Abstracts

20/4

Cognate discovery to bootstrap lexical resources

Closely related languages often contain words that come from the same common ancestor e.g. *hound* in English and *hund* in German. These are termed cognates and can be used to bootstrap lexical resource creation in a low resource language. This project is aimed at automatically discovering cognates between a closely related language pair (e.g. Hindi-Marathi or Hindi-Punjabi). The cognate identification problem has received some attention in the literature (Singh and Surana, 2007; Hauer and Kondrak, 2011), but wordlists used have been relatively small, whereas in this task, we would like to analyse a large part of the vocabulary for both languages.

After cognate identification has taken place, the second step would include the linguistic analysis of cognates. Particularly, we would like to distinguish between cognates that are semantically similar or dissimilar. This division would help in identifying whether lexical resources can be successfully shared between Hindi and the closely related language pair.

References

Singh A. K and Surana H. 2007 Study of cognates among South Asian languages for the purpose of building lexical resources. Journal of Language Technology. Vol 8:78-82

Bradley Hauer and Grzegorz Kondrak. 2011. Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists. The 5th International Joint Conference on Natural Language Processing (IJCNLP 2011).

http://asjp.clld.org/

http://ielex.mpi.nl/

Extracting a Tree-Adjoining Grammar from the Hindi Treebank

The Hindi Treebank (Bhatt et. al, 2009) contains manually annotated parse trees for 400,000 words in a dependency syntax representation. While a dependency representation is useful for NLP applications, Tree-Adjoining Grammars are more linguistically expressive and have been used for a number of applications, including parsing and machine translation as well as natural language generation. This work will contribute towards building a Tree-Adjoining Grammar (TAG) for Hindi.

Bhatt et. al. (2012) describe a detailed algorithm to carry out both of these tasks. There are two primary challenges in this work: the first is to generate phrase structure trees from the dependency representation and second, extract elementary trees (the primitive structures of TAG) from the treebank.

References

Bhatt, Rajesh, Owen Rambow, and Fei Xia. 2012. Creating a Tree-Adjoining Grammar from a Multilayer Treebank. In Proceedings of the 11th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+11), pages 162–170.

Bhatt, Rajesh, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Sharma, and Fei Xia. 2009. A Multi-Representational and Multi-Layered Treebank for Hindi/Urdu. In In the Proceedings of the Third Linguistic Annotation Workshop held in conjunction with ACL-IJCNLP 2009.

An eye tracking experiment for processing of orthographic information

This project is aimed at uncovering the mapping between spoken words and their orthographic counterparts using eye fixations in a visual world paradigm. This is an eye-tracking study that examines the time course of fixations on visually presented words in response to a spoken instruction. Following Salverda and Tannenhaus (2010), participants see four words on a screen, consisting of a target, a competitor and two distractors. The degree of overlap between the target and competitor is varied based on phonology (*bead* and *bean*) or orthography (*bead* and *bear*).

After listening to an instruction, (e.g. *Click on the word bead*) the participant's looks towards a word are recorded to check whether phonology or orthography

The first step would be to replicate the Salverda and Tannenhaus experiment using English data. Rather than native speakers of English, participants will

consist of two groups of more or less proficient English speaking bilinguals. Other variations are also possible. Additionally, this work can also be extended to Indic scripts like Devanagari (or others), such that reading experiments (e.g. Jyotsna and Vaid, 2002) can be replicated within the visual world paradigm.

References

Salverda A. and Tannenhaus M.K. 2010.Tracking the Time Course of Orthographic Information in Spoken-Word Recognition. Journal of Experimental Psychology:Learning, Memory, and Cognition. 2010, Vol. 36, No. 5, 1108–1117

Jyotsna Vaid and Ashum Gupta. 2002. Exploring Word Recognition in a Semi-Alphabetic Script: The Case of Devanagari. Brain and Language: 81,679–690