Two methods for automatic identification of cognates.

Taraka Rama[*1], Prasant Kolachina[†2], and Sudheer Kolachina[‡3]

[1]University of Gothenburg
[2]International Institute of Information Technology, Hyderabad
[3]Massachusetts Institute of Technology

1 Introduction

Cognate identification is a important task in historical linguistics for the purpose of establishing genealogical relationships between languages (Campbell, 2004). Cognates are identified through regular sound correspondences between words, from supposedly related languages, having a similar surface and semantic forms. However, not all cognate pairs are equally similar. In other words, cognacy judgment is not a binary decision but a finely graded one. Consider a cognate pair, German and English, *hund* ∼ *hound* which reveals itself to be cognate through visual inspection. The cognacy similarity score for such a pair should be very high. Now, consider a cognate pair, Sanskrit to English, *chakra* ∼ *wheel* whose similarity is not revealed through visual inspection. Such a cognate pair should have lower similarity score since the regular sound processes have affected the original proto-word to show divergent surface yet, related forms in the distantly related languages. In the rest of the paper, we propose two methods in section 2 for the purpose of cognate identification. We describe the multi-lingual dataset in section 3 and validate our methods in section 4.

2 Methods

Automatic detection of sound correspondences is the crucial step in cognate identification. Sound correspondences can be extracted using the alignments obtained from Levenshtein distance (LD; Levenshtein 1965). LD is defined as the minimum number of insertion, deletion and substitution operations required to transform a string into another with all the operation costs set to 1. However, the alignments generated through LD might not be linguistically meaningful. For instance, the alignments between the words for ashes: Catalan 'sendra' and Italian 'tenere' would be 's' : 't', 'e' : 'e', 'n' : 'n', 'd': 'e', and 'a' and 'e'. The phoneme segment pair 'd': 'e' is linguistically implausible, since a consonant cannot align with a vowel.

Wieling et al. (2009) alleviate this problem, for the classification of Bulgarian dialectal data, by introducing an additional constraint (VC-constraint) that vowels cannot align with consonant and vice-versa. Their approach is summarized as follows:

1. Employ the VC-constraint LD to align a word pair and extract all possible phoneme segment pairs for a pair of languages.

2. The similarity of a segment pair is computed using Pair-wise Mutual Information (PMI; Church and Hanks 1990) which is defined as $\log p(x,y) - \log p(x) - \log p(y)$.

[*]taraka.rama.kasicheyanula@gu.se
[†]prasant@research.iiit.ac.in
[‡]sudheer@mit.edu

3. The segment pair similarity is converted into a distance score, in the range of $[0,1]$, through the formula

$$\frac{max_{pmi} - pmi}{max_{pmi} - min_{pmi}} \tag{1}$$

4. The pair-wise item LD is computed using the segment pair distances obtained from step 3.

Steps $1-5$ are repeated until there is no change in the segment pairs between two successive iterations. The final iteration of the above algorithm yields a list of segment pair distances.

However, the VC-constrained LD operates at a single segment level i.e. the method always operates on a single segment pair. This method, when extended to multiple length segments allows alignment between segments of length greater than 1. Bergsma and Kondrak (2007) employ the idea of multiple length segments to train a linear classifier for the automatic identification of cognates from bi-text data. They align a word pair using the basic LD and extract adjacent segment pairs. The maximum length of a segment pair is limited to 3 in their experiments. This approach is expected to identify word pairs which need not be genetically related but are borrowings. The same authors also identify their multiple segment approach similar to that of "phrases" in Statistical Machine Translation (SMT).

Pursuing the idea of "phrases", the class of generative alignment models commonly referred to in literature on SMT as IBM models (Brown et al., 1993) can be used to generate alignments across multiple length segments. These models are used to align words between translations across a language pair, and are naturally designed to generate alignments between multiple length segments across two languages, unlike the traditional LD method. The IBM models utilize information such as frequency counts and co-occurrence counts across the word lists to generate alignments, using minimal linguistic information. We extend the same approach to automatically align multiple length segments in a word pair across two languages.

For any given pair of languages, the word pairs for identical concepts are extracted to create a bilingual word list for the language pair. Each of the word pairs are aligned using the IBM models to extract multiple length segment pairs. We compute a PMI-based segment distance score for each of the multiple segment pairs using the normalization formula given in equation 1. In our experiments, we limit the maximum length of a segment in the pair to 2. The alignments are obtained using the implementation of publicly available IBM models available in Moses (Koehn et al., 2007). The toolkit additionally provides multiple heuristic algorithms to extract high quality alignments generated from the IBM models. We use all of these heuristics to extract segment pairs prior to the computation of the PMI score for each segment pair.

We observed that the original IPA transcribed data has fine distinctions such as vowel length and primary stress. We ignored the vowel length distinction and stress pattern. Further, the IPA symbols are mapped to a reduced sound class alphabet consisting of 21 symbols; 15 consonant classes and 6 vowel classes (given in List 2012) to encounter symbol sparsity. In all our experiments, k was set to 1.

The contributions of this paper is threefold:

1. We apply the linguistically motivated Levenshtein distance to the task of cognate identification on three different datasets given in List (2012).

2. We apply a popular SMT technique to align phoneme segments between semantically equivalent word pairs and use the segment pair distance to compute the LD between a word pair.

3. We introduce a new evaluation measure to quantify the performance of the two measures.

3 Dataset and Evaluation Measures

Language Group	Number of languages	Number of items
Indo-European (IE)	20	207
Germanic (GER)	7	110
Uralic (URL)	21	110

Table 1: Number of languages and items in the three language groups.

The dataset, in table 1 also contains the cognacy judgments for a item between a pair of languages. In a language pair, the pair-wise item distances are compared to the gold standard cognate judgments using point-biserial correlation (a special case of Pearson's r). In each iteration, we compute the average cognate identification accuracy by taking the average of the correlation for all language pairs. The improvement of the average correlation between two successive iterations is measured through a paired *t-test* with significance level set at 0.05.

4 Results

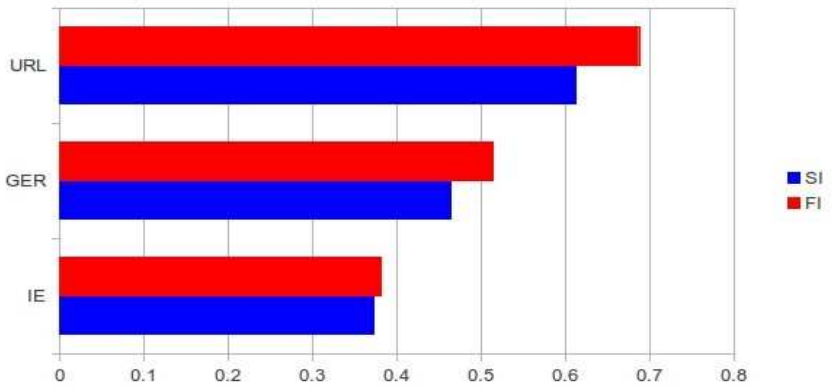

Figure 1: Results of VC-constrained LD

Figure 1 shows the improvement in the average correlation between the starting iteration (SI) and the final iteration (FI) for each language group. The starting iteration makes use of the basic VC-constrained levenshtein distance. The subsequent iterations make use of the PMI-based segment pair distances to compute the LD for a word pair. Table 2 shows the number of iterations the algorithm required to converge as well as the number of statistically significant iterations.

Language group	Number of iterations	Significant iterations
IE	2	2
GER	6	2
URL	4	4

Table 2: Number of iterations and significant iterations for each language group. The significant iterations are always less than or equal to the number of iterations to converge.

Figure 2 shows the results of SMT derived segment pair distances in computing the LD for a word pair. The result for Uralic language group is comparable to the result of the VC-constrained LD. The results for Indo-European and Germanic datasets are lower than the results for VC-constrained LD. It has to be noted that the algorithms are

not directly comparable. VC-constrained LD merges the segment pairs generated from all language pairs and then computes the PMI-based distance score for a segment pair. Whereas, the SMT-based alignments are generated independently for each language pair and the PMI-based segment distances are also computed independently for each language pair.

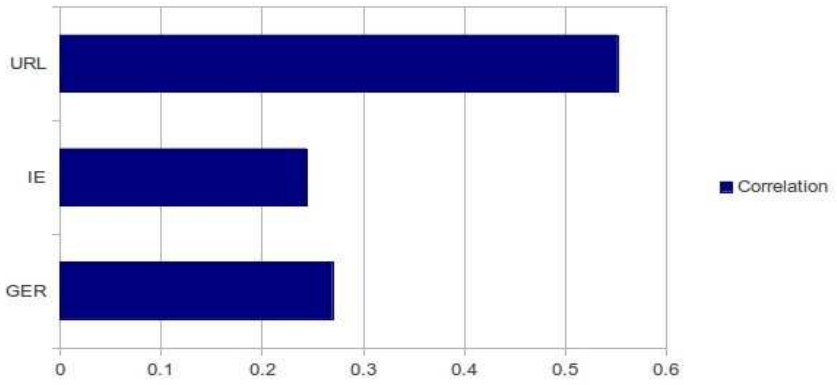

Figure 2: Results of SMT derived segment pair distances. Each bar in the figure shows the average agreement of the method between the pair-wise distances and the gold standard.

5 Conclusion

In this abstract, we described and applied two statistically driven algorithms for the task of cognate identification to three multi-lingual datasets. The initial results suggest that both the approaches are worth pursuing and can be applied to the four other language groups' datasets listed in List (2012). As a future work, we propose that the VC-constrained LD be used for computing the segment pair distances for each language pair. Also, the SMT based segment distances should be computed for the overall language pairs for a direct comparison between the two methods.

References

Shane Bergsma and Grzegorz Kondrak. Alignment-based discriminative string similarity. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 656, 2007.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. Mathematics of statistical machine translation: Parameter estimation. pages 19(2):263–311, 1993.

L. Campbell. *Historical Linguistics: An Introduction*. MIT Press, 2004.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. ISSN 0891-2017.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics, 2007.

VI Levenshtein. Binary codes capable of correcting spurious insertions and reversals. *Cybernetics and Control Theory*, 10:707–710, 1965.

Johann-Mattis List. Lexstat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France, April 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W12-0216`.

Martijn Wieling, Jelena Prokić, and John Nerbonne. Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 26–34. Association for Computational Linguistics, 2009.