other common resources for WSD. The Senseval-3 English all words corpus consists of approximately 5,000 words of running text from two Wall Street Journal articles and one excerpt from the Brown corpus. It contains a total of 2,212 words tagged with the Wordnet senses.

Table 1 shows some statistics from the Semcor 2.0 and the Senseval-3 English all words joint corpora. The statistcs indicate that: (i) the available training corpora are very small, smaller than supposed. Just 21% of the nouns of the corpora are polysemic; (ii) the corpora are very unbalanced. The majority of the examples correspond to the first sense of each noun. The rest of the sense has on average less than five examples.

Table 1. Some statistics from Semcor plus Senseval-3 English all words

Sense	n-secmic Nouns	Average number of examples
1	9082	13.51
2	1368	4.61
3	544	3.68
4	228	3.55
5	117	3.24
6	59	2.74
7	43	3.52
8	22	3.13
9	8	3.17
10	4	2.33
>10	11	1.75

3 Experimental Results

3.1 Experimental Setup

Learning Methods. Naïve Bayes, decision tables, LWL –locally weighted learning–, SVM –support vector machines–,and KNN.

[1] http://www.cs.unt.edu/~rada/downloads.html#semcor
[2] http://www.senseval.org/

Test Set. 10 nouns from the Semcor corpus (refer to table 2). The selection of these nouns was based on two criteria: (i) different number of average examples per sense, and (ii) a more or less balanced distribution of the examples.

Evaluation. It was based on the precision measure (i.e., the porcentage of correctly classified word senses), and on the technique of ten-cross fold validation.

Table 2. Statistics from the test set

Noun	Senses	Examples	Average examples per sense	Distribution of the examples per sense
adult	2	10	5.0	[5 5]
Link	2	10	5.0	[5 5]
formation	5	18	3.6	[4 3 4 4 3]
Dirt	2	20	10.0	[10 10]
stone	3	25	8.3	[8 8 9]
Hope	4	46	11.5	[16 15 14 1]
discussion	2	49	24.5	[27 22]
activity	3	92	30.7	[43 36 13]
plant	2	99	49.5	[63 36]
experience	3	125	41.7	[51 47 27]
state	4	200	50.0	[26 116 21 37]
thing	10	271	27.1	[52 40 32 27 24 20 28 27 17 4]

3.2 Results

Each classifier was tested over the set of selected nouns, and trained using context windows of different sizes (of 4, 6, and 8 words around the noun). Table 3 shows the obtained results. These results demostrate that, even when the classifiers had a similar average precision, their behaviour is altered depending on the training conditions.

The results indicate the following: (i) The size of the context window – number of neighboring words used on the training process – has minor effects on the output average precision; (ii) It seems that for the nouns having few examples most classifiers worked better considering more contextual information; (iii) The Naïve Bayes classifier emerged as the most adequate method for disambiaguting the nouns having few training examples per sense.

In addition, we observed that the majority of the used classifiers had a poor performace when dealing with high polysemic nouns.

4 Conclusions

In this paper we analyzed the coverage and example distribution of the Semcor and Senseval-3 English all word joint corpora. Our results are worrying: the available training corpora is smaller that supossed and unbalanced. This condition greatly affects the performance of most classifiers.

The majority of the supervised methods requiered several examples in order to construct an “accurate” classifier for WSD. According to our results, the Naïve Bayes

algorithm outperforms the others on the disambiguation of nouns having few examples. We consider that this is because it compensates the lack of training examples using more contextual information.

Currently we are studying the performance of the classifiers disambiguating a selection of verbs and adjectives from the Semcor corpus. We believe that this kind of analysis will facilitate the selection of the more appropriate classifier for disambiguating a word depending on its characteristics, which probably would have important repercussions on the construction of hybrid systems for WSD.

Table 3. Peformance of different classifiers on WSD

Classifier	N. Bayes			D.T.			LWL			SVM			KNN; K=1		
W. size	2	4	6	2	4	6	2	4	6	2	4	6	2	4	6
Adult	.40	**.60**	.50	.40	.40	.10	.50	.40	.40	.30	.40	**.60**	.40	.50	.50
Link	.60	**.80**	**.80**	.60	.30	.30	.40	.30	.50	.20	.50	.50	.30	.70	.60
Formation	.38	.61	**.66**	.11	.05	.05	.38	.16	.22	.33	.16	.16	.28	.28	.28
Dirt	**.80**	.70	.60	.70	.70	.70	**.80**	.75	.75	.65	.75	**.80**	.65	.60	.55
Stone	.60	**.64**	**.64**	.36	.40	.40	.44	.44	.48	.48	.44	.48	.48	.48	.48
hope	.37	.39	.37	.37	.34	.32	.34	.32	.32	**.54**	.45	.39	.33	.30	.22
discussion	.59	.63	**.69**	.53	.51	.61	.49	.49	.53	.55	.57	.57	.57	.63	.55
activity	.57	.59	.51	.50	.45	.46	.56	.48	.47	.59	.56	**.60**	**.60**	.61	.48
plant	**.68**	.59	.56	.60	.61	.60	.63	.64	.66	.57	.62	.59	.63	.64	.66
experiencie	**.52**	.45	.46	.44	.43	.42	.42	.42	.43	.50	.48	.51	.46	.47	.49
state	.65	.66	.64	.68	.65	.65	.68	.68	.68	**.69**	.66	.65	.66	.65	.62
thing	**.29**	.23	.24	.21	.21	.21	.26	.25	.24	.25	.22	.22	.28	.23	.24
Average precision	**.50**	.48	.47	.45	.43	.43	.47	.45	.46	.48	.47	.47	.48	.47	.45

Acknowledgements. We would like to thank CONACyT (43990A-1), R2D2 CICYT (TIC2003-07158-C04-03) and ICT EU-India (ALA/95/23/2003/077-054), as well as the *Secretaría de Estado de Educación y Universidades de España* for partially supporting this work.

References

1. Mihalcea, R., Edmonds, P. (Eds.): Proc. of Senseval-3: The 3rd Int. Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Barcelona, Spain. (2004)
2. Paliouras, G., Karkaletsis, V., Androutsopoulos, I., Spyropoulos, C. D.: Learning Rules for Large-Vocabulary Word Sense Disambiguation: a comparison of various classifiers. Proc. of the 2nd International Conference on Natural Language Processing, Patra, Greece (2000).
3. Snyder, B., Palmer, M.: The English All-Words Task. SENSEVAL-3: Third International Workshop on the evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain (2004).
4. Zavrel, J., Degroeve, S., Kool, A., Daelemans, W., Jokinen, K.: Diverse Classifiers for NLP Disambiguation Tasks: Comparison, Optimization, Combination, and Evolution. Proceedings of the 2nd CEvoLE Workshop "Learning to Behave" (2000).