Alignment-Based Discriminative String Similarity

Shane Bergsma and Grzegorz Kondrak
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada, T6G 2E8
{bergsma,kondrak}@cs.ualberta.ca

Abstract

A character-based measure of similarity is an important component of many natural language processing systems, including approaches to transliteration, coreference, word alignment, spelling correction, and the identification of cognates in related vocabularies. We propose an alignment-based *discriminative* framework for string similarity. We gather features from substring pairs consistent with a character-based alignment of the two strings. This approach achieves exceptional performance; on nine separate cognate identification experiments using six language pairs, we more than double the precision of traditional orthographic measures like Longest Common Subsequence Ratio and Dice's Coefficient. We also show strong improvements over other recent discriminative and heuristic similarity functions.

1 Introduction

String similarity is often used as a means of quantifying the likelihood that two pairs of strings have the same underlying meaning, based purely on the character composition of the two words. Strube et al. (2002) use Edit Distance as a feature for determining if two words are coreferent. Taskar et al. (2005) use French-English common letter sequences as a feature for discriminative word alignment in bilingual texts. Brill and Moore (2000) learn misspelled-word to correctly-spelled-word similarities for spelling correction. In each of these examples, a similarity measure can make use of the recurrent substring pairings that reliably occur between words having the same meaning.

Across natural languages, these recurrent substring correspondences are found in word pairs known as cognates: words with a common form and meaning across languages. Cognates arise either from words in a common ancestor language (e.g. *light/Licht, night/Nacht* in English/German) or from foreign word borrowings (e.g. *trampoline/toranporin* in English/Japanese). Knowledge of cognates is useful for a number of applications, including sentence alignment (Melamed, 1999) and learning translation lexicons (Mann and Yarowsky, 2001; Koehn and Knight, 2002).

We propose an alignment-based, discriminative approach to string similarity and evaluate this approach on cognate identification. Section 2 describes previous approaches and their limitations. In Section 3, we explain our technique for automatically creating a cognate-identification training set. A novel aspect of this set is the inclusion of *competitive counter-examples* for learning. Section 4 shows how discriminative features are created from a character-based, minimum-edit-distance alignment of a pair of strings. In Section 5, we describe our bitext and dictionary-based experiments on six language pairs, including three based on non-Roman alphabets. In Section 6, we show significant improvements over traditional approaches, as well as significant gains over more recent techniques by Ristad and Yianilos (1998), Tiedemann (1999), Kondrak (2005), and Klementiev and Roth (2006).

2 Related Work

String similarity is a fundamental concept in a variety of fields and hence a range of techniques

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 656–663,
Prague, Czech Republic, June 2007.

have been developed. We focus on approaches that have been applied to words, i.e., uninterrupted sequences of characters found in natural language text. The most well-known measure of the similarity of two strings is the Edit Distance or Levenshtein Distance (Levenshtein, 1966): the number of insertions, deletions and substitutions required to transform one string into another. In our experiments, we use Normalized Edit Distance (NED): Edit Distance divided by the length of the longer word. Other popular measures include Dice's Coefficient (DICE) (Adamson and Boreham, 1974), and the length-normalized measures Longest Common Subsequence Ratio (LCSR) (Melamed, 1999), and Longest Common Prefix Ratio (PREFIX) (Kondrak, 2005). These baseline approaches have the important advantage of not requiring training data. We can also include in the non-learning category Kondrak (2005)'s Longest Common Subsequence Formula (LCSF), a probabilistic measure designed to mitigate LCSR's preference for shorter words.

Although simple to use, the untrained measures cannot adapt to the specific spelling differences between a pair of languages. Researchers have therefore investigated adaptive measures that are learned from a set of known cognate pairs. Ristad and Yianilos (1998) developed a stochastic transducer version of Edit Distance learned from unaligned string pairs. Mann and Yarowsky (2001) saw little improvement over Edit Distance when applying this transducer to cognates, even when filtering the transducer's probabilities into different weight classes to better approximate Edit Distance. Tiedemann (1999) used various measures to learn the recurrent spelling changes between English and Swedish, and used these changes to re-weight LCSR to identify more cognates, with modest performance improvements. Mulloni and Pekar (2006) developed a similar technique to improve NED for English/German.

Essentially, all these techniques improve on the baseline approaches by using a set of positive (true) cognate pairs to re-weight the costs of edit operations or the score of sequence matches. Ideally, we would prefer a more flexible approach that can learn positive *or* negative weights on *substring* pairings in order to better identify related strings. One system that can potentially provide this flexibility is a discriminative string-similarity approach to named-entity transliteration by Klementiev and Roth (2006). Although not compared to other similarity measures in the original paper, we show that this discriminative technique can strongly outperform traditional methods on cognate identification.

Unlike many recent generative systems, the Klementiev and Roth approach does not exploit the known positions in the strings where the characters match. For example, Brill and Moore (2000) combine a character-based alignment with the Expectation Maximization (EM) algorithm to develop an improved probabilistic error model for spelling correction. Rappoport and Levent-Levi (2006) apply this approach to learn substring correspondences for cognates. Zelenko and Aone (2006) recently showed a Klementiev and Roth (2006)-style discriminative approach to be superior to alignment-based generative techniques for name transliteration. Our work successfully uses the alignment-based methodology of the generative approaches to enhance the feature set for discriminative string similarity.

3 The Cognate Identification Task

Given two string lists, E and F, the task of cognate identification is to find all pairs of strings (e, f) that are cognate. In other similarity-driven applications, E and F could be misspelled and correctly spelled words, or the orthographic and the phonetic representation of words, etc. The task remains to link strings with common meaning in E and F using only the string similarity measure.

We can facilitate the application of string similarity to cognates by using a definition of cognation not dependent on etymological analysis. For example, Mann and Yarowsky (2001) define a word pair (e, f) to be cognate if they are a translation pair (same meaning) and their Edit Distance is less than three (same form). We adopt an improved definition (suggested by Melamed (1999) for the French-English Canadian Hansards) that does not over-propose shorter word pairs: (e, f) are cognate if they are translations and their LCSR $\geq$ 0.58. Note that this cutoff is somewhat conservative: the English/German cognates *light/Licht* (LCSR=0.8) are included, but not the cognates *eight/acht* (LCSR=0.4).

If two words must have LCSR ≥ 0.58 to be cog-

Foreign Language F	Words $f \in F$	Cognates E_{f+}	False Friends E_{f-}
Japanese (Rômaji)	napukin	napkin	nanking, pumpkin, snacking, sneaking
French	abondamment	abundantly	abandonment, abatement, ... wonderment
German	prozyklische	procyclical	polished, prophylactic, prophylaxis

Table 1: Foreign-English cognates and false friend training examples.

nate, then for a given word $f \in F$, we need only consider as possible cognates the subset of words in E having an LCSR with f larger than 0.58, a set we call E_f. The portion of E_f with the same meaning as f, E_{f+}, are cognates, while the part with different meanings, E_{f-}, are not cognates. The words E_{f-} with similar spelling but different meaning are sometimes called *false friends*. The cognate identification task is, for every word $f \in F$, and a list of similarly spelled words E_f, to distinguish the cognate subset E_{f+} from the false friend set E_{f-}.

To create training data for our learning approaches, and to generate a high-quality labelled test set, we need to annotate some of the $(f, e_f \in E_f)$ word pairs for whether or not the words share a common meaning. In Section 5, we explain our two high-precision automatic annotation methods: checking if each pair of words (a) were aligned in a word-aligned bitext, or (b) were listed as translation pairs in a bilingual dictionary.

Table 1 provides some labelled examples with non-empty cognate and false friend lists. Note that despite these examples, this is not a ranking task: even in highly related languages, most words in F have empty E_{f+} lists, and many have empty E_{f-} as well. Thus one natural formulation for cognate identification is a pairwise (and symmetric) cognation classification that looks at each pair (f, e_f) separately and individually makes a decision:

+(napukin,napkin)
– (napukin,nanking)
– (napukin,pumpkin)

In this formulation, the benefits of a discriminative approach are clear: it must find substrings that distinguish cognate pairs from word pairs with otherwise similar form. Klementiev and Roth (2006), although using a discriminative approach, do not provide their infinite-attribute perceptron with competitive counter-examples. They instead use transliterations as positives and randomly-paired English and Russian words as negative examples. In the following section, we also improve on Klementiev and Roth (2006) by using a character-based string alignment to focus the features for discrimination.

4 Features for Discriminative Similarity

Discriminative learning works by providing a training set of labelled examples, each represented as a set of features, to a module that learns a classifier. In the previous section we showed how labelled word pairs can be collected. We now address methods of representing these word pairs as sets of features useful for determining cognation.

Consider the Rômaji Japanese/English cognates: *(sutoresu,stress)*. The LCSR is 0.625. Note that the LCSR of *sutoresu* with the English false friend *stories* is higher: 0.75. LCSR alone is too weak a feature to pick out cognates. We need to look at the actual character substrings.

Klementiev and Roth (2006) generate features for a pair of words by splitting both words into all possible substrings of up to size two:

sutoresu $\Rightarrow$ { *s, u, t, o, r, e, s, u, su, ut, to, ... su* }
stress $\Rightarrow$ { *s, t, r, e, s, s, st, tr, re, es, ss* }

Then, a feature vector is built from all substring pairs from the two words such that the difference in positions of the substrings is within one:

{*s-s, s-t, s-st, su-s, su-t, su-st, su-tr... r-s, r-s, r-es...*}

This feature vector provides the feature representation used in supervised machine learning.

This example also highlights the limitations of the Klementiev and Roth approach. The learner can provide weight to features like *s-s* or *s-st* at the beginning of the word, but because of the gradual accumulation of positional differences, the learner never sees the *tor-tr* and *es-es* correspondences that really help indicate the words are cognate.

Our solution is to use the minimum-edit-distance alignment of the two strings as the basis for feature extraction, rather than the positional correspondences. We also include beginning-of-word (ˆ) and end-of-word ($) markers (referred to as *boundary* *markers*) to highlight correspondences at those positions. The pair $(sutoresu, stress)$ can be aligned:

```
^ s u t o r e s u $
 \ \  | / / /   /
  ^ s t r e s s $
```

For the feature representation, we only extract substring pairs that are consistent with this alignment.[1] That is, the letters in our pairs can only be aligned to each other and not to letters outside the pairing: { *^-^,^s-^s, s-s, su-s, ut-t, t-t,... es-es, s-s, su-ss...*} We define *phrase* pairs to be the pairs of substrings consistent with the alignment. A similar use of the term "phrase" exists in machine translation, where phrases are often pairs of word sequences consistent with word-based alignments (Koehn et al., 2003).

By limiting the substrings to only those pairs that are consistent with the alignment, we generate fewer, more-informative features. Using more precise features allows a larger maximum substring size L than is feasible with the positional approach. Larger substrings allow us to capture important recurring deletions like the "u" in *sut-st*.

Tiedemann (1999) and others have shown the importance of using the mismatching portions of cognate pairs to learn the recurrent spelling changes between two languages. In order to capture mismatching segments longer than our maximum substring size will allow, we include special features in our representation called *mismatches*. *Mismatches* are phrases that span the entire sequence of unaligned characters between two pairs of aligned end characters (similar to the "rules" extracted by Mulloni and Pekar (2006)). In the above example, *su$-ss$* is a mismatch with "s" and "$" as the aligned end characters. Two sets of features are taken from each mismatch, one that includes the beginning/ending aligned characters as context and one that does not. For example, for the endings of the French/English pair *(économique,economic)*, we include both the substring pairs *ique$:ic$* and *que:c* as features.

One consideration is whether substring features should be binary presence/absence, or the count of the feature in the pair normalized by the length of the longer word. We investigate both of these approaches in our experiments. Also, there is no reason not to include the scores of baseline approaches like NED, LCSR, PREFIX or DICE as features in the representation as well. Features like the lengths of the two words and the difference in lengths of the words have also proved to be useful in preliminary experiments. Semantic features like frequency similarity or contextual similarity might also be included to help determine cognation between words that are not present in a translation lexicon or bitext.

5 Experiments

Section 3 introduced two high-precision methods for generating labelled cognate pairs: using the word alignments from a bilingual corpus or using the entries in a translation lexicon. We investigate both of these methods in our experiments. In each case, we generate sets of labelled word pairs for training, testing, and development. The proportion of positive examples in the bitext-labelled test sets range between 1.4% and 1.8%, while ranging between 1.0% and 1.6% for the dictionary data.[2]

For the discriminative methods, we use a popular Support Vector Machine (SVM) learning package called SVM^{light} (Joachims, 1999). SVMs are maximum-margin classifiers that achieve good performance on a range of tasks. In each case, we learn a linear kernel on the training set pairs and tune the parameter that trades-off training error and margin on the development set. We apply our classifier to the test set and score the pairs by their positive distance from the SVM classification hyperplane (also done by Bilenko and Mooney (2003) with their token-based SVM similarity measure).

We also score the test sets using traditional orthographic similarity measures PREFIX, DICE, LCSR, and NED, an average of these four, and Kondrak (2005)'s LCSF. We also use the log of the edit probability from the stochastic decoder of Ristad and Yianilos (1998) (normalized by the length of the longer word) and Tiedemann (1999)'s highest performing system (Approach #3). Both use only the positive examples in our training set. Our evaluation metric is 11-pt average precision on the score-sorted pair lists (also used by Kondrak and Sherif (2006)).

[1] If the words are from different alphabets, we can get the alignment by mapping the letters to their closest Roman equivalent, or by using the EM algorithm to learn the edits (Ristad and Yianilos, 1998).

[2] The cognate data sets used in our experiments are available at http://www.cs.ualberta.ca/~bergsma/Cognates/

5.1 Bitext Experiments

For the bitext-based annotation, we use publicly-available word alignments from the Europarl corpus, automatically generated by GIZA++ for French-English (Fr), Spanish-English (Es) and German-English (De) (Koehn and Monz, 2006). Initial cleaning of these noisy word pairs is necessary. We thus remove all pairs with numbers, punctuation, a capitalized English word, and all words that occur fewer than ten times. We also remove many incorrectly aligned words by filtering pairs where the pairwise Mutual Information between the words is less than 7.5. This processing leaves vocabulary sizes of 39K for French, 31K for Spanish, and 60K for German.

Our labelled set is then generated from pairs with LCSR $\geq$ 0.58 (using the cutoff from Melamed (1999)). Each labelled set entry is a triple of a) the foreign word f, b) the cognates E_{f+} and c) the false friends E_{f-}. For each language pair, we randomly take 20K triples for training, 5K for development and 5K for testing. Each triple is converted to a set of pairwise examples for learning and classification.

5.2 Dictionary Experiments

For the dictionary-based cognate identification, we use French, Spanish, German, Greek (Gr), Japanese (Jp), and Russian (Rs) to English translation pairs from the Freelang program.[3] The latter three pairs were chosen so that we can evaluate on more distant languages that use non-Roman alphabets (although the Rômaji Japanese is Romanized by definition). We take 10K labelled-set triples for training, 2K for testing and 2K for development.

The baseline approaches and our definition of cognation require comparison in a common alphabet. Thus we use a simple context-free mapping to convert every Russian and Greek character in the word pairs to their nearest Roman equivalent. We then label a translation pair as cognate if the LCSR between the words' Romanized representations is greater than 0.58. We also operate all of our comparison systems on these Romanized pairs.

6 Results

We were interested in whether our working definition of cognation (translations and LCSR $\geq$ 0.58) reflects true etymological relatedness. We looked at the LCSR histogram for translation pairs in one of our translation dictionaries (Figure 1). The trendline suggests a bimodal distribution, with two distinct distributions of translation pairs making up the dictionary: incidental letter agreement gives low LCSR for the larger, non-cognate portion and high LCSR characterizes the likely cognates. A threshold of 0.58 captures most of the cognate distribution while excluding non-cognate pairs. This hypothesis was confirmed by checking the LCSR values of a list of known French-English cognates (randomly collected from a dictionary for another project): 87.4% were above 0.58. We also checked cognation on 100 randomly-sampled, positively-labelled French-English pairs (i.e. translated or aligned and having LCSR $\geq$ 0.58) from both the dictionary and bitext data. 100% of the dictionary pairs and 93% of the bitext pairs were cognate.

Next, we investigate various configurations of the discriminative systems on one of our cognate identification development sets (Table 2). The original Klementiev and Roth (2006) (KR) system can

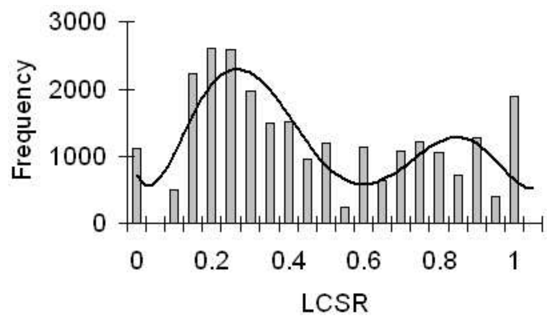

Figure 1: LCSR histogram and polynomial trendline of French-English dictionary pairs.

System	Prec
Klementiev-Roth (KR) L$\leq$2	58.6
KR L$\leq$2 (normalized, boundary markers)	62.9
phrases L$\leq$2	61.0
phrases L$\leq$3	65.1
phrases L$\leq$3 + *mismatches*	65.6
phrases L$\leq$3 + *mismatches* + NED	**65.8**

Table 2: Bitext French-English *development set* cognate identification 11-pt average precision (%).

[3] http://www.freelang.net/dictionary/

	Bitext			Dictionary					
System	Fr	Es	De	Fr	Es	De	Gr	Jp	Rs
PREFIX	34.7	27.3	36.3	45.5	34.7	25.5	28.5	16.1	29.8
DICE	33.7	28.2	33.5	44.3	33.7	21.3	30.6	20.1	33.6
LCSR	34.0	28.7	28.5	48.3	36.5	18.4	30.2	24.2	36.6
NED	36.5	**31.9**	32.3	50.1	**40.3**	23.3	**33.9**	28.2	41.4
PREFIX+DICE+LCSR+NED	**38.7**	31.8	**39.3**	**51.6**	40.1	**28.6**	33.7	22.9	37.9
Kondrak (2005): LCSF	29.8	28.9	29.1	39.9	36.6	25.0	30.5	**33.4**	**45.5**
Ristad & Yanilos (1998)	37.7	32.5	34.6	56.1	46.9	36.9	38.0	52.7	51.8
Tiedemann (1999)	38.8	33.0	34.7	55.3	49.0	24.9	37.6	33.9	45.8
Klementiev & Roth (2006)	61.1	55.5	53.2	73.4	62.3	48.3	51.4	62.0	64.4
Alignment-Based Discriminative	**66.5**	**63.2**	**64.1**	**77.7**	**72.1**	**65.6**	**65.7**	**82.0**	**76.9**

Table 3: Bitext, Dictionary Foreign-to-English cognate identification 11-pt average precision (%).

be improved by normalizing the feature count by the longer string length and including the boundary markers. This is therefore done with all the alignment-based approaches. Also, because of the way its features are constructed, the KR system is limited to a maximum substring length of two ($L \leq 2$). A maximum length of three ($L \leq 3$) in the KR framework produces millions of features and prohibitive training times, while $L \leq 3$ is computationally feasible in the phrasal case, and increases precision by 4.1% over the *phrases* $L \leq 2$ system.[4] Including *mismatches* results in another small boost in performance (0.5%), while using an Edit Distance feature again increases performance by a slight margin (0.2%). This ranking of configurations is consistent across all the bitext-based development sets; we therefore take the configuration of the highest scoring system as our Alignment-Based Discriminative system for the remainder of this paper.

We next compare the Alignment-Based Discriminative scorer to the various other implemented approaches across the three bitext and six dictionary-based cognate identification test sets (Table 3). The table highlights the top system among both the non-adaptive and adaptive similarity scorers.[5] In each language pair, the alignment-based discriminative approach outperforms all other approaches, but the KR system also shows strong gains over non-adaptive techniques and their re-weighted extensions. This is in contrast to previous comparisons which have only demonstrated minor improvements with adaptive over traditional similarity measures (Kondrak and Sherif, 2006).

We consistently found that the original KR performance could be surpassed by a system that normalizes the KR feature count and adds boundary markers. Across all the test sets, this modification results in a 6% average gain in performance over baseline KR, but is still on average 5% below the Alignment-Based Discriminative technique, with a statistically significantly difference on each of the nine sets.[6]

Figure 2 shows the relationship between training data size and performance in our bitext-based French-English data. Note again that the Tiedemann and Ristad & Yanilos systems only use the positive examples in the training data. Our alignment-based similarity function outperforms all the other systems across nearly the entire range of training data. Note also that the discriminative learning curves show no signs of slowing down: performance grows logarithmically from 1K to 846K word pairs.

For insight into the power of our discriminative approach, we provide some of our classifiers' highest and lowest-weighted features (Table 4).

[4]Preliminary experiments using even longer phrases (beyond $L \leq 3$) currently produce a computationally prohibitive number of features for SVM learning. Deploying current feature selection techniques might enable the use of even more expressive and powerful feature sets with longer phrase lengths.

[5]Using the training data and the SVM to weight the components of the PREFIX+DICE+LCSR+NED scorer resulted in negligible improvements over the simple average on our development data.

[6]Following Evert (2004), significance was computed using Fisher's exact test (at $p = 0.05$) to compare the n-best word pairs from the scored test sets, where n was taken as the number of positive pairs in the set.

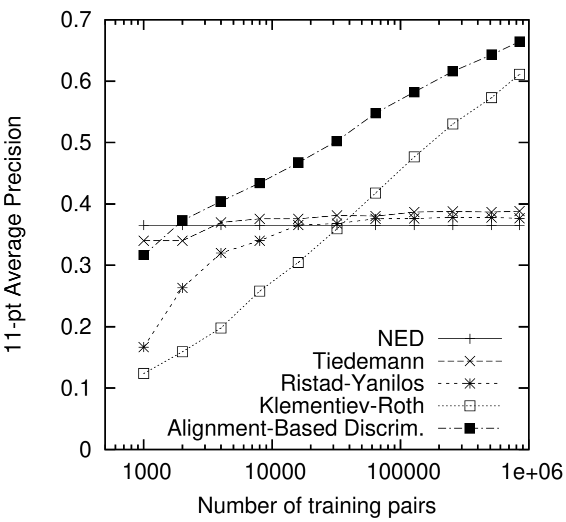

Figure 2: Bitext French-English cognate identification learning curve.

Lang.	Feat.	Wt.	Example
Fr (Bitext)	ées-ed	+8.0	vérifiées:verified
Jp (Dict.)	ru-l	+5.9	penaruti:penalty
De (Bitext)	k-c	+5.5	kreativ:creative
Rs (Dict.)	irov-_	+4.9	motivirovat:motivate
Gr (Dict.)	f-ph	+4.1	symfonia:symphony
Gr (Dict.)	kos-c	+3.3	anarchikos:anarchic
Gr (Dict.)	os$-y$	-2.5	*anarchikos:anarchy*
Jp (Dict.)	ou-ou	-2.6	*handoutai:handout*
Es (Dict.)	_-un	-3.1	*balance:unbalance*
Fr (Dict.)	er$-er$	-5.0	*former:former*
Es (Bitext)	mos-s	-5.1	*toleramos:tolerates*

Table 4: Example features and weights for various Alignment-Based Discriminative classifiers (Foreign-English, negative pairs in *italics*).

Note the expected correspondences between foreign spellings and English (*k-c*, *f-ph*), but also features that leverage derivational and inflectional morphology. For example, Greek-English pairs with the adjective-ending correspondence *kos-c*, e.g. *anarchikos:anarchic*, are favoured, but pairs with the adjective ending in Greek and noun ending in English, *os$-y$*, are penalized; indeed, by our definition, *anarchikos:anarchy* is not cognate. In a bitext, the feature *ées-ed* captures that feminine-plural inflection of past tense verbs in French corresponds to regular past tense in English. On the other hand, words ending in the Spanish first person plural verb suffix *-amos* are rarely translated to English words ending with the suffix *-s*, causing *mos-s* to be penalized. The ability to leverage negative features, learned from appropriate counter examples, is a key innovation of our discriminative framework.

Gr-En (*Dict.*)	Es-En (*Bitext*)
alkali:alkali	agenda:agenda
makaroni:macaroni	natural:natural
adrenalini:adrenaline	márgenes:margins
flamingko:flamingo	hormonal:hormonal
spasmodikos:spasmodic	radón:radon
amvrosia:ambrosia	higiénico:hygienic

Table 5: Highest scored pairs by Alignment-Based Discriminative classifier (negative pairs in *italics*).

Table 5 gives the top pairs scored by our system on two of the sets. Notice that unlike traditional similarity measures that always score identical words higher than all other pairs, by virtue of our feature weighting, our discriminative classifier prefers some pairs with very characteristic spelling changes.

We performed error analysis by looking at all the pairs our system scored quite confidently (highly positive or highly negative similarity), but which were labelled oppositely. Highly-scored false positives arose equally from 1) actual cognates not linked as translations in the data, 2) related words with diverged meanings, e.g. the error in Table 5: *makaroni* in Greek actually means *spaghetti* in English, and 3) the same word stem, a different part of speech (e.g. the Greek/English adjective/noun *synonymos:synonym*). Meanwhile, inspection of the highly-confident false negatives revealed some (often erroneously-aligned in the bitext) positive pairs with incidental letter match (e.g. the French/English *recettes:proceeds*) that we would not actually deem to be cognate. Thus the errors that our system makes are often either linguistically interesting or point out mistakes in our automatically-labelled bitext and (to a lesser extent) dictionary data.

7 Conclusion

This is the first research to apply discriminative string similarity to the task of cognate identification. We have introduced and successfully applied an alignment-based framework for discriminative similarity that consistently demonstrates improved performance in both bitext and dictionary-based cognate identification on six language pairs. Our improved approach can be applied in any of the diverse applications where traditional similarity measures like Edit Distance and LCSR are prevalent. We have also made available our cognate identification data sets, which will be of interest to general string similarity researchers.

Furthermore, we have provided a natural framework for future cognate identification research. Phonetic, semantic, or syntactic features could be included within our discriminative infrastructure to aid in the identification of cognates in text. In particular, we plan to investigate approaches that do not require the bilingual dictionaries or bitexts to generate training data. For example, researchers have automatically developed translation lexicons by seeing if words from each language have similar frequencies, contexts (Koehn and Knight, 2002), burstiness, inverse document frequencies, and date distributions (Schafer and Yarowsky, 2002). Semantic and string similarity might be learned jointly with a co-training or bootstrapping approach (Klementiev and Roth, 2006). We may also compare alignment-based discriminative string similarity with a more complex discriminative model that learns the alignments as latent structure (McCallum et al., 2005).

Acknowledgments

We gratefully acknowledge support from the Natural Sciences and Engineering Research Council of Canada, the Alberta Ingenuity Fund, and the Alberta Informatics Circle of Research Excellence.

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.

Mikhail Bilenko and Raymond J. Mooney. 2003. Adaptive duplicate detection using learnable string similarity measures. In *KDD*, pages 39–48.

Eric Brill and Robert Moore. 2000. An improved error model for noisy channel spelling correction. In *ACL*. 286–293.

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

Thorsten Joachims. 1999. Making large-scale Support Vector Machine learning practical. In *Advances in Kernel Methods: Support Vector Machines*, pages 169–184. MIT-Press.

Alexandre Klementiev and Dan Roth. 2006. Named entity transliteration and discovery from multilingual comparable corpora. In *HLT-NAACL*, pages 82–88.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquistion*.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *NAACL Workshop on Statistical Machine Translation*, pages 102–121.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL*, pages 127–133.

Grzegorz Kondrak and Tarek Sherif. 2006. Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *COLING-ACL Workshop on Linguistic Distances*, pages 37–44.

Grzegorz Kondrak. 2005. Cognates and word alignment in bitexts. In *MT Summit X*, pages 305–312.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

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

Andrew McCallum, Kedar Bellare, and Fernando Pereira. 2005. A conditional random field for discriminatively-trained finite-state string edit distance. In *UAI*. 388–395.

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

Andrea Mulloni and Viktor Pekar. 2006. Automatic detection of orthographic cues for cognate recognition. In *LREC*, pages 2387–2390.

Ari Rappoport and Tsahi Levent-Levi. 2006. Induction of cross-language affix and letter sequence correspondence. In *EACL Workshop on Cross-Language Knowledge Induction*.

Eric Sven Ristad and Peter N. Yianilos. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *CoNLL*, pages 207–216.

Michael Strube, Stefan Rapp, and Christoph Müller. 2002. The influence of minimum edit distance on reference resolution. In *EMNLP*, pages 312–319.

Ben Taskar, Simon Lacoste-Julien, and Dan Klein. 2005. A discriminative matching approach to word alignment. In *HLT-EMNLP*, pages 73–80.

Jörg Tiedemann. 1999. Automatic construction of weighted string similarity measures. In *EMNLP-VLC*, pages 213–219.

Dmitry Zelenko and Chinatsu Aone. 2006. Discriminative methods for transliteration. In *EMNLP*, pages 612–617.