Cognates and Word Alignment in Bitexts

Grzegorz Kondrak
Department of Computing Science
University of Alberta
221 Athabasca Hall
Edmonton, AB, Canada T6G 2E8
kondrak@cs.ualberta.edu

Abstract

We evaluate several orthographic word similarity measures in the context of bitext word alignment. We investigate the relationship between the length of the words and the length of their longest common subsequence. We present an alternative to the longest common subsequence ratio (LCSR), a widely-used orthographic word similarity measure. Experiments involving identification of cognates in bitexts suggest that the alternative method outperforms LCSR. Our results also indicate that alignment links can be used as a substitute for cognates for the purpose of evaluating word similarity measures.

1 Introduction

It has been shown that the quality of word alignment in bitexts can be improved if the actual orthographic form of words is considered (Kondrak et al., 2003). Words that look or sound similar are more likely to be mutual translations than words that exhibit no similarity. The explanation of of this phenomenon lies in the fact that orthographic similarity is a hallmark of cognates. Because of their common origin, cognates normally coincide both in form *and* meaning.

The concept of cognates is largely language-independent. In the context of machine translation, cognates encompass not only genetic cognates (e.g. *name/nom*), but also borrowings (e.g. *computer/komputer*) and proper names. Even non-lexical types, such as numbers and punctuation, are sometimes included as well. While genetic cognates occur only between related languages, other kinds of cognates can be found in virtually any bitext. If languages use different scripts, the identification of cognates must be preceded by a transliteration or transcription process.

In the context of bitexts, cognates have been employed in several bitext-related tasks, including sentence alignment (Simard et al., 1992; Church, 1993; McEnery and Oakes, 1996; Melamed, 1999), inducing translation lexicons (Mann and Yarowsky, 2001; Koehn and Knight, 2001), and improving statistical machine translation models (Al-Onaizan et al., 1999; Kondrak et al., 2003). All those applications depend on an effective method of identifying cognates, which performs a well-defined task: given two words from different languages, compute a numerical score reflecting the likelihood that the words are cognates.

In this paper, we focus on identifying cognates on the basis of their orthographic similarity in the context of word alignment. So far, few comparisons of similarity measures have been published (Brew and McKelvie, 1996; McEnery and Oakes, 1996), We evaluate several measures of orthographic similarity, with the emphasis on the measures based on computing the longest common subsequence length.

Many word similarity/distance measures (including the ones based on letter n-grams, the longest common subsequence, edit distance, and Hidden Markov Models) compute an overall score that is biased towards shorter or longer words. A solution that is often adopted is to normalize the score by the average or the maximum of word lengths (Brew and McKelvie, 1996; Nerbonne and Heeringa, 1997; Melamed, 1999). Although such a straightforward

normalization method is preferable to using an unnormalized score, it is not necessarily optimal. In this paper, we investigate the possibility of deriving an alternative normalization formula by analyzing the behaviour of randomly generated strings. We present the results of experiments involving identification of cognates in bitexts that suggest that the alternative normalization method performs better than the straightforward normalization. Although our focus is on LCS-based measures, a similar approach could be applicable to other types of measures as well.

Although cognates are well suited for the purpose of evaluating word similarity measures, manual identification of cognates is expensive. Can manually or automatically aligned corpora be used as a substitute for cognates? Another objective of our experiments with bitexts is to investigate how closely the cognation relationship correlates with word alignment links.

2 Orthographic similarity

The simplest methods of identifying cognates on the basis of orthographic similarity are binary. A baseline-type approach is to accept as cognates only words that are identical (henceforth referred to as IDENT). Simard (1992) employs a slightly more flexible condition: the identity of the first four letters. The method can be generalized to yield a non-binary coefficient in the $[0, 1]$ range by dividing the length of the longest common prefix by the length of the longer of the two words (henceforth PREFIX). For example, applying PREFIX to *colour* and *couleur* yields $\frac{2}{7}$, because their longest common prefix is "co-".

Dice's similarity coefficient (DICE), originally developed for the comparison of biological specimens, was first used to compare words by Adamson and Boreham (1974). DICE is determined by the ratio of the number of shared character bigrams to the total number of bigrams in both words:

$$DICE(w_1, w_2) = \frac{2 \cdot |bigrams(w_1) \cap bigrams(w_2)|}{|bigrams(w_1)| + |bigrams(w_2)|}$$

where *bigrams(w)* is a multi-set of character bigrams in w. For example, applying DICE to *colour* and *couleur* yields $\frac{6}{11}$ because three bigrams are shared: *co*, *ou*, and *ur*.

Melamed (1999) detects orthographic cognates by thresholding the Longest Common Subsequence Ratio (LCSR). The LCSR of two words is computed by dividing the length of their longest common subsequence (LCS) by the length of the longer word:

$$LCSR(w_1, w_2) = \frac{|LCS(w_1, w_2)|}{max(|w_1|, |w_2|)}.$$

For example, LCSR(*colour*,*couleur*) = $\frac{5}{7}$, as their longest common subsequence is "c-o-l-u-r". Brew and McKelvie (1996) propose a variation in which the denominator is the average of both word lengths.

The orthographic measures described in this section disregard the fact that alphabetic symbols express actual sounds, instead employing a binary identity function on the level of character comparison. While such measures seem to be preferred in practice, a number of more complex approaches have been proposed, including phonetic-based methods, which take advantage of the phonetic characteristics of individual sounds in order to estimate their similarity (Kessler, 1995; Nerbonne and Heeringa, 1997; Kondrak, 2000), and HMM-based methods, which build a similarity model on the basis of training data (Mann and Yarowsky, 2001; Mackay and Kondrak, 2005). Although such methods fall outside the scope of this paper, the problem of length normalization applies to them as well.

3 Normalization

In this section, we investigate the relationship between the similarity score and the word length, focusing on the methods based on the length of the longest common subsequence. We chose LCSR not only because it's a method of choice in several bitext-oriented papers (Melamed, 1999; Brew and McKelvie, 1996; Tiedemann, 1999), but also because of its particularly transparent way of computing the similarity score.

The reason for dividing the LCS length by the word length is to avoid bias towards longer words. For example, the length of the LCS is 4 in both *ideas/idées* and *vegetables/victimes*. Obviously the former is more orthographically similar. However, the division by word length introduces an opposite bias, albeit less pronounced, towards shorter words. For example, the LCSR of both *saw/osa*

and *jacinth/hyacinthe* is $\frac{2}{3}$. Moreover, when the words being compared are completely identical, the LCSR score is always one, regardless of their length. This seems counter-intuitive because longer identical words are more likely to be related. We would like to have a principled method to take length of words into account, without bias in either direction.

A variety of methods have been proposed for a related problem of computing the statistical significance of alignment scores in bioinformatics (Waterman and Vingron, 1994). However, those methods are approximate in nature, and since they were designed for long DNA and protein sequences, they are not directly applicable to short strings like words.

One possible approach would be to compare the likelihood predicted by two different models: one for the related (cognate) words, and another for unrelated (random) words. The random model is described in the following section. For the cognate model, we could assume that two words originate from a single ancestor word, and that the letters in either word can independently change or disappear with a given probability, which could be estimated empirically from cognate training data. However, our experiments suggest that the inclusion of the cognate model does not improve the overall cognate identification accuracy. Since our objective is the ranking of the candidate pairs rather than binary classification, we focus exclusively on the random model.

3.1 The random model

In order to derive a normalization formula, we consider the probability of seeing a given number of matches among a given number of letters in a random model. We adopt a simplifying assumption that letters are randomly and independently generated from a given distribution of letter frequencies. Let p be the probability of a match of two randomly selected letters. For the uniform distribution of letters $p = \frac{1}{t}$, where t is the size of the alphabet. Otherwise, $p = \sum_{i=1}^{t} p_i^2$, where p_i is the probability of the i-th letter occurring in a string, which can be estimated empirically from data. We are interested in estimating the probability that the longest common subsequence of two strings of length m and n has the length of k, denoted as $P(L_{m,n} = k)$. Without loss of generality, we assume that $m \leq n$.

It is relatively easy to come up with the exact probability for some special cases. If $k = m = n$, the probability clearly is equal to p^n. Assuming a uniform letter distribution, the following formula gives exact probability value for $k = 0$:

$$P(L_{m,n} = 0) = \frac{1}{t^{n+m}} \sum_{i=1}^{\max(n,t)} \binom{t}{i} S(n,i)(t-i)^m$$

where t is the alphabet size, and $S(n,i)$ is the Stirling number of the second kind (the number of ways of partitioning a set of n elements into i nonempty sets). The formula is derived by dividing the number of string pairs of length n and m that have no common letters by the total number of string pairs of length n and m. Unfortunately, the case of $k = 0$ is of little use for the purpose of cognate identification.

It is unlikely that an exact analytical formula for the LCS distribution exists in the general case. Such a formula would provide the precise value of the so-called Chvatal-Sankoff constant (Chvatal and Sankoff, 1975), which has been an open problem for three decades. The exact expected value and the variance of the distribution are not known either.

In the absence of an exact formula, one possible approach is to estimate the probability distribution by sampling. However, sampling may not be reliable for the probability values that are very small. Unfortunately, this includes cases when the number of matches is close to the word length, which are the most interesting ones from the point of view of cognate identification. Furthermore, sampling results are specific for a given alphabet size and letter frequencies. Therefore, we focus our efforts on finding an approximate formula for $P(L_{m,n} = k)$.

3.2 Approximation

The first of two formulas that we propose in this section aims at providing a lower bound for the probability $P(L_{m,n} < k)$ by making a simplifying independence assumption.[1] There are exactly $\binom{n}{k}\binom{m}{k}$ possible pairs of subsequences of length k. The probability that there is a mismatch between a pair of randomly selected subsequences of length k is

[1]The formula was suggested by Daniel J. Lizotte (private communication).

$1 - p^k$. The simplifying assumption is that the mismatches between pairs of subsequences are independent of each other. Then, for $1 \leq k \leq m$,

$$P(L_{m,n} < k) \geq (1 - p^k)^{\binom{m}{k}\binom{n}{k}}$$

If we adopt the above lower bound as our approximation, then, for $1 \leq k < m$,

$$\begin{aligned} P(L_{m,n} = k) = & \quad (1) \\ = & \quad P(L_{m,n} < k+1) - P(L_{m,n} < k) \\ \approx & \quad (1 - p^{k+1})^{\binom{m}{k+1}\binom{n}{k+1}} - (1 - p^k)^{\binom{m}{k}\binom{n}{k}} \end{aligned}$$

In particular,

$$P(L_{m,n} = 0) = P(L_{m,n} < 1) \geq (1 - p)^{mn}$$

and

$$P(L_{n,n} = n) = 1 - P(L_{m,n} < n) = p^n$$

Formula 1 it is not exact in general. For example, for $k = 0$ and $t = n = m = 2$, formula 1 predicts $\frac{1}{16}$, whereas the formula given in Section 3.1 correctly yields $\frac{1}{8}$.

The second formula calculates the expected number ET_k of pairs of k-letter subsequences, one from each of the two words, that are identical.[2] The probability that a pair of randomly selected sequences of length k match perfectly is p^k, and there are $\binom{n}{k}\binom{m}{k}$ such pairs. Therefore,

$$ET_k = \binom{n}{k}\binom{m}{k}p^k \quad (2)$$

When ET_k is very small, we may take its value as an approximation of $P(L_{m,n} = k)$. In particular, $ET_n = p^n = P(L_{n,n} = n)$, for $m = n$.

For values of k that are close to $\max(m, n)$, both formulas are good approximations to $P(L_{m,n} = k)$, and return very similar values. However, as the k/n ratio decreases, the approximations become less precise until they become completely unreliable. For certain small ratios, formula 1 actually produces negative values, while formula 2 yields values exceeding 1. However, since our objective is cognate identification, we are mainly interested in words that exhibit higher ratios of k/n.

[2]The second formula was suggested by Robert B. Israel (private communication). Steele (1982) proposed a similar formula as an alternative to the length of the longest common subsequence for measuring genetic proximity of DNA strings.

3.3 An alternative similarity measure

Either of the formulas porpoised in Section 3.2 can serve a basis of a similarity measure. In this section, we define a similarity measure based on formula 2. We chose formula 2 because it is simpler and appears more robust against numerical underflow problems in the case of very long words.

Recall that $LCSR(w_1, w_2) = k/n$, where $k = |LCS(w_1, w_2)|$ and $n = max(|w_1|, |w_2|)$. We define a new measure celled the Longest Common Subsequence Formula (LCSF):

$$LCSF(w_1, w_2) = max(-log(\binom{n}{k}\binom{n}{k}p^k), 0)$$

In the above definition, the length m of the shorter word, which appears in formula 2, is replaced by the length n of the longer word. This introduces a desirable bias against words of different lengths, which are less likely to represent cognates.

In practice, computing LCSF is as fast as LCSR. Since p is constant, the values of LCSF depend only on k and n. Using the dynamic programming principle, the values can be computed incrementally and stored in a two-dimensional array. The same array is then re-used for calculating the similarity of all pairs of words.

4 Evaluation

In order to objectively evaluate similarity measures, we need a gold standard that classifies pairs of words as either similar or dissimilar. However, word similarity is not a binary notion. At most, we can say that some word pairs are more similar than others, but such a judgment is necessarily subjective. Instead, we can use the notion of *cognation* as a substitute for similarity. Cognation is a binary notion, which in the vast majority of cases can be objectively established by human experts. Our assumption is that most cross-language cognates are orthographically similar mutual translations, and vice-versa.

4.1 Cognates vs. alignment links

Manual identification of cognates, although feasible for small data sets, is expensive. Another possibility is using bitext word alignment as the gold standard. Although only a fraction of alignments correspond to orthographically similar words, we hypoth-

esize that the proportion of orthographically similar words among the words that are aligned is much higher than among words that are not aligned. The confirmation of the above hypothesis was one of the objectives of our experiments. Another objective of our experiments was to investigate whether manually aligned bitexts can be substituted with automatically aligned bitexts for the purpose of evaluating word similarity measures. The former are relatively small and expensive to create, while the latter are easily available, but have lower alignment accuracy.

It must be noted that the relationship between cognates co-occurring in aligned sentences and word alignment links is not completely straightforward. First of all, the majority of aligned tokens are not cognates — their surface forms provide no clue that they are in fact translations. When word alignment links are used to evaluate similarity measures, such pairs constitute the majority of false negatives. Fortunately, this problem uniformly affects all measures. On the other hand, not all co-occurring cognates correspond to word alignment links. Quite often, a word is cognate with several words in the corresponding sentence, but it is correctly aligned with just one of them. This phenomenon occurs for instance when there are multiple occurrences of the same word within a sentence. The resulting false positives make it virtually impossible to achieve 100% precision even at low recall levels.

4.2 Evaluation methodology

Our approach to evaluating a word similarity measure on the basis of a word aligned bitext is to treat each set of aligned sentences as bags of words, and compute the similarity of each possible pairing of words. Each pair is evaluated against a gold standard, which is either a list of cognate pairs or a list of alignment links. If the method is binary, it will divide the set of word pairings into likely cognates and unlikely cognates. However, most of the methods do not produce a binary cognation decision. This provides a flexibility of adapting the similarity threshold to the precision required by a specific application. In such a case, the pairs are sorted in the descending order of their similarity value. The true positives should be dense at the top of the list and become less frequent as we move down the list. We measure precision at various cutoff levels in the sorted list. If the total number of true positives in the bitext is known, we set the cutoff levels to correspond to specific recall levels; otherwise, we set them in relation to absolute numbers of true positives.

4.3 Data

In our experiments, we used three different, manually aligned bitexts (Table 1). They contain on average between two and five cognate pairs per sentence.

	Blinker	Hansards	Romanian
Sentences	250	500	248
Tokens (English)	7510	7937	5638
Tokens (other)	8191	8740	5495
Cognates	967	1080	1216
Alignment links	10097	4435	6201

Table 1: Breakdown of the bitexts used in the experiments.

The Blinker bitext (Melamed, 1998) is a word-aligned French-English bitext containing 250 Bible verse pairs. The bitext is somewhat unusual for several reasons: it is not a continuous text, both parts are translations from a third language, and its style is not representative of modern usage. Nevertheless, the alignments are of high quality, produced by several annotators. In our experiments, we used the alignments of a single annotator (A1).

We manually identified all cognate word pairs in the Blinker bitext. The cognation judgments were made on the basis of information in etymological dictionaries. For words to be classified as cognate, their roots, not just affixes, must share a common origin. In the case of compound words with multiple roots, a single cognate root was considered sufficient. In total, there are 967 cognate pairs, which can be classified as genetic cognates (10%), borrowings (57%), and proper names (43%). 84% of the cognate pairs correspond to word alignment links. The total number of *distinct* cognate pairs is 584.

The second bitext used in our experiments is a manually aligned sample from the Hansards - proceedings of the Canadian parliament (Och and Ney, 2000b). The alignment links are classified either as sure (S) or probable (P). In our experiments we used only the S links. The number of cognates in the bi-

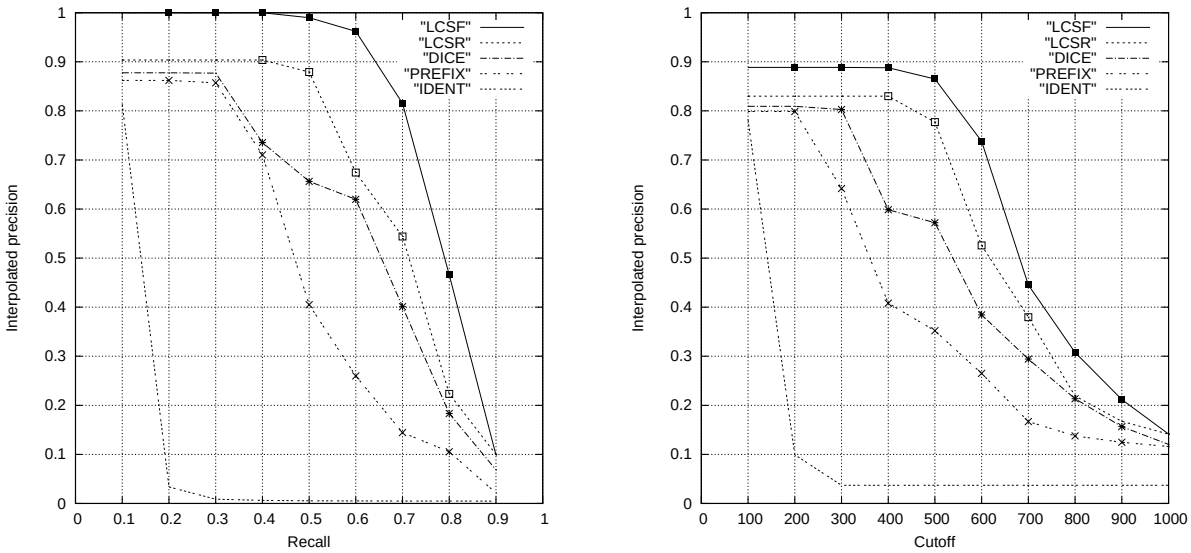

Figure 1: Performance of various similarity measures on the Blinker bitext evaluated against actual cognate pairs (left), and manually annotated word alignment links (right).

text was estimated by counting all cognate pairs in 25 randomly selected sentences, and extrapolating the result.

The third bitext is a manually aligned Romanian-English corpus containing newspaper-style text (Mihalcea and Pedersen, 2003). The estimate of the number of cognates in the bitext was based on 31 randomly selected sentences.

4.4 Results

Figures 1, 2, and 3 contain plots that compare the interpolated precision values corresponding to various measures on the Blinker, Hansards, and the Romanian-English bitexts, respectively. In all three cases, we calculated the precision against a list of manually identified word alignment links. In addition, we also calculated the precision against a complete list of cognate pairs in the Blinker bitext, and against a list of machine-generated links obtained with the Giza statistical machine translation package (Och and Ney, 2000a) in the Hansards sample. Since orthographic similarity by itself can only identify a fraction of word alignment links, we used fixed cutoff levels instead of recall in all plots except the first one. The cutoff levels correspond to absolute numbers of correctly identified alignment links.

We performed statistical significance tests for all pairs of similarity measures at various cutoff levels. Following the method proposed by Evert (2004), we applied Fisher's exact test to counts of word pairs that are accepted by only one of two similarity measures. The results of the significance tests are incorporated into the plots in a compact way: the points that are superimposed on the plot curves indicate cases where the corresponding measure achieves a significantly higher precision than the measure immediately below it (at 95% confidence level). As expected, the minimum spread necessary for statistical significance decreases as the number of true positives increases.

There are three distinct phases that can be identified in the plots. The first phase, characterized by relatively high precision, corresponds to the set of easily identifiable cognate pairs and proper names. In the second phase, there is a perceivable drop in precision corresponding to the set of cognate pairs that are more difficult to distinguish from accidental similarities. Finally, in the third phase, almost all cognates have been identified, and the measures converge to a random baseline level. This third phase is nearly absent from the first plot of Figure 1 because its x axis corresponds to cognate pairs rather than to word alignment links.

The results imply a fairly consistent ranking of the tested similarity measures. The IDENT measure almost immediately reaches the random base-

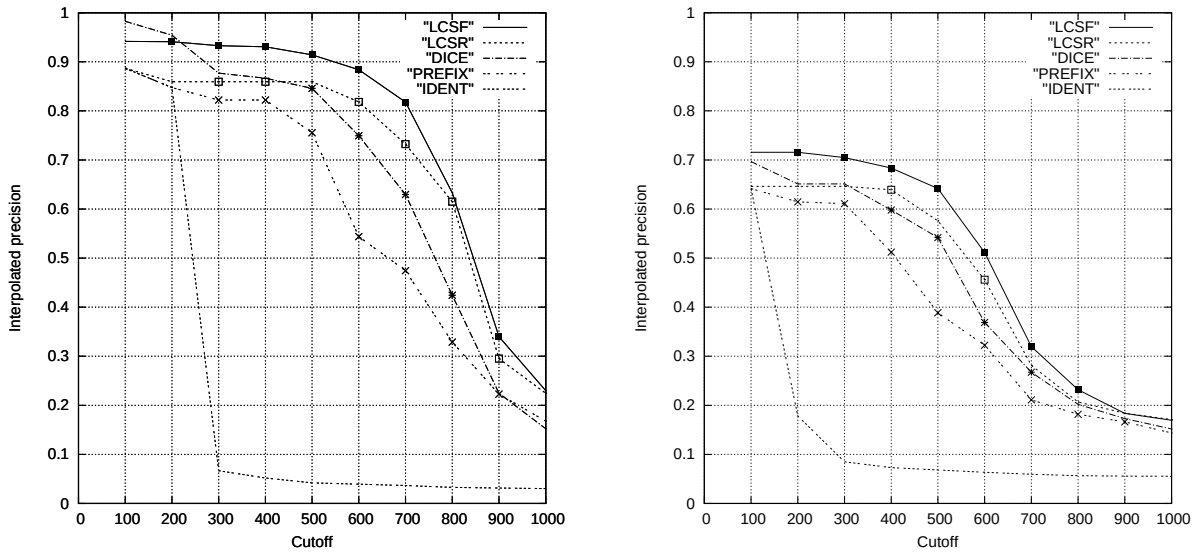

Figure 2: Performance of various similarity measures on the Hansards sample evaluated against manually annotated (left), and automatically generated (right) word alignment links.

line level. PREFIX is comparable to other measures during the first phase, but then its performance drops off sharply. Although LCSR has worse (non-linear) time complexity than DICE, in many cases it is not significantly better, and is actually worse on the Romanian-English corpus. LCSF, the newly proposed measure based on the longest common subsequence length, consistently outperforms LCSR, and the differences are generally statistically significant, except in the third phase.

Apart from the comparison of specific measures, the overall concordance between both pairs of plots in Figures 1 and 2 suggests that for the purpose of evaluating word similarity measures, cognate links can be substituted with word alignment links, even when the links are automatically generated.

5 Conclusion

We have presented an alternative to the longest common subsequence ratio (LCSR), a widely-used orthographic word similarity measure. Experiments involving identification of cognates in bitexts suggest that the alternative method outperforms LCSR. We have also evaluated several other orthographic word similarity measures in the context of bitext word alignment. Our results indicate that alignment links could be used as a substitute for cognates for the purpose of evaluating word similarity measures.

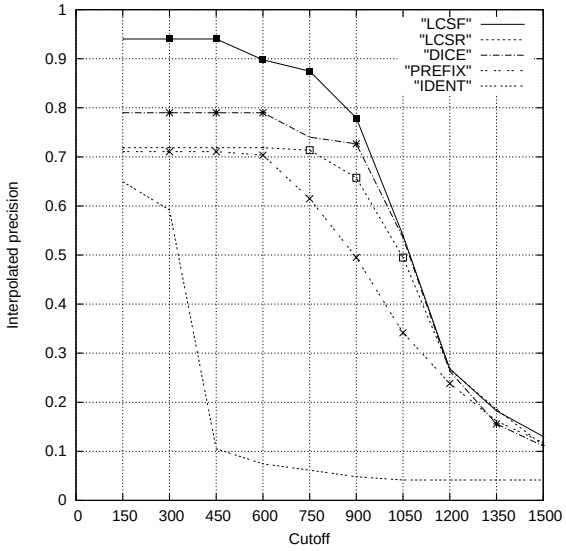

Figure 3: Performance of various similarity measures on the Romanian-English bitext evaluated against manually annotated word alignment links.

In the future, we would like to extend our length normalization approach to edit distance and other similarity/distance measures. We also plan to experiment with incorporating cognate identification directly into statistical machine translation models as an additional feature function within the maximum entropy framework (Och and Ney, 2002).

Acknowledgments

Thanks to Colin Cherry, Robert B. Israel, and Daniel J. Lizotte for their comments and suggestions. This research was supported by Natural Sciences and Engineering Research Council of Canada.

References

George W. Adamson and Jillian Boreham. 1974. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, 10:253–260.

Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. 1999. Statistical machine translation. Technical report, Johns Hopkins University.

Chris Brew and David McKelvie. 1996. Word-pair extraction for lexicography. In *Proceedings of the 2nd International Conference on New Methods in Language Processing*, pages 45–55.

Kenneth W. Church. 1993. Char_align: A program for aligning parallel texts at the character level. In *Proceedings of ACL-93*, pages 1–8.

Vaclav Chvatal and David Sankoff. 1975. Longest common subsequences of two random sequences. *Journal of Applied Probability*, 12:306–315.

Stefan Evert. 2004. Significance tests for the evaluation of ranking methods. In *Proceedings of COLING-2004*, pages 945–951.

Brett Kessler. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of EACL-95*, pages 60–67.

Philipp Koehn and Kevin Knight. 2001. Knowledge sources for word-level translation models. In *Proceedings of EMNLP-2001*, pages 27–35.

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

Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of NAACL 2000*, pages 288–295.

Wesley Mackay and Grzegorz Kondrak. 2005. Computing word similarity and identifying cognates with Pair Hidden Markov Models. In *Proceedings of CoNLL-2005*, pages 40–47.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL 2001*, pages 151–158.

Tony McEnery and Michael Oakes. 1996. Sentence and word alignment in the CRATER Project. In J. Thomas and M. Short, editors, *Using Corpora for Language Research*, pages 211–231. Longman.

I. Dan Melamed. 1998. Manual annotation of translational equivalence: The Blinker project. Technical Report IRCS #98-07, University of Pennsylvania.

I. Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT/NAACL Workshop on Building and Using Parallel Texts*.

John Nerbonne and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Proceedings of SIGPHON-97: 3rd Meeting of the ACL Special Interest Group in Computational Phonology*.

Franz Josef Och and Hermann Ney. 2000a. A comparison of alignment models for statistical machine translation. In *Proceedings of COLING 2000*.

Franz Josef Och and Hermann Ney. 2000b. Improved statistical alignment models. In *Proceedings of ACL-2000*, pages 440–447.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL-2002*, pages 295–302.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81.

J. Michael Steele. 1982. Long common subsequences and the proximity of two random strings. *SIAM Journal on Applied Mathematics*, 42(4):731–737.

Jörg Tiedemann. 1999. Automatic construction of weighted string similarity measures. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Michael S. Waterman and Martin Vingron. 1994. Sequence comparison significance and Poisson approximation. *Statistical Sciences*, 9:367–381.