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 the languages from these regions. This project is aimed at automatically discovering cognates between closely related language pairs like Hindi-Marathi or Hindi-Punjabi, in a scalable manner, which would assist the task of lexical resource creation.**

I. Introduction

Cognates are words across different languages that are known to have originated 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 [1]. Cognates are not always revealingly similar and can change substantially over time such that they do not share form similarity. Cognate information has been successfully applied to NLP tasks, like sentence alignment [2] in bitexts and improving statistical machine translation models [3].

II. Previous Work

Hauer and Kondrak [4] use a number of basic word similarity measures along with features that encode affinity between language pairs, as input to a SVM classifier for cognate prediction. T. Rama [1] uses a string kernel based approach for automatic cognate identification. By identifying all common subsequences between the word pairs and weighting them by their gaps in the strings, he shows that subsequence based features outperform word similarity measures.

The input data to the models for this task include dictionaries, multilingual word lists, and bitexts. We have used the IELex Database, which is an Indo-European lexical database derived from multiple sources. It has over 34,000 lexical items from 163 languages and information of 5,000 cognate sets. 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.

III. Models

We have primarily worked with the gap weighted subsequence model [1]. The model defines a subsequence vector $\Phi(s)$ for any string s based on the subsequences u in s and the gaps between them. The common subsequence vector $\Phi(s_1, s_2)$ for 2 strings can be defined as,

$$\Phi_1(s_1, s_2) = \{\phi_u(s_1) + \phi_u(s_2); \forall u \text{ present in } s_1 \text{ and } s_2\}$$

$$\Phi_2(s_1, s_2) = \{\phi_u(s_1) + \phi_u(s_2); \forall u \text{ present in } s_1 \text{ or } s_2\}$$

The first equation here represents a *Multiplicative* model and the second an *Additive* model. The original paper uses the *Multiplicative* model.

IV. Results & Analysis

Performance over different POS categories : We divide the samples into Nouns, Adjectives and Others categories. It is found the the models perform significantly poorly on Others category as compared to Nouns and Adjectives.

Performance over different Concepts : A closer look at the individual concepts shows that the recall varies from as high as 80% for some meanings like 'CHILD', 'TOOTH', 'LAKE' to as low as 5% for concepts like 'WHEN', 'WHERE', 'WHAT'. It was realised that the number of distinct cognate classes in the dataset is on average less for concepts that perform poorly for the model. Such concepts have large variations of sounds within a class of cognates.

Transcription of text : Different transcriptions of the data affects the performance of the model. Using a finer transcription like IPA gives higher f-scores than Romanised IPA. However, a finer character means that there are less common subsequences. This results in sparser features and information is lost in the *Multiplicative* model where we ignore non-common subsequences.

Analysis of Additive Model : The *Additive* model was found to give very poor results as compared to the *Multiplicative* model. It over-fit on the training data, despite varying the regularization penalty to high values. It was observed that it tends to perform better on meanings with a high positive sample bias in the data.

Hybrid Model : We implemented a hybrid model between the *Additive* and *Multiplicative* model which was formed by stealing weight from positive samples in the latter model and distributing it amongst the positive samples in the former. It was observed that as the model transforms from *Multiplicative* to *Additive*, the F-score increases first and then decreases drastically after a threshold.

V. Future Work

We have seen that semantics of the words influence the behavior and performance of our models. All the previous works on cognate identification mainly use phonetic or orthographic features of words for cognate judgment. We would like to introduce semantic information by utilizing the word embedding features. Word embeddings are task independent features that are arranged such that their structure captures some semantic relationships between the words. Along with using the word embedding features in our model, we would also like to move towards a neural network based model for classification of cognates, by utilizing recurrent networks like RNNs and LSTMs to encode the input words at the character level. Once our models have been trained in the multilingual setting, we would like to apply it specifically to the domain of Hindi-Marathi using the sentence aligned corpus and evaluate the cognate pairs that we are able to discover.

REFERENCES

[1] 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.

[2] M. Simard, G. F. Foster, and P. Isabelle, "Using cognates to align sentences in bilingual corpora," in *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*, pp. 1071–1082, IBM Press, 1993.

[3] G. Kondrak, D. Marcu, and K. Knight, "Cognates can improve statistical translation models," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pp. 46–48, Association for Computational Linguistics, 2003.

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