Processing of Linguistics knowledge for Annotation

Sree Ganesh Thottempudi
sganeshhcu@gmail.com
NTNU

Abstract

This paper investigates the effect of the Linguistic knowledge on the accuracy of stochastic machines. Presently, annotation techniques are purely based on supervised and unsupervised methods. Accuracy of these results are significant. However, they can be greatly improved by including Linguistic knowledge to these methods. Our main hypothesis is, adding the linguistic knowledge to these supervised and unsupervised machines can improve the accuracy of annotation. In this paper we are dealing with Telugu (ISO639:te1) language annotation. Extracted linguistic annotation rules from TypeCraft pre annotated data supports our hypothesis. These rules are mainly classified in to two categories based on their function. 1. Morpho-syntactic annotation rules 2. Semantic annotation rules. These annotation rules work on every ambiguity clause. We have tested these two kinds of rules on IIIT Telugu annotated data which was annotated by unsupervised method.

Key words: Corpus annotation, TypeCraft, Language documentation, Rule based methods, Stochastic methods

TypeCraft: TypeCraft (TC) is a multi-lingual on-line database where of linguistically-annotated natural language text, embedded in a collaboration and information tool. TC allows users to annotated natural language data with morpho-syntactic and semantic tags. User can share his data with public. It allows the user to import raw text for annotation and export annotated data to MS Word, OpenOffice.org, LaTeX or XML for further use (Beermann and Prange, 2006).
Cutting et al. (1992) proposed a set of desiderata for a good annotator. They are: Robustness, Efficiency, Accuracy, Tunability and Reusability. If we observe the field many people are concentrating on *accuracy* and *reusability*. Since many people are implementing supervised and unsupervised methods to achieve good accuracy in annotation. It has been observed that these two methods have some disadvantages. Supervised methods required large amount of pre annotated data which is not available for many Indian languages (In particular, for Telugu). Unsupervised methods need to build word clumps. Lacking of "fine distinctions found in the carefully designed tag set used in the supervised methods" (Sanchez, 1995) is a drawback in that.

The approached we followed faced many problems: lack of standard pre-annotated data and tag set for Telugu and so on. The main idea we are going to propose is by adding the linguistic knowledge to these supervised and unsupervised machines can improve the accuracy of annotation. This was fulfilled partially. The methodology we followed was:

1) Build a set of annotating rules
2) Build a small standard training data from TypeCraft
3) Manual rules application on training data
4) Automatic refine of manual rules on the training data
5) Apply the refined rules to test data and evaluation of results.

We tested our method on IIIT Telugu data. We observed a significant drop down in error rate.
Final conclusion is: rules derived from deeply pre-annotated data (with TC) definitely effect the accuracy of agglutinative languages like Telugu.

References:

Beermann, Dorothee and Atle Prange. 2006. "Annotating and archiving Natural Language Paradigms online", presentation at the Texas Linguistic Society Annual Conference. Austin, Texas. 2006.
Cutting.D, j.Pedersen, P.Sibum. 1992. “A practical parts-of-speech tagger” proc. Of 3rd Applied natural language conference.
Sanchez Leon, Nieto Serrano AF. 1995. “Development of a Spanish version of the xerox tagger”. Cmplg/
9505035.