Cognate Discovery to Bootstrap Lexical Resources

Shantanu Kumar, Dept. of Electrical Engineering, IIT Delhi

***Abstract*—The high level of linguistic diversity in South Asia poses the challenge of building lexical resources across these languages. This project is aimed at automatically discovering cognates between closely related language pairs (e.g. Hindi-Marathi or Hindi-Punjabi) in a scalable manner. We would like to analyze a large part of the vocabulary for both languages, as opposed to small word lists used in previous works. We also aim to do a linguistic analysis over the identified cognates to conclude whether lexical resources can be successfully shared between Hindi and related languages.**

I. Introduction

Cognates are words across different languages that originate from the same word in a common ancestral language. For example, the English word '*Night*' and the German '*Nacht*', both meaning *night* and English '*Hound*' and German '*Hund*', meaning *dog* are cognates whose origin can be traced back to Proto-Germanic. Cognates are not always revealingly similar and can change substantially over time such that they do not share form similarity.

Automatic cognate identification, in NLP, refers to the application of string similarity algorithms with machine learning algorithms for determining if a given word pair is cognate or not. Identification of cognates is essential in historical linguistics, and cognate information has been successfully applied to NLP tasks, like sentence alignment and statistical machine translation. It can also be used to bootstrap lexical resource creation for a language with low resources by finding parallels in related rich resource languages.

II. Previous Work

The approaches developed for the task of cognate identification are usually based on combination of different similarity measures between a pair of words as features to a linear classifier. These include orthographic, phonetic and semantic similarity. The objective can be finding pairs of cognates among two related languages, or finding groups of cognates among multiple languages.

Hauer and Kondrak [1] incorporate a number of diverse word similarity measures that are manually identified along with features that encode the degree of affinity between pairs of languages as input to a SVM classifier. The authors employ binary language-pair feature that is used to weigh the language distance and assist the task of semantic clustering of cognates.

T. Rama [2] uses a string kernel based approach for automatic cognate identification. By identifying all common subsequences of a fixed length between the word pairs and using that as input features to the linear classifier, they show that subsequence based features outperform word similarity measures.

III. Datasets

The input data to the models for the task of cognate identification include dictionaries, multilingual word lists, and bitexts. But the word lists that have been used in all the works so far have been relatively small, probably due to the difficulty/disputability in cognate judgments.

The freely available Indo-European Dataset (Dyen et al., 1992) is the most commonly used dataset for cognate identification. It provides 16,520 lexical items for 200 concepts and 84 language varieties. It provides a unique Cognate Class Number to each word. The dataset is transcribed in a broad romanized phonetic alphabet.

We would also use the TDIL Hindi-Marathi sentence-aligned corpus as the testing data for our final model. This dataset would provide a large part of the vocabulary from the both the languages to search for cognates.

IV. Solution Approach

All the previous works on cognate identification mainly use phonetic or orthographic features of words for this judgment. However when applying the model in a realistic setting, on larger portions, where the words are not aligned or grouped by meaning, some sort of semantic information would be needed by the classifier. We propose to introduce this semantic information by utilizing the word embedding features.

Word embeddings are representation features where the words in the vocabulary are represented as points in a low dimensional space as compared to the vocabulary size. These are learnt by unsupervised approached using deep learning models. They are task independent features that are arranged such that their structure captures some sort of semantic relationships between the words. We propose to use the multilingual word embeddings Polyglot [3] that provide word vector embeddings for 116 languages over a rich vocabulary. These are trained on the processed Wikipedia text dumps of the various languages. Once the model is trained in the multilingual setting, we would like to apply it specifically to the domain of Hindi-Marathi and observe the cognate pairs that the model is able to identify.

References

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

[2] T. Rama, "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*, pp. 1227–1231, 2015.

[3] R. Al-Rfou, B. Perozzi, and S. Skiena, "Polyglot: Distributed word representations for multilingual nlp," in *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, (Sofia, Bulgaria), pp. 183–192, Association for Computational Linguistics, August 2013.