Similarity Patterns in Words

Grzegorz Kondrak
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada, T6G 2E8
gkondrak@ualberta.ca

Abstract

Words are important both in historical linguistics and natural language processing. They are not indivisible abstract atoms; much can be gained by considering smaller units such as morphemes, phonemes, syllables, and letters. In this presentation, I attempt to sketch the similarity patterns among a number of diverse research projects in which I participated.

1 Introduction

Languages are made up of words, which continuously change their form and meaning. Languages that are related contain cognates — reflexes of proto-words that survive in some form in the daughter languages. Sets of cognates regularly exhibit recurrent sound correspondences. Together, cognates and recurrent sound correspondences provide evidence of a common origin of languages.

Although I consider myself more a computer scientist than a linguist, I am deeply interested in words. Even though many NLP algorithms treat words as indivisible abstract atoms, I think that much can be gained by considering smaller units: morphemes, phonemes, syllables, and letters. Words that are similar at the sub-word level often exhibit similarities on the syntactic and semantic level as well. Even more important, as we move beyond written text towards speech and pronunciation, the make-up of words cannot be ignored anymore.

I commenced my NLP research by investigating ways of developing computer programs for various stages of the language reconstruction process (Kondrak, 2002a). From the very start, I aimed at proposing language-independent solutions grounded in the current advances in NLP, bioinformatics, and computer science in general. The algorithms were evaluated on authentic linguistic data and compared quantitatively to previous proposals. The projects directly related to language histories still form an important part of my research. In Section 2, I refer to several of my publications on the subject, while in Section 3, I focus on other NLP applications contributions that originate from my research on diachronic linguistics.

2 Diachronic NLP

The comparative method is the technique applied by linguists for reconstructing proto-languages. It consists of several stages, which include the identification of cognates by semantic and phonetic similarity, the alignment of cognates, the determination of recurrent sound correspondences, and finally the reconstruction of the proto-forms. The results of later steps are used to refine the judgments made in earlier ones. The comparative method is not an algorithm, but rather a collection of heuristics, which involve intuitive criteria and broad domain knowledge. As such, it is a very time-consuming process that has yet to be accomplished for many language families.

Since the comparative method involves detection of regularities in large amounts of data, it is natural to investigate whether it can be performed by a computer program. In this section, I discuss methods for implementing several steps of the comparative method that are outlined above. The ordering of projects is roughly chronological. For an article-length summary see (Kondrak, 2009).

Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 49–53,
Avignon, France, April 23 - 24 2012.

2.1 Alignment

Identification of the corresponding segments in sequences of phonemes is a necessary step in many applications in both diachronic and synchronic phonology. ALINE (Kondrak, 2000) was originally developed for aligning corresponding phonemes in cognate pairs. It combines a dynamic programming alignment algorithm with a scoring scheme based on multi-valued phonetic features. ALINE has been shown to generate more accurate alignments than comparable algorithms (Kondrak, 2003b).

Bhargava and Kondrak (2009) propose a different method of alignment, which is an adaptation of Profile Hidden Markov Models developed for biological sequence analysis. They find that Profile HMMs work well on the tasks of multiple cognate alignment and cognate set matching.

2.2 Phonetic Similarity

In many applications, it is necessary to algorithmically quantify the similarity exhibited by two strings composed of symbols from a finite alphabet. Probably the most well-known measure of string similarity is the edit distance, which is the number of insertions, deletions and substitutions required to transform one string into another. Other measures include the length of the longest common subsequence, and the bigram Dice coefficient. Kondrak (2005b) introduces a notion of n-gram similarity and distance, and shows that edit distance and the length of the longest common subsequence are special cases of n-gram distance and similarity, respectively.

Another class of similarity measures are specifically for phonetic comparison. The ALINE algorithm chooses the optimal alignment on the basis of a similarity score, and therefore can also be used for computing phonetic similarity of words. Kondrak (2001) shows that it performs well on the task of cognate identification.

The above algorithms have the important advantage of not requiring training data, but they cannot adapt to a specific task or language. Researchers have therefore investigated adaptive measures that are learned from a set of training pairs. Mackay and Kondrak (2005) propose a system for computing string similarity based on Pair HMMs. The parameters of the model are automatically learned from training data that consists of pairs of strings that are known to be similar.

Kondrak and Sherif (2006) test representatives of the two principal approaches to computing phonetic similarity on the task of identifying cognates among Indoeuropean languages, both in the supervised and unsupervised context. Their results suggest that given a sufficiently large training set of positive examples, the learning algorithms achieve higher accuracy than manually-designed metrics.

Techniques such as Pair HMMs improve on the baseline approaches by using a set of similar words to re-weight the costs of edit operations or the score of sequence matches. A more flexible approach is to learn from both positive and negative examples of word pairs. Bergsma and Kondrak (2007a) propose such a discriminative algorithm, which achieves exceptional performance on the task of cognate identification.

2.3 Recurrent Sound Correspondences

An important phenomenon that allows us to distinguish between cognates and borrowings or chance resemblances is the regularity of sound change. The regularity principle states that a change in pronunciation applies to sounds in a given phonological context across all words in the language. Regular sound changes tend to produce recurrent sound correspondences of phonemes in corresponding cognates.

Although it may not be immediately apparent, there is a strong similarity between the task of matching phonetic segments in a pair of cognate words, and the task of matching words in two sentences that are mutual translations. The consistency with which a word in one language is translated into a word in another language is mirrored by the consistency of sound correspondences. Kondrak (2002b) proposes to adapt an algorithm for inducing word alignment between words in bitexts (bilingual corpora) to the task of identifying recurrent sound correspondences in word lists. The method is able to determine correspondences with high accuracy in bilingual word lists in which less than a third the word pairs are cognates.

Kondrak (2003a) extends the approach to the identification of complex correspondences that involve groups of phonemes by employing an algorithm designed for extracting non-compositional compounds from bitexts. In experimental evaluation against a set of correspondences manually

identified by linguists, it achieves approximately 90% F-score on raw dictionary data.

2.4 Semantic Similarity

Only a fraction of all cognates can be detected by analyzing Swadesh-type word lists, which are usually limited to at most 200 basic meanings. A more challenging task is identifying cognates directly in bilingual dictionaries, which define the meanings of words in the form of glosses. The main problem is how to quantify semantic similarity of two words on the basis of their respective glosses.

Kondrak (2001) proposes to compute similarity of glosses by augmenting simple string-matching with a syntactically-informed keyword extraction. In addition, the concepts mentioned in glosses are mapped to WordNet synsets in an attempt to account for various types of diachronic semantic change, such as generalization, specialization, and synechdoche.

Kondrak (2004) presents a method of combining distinct types of cognation evidence, including the phonetic and semantic similarity, as well as simple and complex recurrent sound correspondences. The method requires no manual parameter tuning, and performs well when tested on cognate identification in the Indoeuropean word lists and Algonquian dictionaries.

2.5 Cognate Sets

When data from several related languages is available, it is preferable to identify cognate sets simultaneously across all languages rather than perform pairwise analysis. Kondrak et al. (2007) apply several of the algorithms described above to a set of diverse dictionaries of languages belonging to the Totonac-Tepehua family in Mexico. They show that by combining expert linguistic knowledge with computational analysis, it is possible to quickly identify a large number of cognate sets within the family, resulting in a basic comparative dictionary. The dictionary subsequently served as a starting point for generating lists of putative cognates between the Totonacan and Mixe-Zoquean families. The project eventually culminated in a proposal for establishing a super-family dubbed Totozoquean (Brown et al., 2011).

Bergsma and Kondrak (2007b) present a method for identifying sets of cognates across groups of languages using the global inference framework of Integer Linear Programming. They show improvements over simple clustering techniques that do not inherently consider the transitivity of cognate relations.

Hauer and Kondrak (2011) present a machine-learning approach that automatically clusters words in multilingual word lists into cognate sets. The method incorporates a number of diverse word similarity measures and features that encode the degree of affinity between pairs of languages.

2.6 Phylogenetic Trees

Phylogenetic methods are used to build evolutionary trees of languages given data that may include lexical, phonological, and morphological information. Such data rarely admits a perfect phylogeny. Enright and Kondrak (2011) explore the use of the more permissive conservative Dollo phylogeny as an alternative approach that produces an output tree minimizing the number of borrowing events directly from the data. The approach which is significantly faster than the more commonly known perfect phylogeny, is shown to produce plausible phylogenetic trees on three different datasets.

3 NLP Applications

In this section, I mention several NLP projects which directly benefitted from insights gained in my research on diachronic linguistics.

Statistical machine translation in its original formulation disregarded the actual forms of words, focusing instead exclusively on their co-occurrence patterns. In contrast, Kondrak et al. (2003) show that automatically identifying orthographically similar words in bitexts can improve the quality of word alignment, which is an important step in statistical machine translation. The improved alignment leads to better translation models, and, consequently, translations of higher quality.

Kondrak (2005a) further investigates **word alignment** in bitexts, focusing on on identifying cognates on the basis of their orthographic similarity. He concludes that word alignment links can be used as a substitute for cognates for the purpose of evaluating word similarity measures.

Many hundreds of drugs have names that either look or sound so much alike that doctors, nurses and pharmacists sometimes get them confused, dispensing the wrong one in errors that may

injure or even kill patients. Kondrak and Dorr (2004) apply anumber of similarity measures to the task of identifying **confusable drug names**. They find that a combination of several measures outperforms all individual measures.

Cognate lists can also assist in **second-language learning**, especially in vocabulary expansion and reading comprehension. On the other hand, the learner needs to pay attention to *false friends*, which are pairs of similar-looking words that have different meanings. Inkpen et al. (2005) propose a method to automatically classify pairs of words as cognates or false friends, with focus on French and English. The results show that it is possible to achieve very good accuracy even without any training data by employing orthographic measures of word similarity.

Transliteration is the task of converting words from one writing script to another. Transliteration mining aims at automatically constructing bilingual lists of names for the purpose of training transliteration programs. The task of detecting phonetically-similar words across different writing scripts is quite similar to that of identifying cognates, Sherif and Kondrak (2007) applies several methods, including ALINE, to the task of extracting transliterations from an English-Arabic bitext, and show that it performs better than edit distance, but not as well as a bootstrapping approach to training a memoriless stochastic transducer. Jiampojamarn et al. (2009) employ ALINE for aligning transliterations from distinct scripts by mapping every character to a phoneme that is the most likely to be produced by that character. They observe that even such an imprecise mapping is sufficient for ALINE to produce high quality alignments.

Dwyer and Kondrak (2009) apply the ALINE algorithm to the task of **grapheme-to-phoneme conversion**, which is the process of producing the correct phoneme sequence for a word given its orthographic form. They find ALINE to be an excellent substitute for the expectation-maximization (EM) algorithm when the quantity of the training data is small.

Jiampojamarn and Kondrak (2010) confirm that ALINE is highly accurate on the task of **letter-phoneme alignment**. When evaluated on a manually aligned lexicon, its precision was very close to the theoretical upper bound, with the number of incorrect links less than one in a thousand.

Lastly, ALINE has also been used for the **mapping of annotations**, including syllable breaks and stress marks, from the phonetic to orthographic forms (Bartlett et al., 2008; Dou et al., 2009).

4 Conclusion

The problems involved in language reconstruction are easy to state but surprisingly hard to solve. As such, they lead to the development of new methods and insights that are not restricted in application to historical linguistics. Although the goal of developing a program that performs a fully automatic reconstruction of a proto-language has yet to been attained, the research conducted towards this goal has been, and is likely to continue to influence other areas of NLP.

Acknowledgments

This paper refers to research projects that were conducted jointly with the following colleagues: Susan Bartlett, David Beck, Shane Bergsma, Aditya Bhargava, Cecil Brown, Colin Cherry, Philip Dilts, Bonnie Dorr, Qing Dou, Elan Dresher, Ken Dwyer, Jessica Enright, Oana Frunza, Bradley Hauer, Graeme Hirst, Diana Inkpen, Sittichai Jiampojamarn, Kevin Knight, Wesley Mackay, Daniel Marcu, and Tarek Sherif.

References

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2008. Automatic syllabification with structured SVMs for letter-to-phoneme conversion. In *Proceedings of ACL-08: HLT*, pages 568–576.

Shane Bergsma and Grzegorz Kondrak. 2007a. Alignment-based discriminative string similarity. In *Proceedings of ACL*, pages 656–663.

Shane Bergsma and Grzegorz Kondrak. 2007b. Multilingual cognate identification using integer linear programming. In *Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons*, pages 11–18.

Aditya Bhargava and Grzegorz Kondrak. 2009. Multiple word alignment with Profile Hidden Markov Models. In *Proceedings of the Student Research Workshop at NAACL-HLT*, pages 43–48.

Cecil H. Brown, David Beck, Grzegorz Kondrak, James K. Watters, and Søren Wichmann. 2011. Totozoquean. *International Journal of American Linguistics*, 77(3):323–372, July.

Qing Dou, Shane Bergsma, Sittichai Jiampojamarn, and Grzegorz Kondrak. 2009. A ranking approach

to stress prediction for letter-to-phoneme conversion. In *Proceedings of ACL-IJCNLP*, pages 118–126.

Kenneth Dwyer and Grzegorz Kondrak. 2009. Reducing the annotation effort for letter-to-phoneme conversion. In *Proceedings of ACL-IJCNLP*, pages 127–135.

Jessica Enright and Grzegorz Kondrak. 2011. The application of chordal graphs to inferring phylogenetic trees of languages. In *Proceedings of IJCNLP 2011: 5th International Joint Conference on Natural Language Processing*, pages 545–552.

Bradley Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of IJCNLP 2011: 5th International Joint Conference on Natural Language Processing*, pages 865–873.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Identification of cognates and false friends in french and english. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2005)*, pages 251–257.

Sittichai Jiampojamarn and Grzegorz Kondrak. 2010. Letter-phoneme alignment: An exploration. In *Proceedings of ACL*, pages 780–788.

Sittichai Jiampojamarn, Aditya Bhargava, Qing Dou, Kenneth Dwyer, and Grzegorz Kondrak. 2009. DirecTL: a language-independent approach to transliteration. In *Named Entities Workshop: Shared Task on Transliteration*, pages 28–31.

Grzegorz Kondrak and Bonnie Dorr. 2004. Identification of confusable drug names: A new approach and evaluation methodology. In *Proceedings of COLING 2004: 20th International Conference on Computational Linguistics*, pages 952–958.

Grzegorz Kondrak and Tarek Sherif. 2006. Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *Proceedings of the COLING-ACL Workshop on Linguistic Distances*, pages 43–50.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of HLT-NAACL*, pages 46–48. Companion volume.

Grzegorz Kondrak, David Beck, and Philip Dilts. 2007. Creating a comparative dictionary of Totonac-Tepehua. In *Proceedings of the ACL Workshop on Computing and Historical Phonology (9th Meeting of SIGMORPHON)*, pages 134–141.

Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of NAACL 2000: 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 288–295.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Proceedings of NAACL 2001: 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 103–110.

Grzegorz Kondrak. 2002a. *Algorithms for Language Reconstruction*. Ph.D. thesis, University of Toronto.

Grzegorz Kondrak. 2002b. Determining recurrent sound correspondences by inducing translation models. In *Proceedings of COLING 2002: 19th International Conference on Computational Linguistics*, pages 488–494.

Grzegorz Kondrak. 2003a. Identifying complex sound correspondences in bilingual wordlists. In *Proceedings of CICLing 2003: 4th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 432–443.

Grzegorz Kondrak. 2003b. Phonetic alignment and similarity. *Computers and the Humanities*, 37(3):273–291.

Grzegorz Kondrak. 2004. Combining evidence in cognate identification. In *Proceedings of Canadian AI 2004: 17th Conference of the Canadian Society for Computational Studies of Intelligence*, pages 44–59.

Grzegorz Kondrak. 2005a. Cognates and word alignment in bitexts. In *Proceedings of MT Summit X: the Tenth Machine Translation Summit*, pages 305–312.

Grzegorz Kondrak. 2005b. N-gram similarity and distance. In *Proceedings of SPIRE: the 12th International Conference on String Processing and Information Retrieval*, pages 115–126.

Grzegorz Kondrak. 2009. Identification of cognates and recurrent sound correspondences in word lists. *Traitement automatique des langues et langues anciennes*, 50(2):201–235, October.

Wesley Mackay and Grzegorz Kondrak. 2005. Computing word similarity and identifying cognates with Pair Hidden Markov Models. In *Proceedings of CoNLL-2005: 9th Conference on Computational Natural Language Learning*, pages 40–47.

Tarek Sherif and Grzegorz Kondrak. 2007. Bootstrapping a stochastic transducer for arabic-english transliteration extraction. In *Proceedings of ACL 2007: 45th Annual Meeting of the Association for Computational Linguistics*, pages 864–871.