Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

STRING SIMILARITY MEASURES AND PAM-LIKE MATRICES FOR COGNATE IDENTIFICATION

Antonella Delmestri and Nello Cristianini

October 2010

Technical Report # DISI-10-054

Also: accepted for publication in *Bucharest Working Papers in Linguistics*, vol. XII, no. 2, 2010.

STRING SIMILARITY MEASURES AND PAM-LIKE MATRICES FOR COGNATE IDENTIFICATION

Antonella Delmestri, Nello Cristianini

Abstract

We present a new automatic learning system for cognate identification. We design a linguistic-inspired substitution matrix to align sensibly our training dataset. We introduce a PAM-like technique, similar to the one successfully used in biological sequence analysis, in order to learn substitution parameters. We propose a novel family of parameterised string similarity measures and we apply them together with the PAM-like matrices to the task of cognate identification. We train and test our proposal on standard datasets of Indo-European languages in orthographic format based on the Latin alphabet, but it could easily be adapted to datasets using any other alphabet, including the phonetic alphabet if data was available. We compare our system with other models reported in the literature and the results show that our method outperforms both orthographic and phonetic approaches formerly presented, increasing the accuracy by approximately 5%.

Keywords: Cognate identification, substitution matrices, string similarity measures.

1. Introduction

Language is a defining feature that distinguishes modern humans from all the other species, is a carrier of culture and plays a key role in communication. The analogy of language evolution with species evolution, predicted by Charles Darwin (1859) in his "*On the Origin of Species*" has aroused a growing interest in the scientific community following the extraordinary progress of computational molecular biology in the field of genomes. Bioinformatics techniques are now applied to the field of natural language processing where they are making significant contributions and presenting exciting opportunities for further investigation.

Natural languages that originate from a common ancestor are genetically related, words are the backbone of any natural language and *cognates* are words sharing the same ancestor and etymology. Therefore cognate identification represents the foundation for discovering the evolutionary history of languages. However, cognate recognition has proved to be useful not only in historical linguistics, but also in very diverse fields of natural language processing. Applications that benefit from cognate identification include lexicography (Brew and McKelvie 1996), parallel bilingual corpora processing, such as sentence alignment (Melamed 1999), word alignment (Tiedemann 1999; Kondrak 2005) and lexicon translation (Mann and Yarowsky 2001), statistical machine translation (Kondrak et al. 2003), and confusable drug name detection (Kondrak and Dorr 2004).

In historical linguistics, cognates are also called *strict or genetic cognates* as they derive from a "*vertical*" transmission and they do not include borrowings. *Borrowings or loans* are words borrowed from other languages through a "*horizontal*" transmission and for this reason do not follow the same phonological changes that occur over time. In many disciplines of natural language processing, the term cognates or *broad cognates* has a wider meaning and also includes borrowings.

When relatedness between cognates have to be evaluated, the methodologies applied can be either *orthographic*, where cognates are analysed in their writing form of graphemes, or *phonetic*, where cognates have to be represented in a phonetic notation in order to be

examined. The orthographic approach relies on the fact that alphabetic character correspondences represent in some way sound correspondences, as sound changes leave traces in the orthography. However, it does not require any phonetic transcription, whose attainment is still a very time consuming and challenging task. On the other hand, in evaluating word relatedness, phonetic methods depend on phonetic transcriptions of texts, but benefit from the phonetic characteristics and features of phonemes that can be decomposed into vectors of phonetic attributes. Even if for the task of cognate identification a phonetic approach is supposed to be more accurate than an orthographic one for its understanding of phonetic changes, the debate is still open and a comparative evaluation of several recent results seems to prove the opposite (Mackay and Kondrak 2005; Kondrak and Sherif 2006; Kondrak 2009).

Another differentiating feature between methods applied to the assessment of word relatedness is the capacity to adapt to different contexts, and, based on that, evaluation systems can be either *static* or *active*. A *static system* is based on manually designed and incorporated knowledge, does not require any supervision and is not able to learn by processing data. On the other hand, an *active system* has the capacity to learn and adjust, but may need supervision.

Several different approaches to the cognate identification problem have been proposed and orthographic or phonetic methodologies have been applied as well as learning algorithms or manually-designed procedures. In this paper we consider some authoritative methods proposed in the literature and compare them with our novel system.

The remainder of the paper is organized as follows. Section 2 introduces alignments and substitution matrices to the task of word relatedness. Section 3 proposes a new learning system, including a linguistic-inspired substitution matrix to align the training dataset, a PAM-like technique to produce scoring schemes and a novel family of string similarity measures to score the similarity between strings. Section 4 describes the experimental design including the datasets and the evaluation methodology used. Section 5 presents the results of this study and compares them with others reported in the literature. Finally, Section 6 reports the conclusions reached by this investigation and our future plans.

2. Word relatedness

Cognate words can be studied by string matching techniques and cognate recognition represents a typical *inexact string matching* problem (Gusfield 1997). By adopting this approach to determine the relatedness of two strings, it is possible to either measure their distance, evaluating how distant the two strings are from each other, or to measure their similarity, calculating instead how similar the two strings are. The distance method leads to a minimisation problem because it aims to find the minimum distance between two strings, while the similarity method guides towards a maximisation problem as its target is to find the maximum similarity between two strings.

2.1 Alignments

The task of calculating the distance or the similarity between two strings is closely related to the job of finding an optimal alignment between the two strings: dynamic programming algorithms can perform both tasks (Gusfield 1997). *Global* or *local alignment* algorithms, widely used in biological sequence analysis where strings are generally addressed as sequences, usually consist of a scoring system, that reports distances or similarities between the characters of the alphabet employed, and a procedure that finds the optimal alignment.

Even if the small length of the cognate words could make global alignment apparently more appropriate, local alignment can be useful in order to focus on the word roots, disregarding inflectional and derivational affixes (Kondrak 2000). Local alignment is only appropriate under the similarity approach. The dynamic programming algorithm for solving the problem of *global sequence alignment* is known as the *Needleman-Wunsch algorithm* (Needleman and Wunsch 1970), but the more efficient version generally used was introduced by Gotoh (1982). The dynamic programming algorithm for solving the problem of *local sequence alignment* is called the *Smith-Waterman algorithm* (Smith and Waterman 1981), but the more efficient version generally used is again the one proposed by Gotoh (1982).

2.2 Substitution matrices

Substitution matrices or *scoring matrices* are widely used in bioinformatics in the context of protein or nucleic acids sequence alignments. The significance of the resulting alignment depends greatly on the chosen scoring scheme, which is generally symmetric and whose choice must be determined by the type of application (Gusfield 1997).
Given an alphabet $\mathcal{A}$ with $|\mathcal{A}| \geq 2$, each character of $\mathcal{A}$ is more or less likely to transform into several other characters over time. A *substitution matrix* $|\mathcal{A}|$-by-$|\mathcal{A}|$ over $\mathcal{A}$ represents the rates at which each character of $\mathcal{A}$ may change into another character of $\mathcal{A}$. These rates in principle can be costs, when they signify distances, or can be scores, when they signify similarities. Ideally, substitution matrices should reflect the true probabilities of mutations occurring through a period of evolution and should contain values proportional to these probabilities.
There are many different ways to construct a substitution matrix, but the general approach is to assemble a large sample of verified pairwise alignments, or multiple sequence alignments, and derive the values using a probabilistic model. Ideally, the values in the matrix should reveal the phenomena that the alignments try to represent. The target is to assign a rate to the alignments that gives a measure of the relative likelihood that the sequences are related as opposed to being unrelated (Durbin et al. 1998). To compare these two hypotheses, the *log-odds-ratio* is considered, that is the logarithm of the ratio of the probability that the sequences are associated as opposed to being random. The choice of the logarithm base is generally not important. In the related or *match model*, aligned pairs of residues occur with a joint probability, and the probability for the whole alignment is the product of these joint probabilities. In the unrelated or *random model*, the probability of the two sequences is just the product of the probabilities of each character, because the model assumes that each character occurs independently. When properly arranged, these log-odds-ratios, that may be scaled and rounded, constitute the substitution matrix. Ideally, if the similarity approach is adopted, positive and negative scores should indicate respectively conservative and non conservative substitutions. Indeed, when two characters are expected to be aligned together in related sequences more often than to occur by chance, then the odds-ratio is greater than one and the score is positive. It is worth noting that the rates of identical character substitutions are inversely proportional to their occurrences because the rarer the character is, the smaller the likelihood to find two of them aligned by chance.

3. A new learning system

In order to study word relatedness, we have decided to choose the similarity approach which is the standard in biological sequence analysis and frequently used in natural language processing. Similarity allows local alignment, as well as global alignment, to be performed

and it leads to the maximisation problem of finding the highest scoring alignment of the two words. We have developed this new learning system utilising orthographic data based on the Latin alphabet, but our proposal may easily be adapted to any alphabetic system, including the phonetic alphabet.

3.1 A linguistic-inspired substitution matrix

In order to generate automatically a sensibly aligned training dataset, we have produced a linguistic-inspired substitution matrix based on knowledge of orthographic changes in the Indo-European languages. We have considered the 26 letters of the Latin alphabet and we have prepared a symmetric 26-by-26 matrix that contains a-priori likelihood of transformation between each character of the alphabet into another. We have given a value of 2 to all the elements of the main diagonal, because it is likely that a character preserves itself. We have assigned 0 values to all the character transformations considered "possible", a score of -3 to all the character transformations considered "impossible" and a gap penalty of -1 for insertion and deletion (*indels*), to avoid possible overlaps between two indels and an "impossible" match. We have tried to represent in this linguistic-inspired matrix the traces that systematic sound changes left in written languages. Vowel shift chains, consonant shift chains including Grimm's and Verner's laws, Centum-Satem division, rhotacism, assimilation, dissimilation, lenition, fortition and L-vocalisation have been considered. We have used this substitution matrix to perform global pairwise alignment on cognate pairs by the Needleman-Wunsch algorithm (Needleman and Wunsch 1970; Gotoh 1982), which is the standard for global sequence alignment. If more than one optimal alignment has been found, one alignment has been chosen through an alternate trace back (↖←↑ | ↖↑← | ←↖↑ | ←↑↖ | ↑←↖ | ↑↖←) with the aim of assuring a more balanced learning process by avoiding possible bias due to always giving priority to the same conditional predicates.

3.2 PAM matrices

We have investigated Point Accepted Mutation (PAM) matrices that have been the standard and sole substitution matrices for amino acid alignments up until the advent of BLOSUM matrices (Henikoff and Henikoff 1992). The term PAM refers to a family of amino acid substitution matrices, developed by Margaret Dayhoff et al. (1968; 1972; 1978), which encode and summarise expected evolutionary changes of amino acids. An accepted point mutation in a protein is a replacement of one amino acid by another that has been accepted by natural selection and passed on to its progeny. The name PAM is also used as a measure unit to express the evolutionary divergence between two amino acid sequences. In this way, a PAM0 matrix coincides with the identity matrix where each character is considered maximally similar to itself, but not able to transform into any other character. The foundation of Dayhoff and co-workers approach is to obtain substitution rates from global alignments between closely related proteins and then to infer from this data longer evolutionary divergences. A mutation probability matrix is calculated from comparisons of sequences with no more than 1% divergence and all the PAM matrices are extrapolated from it. This approach assumes that the frequencies of the amino acids remain constant over time and that the mutational process causing replacements in an interval of 1 PAM unit operate the same for longer periods (Gusfield 1997).

3.3 PAM-like matrices

Due to the lack of supervised and organised databases of cognate words and to the small length of words compared with the length of biological sequences, we have been forced to differentiate partially our method, from the one Margaret Dayhoff and co-workers used to create the PAM matrices for biological sequence analysis. Their starting point was to identify a group of protein families where each pair of sequences showed amino acid diversity up to 15% and from them they built hypothetical phylogenetic trees with the parsimony method (Dayhoff and Eck 1968). The group of cognate families showing up to 15% of identity that we have been able to extract from our dataset has been completely useless because it was composed of a few families of nearly identical words where the only mismatches were due to *indels*. Increasing the identity threshold up to 25% or 35% has not produced any substantial improvement. For example, the cognate words Italian *fiore* and French *fleur,* that are clearly closely related, present a diversity of 80% as 4 letters out of 5 are different. We have decided to use the whole dataset available and due to the small size of the cognate families we have compared the cognate words with each other and not with their hypothetical ancestors. We have then followed the Dayhoff method to produce a family of PAM-like matrices based on a non symmetric matrix M of mutation probabilities. Firstly, a matrix A of accepted point mutation has been calculated ignoring the evolutionary direction meaning that A(i,j) and A(j,i) were incremented every time character $\mathcal{A}_i$ was replaced by $\mathcal{A}_j$ or vice-versa. Then the relative mutability m(j) of each character $\mathcal{A}_j$ has been calculated as the ratio of observed changes to the frequency of occurrence. Finally, M has been calculated as follows:

$$M(i,j) = \frac{\mu * m(j) * A(i,j)}{\sum_i A(i,j)} \qquad \forall\, i \neq j \qquad (1)$$

$$M(i,i) = 1 - \mu * m(i) \qquad \forall\, i \qquad (2)$$

where M(i,j) contains the probability that character $\mathcal{A}_j$ mutates to character $\mathcal{A}_i$ in 1 PAM unit and μ is a proportionality constant we set to 1. To generate scoring matrices suitable for longer periods of time, we have produced matrices M^n by multiplying matrix M by itself *n* times that gives the probability that any particular character mutates to another one in *n* PAM units. Each PAMn matrix was obtained by the following log-odds-ratios where f(i) and f(j) are the observed frequencies of character $\mathcal{A}_i$ and $\mathcal{A}_j$ normalized respectively by the number of all mutations.

$$\text{PAMn(i, j)} = 10 * log_{10} \frac{f(j) * M^n(i,j)}{f(i) * f(j)} = 10 * log_{10} \frac{M^n(i,j)}{f(i)} \qquad (3)$$

We have not scaled the values in the PAM-like matrices and we have left the final scores with two decimal numbers to preserve accuracy. Because we have not limited the identity percentage within the cognate family considered for the training, ten PAM-like matrices have shown to be sufficient for modelling the divergence time of the languages considered.

3.4 A family of parameterised string similarity measures

We have proposed a family of parameterised string similarity measures obtained through different normalisations of a generic similarity rating algorithm score. In doing so, our aim has been to take into account the similarity of each string with itself in order to eliminate, or at least reduce, the bias due to different string length. Indeed, alignments of two identical strings

do not have a constant rate under the similarity approach because the score depends on the length of the string but also on the substitution rates of the characters involved.
Given two strings, S_1 and S_2, and a generic similarity rating algorithm *AL*, we have defined the family of string similarity measures reported in Table 1. The similarity measure sim_1 normalises the rate of a similarity scoring algorithm *AL* applied to calculate the similarity of S_1 with S_2 by the *arithmetic mean* of the rates given by the same algorithm applied to calculate the similarity of each string with itself. The similarity measure sim_2 does the same but normalises the rate by the *weighted arithmetic mean* that considers also the length of the two strings. The similarity measures sim_3 and sim_4 employ a normalisation by using the *geometric mean* and the *weighted geometric mean* respectively, while sim_5 and sim_6 normalise by the *harmonic mean* and the *weighted harmonic mean*. The *Heronian mean* is used to normalise the rate in sim_7, the *root mean square* is utilised in sim_8 and the *contra-harmonic mean* is employed in sim_9. Following the idea of considering the similarity of each string with itself in calculating string similarity, other similarity measures may be added to this family.
We have used these new similarity measures with the *Needleman-Wunsch* algorithm (Needleman and Wunsch 1970; Gotoh 1982) for global alignment and with the *Smith-Waterman* algorithm (Smith and Waterman 1981; Gotoh 1982) for local alignment, but the new measures may be used with any other similarity rating algorithm.

String similarity measures	***Normalised by***
$sim_1(S_1,S_2,AL)=\frac{2*AL(S_1,S_2)}{AL(S_1,S_1)+AL(S_2,S_2)}$	*Arithmetic Mean*
$sim_2(S_1,S_2,AL)=\frac{(\text{len}(S_1)+\text{len}(S_2))*AL(S_1,S_2)}{((\text{len}(S_1)*AL(S_1,S_1)+\text{len}(S_2)*AL(S_2,S_2))}$	*Weighted Arithmetic Mean*
$sim_3(S_1,S_2,AL)=\frac{AL(S_1,S_2)}{\sqrt{AL(S_1,S_1)*AL(S_2,S_2)}}$	*Geometric Mean*
$sim_4(S_1,S_2,AL)=\frac{AL(S_1,S_2)}{\sqrt[len(S_1)+len(S_2)]{AL(S_1,S_1)^{len(S_1)}*AL(S_2,S_2)^{len(S_2)}}}$	*Weighted Geometric Mean*
$sim_5(S_1,S_2,AL)=\frac{(AL(S_1,S_1)+AL(S_2,S_2))*AL(S_1,S_2)}{2*AL(S_1,S_1)*AL(S_2,S_2)}$	*Harmonic Mean*
$sim_6(S_1,S_2,AL)=\frac{(len(S_1)*AL(S_2,S_2)+len(S_2)*AL(S_1,S_1))*AL(S_1,S_2)}{(len(S_1)+len(S_2))*AL(S_1,S_1)*AL(S_2,S_2)}$	*Weighted Harmonic Mean*
$sim_7(S_1,S_2,AL)=\frac{3*AL(S_1,S_2)}{AL(S_1,S_1)+\sqrt{AL(S_1,S_1)*AL(S_2,S_2)}+AL(S_2,S_2)}$	*Heronian Mean*
$sim_8(S_1,S_2,AL)=\frac{AL(S_1,S_2)}{\sqrt{(AL(S_1,S_1)^2+AL(S_2,S_2)^2/2}}$	*Root Mean Square*
$sim_9(S_1,S_2,AL)=\frac{(AL(S_1,S_1)+AL(S_2,S_2))*AL(S_1,S_2)}{AL(S_1,S_1)^2+AL(S_2,S_2)^2}$	*Contraharmonic Mean*

Table 1 - A family of parameterised string similarity measures

4. Experimental design

We have designed our experiments with the aim of generating an automatic system able to learn meaningful information, such as traces of sound correspondences left in the words orthography, and to apply it to the task of cognate identification.

4.1 Datasets

In order to develop our system, we have employed a training dataset and a test dataset with no intersection in their language sets.
The *training dataset* for our learning system has been extracted from the *Comparative Indo-European Database* by Dyen et al. (1992). This corpus contains 200-word Swadesh lists (Swadesh 1952) of universal, non cultural and stable meanings from eighty-four contemporary Indo-European speech varieties. In it, each word is presented in orthographic format without diacritics, using the 26 letters of the Roman alphabet. The data are grouped by meaning and cognateness, which is reported as certain or doubtful. From all the languages available, we have considered three Romance languages (Italian, Portuguese and Spanish) and three Germanic languages (Dutch, Danish and Swedish) to have a balanced training dataset able to learn traces of sound correspondences of most of the language families of which the test dataset is made, but contemporarily avoiding any overlap between the languages of the training and test datasets. From this group of six languages, we have extracted about 650 cognate pairs by considering only the word pairs reported by Dyen et al. (1992) as certain cognates and only the first cognate pair, if more words were provided for the same meaning in the same language. We have corrected a few evident errors. We have then automatically aligned these word pairs as described in Section 3.1.
The *test dataset* consists of the orthographic form of the 200-word Swadesh lists (Swadesh 1952) of English, German, French, Latin and Albanian provided by Kessler (2001) enhanced with his cognateness information. We have discovered two inconsistencies[1] related to the cognation of two French - German word pairs, as the author has confirmed. To make our results properly comparable with others reported in the literature (Mackay and Kondrak 2005; Kondrak and Sherif 2006) where the same test datasets have been used, we have decided not to correct the mentioned errors.

4.2 Evaluation methodology

Cognate identification is an excellent method of measuring the ability of a word similarity evaluation system. We have examined pairs of words belonging to different languages but having the same meaning, for which the cognateness is known information. Ten language pairs deriving from the combination of the five languages present in the test dataset have been considered.
We have produced two families of PAM-like matrices, one based on the Roman alphabet and one on its extension with gap, as proposed in Section 3.3. The two learning models, trained on the 6 language dataset described in Section 4.1, have been named respectively DAY6 and DAY6b.
We have employed these families of PAM-like matrices to align and rate the word pairs of the test dataset with the basic sequence alignment algorithms (Needleman and Wunsch 1970; Smith and Waterman 1981; Gotoh 1982) and the family of parameterised similarity measures proposed in Section 3.4. For the model based on the Roman alphabet, a unary gap penalty has been applied in the alignment algorithms. Our aim has been to assign a score to each word pair that represents how likely the words are to be cognates. The calculated rates, which are relative to each other and do not reflect any universal scale, have then been ordered. When more word pairs have showed the same rate, the alphabetic order has been considered as well, to avoid random results and make the experiments reproducible. We have expected to find

[1] The Latin word *folium*, meaning "leaf", is reported to be cognate with the French *feuille* and the German *Blatt*, but the latter two are not reported as cognates with each other. The same happens to the Latin word *collum*, meaning "neck" with the French *cou* and German *Hals*.

high density of true cognates at the top of the list and low density of true cognates at its bottom. To appraise our string similarity system on the task of cognate identification, we have not used a score threshold that may be influenced by the type of application, the method used and the degree of language relatedness (Kondrak 2009). Instead, we have borrowed from the field of *Information Retrieval*, a measure designed specifically to evaluate rankings, the *11-point interpolated average precision* (Manning and Schütze 1999). For each level of recall $R \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$, it calculates the interpolated precision, which is the highest precision found for any recall level $R' \geq R$, and then it averages these eleven values. This measure has also been frequently used by other systems in the field of cognate recognition (Mackay and Kondrak 2005; Kondrak and Sherif 2006) with which we wanted to make our results properly comparable. For the same reason, we have not distinguished between cognates and borrowings.

5. Experimental results

We have employed the Needleman-Wunsch algorithm (NW) for global alignment and the Smith-Waterman algorithm (SW) for local alignment with the novel family of string similarity measures to evaluate the performance of our cognate identification system. For each PAM-like matrix and for each similarity measure, we have computed the 11-point interpolated average precision for each of the ten language pairs of our test dataset and then we have calculated their average, standard deviation, variance and median. The two models DAY6 and DAY6b achieve very good results especially when employing local alignment with the Smith-Waterman algorithm, even if the difference when using global alignment is not significant. DAY6b, that utilises the Roman alphabet extended with gap, achieves the best results suggesting that the system is also able to learn appropriate gap penalties. Figure 1 shows the results produced by the first ten PAM-like matrices of DAY6b, using NW and SW respectively. As the identity matrix can be considered as a PAM matrix at 0 evolutionary distance, it has been included for completeness.

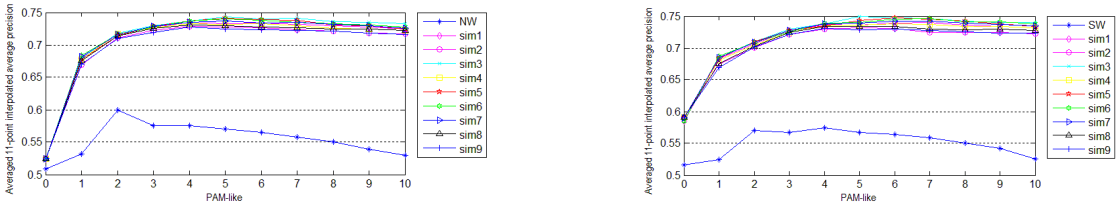

Figure 1 - Averaged 11-point interpolated average precision for DAY6b using NW and SW

The PAM-like matrices PAM4, PAM5, PAM6 and PAM7 produce the higher averaged 11-point interpolated average precisions for all the family of similarity measures. All the similarity measures proposed perform consistently well and outperform the basic algorithms on which they are based.

5.1 Related works

Mackay (2004) on the task of cognate identification followed the orthographic approach and developed a suite of Pair Hidden Markov Model (PHMM) variations and training algorithms based on a model originally presented by Durbin et al. (1998). The training dataset consisted of about 120,000 word pairs extracted from the *Comparative Indo-European Database* by Dyen et al. (1992). A development dataset was used to determine several parameters of the models. *Mackay and Kondrak* (2005) tested this system on the dataset proposed by Kessler

(2001), that also provides word phonetic transcriptions, and they compared it with ALINE (Kondrak 2000). This is an algorithm for phonetic sequence alignment which incorporates linguistic knowledge. Mackay and Kondrak tested the PHMMs also against the Levenshtein distance with Learned Weights (LLW) method, formerly proposed by Mann and Yarowsky (2001) in the task of lexicon translation. LLW learned the costs for edit operations from the same orthographic training dataset using a stochastic transducer. Mackay and Kondrak showed that all the PHMMs outperformed the other methods in terms of 11-point interpolated average precision and the one which produced the better results will be called hereinafter, simply, PHMM.

Kondrak and Sherif (2006) working on orthographic data developed four different models of a Dynamic Bayesian Net previously proposed by *Filali and Bilmes* (2005) in the field pronunciation classification. In order to train their system on the task of cognate recognition, Kondrak and Sherif extracted from the *Comparative Indo-European Database* by Dyen et al. (1992) about 180,000 word pairs. They used them twice to enforce the symmetry of the scoring and they built up a development dataset to set-up the parameters of their system. They also evaluated a group of other phonetic and orthographic algorithms, including ALINE (Kondrak 2000), LLW (Mann and Yarowsky 2001), and PHMM (Mackay and Kondrak 2005), and tested them on the dataset proposed by Kessler (2001). One of the DBN, called hereinafter only DBN, outperformed in terms of 11-point interpolated average precision all the other systems including PHMM, but not significantly.

Kondrak (2009) investigated identification of cognates and recurrent sound correspondences testing several phonetic methods on the test dataset provided by Kessler (2001). His best result was achieved combining ALINE (Kondrak 2000) with a sound correspondence-based method trained using a six languages development dataset. This dataset was extracted from the orthographic *Comparative Indo-European Database* by Dyen et al. (1992) and then manually transcribed into a phonetic notation. This system improved the performance of ALINE, but did not outperform in terms of 11-point interpolated average precision the orthographic PHMM and DBN previously described.

All the results presented in this section are quite remarkable because they suggest that orthographic learning models can outperform systems specifically designed for the task of phonetic alignment, like ALINE (Kondrak 2000) and its variations (Kondrak 2009), given enough training data.

5.2 Comparison

Both our models, DAY6 and DAY6b, when using global alignment as well as local alignment, consistently outperform ALINE (Kondrak 2000), PHMM (Mackay and Kondrak 2005) and DBN (Kondrak and Sherif 2006) in terms of 11-point interpolated average precision in the task of cognate identification. Table 2 shows the proportion of cognate per language pair and a comparison of all the systems considered including our best results produced by DAY6 and DAY6b when utilising NW and SW. It does not include the method proposed by Kondrak (2009) as only the averaged 11-point interpolated average precision, 0.681, was reported. We have used as a baseline NEDIT, the edit distance with unitary costs (Levenshtein 1966; Gusfield 1997) normalised by the length of the longer string. The 11-point interpolated average precision achieved by ALINE (Kondrak 2000), PHMM (Mackay and Kondrak 2005) and DBN (Kondrak and Sherif 2006) is reported as in the literature.

DAY6b using local alignment produces an averaged 11-point interpolated average precision approximately 5% higher than DBN and PHMM, 18% higher than ALINE and 28% higher than NEDIT. Not only the average of the 11-point interpolated average precision of our sample is higher, but also the standard deviation and variance are much lower, suggesting that

our system is also more consistent in its performance across the different language pairs. This is confirmed by a higher median which indicates the central tendency. When comparing the results produced by PHMM and DBN with each other, it is interesting to observe that while the average of the 11-point interpolated average precision of PHMM and DBN are very close, DBN's standard deviation and variance are much lower, showing a better data distribution.

Languages		Cognate proportion	NEDIT	ALINE	PHMM	DBN	DAY6 NW	DAY6 SW	DAY6b NW	DAY6b SW
English	German	0.590	0.907	0.912	0.930	0.927	0.932	0.937	0.929	0.934
French	Latin	0.560	0.921	0.862	0.934	0.923	0.927	0.930	0.921	0.924
English	Latin	0.290	0.703	0.732	0.803	0.822	0.826	0.833	0.823	0.826
German	Latin	0.290	0.591	0.705	0.730	0.772	0.741	0.759	0.770	0.772
English	French	0.275	0.659	0.623	0.812	0.802	0.811	0.815	0.836	0.830
French	German	0.245	0.498	0.534	0.734	0.645	0.763	0.776	0.796	0.788
Albanian	Latin	0.195	0.561	0.630	0.680	0.676	0.685	0.683	0.690	0.721
Albanian	French	0.165	0.499	0.610	0.653	0.658	0.636	0.607	0.607	0.625
Albanian	German	0.125	0.207	0.369	0.379	0.420	0.508	0.519	0.553	0.552
Albanian	English	0.100	0.289	0.302	0.382	0.446	0.463	0.487	0.503	0.518
Average		0.284	0.584	0.628	0.704	0.709	0.729	0.735	0.743	**0.749**
Standard deviation		0.168	0.231	0.193	0.194	0.176	0.159	0.158	0.149	**0.144**
Variance		0.260	0.054	0.037	0.038	0.031	0.025	0.025	0.022	**0.021**
Median		0.284	0.576	0.627	0.732	0.724	0.752	0.768	0.783	**0.780**

Table 2 - 11-point interpolated average precision of several methods

We have used the same source for the training dataset and the same test dataset that Kondrak and co-workers have used in the design of PHMM (Mackay and Kondrak 2005) and DBN (Kondrak and Sherif 2006). However, there are several aspects that differentiate considerably our learning approach, including the dimension of the training dataset used, its quality and its meaningfulness. In fact we have employed less than 1% of the data they utilised and we have considered only word pairs reported by Dyen et al. (1992) as certain cognates. Moreover, we have automatically aligned the cognate pairs from which the system has to learn, using a substitution matrix that incorporates some linguistic knowledge in an attempt to generate a meaningful training dataset. It is also worth noting that our system accommodates quite well the Albanian language that makes the test dataset challenging. In fact Albanian constitutes its own branch in the Indo-European language family and it is not part of the language families with which our system has been trained.

6. Conclusion

We have developed a learning system for the task of cognate identification and we have shown its superior performance when compared with the best phonetic and orthographic systems previously proposed in the literature. Our results reinforce the hypothesis that orthographic learning systems may recognise traces of sound correspondences left in the words orthography and can perform better than phonetic static models. This idea is very encouraging considering that phonetic transcriptions are very difficult to produce and frequently performed by hand with the consequent loss of time and the possible lack of accuracy and uniformity. Our PAM-like matrices, together with our new family of similarity measures, may help to identify distant relationships between languages, where controversies still exist, and to analyse less studied language families.

Our future objective is to continue investigating substitution matrices for the tasks of cognate recognition and word similarity. In particular, we would like to study the influence of the

training dataset dimension on our system performance. Another step forward would be to apply our methodology to the investigation of language evolution.

Acknowledgments

We would like to thank Ana Fortun for her contribution to the linguistic-inspired substitution matrix, Grzegorz Kondrak for providing his version of the test dataset and Brett Kessler for commenting on his lists.

Antonella Delmestri
Department of Information Engineering and Computer Science, University of Trento, Italy.
E-mail: antonella.delmestri@gmail.com

Nello Cristianini
Intelligent Systems Laboratory, University of Bristol, U.K.
E-mail: nello@support-vector.net

References

Brew, C., McKelvie, D. 1996. Word-pair extraction for lexicography. In *Proceedings of the 2nd International Conference on New Methods in Language Processing (NEMLP)*, Ankara, Turkey, 45-55.

Darwin, C.R. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London, U.K.: John Murray.

Dayhoff, M.O., Eck, R.V. 1968. A Model of Evolutionary Change in Proteins. *Atlas of Protein Sequence and Structure 1967-1968*, 3: 33-41. National Biomedical Research Foundation, Silver Spring, Maryland.

Dayhoff, M.O., Eck, R.V., Park, C.M. 1972. A Model of Evolutionary Change in Proteins. *Atlas of Protein Sequence and Structure*, 5: 89-99. National Biomedical Research Foundation, Washington, D.C., U.S.A.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. 1978. A Model of Evolutionary Change in Proteins. *Atlas of Protein Sequence and Structure*, 5(3): 345-352. National Biometical Research Foundation, Washington, D.C., U.S.A.

Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G. 1998. *Biological Sequence Analysis*. Cambridge, U.K.: Cambridge University Press.

Dyen, I., Kruskal, J.B., Black, P. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5).

Filali, K., Bilmes, J. 2005. A Dynamic Bayesian Framework to Model Context and Memory in Edit Distance Learning: An Application to Pronunciation Classification. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann-Arbor, Michigan, U.S.A., 338-345.

Gotoh, O. 1982. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3): 705-708.

Gusfield, D. 1997. *Algorithms on Strings, Trees and Sequences*. New York, U.S.A.: Cambridge University Press.

Henikoff, S., Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89: 10915–10919.

Kessler, B. 2001. *The Significance of Word Lists*. Stanford, California, U.S.A.: CSLI Publications.

Kondrak, G. 2000. A New Algorithm for the Alignment of Phonetic Sequences. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000)*, Seattle, Washington, U.S.A., 288-295. Morgan Kaufmann Publishers Inc., San Francisco, California, U.S.A.

Kondrak, G. 2005. Cognates and Word Alignment in Bitexts. In *Proceedings of the 10th Machine Translation Summit (MT Summit X)*, Phuket, Thailand, 305-312.

Kondrak, G. 2009. Identification of Cognates and Recurrent Sound Correspondences in Word Lists. *Traitement automatique des langues*, 50(2): 201-235.

Kondrak, G., Dorr, B.J. 2004. Identification of Confusable Drug Names: A New Approach and Evaluation Methodology. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, 952-958.

Kondrak, G., Marcu, D., Knight, K. 2003. Cognates Can Improve Statistical Translation Models. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003) companion volume*, Edmonton, Alberta, Canada, 46-48.

Kondrak, G., Sherif, T. 2006. Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification. In *Proceedings of the COLING-ACL 2006 Workshop on Linguistic Distances*, Sydney, Australia, 43-50.

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

Mackay, W. 2004. *Word similarity using Pair Hidden Markov Models*. Master's thesis. University of Alberta.

Mackay, W., Kondrak, G. 2005. Computing word similarity and identifying cognates with Pair Hidden Markov Models. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, Ann Arbor, Michigan, U.S.A., 40–47.

Manning, C.D., Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts, U.S.A.: The Massachusetts Institute of Technology (MIT) Press.

Mann, G.S., Yarowsky, D. 2001. Multipath Translation Lexicon Induction via Bridge Languages. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, Pittsburgh, Pennsylvania, U.S.A., 151-158.

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

Needleman, S.B., Wunsch, C.D. 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.

Smith, T.F., Waterman, M.S. 1981. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147(1): 195-197.

Swadesh, M. 1952. Lexico-Statistic Dating of Prehistoric Ethnic Contacts. *Proceedings of the American Philosophical Society*, 96(4): 452-463.

Tiedemann, J. 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 (EMNLP/VLC-99)*, College Park, Maryland, U.S.A., 213-219.