Cognate Identification to Bootstrap Lexical Resources

Shantanu Kumar

Problem Statement

The project aims at automatically identifying cognate pairs between closely related South Asian language pairs like Hindi and Marathi and the subsequent linguistic analysis of these cognate pairs.

Previous Work

There have been works that have tried to use machine learning techniques in the domain of cognate identification. These works try to incorporate various types of features based on Orthographic and Phonetic similarity.

In [2], Hauer et al used the feature based SVM classifier to identify cognate pairs. They used *Word Similarity Measures* as well as *Language Similarity* features, which incorporate a sense of relatedness between the language pair, as input to their classifier. Further on the basis of this classification, they group the words into different clusters to create cognate groups.

T. Rama [3], used string sub-sequence based features for classification of cognate pairs. The feature vector for any word is a vector of weights. Each weight corresponds to a n-length character sequence which can be present in the string as a sub-sequence. These weights are varied depending upon the frequency of the sub-sequence and also the gaps and the length over which the sub-sequence is present in the string.

Datasets

- Dyen et al [1], provide a Indo-European dataset with 200 concepts and 84 languages. The words are labeled with a unique Cognate Class Number (CCN). It is has been used in both [2] [3] for the task of cognate identification.
- ASJP provides 40 item word lists for many languages but the cognate information is limited to only a few selected languages.
- The Indo-European Lexical Cognacy Database contains cognacy judgements for 163 languages and 225 meanings. The words are labelled into 5013 Cognate sets.
- The Europarl Parallel Corpus contains aligned sentences for various language pairs like English-French and English-German, for the task of machine translation. The dataset is not meant for the task of cognate identification but the aligned texts can be used for training word vectors in the languages.

Workflow and Timeline

Implmenting Baselines

We shall start by implementing the Word Similarity Features based classifier by [2] and String Sub-sequence Features based classifier by [3] which shall provide us with baseline models on the Indo-European dataset by Dyen et al.

Word Vector Features

Most works for cognate identification use Orthographic and Phonetic similarity features for classification. In addition to improving these features, incorporating semantic information during classification should also help to boost the model. Word vector embeddings can be used as features to represent this semantic information related to the words. Since word vectors are vector representations of the vocabulary in a low dimension space, they include some level of semantic structure in their arrangement.

Domain Adaptation

The various wordlists based datasets provide a starting point for guessing cognate pairs in the sense that the provide words of the same *meaning* from different languages. But in a practical setting we do not have this starting point. Given a large piece of text from an unknown language to compare with a known language, we would not know which cross language word pairs to test for cognates. Word vectors can again be used here to align the text (given some initial seed) and find the closest related words which can be further tested using a classifier for cognates.

We can try to train word vectors for different languages together in a constraint model as a semi-supervised method for revealing cognate information. For example, we can try to train word vectors for a pair of languages (for which we have a small set of labeled cognate pairs) together while constraining the known cognate pairs to have the same vector. This can probably lead to a vector representation of the vocabularies where all cognate pairs have the same vector representation or maybe are the closest neighbors of each other.

These ideas can be used to adapt our models to the domain of South Asian languages like Marathi-Hindi and Punjabi-Hindi. Using the information mined on a smaller annotated dataset, we can try to apply the model on a larger set in this semi-supervised fashion, from which any promising results can be evaluated manually.

References

[1] Dyen, I., Kruskal, J. B., and Black, P. An indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical society 82*, 5 (1992), iii–132.

[2] Hauer, B., and Kondrak, G. Clustering semantically equivalent words into cognate sets in multilingual lists. In *IJCNLP* (2011), Citeseer, pp. 865–873.

[3] Rama, T. Automatic cognate identification with gap-weighted string subsequences. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, May 31–June 5, 2015 Denver, Colorado, USA* (2015), pp. 1227–1231.

[4] Singh, A. K., and Surana, H. Study of cognates among south asian languages for the purpose of building lexical resources. In *Proceedings of National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing, Mumbai, India* (2007), Citeseer.