Information Retrieval and Text Categorization with Semantic Indexing[⋆]

Paolo Rosso, Antonio Molina, Ferran Pla,
Daniel Jiménez, 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:

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

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:

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

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.