Using Prior Domain Knowledge to Build Robust HMM-Based Semantic Tagger Trained on Completely Unannotated Data

Kinfe T. Mengistu, M. Hannemann,
T. Baum and A. Wendemuth

Cognitive Systems Group, Otto-von-Guericke University

09.07.2008

Summary

In this paper, an approach towards spoken language understanding is described.
We present a robust HMM-based statistical semantic tagging model with the following main features:

- the model is trained on completely unannotated data,
- the approach relies mainly on prior domain knowledge to counterbalance the lack of semantically annotated data,
- the proposed approach encodes longer context by grouping strongly related semantic concepts together into cohesive units known as super-concepts,
- each super-concept is modeled as a sub-network in the HMM.

Summary

In general, the approach results in a model that:

- offers high ambiguity resolution power,
- outputs semantically rich information,
- requires relatively low human effort.

Moreover, the model is robust in that:

- it can parse utterances that contain unseen transitions and out-of-vocabulary words (OOVs), and
- it could correctly label a significant amount of OOVs using the encoded contextual information.

The performance of the resulting models, as evaluated on two different corpora in two application domains (in two languages), is:

- 97.01% for airline travel planning (in English), and
- 97.11% for train-inquiries system (in German)

in F-measure on 1000-utterance test-sets.