Regular Patterns - Probably Approximately Correct Language Model

Octavian Popescu
Fondazione Bruno Kessler (FBK-irst), Italy

Almost any word in natural language has a great potential of expressing different meanings. However, in certain contexts, this potential is limited up to the point that one and only one sense is possible. When this happens, we are not dealing with an individual phenomenon, but, rather, all the words in that context have their own meaning potential limited.

In this talk, we properly define such meaning restricting contexts, analyze their properties and propose an automatic procedure for their identification in large corpora. We show that these contexts are patternable and that the words are completely disambiguated. We therefore call such contexts sense discriminative patterns (SDP). By comparing minimally different SDPs, we discover a set of lexical semantic features that are used in devising a learning algorithm.

The form of patterns is regular, they are generated by a finite state automaton. Inducing the form of the grammar from annotated examples and finding the right generalization level is done using Angluin Algorithm. The patterns contain the syntactic and lexical information which is relevant for sense disambiguation, so they are SDPs. The patterns are minimally self-sufficient, thus the senses of the words matched by a pattern are in mutual disambiguation relationship. The disambiguation process of the meanings of all slots is sequential, identifying the meaning of one slots leads to the identification of the meaning of all slots. We call this relationship between the senses of the words which are caught in a pattern, chain clarifying relationship, CCR.

The main problem that needs to be addressed is the fact that pattern acquisition is very sensitive to errors. On the basis of the PAC-learning technique, we have developed a technique that produces an approximately correct grammar, having a high probability to be correct in spite of the noisy examples. We restrict the type of patterns that could be learned and we construct hypotheses which are statistically tested against large sample using the statistical query model for learning new patterns.

We will also present the applications of SDPs to various meaning related natural language processing tasks, like word sense disambiguation, textual entailment and meaning preserving translation.