Automatic Induction of FrameNet Lexical Units

Abstract

Many attempts to integrate FrameNet in NLP systems have so far failed because of its limited coverage. In this work, we investigate the applicability of data-driven models on the task of *lexical unit (LU) induction.* Formally, we define LU induction as the task of assigning a LU not present in the FrameNet database (*unknown LU*) to its evoked frame(s). As the number of frames is considerably high, the task is intuitively difficult for a machine to solve. A further complexity regards multiple assignments: lexical units can be ambiguous, i.e. they can be mapped to more than one frame (e.g. *tea* could map both to FOOD and SOCIAL EVENT). We propose two models to solve the task.

The basic idea for the *distributional model* is to induce new LUs by modelling existing frames and unknown LUs in the form of co-occurrence vectors computed over a corpus. The intuition is that the meaning of a LU can be described by the set of textual contexts in which it appears, and that semantically related LUs share similar contexts (*Distributional Hypothesis* (Harris, 1964)). In our setting, the goal is to find a *semantic space model* able to capture the notion of frames – i.e. the properties of *"being characteristic of a frame"*.

The *WordNet-based model* is designed under the hypothesis that meaning aspects evoked by a frame can be detected by jointly considering the WordNet synsets activated by *all* LUs of the frame. We implement this intuition in a weakly supervised model, where each frame is represented as a set of specific subgraphs of the WordNet hierarchy. As every part-of-speech has a separated hierarchy, we build a subgraph for each of them. These sub-graphs assumedly represent lexical semantic properties characterizing the frame. An unknown LU of a given part-of-speech is assigned to the frame whose corresponding subgraph is semantically most similar to one of the senses of the considered LU.

We also combine the two approaches using a simple back-off model with the WordNet approach as defaults, backing-off to the distributional model if no frame can be found. The intuition is that WordNet should guarantee the highest precision in the assignment, while distributional similarity can cover LUs not present in WordNet.

We evaluate our models in a leave-one-out setting over a reference gold standard (FrameNet 1.3), i.e. for each LU we simulate the induction task by executing a leave-one-out procedure: We remove the LU from its original frames and then ask our models to reassign it to the most similar frame(s). We compute (*k-best) accuracy* as the fraction of LUs that are correctly (re-)assigned to the original frame considering the k most similar frames from our model. In this evaluation, our combined model achieves a 1-best accuracy of 43% and a 10-best accuracy of 73%, with coverage of 95%. A qualitative analysis of erroneous predictions shows that the main problems were data sparseness, semantic ambiguity and plausible assignments not present in FrameNet. Our results show that the automatic expansion of FrameNet is viable and promising.