Verb Sense Disambiguation Using Support Vector Machines: Impact of WordNet-Extracted Features

Davide Buscaldi, Paolo Rosso, Ferran Pla,
Encarna Segarra, and Emilio Sanchis Arnal

Dpto. Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia, Valencia, Spain
{dbuscaldi, prosso, fpla, esegarra, esanchis}@dsic.upv.es

Abstract. The disambiguation of verbs is usually considered to be more difficult with respect to other part-of-speech categories. This is due both to the high polysemy of verbs compared with the other categories, and to the lack of lexical resources providing relations between verbs and nouns. One of such resources is WordNet, which provides plenty of information and relationships for nouns, whereas it is less comprehensive with respect to verbs. In this paper we focus on the disambiguation of verbs by means of Support Vector Machines and the use of WordNet-extracted features, based on the hyperonyms of context nouns.

1 Introduction

Word Sense Disambiguation (WSD) is an open problem in the field of Natural Language Processing (NLP). The resolution of lexical ambiguity that appears when a given word in a context has several different meanings is commonly referred as Word Sense Disambiguation. Supervised approaches to WSD usually perform better than unsupervised ones [4]. Results of the recent Senseval-3[1] contest attest this supremacy; moreover, recent results of the application of Support Vector Machines (SVM), a well-known supervised learning technique, to the Word Sense Dismbiguation task seem promising [3].

Some interesting results have been obtained recently in the supervised disambiguation of verbs [1], by using context-extracted features and a multi-class learning architecture. The disambiguation method described in this paper replicates the feature extraction model proposed in [1], with the addition of WordNet [5] extracted features, while using a SVM-based learning architecture. The system was tested over a subset of the Senseval-3 Lexical Sample corpus.

2 Support Vector Machines

The SVM [6] performs optimization to find a hyperplane with the largest margin that separates training examples into two classes. A test example is classified

[1] http://www.senseval.org

A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 192–195, 2006.

depending on the side of the hyperplane it lies in. Input features can be mapped into high dimensional space before performing the optimization and classification. A kernel function can be used to reduce the computational cost of training and testing in high dimensional space. If the training examples are nonseparable, a regularization parameter C ($= 1$ by default) can be used to control the trade-off between achieving a large margin and a low training error. We used the implementation of SVM from Thorsten Joachims [2], *SVM-light*. In order to apply the SVM to the WSD task, each nominal feature with possible values was converted into binary (0 or 1) features. If a nominal feature took the i-th value, then the i-th binary feature was set to 1 and all the other binary features were set to 0. The default linear kernel was used. Since SVMs handle only binary (2-class) classification, we built one binary classifier for each sense class.

3 Disambiguation Model

Given a verb in its sentential context `<verb, sentence>`, the goal is to develop procedures for the automatic labeling of the semantic class it encodes. An important first step is to map the context information of each verb into feature vectors. The following features were selected among the ones proposed in [1]:

– *Word feature*: is the lexical form of each word in a window of size six (three before and three after) surrounding the target verb.
– *Part-of-Speech tag (POS) feature*: is the POS tag of each word in a window of size six surrounding the target verb.
– *Word.POS tag feature*: is the conjunction of each word and its POS tag for each word within a window of size six of the target verb.

Moreover, a feature was added for each noun found in the verb's context: the *l-hyperonym feature*: the hyperonyms extracted from WordNet at depth l, where l indicates the maximum number of levels to be considered, upwards in the WordNet hierarchy.

For instance, let us consider the following POS-tagged sentence: "*Reid/NNP saw/ VBD me/PRP looking/VBG at/IN the/DT iron/NN bars/NNS ./.*", where *looking* is the verb to be disambiguated. Therefore, the following feature vectors are associated to this instance of the verb: (*Reid, saw, me, at, the, iron*) as word feature vector, (*NNP, V BD, PRP, IN,DT,NN*) as the POS feature vector, (*Reid.NNP, saw.VBD, me.PRP, at.IN, the.DT, iron.NN*) as the Word.POS feature vector. Finally, WordNet is used in order to collect the l-hyperonyms of *iron*, the only noun in the context.

Iron has 5 senses in WordNet; for readability reasons, we limit to the first two: *iron, Fe, atomic number 26*: (a heavy ductile magnetic metallic element;), and *iron*: (a golf club). The hyperonyms trees obtained from WordNet for these senses are:

Sense 1: iron, Fe, atomic number 26
⇒ metallic element, metal
⇒ chemical element, element
⇒ substance, matter
⇒ entity

Sense 2: iron
⇒ golf club, golf-club, club
⇒ golf equipment
⇒ sports equipment
⇒ equipment
⇒ instrumentation
⇒ . . .
⇒ *entity*

When 1-hyperonyms are used as features, only the first hyperonym is added to the feature vector; in this case, then only the offsets (numeric IDs that identify in an unique way the WordNet synsets) corresponding to (*metallic element, metal*) and (*golf club, golf-club, club*) are added to the features. Otherwise, if, for instance, 5-hyperonyms are used, therefore all the hyperonyms of sense 1 and the hyperonyms up to *instrumentality, instrumentation* for sense 2 are added to the feature vector. The hyperonyms are extracted from all the senses of the noun, without any assumption on the right sense of that noun in its context.

4 Experiments

The experiments have been carried out over the verbs in the Senseval-3 Lexical Sample corpus; for each verb a training set of xml-tagged sentences is provided together with a smaller test set of sentences in the same format. The averaged number of training samples for each verb is 123.53, roughly doubling the averaged number of test samples (61.81). Eight SVM models were trained for every sense of each verb, one without considering the WordNet extracted features, and the remaining seven with l-hyperonyms features, with $l \in \{1, 2, 3, 4, 5, 6, 7\}$. Therefore, for each sense $s_i(v)$ of verb v, a SVM was trained to classify verb instances of sense $s_i(v)$ against the others.

In the testing phase, the $|s(v)|$ SVMs, where $|s(v)|$ is the number of senses of verb v, are used to classify the verb instance. Although *SVM_light* is not a

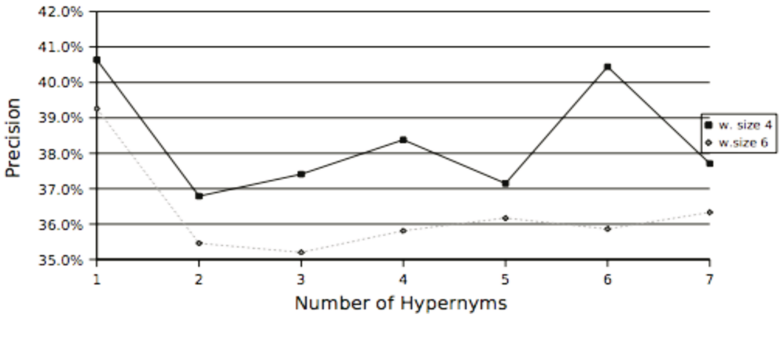

Fig. 1. Results obtained at the different hypernymy levels, considering a window of size 4 and 6

probabilistic SVM, we used its output value w as a confidence weight in the range $[-1, 1]$,$w \in R$, where 1 means 100% confidence in v having the sense $s_i(v)$, and -1 means 100% confidence in v *not* having sense $s_i(v)$. The sense assigned to the verb corresponds to the sense related to the SVM which returns the highest w.

We carried out a study of the window size, obtaining an optimal size of 4 words (2 before and 2 after the verb to disambiguate). The calculated baseline precision, obtained by assigning the most frequent sense to each test instance, is 47.5%. A comparison with Senseval-3 is not feasible, given that the results are not calculated separately for each Part-Of-Speech. The results obtained at the different hypernymy levels are displayed in Fig. 1.

In our experiments the use of all context hypernyms did not allow to improve those obtained by using the base SVM configuration. Nevertheless, we suppose that better results can be achieved if only the hypernyms of the right sense of the context nouns are considered. In fact, we obtained 49.7% in precision by taking into account only the hypernyms from the 3 most common senses of context nouns.

5 Conclusions and Further Work

Our experiments show that although the use of SVMs allowed to obtain better results than the baseline, WordNet-extracted features did not prove so useful. Precision dropped dramatically when using only one hypernym. Size of training samples may be also too small to draw definitive conclusions. Further work may include the possibility to take into account also the distance of context features from the verb (as proposed in [3]) and use a weighting proportional to the depth of the hyperonyms in the hierarchy.

Acknowledgments

We would like to thank R2D2 CICYT (TIC2003-07158-C04-03) and ICT EU-India (ALA/95/23/2003/077-054) research projects for partially supporting this work.

References

1. Girju, R., Roth, D., Sammons, M.: Token-level Disambiguation of VerbNet classes. Proc. of the Interdisciplinary Workshop on Verb Features and Verb Classes, Saarbruckem, Germany (2005)
2. Joachims, T.: Making large-scale SVM Learning Practical. Advances in Kernel Methods. MIT-press, 1999.
3. Lee, Y.K., Ng, H.T: Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources. Proc. of the SENSEVAL-3 workshop. Barcelona, Spain, 2004
4. Mihalcea, R., Moldovan, D.I.: A Method for Word Sense Disambiguation of Unrestricted Text. Proc. of the ACL-99 Conference. Maryland, NY, U.S.A., 1999
5. Miller, G.: WordNet: a lexical database for english. CACM, 38(11):39-41, 1995.
6. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (1995)