Information Retrieval and Text Categorization with Semantic Indexing*

Paolo Rosso, Antonio Molina, Ferran Pla,
Daniel Jimenez, and Vicent Vidal

Dpto. de Sist. Informáticos y Computación, U. Politécnica de Valencia, Spain
{prosso,amolina,fpla,djimenez,vvidal}@dsic.upv.es

Abstract. In this paper, we present the effect of the semantic indexing using *WordNet* senses on the Information Retrieval (IR) and Text Categorization (TC) tasks. The documents have been sense-tagged using a Word Sense Disambiguation (WSD) system based on Specialized Hidden Markov Models (SHMMs). The preliminary results showed that a small improvement of the performance was obtained only in the TC task.

1 WSD with Specialized HMMs

We consider WSD to be a tagging problem. The tagging process can be formulated as a maximization problem using the Hidden Markov Models (HMMs) formalism. Let $\mathcal{S}$ be the set of sense tags considered, and $\mathcal{W}$, the vocabulary of the application. Given an input sentence, $W = w_1, \ldots, w_T$, where $w_i \in \mathcal{W}$, the tagging process consists of finding the sequence of senses ($S = s_1, \ldots, s_T$, where $s_i \in \mathcal{S}$) of maximum probability on the model, that is:

$$\begin{aligned} \widehat{S} &= \arg\max_S P(S|W) \\ &= \arg\max_S \left(\frac{P(S) \cdot P(W|S)}{P(W)} \right) ; \, S \in \mathcal{S}^T \end{aligned} \tag{1}$$

Due to the fact that the probability $P(W)$ is a constant that can be ignored in the maximization process, the problem is reduced to maximizing the numerator of equation 1. To solve this equation, the Markov assumptions should be made in order to simplify the problem. For a first-order HMM, the problem is reduced to solving the following equation:

$$\arg\max_S \left(\prod_{i:1\ldots T} P(s_i|s_{i-1}) \cdot P(w_i|s_i) \right) \tag{2}$$

The parameters of equation 2 can be represented as a first-order HMM where each state corresponds to a sense s_i, $P(s_i|s_{i-1})$ representing the transition probabilities between states and $P(w_i|s_i)$ representing the probability of emission

* This work was supported by the Spanish Research Projects CICYT TIC2000-0664-C02 and TIC2003-07158-C04-03. We are grateful to E. Ferretti for sense-tagging the data.

of words, w_i, in every state, s_i. The parameters of this model are estimated by maximum likelihood from semantic annotated corpora using an appropriate smoothing method (e.g. linear interpolation).

The HMM approach above presented cannot include different kinds of available linguistic information which can be useful in solving WSD. In particular, the *SemCor* corpus was used to learn the models. It provided the following input features: words ($\mathcal{W}$), lemmas ($\mathcal{L}$) and the corresponding POS tags ($\mathcal{P}$). Therefore, in the formulation presented above, the input vocabulary ($\mathcal{W}$) can be redefined as $\mathcal{I} = \mathcal{W} \times \mathcal{L} \times \mathcal{P}$. Then, an input sentence will be a sequence of tuples of words, lemmas and POS.

In order to incorporate this kind of information to the model we used Specialized HMMs [4]. Basically, a SHMM consists of changing the topology of the HMM in order to get a more accurate model which includes more information. This is done by means of an initial step previous to the learning process. It consists of the redefinition of the input vocabulary and the output tags. Therefore, no changes are needed in the usual HMM learning task. This redefinition is done by means of two processes: the *selection* process, which is applied to the input vocabulary, and the *specialization* process, which redefines the output tags.

The aim of the *selection* process is to choose which input features are relevant to the task. This process applies a determined *selection criterion* to $\mathcal{I}$ that produces a new input vocabulary ($\hat{\mathcal{I}}$). This new vocabulary consists of the concatenation of the relevant features selected. The *selection criteria* used in this work is as follows: if a word has a sense in *WordNet* we concatenate the lemma and the POS (Part-Of-Speech) associated to the word as input vocabulary. For non-content words (i.e., words without meaning), we only consider their lemma as input.

The *specialization process* allows for the codification of certain information into the context (i.e., into the states of the model). It consists of redefining the output tag set by adding information from the input. This redefinition produces some changes in the model topology, in order to allow the model to better capture some contextual restrictions and to get a more accurate model. The application of a *specialization criterion* to $\mathcal{S}$ produces a new output tag set ($\hat{\mathcal{S}}$), whose elements are the result of the concatenation of some relevant input information to the original output tags.

In the WSD system used here, we defined the output semantic tag set by considering certain statistical information which was extracted from the annotated corpora. In the *SemCor* corpus, each annotated word is tagged with a *sense_key* which has the form *lemma%lex_sense*. In general, we considered the *lex_sense* field of the *sense_key* associated to each lemma as the semantic tag in order to reduce the size of the output tag set. This does not lead to any loss of information because we can obtain the *sense_key* by concatenating the lemma to the output tag. For certain frequent lemmas, we can specialize their output tags to produce a more fine-grained semantic tag (which is equivalent to the *sense_key*). These choices were made experimentally by taking into account a set of frequent lemmas, which were extracted from the Semcor corpus.

The evaluation of the WSD system was previously carried out on the *Semcor* corpus (73.3% of precision) and on the English all-word task of the *Senseval-2* competition (60.2% of precision) [4].

2 Semantic IR and TC: Experimental Results

The classical vector space model for IR was shown by Gonzalo [1] to give better results if WordNet synsets are chosen as the indexing space instead of terms (up to 29% improvement in the experimental results was obtained for a manually disambiguated test collection derived from the SemCor corpus). Therefore, in our research work, we decided to represent each document through a vector of relevant synsets instead of a vector of relevant terms. The disambiguation of the meaning of each term was obtained using SHMMs.

When searching for a document, it could be often useful to previously group, or cluster, the documents of the collection. Therefore, the IR task was initially carried out employing the Bisecting-Spherical K-Means clustering technique. Its algorithm tries to join the advantages of the bisecting K-Means algorithm with the advantages of a modified version of the Spherical K-Means [3]. The corpus used for the experiments contains articles from the *1963 Times Magazine* [1]. Query statistics were also obtained for the query collection, formed by a total of 83 queries with an average of 15 words and one line per query. The same experiments were also carried out using the Singular Value Decomposition (SVD) technique of the Latent Semantic Indexing (LSI) model, in order to understand better the influence of semantics in the IR task. For both clustering and LSI models a worse precision was obtained when semantics was taken into account: 42.41% (sense indexing) vs. 63.58% (term indexing) for the clustering technique, and 51.72% (sense indexing) vs. 67.95% (term indexing) for the SVD technique [2]. This could be due to the length of the queries because such long queries implicitly have a disambiguation effect. At the moment of writing this paper, some experiments have been carrying out using the *TREC document collection* [2] in which queries are shorter on average.

The K Nearest Neighbours (K-NN) is the technique we used in the TC task. The TC was performed using the K-NN method provided by the Rainbow system [3], with the value for the parameter K which was established as 30. Different experiments were carried out on the *20 Newsgroups* corpus [4] for the semantic TC task. This corpus contains about 20,000 news messages from 20 UseNet discussion groups (i.e., categories) that were sent in 1993. The task consisted of predicting which group each test document was sent to. The training set was composed of 16,000 documents (the first 800 ones of each category), whereas the

[1] Available at ftp.cs.cornell.edu/pub/smart/time/

[2] Text Retrieval Conference document collection; at www.trec.nist.gov

[3] The Bow: A toolkit for Statistical Language Modelling, Text Retrieval, Classification and Clustering is available at `www.cs.cmu.edu/~mccallum/bow/`

[4] Available at www.ai.mit.edu/~jrennie

other 3,997 documents were used as test set. The introduction of semantics allowed for a small improvement of the precision: 79% (sense indexing) vs. 77.68% (term indexing). Figure 1 shows the comparison of the error percentage obtained with and without the introduction of the semantics with respect to the size of the vocabulary.

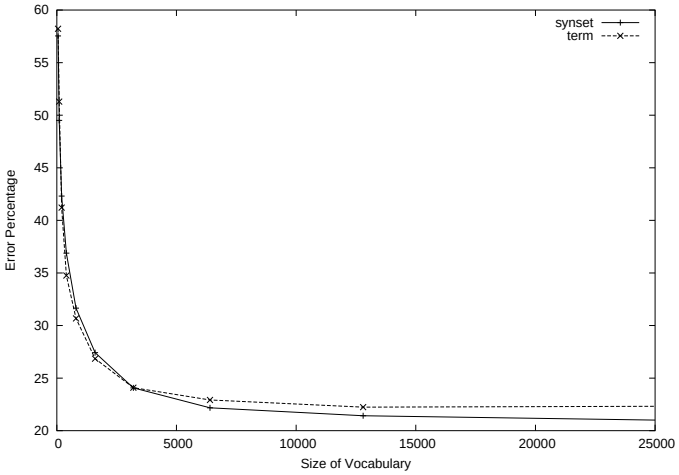

Fig. 1. Text Categorization: term vs. sense indexing

As further work, the two vector representations of each document should be combined, in order to take into account with different weights, terms and WordNet synsets at the same time.

References

1. J. Gonzalo, F. Verdejo, I. Chugur, J. Chigarrán. Indexing with WordNet Synsets can improve Text Retrieval. In: Proc. of the Workshop on Usage of WordNet for NLP, 1998.
2. D. Jiménez, E. Ferretti, V. Vidal, P. Rosso, C.F. Enguix. The Influence of Semantics in IR using LSI and K-Means Clustering Techniques. In: Proc. of the Workshop on Conceptual Information Retrieval and Clustering of Documents, ACM Int. Conf. on Information and Communication Technologies, 2003.
3. D. Jiménez, V. Vidal, C.F. Enguix. A Comparison of Experiments with the Bisecting-Spherical K-Means Clustering and SVD Algorithms. In: Proc. of JOTRI, 2002.
4. A. Molina, F. Pla, E. Segarra. A Hidden Markov Model Approach to Word Sense Disambiguation. In: Proc. of VIII Conf. Iberoamericana de Inteligencia Artificial (IBERAMIA2), Sevilla, Spain, 2002.