Training Parsers on Incompatible Treebanks

We consider the problem of training a statistical parser in the situation when there are multiple treebanks available, and these treebanks are annotated according to different linguistic conventions. To address this problem, we present two simple adaptation methods: the first method is based on the idea of using a shared feature representation when parsing multiple treebanks, and the second method on guided parsing where the output of one parser provides features for a second one.

To evaluate and analyze the adaptation methods, we train parsers on treebank pairs in four languages: German, Swedish, Italian, and English. We see significant improvements for all eight treebanks when training on the full training sets. However, the clearest benefits are seen when we consider smaller training sets. Our experiments were carried out with unlabeled dependency parsers, but the methods can easily be generalized to other feature-based parsers.

Bio:
Richard Johansson earned his PhD in computer science at Lund University in 2008. He is currently working at the Language Bank and Centre for Language Technology at the University of Gothenburg. Richard Johansson has made substantial contributions to a wide range of NLP applications, including dependency parsing, semantic role labeling, opinion mining/sentiment analysis, and discourse parsing, among others.

http://www.df.lth.se/~richardj/