Translating Part-of-Speech Tags via Dependency Structure

Adam R. Teichert, Jagadeesh Jagarlamudi and Hal Daumé III
{teichert,jags,hal}@cs.utah.edu
School of Computing
University of Utah
Salt Lake City, Utah 84112

If languages does not vary in arbitrary ways [2], we can expect data in any language to prove helpful for learning in other languages. In this work we explore using the labeled data in one language to perform part-of-speech tagging of data in another language. In specific, we use dependency structure to transfer part-of-speech knowledge from an annotated source corpus to a non-parallel target corpus in another language. This work is in contrast to other notable work in multilingual part-of-speech tagging which uses parallel corpora to transfer part-of-speech knowledge [3]. To accomplish the knowledge transfer, along with the annotated source corpus, we use an unsupervised part-of-speech tagger based on Gibbs sampling inference of a graphical model. Like the model presented in [1], our model also relies on annotations that provide the sentence dependency structure of the data[1].

There are three main aspects to our approach. First, we perform preliminary unsupervised part-of-speech tagging of the target corpus. This stage results in an assignment of cluster ids to each token in the target corpus. Next we find a mapping between the tags of the source tagset and the tagset of the target corpus. Finally, we transfer the counts (for example, the transition counts) from the annotated source corpus to the target corpus via the mapping and use them to improve the tagging of the target corpus and also to give meaning to the unsupervised cluster ids. In practice, however, the second and third components can be accomplished simultaneously.

We assume that the corpus is generated based on the following generative story:

- for each sentence in the source corpus
    1. generate the dependency tree where nodes are labeled using the *target* language part-of-speech tagset
    2. for each node
        (a) generate a source language part-of-speech tag given the target language part-of-speech tag
        (b) generate the source language word given the source language part-of-speech tag

Note that in step (1) the tree is generated with the target language part-of-speech tags and these tags are then mapped to source language part-of-speech tags. This enables us to automatically find a mapping between the tagsets of the target and source corpora. Similarly, the target corpus sentences are generated by first generating the dependency tree with nodes labelled using source language part-of-speech tags. These tags are then mapped to target language tags and subsequently the target language words are generated. We use collapsed Gibbs sampling for the inference which provides the mapping and also gives us a tagging of each corpus using the tagset of the other.

There are two advantages of using the annotated source corpus. First, it allows us to automatically label the target language cluster ids which is often done using a lexicon of possible part-of-speech tags for some or all of the words in the language vocabulary [3]. Secondly, the information from the source language could also help to improve the clustering aspect of part-of-speech tagging on the target corpus.

[1] This of course might not be available in practice, but similar helpful structure might be learned in an unsupervised way. We merely wish to demonstrate how such information can be *u*sed, and we leave the problem of how to *f*ind it to future work.

We present four sets of experiments. First, as a sanity check, we show the quality of tagset mappings within the same language using the same tagset. A correct mapping of a tag to itself is evidence that the mapping stage in our generative model is doing something reasonable. We also show the ability of our model to find a quality mapping between *fine* and *course-grained* tagsets in the same language. Next, we evaluate the ability of our model to transfer part-of-speech tagging knowledge from an annotated corpus to a target corpus in another language when there is no available part-of-speech tag information for the target corpus. Finally, we discuss how the similarity of the languages used in the mapping affects the mapping quality, and we compare results on various language pairs.

References

[1] J.R. Finkel, T. Grenager, and C.D. Manning. The infinite tree. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 272, 2007.

[2] J.H. Greenberg. *Universals of language*. MIT Press Cambridge, MA:, 1966.

[3] T. Naseem, B. Snyder, J. Eisenstein, and R. Barzilay. Multilingual Part-of-Speech Tagging: Two Unsupervised Approaches. *Journal of Artificial Intelligence Research*, 36:1–45, 2009.

Topic: estimation, prediction, and sequence modeling
Preference: oral