Part of Speech Tagging in NLP

Linshan Jiang
Advanced Topics in Machine Learning and Data Mining

Part of speech tagging, the process of assigning a part-of-speech tagger to sequence of words, is an essential task in natural language processing, in that it helps with information retrieval and syntactic parsing. Knowing the syntactic categories of a word provides information not only for itself (e.g. how to pronounce it), but also for its neighbors (e.g. nouns are likely proceeded with adjectives).

POS tagging is hard, since words are ambiguous – a word can correspond to many tags. One of the approaches is HMM (Hidden Markov Mode), a probabilistic sequence model, which maps a sequence of observations to their hidden states (their labels), by computing a probability distribution of possible sequence of hidden states and outputting the best of them. The Viterbi algorithm is one of the constituent algorithms for HMM. In Brants, T. (2000), he presents a state-of-the art implementation of HMM model.

Vanilla HMM imposes many limitations, one of which is it's hard to incorporate a wider variety of features. If we want to encode knowledge about tagging process, for example, the HMM model would be far more complicated. Thus we introduce a second tagging algorithm – MEMM (maximum entropy Markov model), a discriminative model, which makes adding features easier. McCallum, Andrew et al (2000) presents this model in detail.

However, traditional HMM and discriminative models like MEMM all share a problem: the label bias problem – they exclusively run from left to right. Thus some information about future is ignored. If decision of a word can be informed from future tags rather than only previous tags, we can run bidirectional. This is when CRF (Conditional Random Field) comes into play. (see Lafferty, J. D. et al. 2001) Many recent systems are based on CRF model, such as the Twitter POS tagger (Kevin Gimpel, et al. 2011), and tagger for code-mixed social media text (Souvick Ghosh et al. 2016).

Reading Materials:

Fundamentals:

1. Brants, T. (2000). TnT: A statistical part-of-speech tagger. In ANLP 2000, Seattle, WA, pp. 224–231.
2. McCallum, Andrew; Freitag, Dayne; Pereira, Fernando (2000). "Maximum Entropy Markov Models for Information Extraction and Segmentation" (PDF). Proc. ICML 2000. pp. 591–598.
3. Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML 2001, Stanford, CA.

Interesting recent works regarding social media tagger:

1. Kevin Gimpel, et al. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments, In Proceedings of ACL 2011.
2. Souvick Ghosh et al. Part-of-speech Tagging of Code-Mixed Social Media Text, 2016

Spotlight Question:
What problems do you see in POS tagging that makes it HARD?