Combining Information-Weighted Sequence Alignment and Sound Correspondence Models for Improved Cognate Detection

Johannes Dellert
Seminar für Sprachwissenschaft
Universität Tübingen
johannes.dellert@uni-tuebingen.de

Abstract

Methods for automated cognate detection in historical linguistics invariably build on some measure of form similarity which is designed to capture the remaining systematic similarities between cognate word forms after thousands of years of divergence. A wide range of clustering and classification algorithms has been explored for the purpose, whereas possible improvements on the level of pairwise form similarity measures have not been the main focus of research. The approach presented in this paper improves on this core component of cognate detection systems by a novel combination of information weighting, a technique for putting less weight on reoccurring morphological material, with sound correspondence modeling by means of pointwise mutual information. In evaluations on expert cognacy judgments over a subset of the IPA-encoded NorthEuraLex database, the combination of both techniques is shown to lead to considerable improvements in average precision for binary cognate detection, and modest improvements for distance-based cognate clustering.

1 Introduction

The inference of cognate sets, i.e. sets of words which are related by being inherited from a common ancestor, is a central problem of historical linguistics. In the classical comparative method, the detection of cognates is an elementary preprocessing step for establishing sound laws, reconstructing the original forms in proto-languages, and thereby ultimately showing that groups of languages are related by common ancestry. In computational historical linguistics, cognacy-encoded datasets are the most common input format for modern methods of phylogenetic inference as pioneered by Gray and Jordan (2000). While in classical historical linguistics, a sharp distinction is made between cognates connected purely by inheritance, and words which are related due to borrowing, the common usage in the computational field refers to both types of relations as cognacy, with the understanding that the primary distinction is whether words are etymologically related, and the decision whether they are related by inheritance or borrowing will be made at a later refinement stage. Since this paper deals with automated cognate detection, the term will be used in the broader sense.

This paper builds on Dellert and Buch (to appear), where we introduced a segment-wise information content model as a technique for alignment and comparison of IPA strings. Information content is used to focus the string distance judgment on the parts of the word forms that are distinctive with respect to all the other words in the language. The main benefit for cognate detection is that it weakens distorting effects older approaches were faced with when dealing with lexical forms that were not reduced to stems.

In this paper, I take the next step of combining information weighting with the more established idea of inferring specialized segment similarity matrices for each language pair, with the goal of capturing some of the regularity in sound correspondences caused by sound laws, and therefore making cognates at a higher timedepth more similar.

In order to evaluate the resulting form distances independently of the choice of clustering algorithm, I continue to work primarily on the level of pairwise cognacy judgments. This allows performance to

Proceedings of the 27th International Conference on Computational Linguistics, pages 3123–3133
Santa Fe, New Mexico, USA, August 20-26, 2018.

be measured in terms of precision and recall for each threshold that could be used to separate cognates from non-cognates. The average precision across all possible threshold values provides a criterion for assessing the quality of form distances that is independent of the clustering algorithm. To assess whether improvements in average precision do indeed lead to improved cognate sets after inferring clusters from the distances, B-cubed precision and recall scores are compared in a second step.

As in previous work, our recently published NorthEuraLex database (Dellert and Jäger, 2017) is used for the experiments. NorthEuraLex is a data collection effort aiming to cover a large slice of the basic vocabulary (more than 1,000 concepts) across all sufficiently well-documented languages of Northern Eurasia. At the current stage, dozens of languages from the well-studied Uralic and Indo-European language make it a useful source of cross-family lexicostatistical data. The unique advantage of NorthEuraLex among deep-coverage databases is that it provides a full set of (mostly automatically generated) IPA transcriptions which adhere to a single unified set of conventions across more than 100 languages from a range of families. The disadvantage is that expert cognate annotations are currently only available for a small subset. What is available was produced by intersecting NorthEuraLex with the data present in the most recent online version of IELex (Dunn, 2015), a cognacy-annotated database covering about 200 concepts across more than 120 Indo-European languages.

The Java code implementing the entire toolchain described in this paper, from the NorthEuraLex dump file to the different form distances, as well as the evaluation code, is available under a GPL license in a public repository[1] for other researchers to inspect and build upon.

2 Cognate Detection

In its simplest form, cognate detection is the task of deciding for word forms a from a language L_a and b from a language L_b whether a and b can be traced back to a single proto-word from which both forms derive on different paths of regular sound shifts and borrowing.

On the algorithmic level, cognate detection can be approached directly as a binary classification problem. For instance, machine learning algorithms can be applied to training data in the form of pairwise cognate judgments in order to let a system learn to make correct binary cognacy decisions for pairs of words. Early work in this direction is summarized by Rama (2015), who trains a Support Vector Machine (SVM) over a feature representation that encodes shared subsequences, and shows that this feature set outperforms earlier SVM-based attempts. Moving to non-linear classifiers, Rama (2016) trains a convolutional neural network (CNN) on handcrafted representations of ASJP data, a large database by Wichmann et al. (2016) which covers 40 concepts across more than 5,000 languages in an approximate phonetic transcription. The trained network proves to be quite successful at deciding cognacy between pairs of words from families of limited timedepth.

2.1 Form Distances

A variety of string distance measures has been proposed and tested as a basis for cognate detection. Measures proposed can be as simple as the number of shared bigrams, the longest common subsequence, or even just a binary distinction where strings are judged as similar if there is a match of very coarse-grained equivalence classes assigned to the first two consonants, and as dissimilar otherwise (Turchin et al., 2010). Kondrak (2005) systematically evaluates the potential of some of these simple measures, and shows that they achieve acceptable results on the cognate detection task, albeit on small testsets covering only closely related languages.

However, as shown by List et al. (2017b), more elaborate measures are needed to arrive at state-of-the-art performance. The most widely used non-trivial string distance measure is the Levenshtein distance or edit distance (Levenshtein, 1965), which counts the minimal number of elementary editing operations (deletions, insertions, or replacements) needed to transform the one string into the other. The Levenshtein distance on either the orthography or some coarse-grained sound-class model tends to lead to a workable first approximation to phonetic form distance. In this paper, the **normalized edit distance (NED)**, the edit distance divided by the length of the longer string, will be used as a baseline for IPA-encoded data.

[1] `https://github.com/jdellert/iwsa`

Since assessing the usefulness of different sound class schemes would add an additional dimension to our evaluation, we are using NED on full IPA with the expectation that it will perform very badly, as in the absence of a notion of sound similarity, very small differences in pronunciation between e.g. [ɑ] and [ɒ], will be weighted just as much as e.g. the difference between a vowel and a consonant.

The major step improving on edit distance has been to estimate a similarity score for each pair of phonemes, and to count replacement of similar phonemes by only a fraction of a full replacement in distance computation. For instance, when assessing the similarity of English orthographic strings, changing an *o* to a *u* should be much better than changing an *l* to an *n*. This natural extension to the Levenshtein distance leads to the algorithm first presented by Needleman and Wunsch (1970), another dynamic programming approach which maximizes the similarity score between strings by introducing gaps. Variants of the Needleman-Wunsch algorithm are a very popular method for computing string distances in distance-based phylogenetics.

2.2 Clustering

Since the results of binary classification will typically not lead to a consistent result (as the cognacy relation is transitive, whereas binary decisions taken in isolation will typically not be), the output of a system trained for binary classification needs to be combined with an additional clustering stage in order to arrive at a partition into cognate sets. This approach is exemplified by Jäger and Sofroniev (2016), who train an SVM to predict pairwise probabilities of non-cognacy from phonetic distances, overall language distance, and average word length, and combine the result with UPGMA clustering (Sokal and Michener, 1958) to derive the cognate sets.

The **LingPy** system (List et al., 2017a) is an actively maintained and user-friendly suite of tools for computational historical linguistics. One of its core components is the **LexStat** module (List, 2012) which still provides state-of-the-art performance for cognate clustering, and has become the standard benchmark for the more recent cognate detection systems. For instance, Jäger and Sofroniev compare their system to LexStat in terms of the B-Cubed F-score measure of clustering quality, finding that it only slightly outperforms LexStat, an advantage which seems to hinge mostly on improved clustering methods. In LexStat, IPA input sequences are first converted into sound classes and annotated with sonority profiles. Segment similarity scores are then inferred for each language pair by comparing how often segments are mapped to each other when aligning candidate cognate pairs as opposed to semantically unrelated word pairs, and the Needleman-Wunsch algorithm is applied on these scores. The resulting scores are converted into pairwise form distances (**LexStat distances**), and these distances are used to infer cognate sets based on standard clustering algorithms like UPGMA and InfoMap.

Rama et al. (2017) combine distance scores based on pointwise mutual information (PMI) with InfoMap clustering, and evaluate the resulting system on a range of datasets against LexStat distances, as well as an alternative distance score derived from Pair Hidden Markov Models (PHMM). On a range of testsets, they find that PMI scores trained in an unsupervised fashion using online expectation-maximization, in combination with the Needleman-Wunsch algorithm and InfoMap clustering, beat LexStat by a small margin on a number of datasets.

List et al. (2017b) compare the performance of UPGMA clustering on various older form distance measures with UPGMA on LexStat distances as well as the more advanced InfoMap clustering on LexStat distances, finding that the two LexStat-based approaches clearly perform best, whereas InfoMap clustering only leads to a very small improvement over much simpler UPGMA. This means that the influence of the quality of form similarity scores on the quality of automatically inferred cognate sets seems to be much greater than that of the clustering method used on top of them, motivating the focus of this paper on improving and evaluating form similarity scores.

Whichever clustering method is used, all state-of-the-art cognate detection methods are ultimately based on form similarity or distance scores, whether they are computed from weighted edit distances, PHMMs, or stochastic models. The quality of these scores can (and should) be assessed independently of the clustering method on the level of pairwise cognacy judgments, since any improvement on this level is likely to increase the performance of all clustering methods explored in the literature.

There are quite a few interesting alternative approaches to cognate detection for which no implementations are available. Older approaches like Kondrak (2005) have tended to be evaluated only on very small wordlists from closely related languages. Some newer approaches would be interesting to compare because they provide very general models (Berg-Kirkpatrick and Klein, 2011), or even integrate cognate detection with phylogenetic inference and reconstruction in a comprehensive statistical model (Bouchard-Côté et al., 2013), but their source code remains unavailable, and the system descriptions do not contain all the necessary details for exact reimplementations.

3 Alignment-Based Form Distances

In bioinformatics, the Needleman-Wunsch algorithm is used on standardized and well-tested similarity matrices which encode current knowledge about the different probabilities for each nucleotide base to turn into a different one due to mutation. Unfortunately, no such standard matrix exists so far for phonemes, due to the absence of a global inventory of attested sound changes. One might try to derive such a matrix from phonological knowledge, but methods which estimate a distance matrix from large amounts of data consistently fare better in practice than such attempts, due to the impossibility for a human to assign an intuitive meaning to the distance weights. In practice, phoneme similarity distances are therefore always estimated from large datasets.

Estimation of phoneme similarity builds on observation of segment pairs which are likely to be equivalent given the context. Any method which uses dynamic programming to compute string similarity implicitly constructs an **alignment**, i.e. a separation of the two or more aligned strings into columns of equivalent segments. A binary alignment specifies which phonemes are cognate in a pair of cognate word, providing pairs of observations which can be counted and correlated to build models of phoneme distances. The procedure used in the toolchain for this paper to extract phoneme similarity scores from the NorthEuraLex database is detailed in Section 6.

4 Information Weighting for IPA Sequences

This section revisits the information weighting model which we proposed in Dellert and Buch (to appear), and adds some additional examples and considerations to motivate the changes that were necessary to handle pairwise correspondences. Assume we are faced with the task of assessing the closeness of the English word "to freeze" and its German equivalent "gefrieren". The NorthEuraLex IPA representations of the words are [friiz] and [gəfʁiiʁən], respectively. The NED between these forms is 0.667, which is clearly too high for a pair of cognates from closely related languages. Assuming that we use alignment weights, and that our global similarity matrix additionally tells us that [ɾ] and [ʁ] are a good fit, and (optimistically) that sound correspondence detection will have determined that English [z] clearly coincides with German [ʁ] in some contexts, we would still be left with distance of at least 0.444, i.e. only slightly better than, say, the distance between *sink* [sɪŋk] and *song* [sɔŋ].

The reason for the problems is, of course, that there is some additional material in the German form which would traditionally need to be stripped in order to only map the core portion, the stem *frier-*, to *freeze*. If we cannot extract the stems manually because it would require too much time, or because too little is known about the languages in question (which is frequently the case for languages where automated methods might yield new results), a mathematical model is needed to tell us which bits to ignore, and then a way to incorporate this information into the sequence distance computation.

4.1 Gappy Trigram Models

An important criterion for mathematical models of relevance is that the irrelevant material will be predictable. For instance, the infinitive ending *-ć* is present at virtually every Polish verb, so seeing it at the end of a verbal lexeme is completely unsurprising. Put differently, using the information-theoretic notion of surprise as high information, we will generally find the low-information segments to be more justified to ignore when comparing lexical material across languages. The most direct way to model predictability builds on the probability of seeing the item in question given the knowledge we already have. In phonetic strings, the knowledge we have are the surrounding symbols. If the probability of seeing a segment given

the neighboring segments is very high, this implies low information content. These considerations lead to the use of language-specific (gappy) n-gram models for modeling information content.

Depending on the context, I will use the terms **gappy trigram** and **extended bigram** interchangeably for trigrams where one of the three symbols (the "gap") is a wildcard, i.e. can be replaced by any symbol. For instance, the gappy trigram Xbc represents any of the trigrams abc, bbc, cbc, dbc, etc. A gappy trigram aXb with the gap in the middle could also be called a 1-skip-bigram in the terminology of e.g. Guthrie et al. (2006), but the other two types of extended bigrams are not skip bigrams.

Writing c_{abc}, c_{abX}, c_{Xbc}, c_{aXc} for the trigram and extended bigram counts extracted from all word forms of a language L, the **information content** of a segment c in its five-segment context $abcde$ is defined as

$$I_L(c, [ab\_de]) := -\log \left\{ \frac{c_{abc} + c_{bcd} + c_{cde}}{c_{abX} + c_{bXd} + c_{Xde}} \right\}$$

In words, one combines the number of times c was observed together with the two segments before it, the two segments after it, and its immediate neighbors, and compares this number of observations with the analogous count for the gappy bigrams. It is easy to show that these fractions define a probability distribution over segments in the context $[ab\_de]$, which implies that the negative logarithm does indeed define a measure of surprisal in the information-theoretic sense.

To also be able to define information content values for segments at the start and the end of an IPA representation, the word boundary symbol # is used for expanding a string a of length k to the positions a_{-1}, a_{-0}, a_{k+1}, and a_{k+2}. On these padded strings, no special definition is needed for the trigram and extended bigram counts involving peripheral segments. For instance, the counts for the affricate [t͡ɕ] in Polish *dać* [dat͡ɕ] "to give" are $c_{\text{dat͡ɕ}}$ = 13, c_{daX} = 30, $c_{\text{at͡ɕ\#}}$ = 132, $c_{\text{daX\#}}$ = 339, $c_{\text{t͡ɕ\#\#}}$ = 350, and $c_{\text{X\#\#}}$ = 1124. After smoothing, the information content of [t͡ɕ] in this word is I_{pol}(t͡ɕ,[da_##]) = 1.287. For comparison, an information content I_{pol}(d,[##_at͡ɕ]) = 3.306 is inferred for the first segment.

5 Information-Weighted Sequence Alignment

The next step is to define how the segment-wise information content values are used during alignment. Our solution, first presented in Dellert and Buch (to appear), is to use information content in a modified Needleman-Wunsch algorithm which we call **Information-Weighted Sequence Alignment (IWSA)**.

The modified dynamic programming procedure for computing the raw sequence similarity score $sc(a, b) := M(m, n)$ for two IPA strings $a \in L_a$ of length m and $b \in L_b$ of length n is defined by the following recursion:

$$\begin{array}{lcl} M(0,0) & := & 0 \\ M(i,0) & := & M(i-1,0) + w(a_i, \epsilon) \cdot I^2_{L_a,L_a}(a_i, a_i) \\ M(0,j) & := & M(0,j-1) + w(\epsilon, b_j) \cdot I^2_{L_b,L_b}(b_j, b_j) \\ M(i,j) & := & min \left(\begin{array}{c} M(i-1,j-1) + w(a_i, b_j) \cdot I^2_{L_a,L_b}(a_i, b_j), \\ M(i-1,j) + w(a_i, \epsilon) \cdot I^2_{L_a,L_a}(a_i, a_i), \\ M(i,j-1) + w(\epsilon, b_j) \cdot I^2_{L_b,L_b}(b_j, b_j), \end{array} \right) \end{array} \quad (1)$$

For the $w(a_i, b_j)$, any segment similarity score can be used. My choices for these scores, and their motivation, are described in Section 6. $I^2_{L_a,L_b}(a_i, b_j)$ is defined as the quadratic mean of information weights assigned to the segments:

$$I^2_{L_a,L_b}(a_i, b_j) := \sqrt{\frac{I_{L_a}(a_i, [a_{i-2} \ldots a_{i+2}])^2 + I_{L_b}(b_j, [b_{j-2} \ldots b_{j+2}])^2}{2}}$$

In the cases of insertion and deletion, the score combinations are equivalent to the information content of the segment that was matched to a gap. The quadratic mean is used because it remains high for similar segments with equally high information content, while not penalizing alignment of dissimilar low-information segments, but that of high-information with dissimilar low-information segments. These three properties combine into a focus on matching stems (high-information regions), while discounting differences in low-information parts such as frequently recurring affixes.

In Dellert and Buch (to appear), we only used a single globally inferred phoneme similarity matrix for IWSA. The Needleman-Wunsch scores $sc(a, b)$ were normalized through division by the average self-similarity of both word forms, leading to the following formula for deriving form distance values from similarity scores:

$$d(a, b) := 1 - \frac{2 \cdot sc(a, b)}{sc(a, a) + sc(b, b)}$$

When moving beyond the global similarity matrix and inferring pair-specific phoneme similarity scores in the style of LexStat, this definition turns out to be problematic. If we infer a separate model for each language pair, this leads to very high self-similarity scores because all words can be perfectly aligned with themselves. This causes distance values to be close to 1 for any pair of longer words, even if they are actually very similar. On the other hand, simply using the global similarity scores instead leads to negative distance values, because a specialized model will of course lead to higher $sc(a, b)$ values than the global one. I therefore normalize each similarity score by the length of the relevant sequence:

$$d(a, b) := 1 - \frac{2 \cdot \frac{sc(a,b)}{\max\{n,m\}}}{\frac{sc(a,a)}{m} + \frac{sc(b,b)}{n}}$$

The resulting scores will still often lie beyond the interval $[0, 1]$, but at least there are only very few negative distances, and distances can easily be scaled to the interval if necessary.

6 Inferring Phoneme Similarity Scores

We now turn to the problem of inferring good phoneme similarity matrices. Any model which attempts to estimate such matrices will quickly run into problems if too many parameters are to be estimated from too little data. For many language pairs where we would like to estimate correspondences (e.g. members from different branches of the same family), we often need to get by on little more than 100 cognate pairs. List (2012) reduces the number of similarity scores to estimate by internally reducing IPA (which LingPy accepts as input) to 28 equivalence classes, over which it is easy to estimate sound correspondences even based on only 100 or 200 cognate pairs. On the NorthEuraLex data, where about a thousand word pairs are available for each language pair, we can operate directly on the tokenized and simplified IPA representations in NorthEuraLex, which distinguish 105 IPA symbols.

Most work on phoneme similarity has been based on **pointwise mutual information (PMI)**, which is defined as the logarithm of the observed number of co-occurrences of two events divided by the number of co-occurrences we would expect if both events were independent. Pointwise mutual information has successfully been applied in many areas of computational linguistics. Treating the occurrences of segments in IPA strings as observations of a variable, the pointwise mutual information of two segments x and y would be defined as

$$i(x, y) := \log \frac{p(x, y)}{p(x)p(y)} \tag{2}$$

The problem of PMI scores for sound correspondences is that due to the non-random nature of phoneme sequences (e.g. a preference for the consonant-vowel pattern CVCVCV over CCCVVV), there will always be non-zero mutual information between any pair of consonants and vowels, even if the languages the correspondences are inferred for are completely unrelated. It is unclear how one could correct for this effect in an explicit parametrization of the expected probabilities.

The decisive idea which successfully addresses this problem goes back to Kessler (2001). The similarity scores $w(x, y)$ for IPA segments x and y are PMI scores based on the probability $p(x, y)$ of x being aligned with y in cognate pairs based on counts for a large set of likely cognate pairs, compared to an estimate $\hat{p}(x, y)$ of that probability on non-cognate words:

$$w_{glo}(x, y) := \log \frac{p(x, y)}{\hat{p}(x, y)} \tag{3}$$

6.1 Global Similarity Scores

The interesting decisions are now hidden behind the symbols p and $\hat{p}$. In our architecture, we stay within the framework of aligning randomly chosen wordpairs, and only modify the counting procedure in the case where information weights are used. Our implementation of this in the NWD case is very similar to LexStat, except that LingPy uses multiple-sequence alignments instead of only pairwise alignments to increase consistency. Due to the additional weighting factor, it is not obvious how IWSA could be generalized to multiple-sequence alignment in a consistent way, but as we shall see, this disadvantage does not appear to negatively impact performance.

The distribution $\hat{p}(x, y)$ to compare $p(x, y)$ against is derived by randomly sampling as many word pairs (of any meaning) from random language pairs as there are form pairs of identical meaning in the dataset, aligning each pair in the same way that the cognacy candidates are aligned, and then counting the number of times each pair of phonemes occurred in one column in the resulting alignments. 20% of the overall observation mass is redistributed for Laplace smoothing of the phoneme pair distributions.

$\hat{p}(x, y)$ is kept constant throughout each iteration of re-estimating $p(x, y)$ from a refined set of cognate candidates. Cognate candidates are selected based on an NED threshold (< 0.35) in the initial step, and on a threshold on the Needleman-Wunsch scores (< 1.2) for the current matrix in each subsequent iteration. Due to the large amounts of data in NorthEuraLex, three iterations were enough to derive very stable values for $\hat{c}(x, y)$. Also, no special treatment was needed for the gap symbol, which occurs in correspondences whenever insertions or deletions were inferred in the optimal alignment, and models the costs for inserting or deleting the respective phoneme in the inferred matrices.

In the information-weighted case, our implementation diverges from previous approaches in a crucial way. The probabilities are not directly based on counts of the number of times each symbol pair was aligned, but each instance in a candidate cognate pair only counts with its combined information content. Using the notation $al(a, b)$ for the optimal information-weighted alignment of a word pair (a, b) according to the current segment distances, $sc(a, b)$ for the corresponding Needleman-Wunsch score, and $al(a, b).a$ and $al(a, b).b$ to refer to the individual strings (with gap symbols) in which positions can be indexed by subscripts, this way of counting pairs of aligned segments can be written in one expression as follows:

$$c(x, y) := \sum_{L_1, L_2 \in \mathcal{L}} \sum_{\substack{(a,b) \in lex(L_a, L_b), \\ sc(a,b) < 1.2}} \sum_{\substack{1 \leq i \leq \max\{m,n\}, \\ al(a,b).a_i = x, \\ al(a,b).b_i = y}} I^2_{L_a, L_b}(a_i, b_i) \tag{4}$$

I will call the distance measure based on Needleman-Wunsch scores over the resulting global phoneme similarity matrix **Needleman-Wunsch Distance (NWD)**, whereas the variant using information weighting both for inferring the phoneme similarity scores and during alignment will be called **Information-Weighted Distance (IWD)**. NWD can thus be seen as a special case of IWD where all information weights $I_L(c, [ab\_de])$ are set to 1.

6.2 Pairwise Correspondences for NorthEuraLex

In the same way that global similarity scores were estimated, it is possible to infer specialized scores from the data for any pair of languages. These will tend to assign low costs to sound pairs which are equivalent across many alignments, and can therefore be interpreted as encoding some of the sound correspondences the comparative method operates with. For instance, given enough examples such as *water*/*Wasser*, *street*/*Straße*, and *foot*/*Fuß*, the alignment costs of English [t] and German [s] will be rather low, encoding the consequence of a part of the High German consonant shift.

To infer sound correspondence models for each pair of languages (leading to what I will call **local similarity scores**), we repeat the procedure that we used for inferring the global segment similarities on cognate pairs for that language pair. The values for $\hat{p}(x, y)$ are estimated based on 100,000 random word pairs sampled independently of the associated concepts. Like in LexStat, the inferred similarity scores will be based on a mixture of global and local PMI scores, because the local models for some pairs of unrelated languages will otherwise include some very strong correspondences that are only due to

random noise. To keep the mixture interpretable as an information measure, some weighted mean of both scores needed to be chosen. In order to avoid overfitting the method to the dataset, we based our decision for the proportions between local and global PMI scores on inspection of the resulting correspondences for a small subset of language pairs, and not on optimization. It turned out that one of the simplest option, the arithmetic mean of the local and global PMI scores, already resulted in convincing PMI scores for the inspected language pairs, leading the following definition of local similarity scores:

$$w_{L_1,L_2}(x,y) := \frac{w_{glo}(x,y) + \log \frac{p_{L_1,L_2}(x,y)}{\hat{p}_{L_1,L_2}(x,y)}}{2} \tag{5}$$

Again, using information weighting for the observation counts is crucial for getting good information-weighted alignments. If information weighting is not used for the counts, the local similarity scores will be influenced by reoccurring phonological material, leading to spurious low scores which are then applied across the board. For instance, the frequent mapping of the Persian infinitive ending [æn] to Ukrainian [tɪ] will result in an erroneous high local similarity score [æ] and [t], leading to unwanted small form distances between such pairs as Persian [ælæf] and Ukrainian [trɑwɑ] "grass". More generally, information content weighting of counts will diminish the effect of reoccurring morphological material in non-stemmed data.

I will call the extension of IWD by local similarity scores **Information-Weighted Distance with Sound Correspondences (IWDSC)**. Analogously, the combination of NWD with local similarity scores will be called **Needleman-Wunsch Distance with Sound Correspondences (NWDSC)**.

7 Evaluation

7.1 Test Data

To evaluate whether information weighting does indeed help to better separate cognate from non-cognate word pairs, all five methods were applied to a large IELex-based cognate pair testset which is available as part of the supplementary materials[2] to my dissertation (Dellert, 2017). For this testset, the 6,106 word forms which exist in both IELex and NorthEuraLex were manually mapped to each other, allowing the expert judgments contained in IELex to be added as annotations to the corresponding entries in NorthEuraLex. The resulting testset covers 185 concepts across 36 languages. At a size of 100,156 pairwise judgments, this is one of the largest expert-annotated testsets for cognacy detection currently available.

All the methods that are compared in this evaluation were run on the full version of NorthEuraLex 0.9 (121,615 forms covering 1,016 concepts across 107 languages), including the inference of global and pairwise sound similarity scores. The resulting distance values and clusterings were then reduced to the testset for evaluation. The numbers that will be reported are thus computed on the testset, whereas the dataset used for inference was about 20 times larger. Evaluation in previous literature on cognate detection has relied on testsets that are much smaller in terms of the number of concepts covered, and therefore do not include enough data for building high-quality language-specific information models as needed for information-weighted alignment.

7.2 Results

Using the form distances for pairwise cognate detection boils down to picking a single threshold value θ such that a word pair (a, b) is considered a pair of cognates if and only if $d(a, b) \leq \theta$. For any given threshold, each word pair annotated as cognates in the IELex-based gold standard is counted as a true positive if their distance is below the threshold, and as a false negative if it is not. Analogously, pairs which are considered non-cognates due to a distance above the threshold, are classified into true negatives and false positives depending on the gold standard. This makes it possible to compute precision and recall values for each value of θ.

[2] http://www.sfs.uni-tuebingen.de/~jdellert/pubs/jdellert-diss-supplements.tar.gz

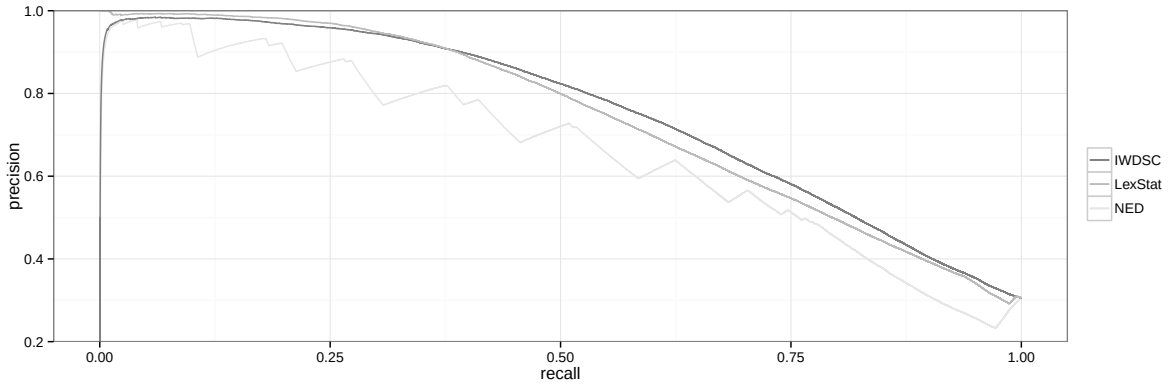

Figure 1: Precision-recall curves for form distance measures in cognacy detection.

As a threshold-independent performance measure, I use the **average precision**, defined as the precision integrated over all possible recall values. To approximate this integral, one simply orders the precision-recall pairs by recall, and then sums up the precision values weighted by the distance between the current recall value and the last, which approximates the area under the precision-recall curve as a sum of rectangles. For illustration, Figure 1 shows the curves for the three most important distance measures in my comparison. In this visualization of the tradeoff between precision and recall, the curve for the NED baseline is jagged and does not drop monotonously. The reason for this is that to ensure a fair comparison, all tied values (of which there are of course many in a distance based on counting edit operations) need to be ordered in the worst-case fashion, i.e. all non-cognate pairs need to be counted as if they had slightly lower distances, followed by the cognate pairs.

Between the smooth curves for LexStat and IWDSC, we observe that in the low-recall range (i.e. for the easy cases), LexStat is slightly better, but its precision decays more quickly with higher recall, showing that information weighting leads to improvement especially for the more difficult instances. The curve already suggests that IWDSC will have a global advantage at desirable recall values of more than 50%, and the average precision values will substantiate this impression.

Table 1 shows the average precision values for all the distance measures discussed in this paper. As already suggested by the curve, IWDSC distances clearly outperform LexStat distances at a gain in average precision of 4.3 percentage points, which in this case corresponds to an average reduction of the false discovery rate by 15.8%.

Method	**NED**	**LexStat**	**NWD**	**NWDSC**	**IWD**	**IWDSC**
Average Precision	0.604	0.727	0.741	0.747	0.764	**0.771**
Maximum F-score	0.599	0.631	0.652	0.654	0.673	**0.679**
Precision at max. F-score	0.639	0.652	0.666	0.660	0.696	**0.706**
Recall at max. F-score	0.564	0.611	0.639	0.648	0.652	**0.654**

Table 1: Quantitative comparison of form distance methods on pairwise cognacy testset.

To evaluate whether this advantage carries over to clustering results, and to maintain comparability with the previous literature, we now turn to the evaluation in terms of B-Cubed F-score for the full cognate detection task, i.e. in terms of the cognate clusters inferred over the testset. For this, the IWDSC values were squared to enhance the differences in the upper range, and then divided by 1.5 (with cutoff at the same value) to scale them to the interval $[0, 1]$ before clustering. To ensure a fair comparison, the main clustering parameter (a single threshold value) was varied with a step size of 0.05 in order to find a near-optimal value for the testset without overfitting. This emulates the usage of LingPy in practice, where on a new dataset, the user is encouraged try other threshold values instead of the default 0.6.

As the numbers in Figure 2 show, the advantage of IWDSC turns out to be far less prominent on the full task than it was for the pairwise cognacy judgments. In combination with UPGMA, IWDSC

does lead to better cognate clusters than LexStat, but only by about 1 percentage point in B-Cubed F-score. Regarding the comparison of clustering algorithms, the NorthEuraLex dataset does not confirm the superiority of InfoMap over much simpler UPGMA found by List et al. (2017b) on their testsets.

Method	Best Threshold	Precision	Recall	F-score
LexStat + InfoMap	0.70	0.757	0.612	0.677
LexStat + UPGMA	0.80	0.763	0.616	0.681
IWDSC + UPGMA	0.75	0.790	0.613	**0.690**

Table 2: B-Cubed measures for LexStat and IWDSC on the NorthEuraLex/IELex testset.

7.3 Discussion

To understand which of the differences between IWDSC and LexStat contributes most to the improvement in average precision, we consider the result figures for the other measures in Table 1. We see that about a fourth of the difference already appears in the comparison between LexStat and our implementation of NWD. This is likely due to the difference in the segment model (with full IPA vs. a reduction to equivalence classes in LexStat), since a test for the only other major difference (the normalization by length during the score-to-distance transformation) did not lead to a lower average precision score.

The performance gain from pairwise segment similarity matrices is of almost exactly the same size for NWD and IWD. This increases my confidence that the small improvement by little more than half a percentage point is a meaningful difference, establishing that there is a gain due to sound correspondences that is orthogonal to and can be combined with the improvements coming from information weighting.

Information weighting alone appears to account for about half of the difference between LexStat and IWDSC, a difference which persists whether we additionally infer sound correspondences (NWDSC vs. IWDSC) or not (NWD vs. IWD). This orthogonality to other improvements makes it a useful technique for inclusion in cognate detection systems.

The small improvement for the full cognacy task is of a similar size to the ones reported by Jäger and Sofroniev (2016) and List et al. (2017b), once more confirming the difficulty to achieve substantial gains over LexStat. To explain how the much more pronounced gain in average precision fits into this picture, the most obvious hypothesis is that while better for pairwise comparisons, the IWDSC distances are not more consistent within cognate sets than LexStat distances, making it difficult to transfer the clear advantage in pairwise judgments to the level of clusters.

8 Conclusion and Future Work

This paper has shown that information weighting and pair-specific phoneme similarity matrices can be used independently to increase the quality of pairwise form distances for automated cognate detection, and that a combination of both techniques improves the Needleman-Wunsch scores for this task.

A large part of the improvement appears due to the use of information weights already for counting the observations that the similarity matrices are based on, whereas the advantage of modeling sound correspondences in the form of local similarity scores turned out to lead to only small additional gains.

The reasons for the small size of the improvement on the full cognate detection task as opposed to the quality of pairwise form distances, and possible remedies, will be a focus of future work. Also, it will be interesting to apply IWSA to other language families as soon as cognacy-annotated databases for other subsets of the lexical data covered by NorthEuraLex (or possible future databases of similar coverage) become available.

Acknowledgements

This research has been supported by the ERC Advanced Grant 324246 EVOLAEMP, which is gratefully acknowledged. The author also wishes to thank three anonymous reviewers for their helpful comments and suggestions, especially the request to also evaluate IWDSC in combination with clustering.

References

Taylor Berg-Kirkpatrick and Dan Klein. 2011. Simple Effective Decipherment via Combinatorial Optimization. In *EMNLP 2011*, pages 313–321. ACL.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 10.1073/pnas.1204678110.

Johannes Dellert and Armin Buch. to appear. A New Approach to Concept Basicness and Stability as a Window to the Robustness of Concept List Rankings. *Language Dynamics and Change*.

Johannes Dellert and Gerhard Jäger. 2017. NorthEuraLex (version 0.9). http://northeuralex.org.

Johannes Dellert. 2017. *Information-Theoretic Causal Inference of Lexical Flow*. Ph.D. thesis, University of Tübingen.

Michael Dunn. 2015. Indo-European Lexical Cognacy Database. http://ielex.mpi.nl/.

Russell D Gray and Fiona M Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature*, 405(6790):1052–1055.

David Guthrie, Ben Allison, W. Liu, Louise Guthrie, and Yorick Wilks. 2006. A Closer Look at Skip-gram Modelling. In *LREC-2006*, Genoa, Italy.

Gerhard Jäger and Pavel Sofroniev. 2016. Automatic cognate classification with a Support Vector Machine. Proceedings of the 13th Conference on Natural Language Processing (KONVENS).

Brett Kessler. 2001. *The significance of word lists. Statistical tests for investigating historical connections between languages.* CSLI Publications, Stanford.

Grzegorz Kondrak. 2005. N-gram similarity and distance. In *International Symposium on String Processing and Information Retrieval*, pages 115–126. Springer.

V. I. Levenshtein. 1965. Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov. *Dokladi Akademij Nauk SSSR*, 163(4):845848.

Johann-Mattis List, Simon Greenhill, and Robert Forkel. 2017a. LingPy. A Python library for quantitative tasks in historical linguistics.

Johann-Mattis List, Simon J Greenhill, and Russell D Gray. 2017b. The Potential of Automatic Word Comparison for Historical Linguistics. *PloS one*, 12(1):e0170046.

Johann-Mattis List. 2012. LexStat: Automatic Detection of Cognates in Multilingual Wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France. ACL.

Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Taraka Rama, Johannes Wahle, Pavel Sofroniev, and Gerhard Jäger. 2017. Fast and unsupervised methods for multilingual cognate clustering. *CoRR*, abs/1702.04938.

Taraka Rama. 2015. Automatic cognate identification with gap-weighted string subsequences. In *Proceedings of NAACL HLT 2015, May 31 June 5, 2015 Denver, Colorado, USA*, pages 1227–1231.

Taraka Rama. 2016. Siamese convolutional networks based on phonetic features for cognate identification. *CoRR*, abs/1605.05172.

Robert R. Sokal and Charles D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Peter Turchin, Ilja Peiros, and Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship*, 3:117–126.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. 2016. The ASJP Database (version 17). http://asjp.clld.org/.