Efficient Interactive Annotation with Cascaded Markov Models

Thorsten Brants
Computational Linguistics
Saarland University
66041 Saarbrücken, Germany
thorsten@coli.uni-sb.de

Abstract

During the creation of the NEGRA corpus[1] (Skut et al., 1997), we developed very efficient interactive annotation tools. An easy-to-use graphical tool, Annotate (Plaehn and Brants, 2000), is used to manipulate syntactic structures. Annotate interacts with a part-of-speech tagger and a parser running in the background, thus facilitating rapid semi-automatic corpus annotation.

In our talk, we will present the underlying model of the parser. It generates the structure incrementally. Each increment is based on the existing partial annotation and the annotator can either accept or reject the new increment.

The parser uses the part-of-speech tagger TnT (Brants, 2000)[2] and *Cascaded Markov Models* (Brants, 1999). Each layer of the syntactic structure is represented by a separate Markov Model, starting with the part-of-speech layer. Higher layers are a generalization of the part-of-speech tagging layer. For the purpose of annotation, the tagger and the parser both run in interactive mode.

First, a sequence of part-of-speech tags is generated on-line for a given sentence. In order to facilitate error detection, the tagger generates alternative tags together with their probabilities $P(t_1)$ and $P(t_2)$, giving rise to a reliability measure. If the distance (their quotient) is large, the best tag is simply added to the annotation. If their distance is small, the human annotator is asked for confirmation. We choose a threshold θ and classify the best tag t_1:

$$\text{reliable, if } \frac{P(t_1)}{P(t_2)} \geq \theta, \text{unreliable, if } \frac{P(t_1)}{P(t_2)} < \theta.$$

θ is chosen empirically such that the expected accuracy of all reliable cases is above 99%.

For part-of-speech tagging in the NEGRA corpus, approx. 85% of the tags are classified as reliable with $\theta = 100$. These have an accuracy of 99.2%. Such a level of accuracy is usually very hard to achieve even for human annotators. The remaining 15% are classified as unreliable with an accuracy of 83.0%. The human annotator mainly needs to concentrate on these 15%, which speeds up the process enormously.

Having finished the part-of-speech annotation, it is sent to the parser. The parser builds a lattice with all phrase hypotheses for layer 1 of the resulting structure. The best phrase is selected according to the γ probabilities assigned by the Markov model. We present the structure corresponding to

$$\operatorname*{argmax}_{i,j,e} \gamma_{i,j}(e)$$

(representing the probability of a phrase with label e starting at position i and ending at j). If the user accepts the phrase, all hypotheses that are not compatible with this phrase are removed from the lattice. If the user rejects it, this phrase is removed from the lattice. Probabilities are re-calculated, and again the best hypothesis is displayed to the annotator.

The presented approach is very flexible and trainable for different languages, tagsets, and annotation schemes.

[1] For availability, please check http://www.coli.uni-sb.de/sfb378/negra-corpus

[2] cf. http://www.coli.uni-sb.de/~thorsten/tnt

References

Thorsten Brants. 1999. Cascaded Markov models. In *Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics EACL-99*, Bergen, Norway.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing ANLP-2000*, Seattle, WA.

Oliver Plaehn and Thorsten Brants. 2000. Annotate – an efficient interactive annotation tools. In *Proceedings of the Sixth Conference on Applied Natural Language Processing ANLP-2000*, Seattle, WA.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*, Washington, DC.