Automatic Discrimination between Cognates and Borrowings

Alina Maria Ciobanu, Liviu P. Dinu
Faculty of Mathematics and Computer Science, University of Bucharest
Center for Computational Linguistics, University of Bucharest
alina.ciobanu@my.fmi.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

Identifying the type of relationship between words provides a deeper insight into the history of a language and allows a better characterization of language relatedness. In this paper, we propose a computational approach for discriminating between cognates and borrowings. We show that orthographic features have discriminative power and we analyze the underlying linguistic factors that prove relevant in the classification task. To our knowledge, this is the first attempt of this kind.

1 Introduction

Natural languages are living eco-systems. They are subject to continuous change due, in part, to the natural phenomena of language contact and borrowing (Campbell, 1998). According to Hall (1960), there is no such thing as a "pure language" – a language "without any borrowing from a foreign language". Although admittedly regarded as relevant factors in the history of a language (McMahon et al., 2005), borrowings bias the genetic classification of the languages, characterizing them as being closer than they actually are (Minett and Wang, 2003). Thus, the need for discriminating between cognates and borrowings emerges. Heggarty (2012) acknowledges the necessity and difficulty of the task, emphasizing the role of the "computerized approaches".

In this paper we address the task of automatically distinguishing between borrowings and cognates: given a pair of words, the task is to determine whether one is a historical descendant of the other, or whether they both share a common ancestor. A *borrowing* (also called *loanword*), is defined by Campbell (1998) as a "lexical item (a word) which has been 'borrowed' from another language, a word which originally was not part of the vocabulary of the recipient language but was adopted from some other language and made part of the borrowing language's vocabulary". The notion of *cognate* is much more relaxed, and various NLP tasks and applications use different definitions of the cognate pairs. In some situations, cognates and borrowings are considered together, and are referred to as *historically connected words* (Kessler, 2001) or denoted by the term *correlates* (Heggarty, 2012; McMahon et al., 2005). In some tasks, such as statistical machine translation (Kondrak et al., 2003) and sentence alignment, or when studying the similarity or intelligibility of the languages, cognates are seen as words that have similar spelling and meaning, their etymology being completely disregarded. However, in problems of language classification, distinguishing cognates from borrowings is essential. Here, we account for the etymology of the words, and we adopt the following definition: two words form a cognate pair if they share a common ancestor and have the same meaning. In other words, they derive directly from the same word, have a similar meaning and, due to various (possibly language-specific) changes across time, their forms might differ.

2 Related Work

In a natural way, one of the most investigated problems in historical linguistics is to determine whether similar words are related or not (Kondrak, 2002). Investigating pairs of related words is very useful not only in historical and comparative linguistics, but also in the study of language relatedness (Ng et al., 2010), phylogenetic inference (Atkinson et al., 2005) and in identifying how and to what extent languages changed over time or influenced each other.

Most studies in this area focus on automatically identifying pairs of cognates. For measuring the orthographic or phonetic proximity of the cognate candidates, string similarity metrics (Inkpen

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 431–437,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

et al., 2005; Hall and Klein, 2010) and algorithms for string alignment (Delmestri and Cristianini, 2010) have been applied, both in cognate detection (Koehn and Knight, 2000; Mulloni and Pekar, 2006; Navlea and Todirascu, 2011) and in cognate production (Beinborn et al., 2013; Mulloni, 2007). Minett and Wang (2003) focus on identifying borrowings within a family of genetically related languages and propose, to this end, a distance-based and a character-based technique. Minett and Wang (2005) address the problem of identifying language contact, building on the idea that borrowings bias the lexical similarities among genetically related languages.

According to the regularity principle, the distinction between cognates and borrowings benefits from the regular sound changes that generate regular phoneme correspondences in cognates (Kondrak, 2002). In turn, sound correspondences are represented, to a certain extent, by alphabetic character correspondences (Delmestri and Cristianini, 2010).

3 Our Approach

In light of this, we investigate whether cognates can be automatically distinguished from borrowings based on their orthography. More specifically, our task is as follows: given a pair of words in two different languages (x, y), we want to determine whether x and y are cognates or if y is borrowed from x (in other words, x is the etymon of y).

Our starting point is a methodology that has previously proven successful in discriminating between related and unrelated words (Ciobanu and Dinu, 2014b). Briefly, the method comprises the following steps:

1) Aligning the pairs of related words using a string alignment algorithm;
2) Extracting orthographic features from the aligned words;
3) Training a binary classifier to discriminate between the two types of relationship.

Lang.	**Cognates**			**Borrowings**		
	len$_1$	**len$_2$**	**edit**	**len$_1$**	**len$_2$**	**edit**
It-Ro	7.95	8.78	0.26	7.58	8.41	0.29
Es-Ro	7.91	8.33	0.26	5.78	6.14	0.52
Pt-Ro	7.99	8.35	0.28	5.35	5.42	0.52
Tr-Ro	7.35	6.88	0.31	6.49	6.09	0.44

Table 2: Statistics for the dataset of related words. Given a pair of languages *(L_1, L_2)*, the **len$_1$** and **len$_2$** columns represent the average word length of the words in L_1 and L_2. The **edit** column represents the average normalized edit distance between the words. The values are computed only on the training data, to keep the test data unseen.

To align the pairs of related words, we employ the Needleman-Wunsch global alignment algorithm (Needleman and Wunsch, 1970), which is equivalent to the weighted edit distance algorithm. We consider words as input sequences and we use a very simple substitution matrix[1], which assigns equal scores to all substitutions, disregarding diacritics (e.g., we ensure that *e* and *è* are matched). As features, we use characters n-grams extracted from the alignment[2]. We mark word boundaries with $ symbols. For example, the Romanian word *funcţie* (meaning *function*) and its Spanish cognate pair *función* are aligned as follows:

```
$ f u n c ţ i e - $
$ f u n c - i ó n $
```

The features for n = 2 are:

```
$f≻$f, fu≻fu, un≻un, nc≻nc, cţ≻c-,
ţi≻-i, ie≻ió, e-≻ón, -$≻n$.
```

For the prediction task, we experiment with two models, Naive Bayes and Support Vector Machines. We extend the method by introducing additional linguistic features and we conduct an analysis on their predictive power.

4 Experiments and Results

In this section we present and analyze the experiments we run for discriminating between cognates and borrowings.

4.1 Data

Our experiments revolve around Romanian, a Romance language belonging to the Italic branch of the Indo-European language family. It is surrounded by Slavic languages and its relationship with the big Romance kernel was difficult. Its geographic position, at the North of the Balkans, put

[1] In our future work, we intend to also experiment with more informed language-specific substitution matrices.

[2] While the original methodology proposed features extracted around mismatches in the alignment, we now compare two approaches: 1) features extracted around mismatches, and 2) features extracted from the entire alignment. The latter approach leads to better results, as measured on the test set.

Lang.	Borrowings			Cognates			
IT-RO	baletto	→	balet (ballet)	vittoria	-	victorie (victory)	↑ victoria (LAT)
PT-RO	selva	→	selvă(selva)	instinto	-	instinct (instinct)	↑ instinctus (LAT)
ES-RO	machete	→	macetă (machete)	castillo	-	castel (castle)	↑ castellum (LAT)
TR-RO	tütün	→	tutun (tobacco)	aranjman	-	aranjament (arrangement)	↑ arrangement (FR)

Table 1: Examples of borrowings and cognates. For cognates we also report the common ancestor.

it in contact not only with the Balkan area, but also with the vast majority of Slavic languages. Political and administrative relationships with the Ottoman Empire, Greece (the Phanariot domination) and the Habsburg Empire exposed Romanian to a wide variety of linguistic influences. We apply our method on four pairs of languages extracted from the dataset proposed by Ciobanu and Dinu (2014c):

- Italian - Romanian (IT-RO);
- Portuguese - Romanian (PT-RO);
- Spanish - Romanian (ES-RO);
- Turkish - Romanian (TR-RO).

For the first three pairs of languages, which are formed of *sister languages*[3], most cognate pairs have a Latin common ancestor, while for the fourth pair, formed of languages belonging to different families (Romance and Turkic), most of the cognate pairs have a common French etymology, and date back to the end of the 19th century, when both Romanian and Turkish borrowed massively from French. In Table 1 we provide examples of borrowings and cognates.

The dataset contains borrowings[4] and cognates that share a common ancestor. The words (and information about their origins) were extracted from electronic dictionaries and their relationships were determined based on their etymology. We use a stratified dataset of 2,600 pairs of related words for each pair of languages. In Table 2 we provide an initial analysis of our dataset. We report statistics regarding the length of the words and the edit distance between them. The difference in length between the related words shows what operations to expect when aligning the words. Romanian words are almost in all situations shorter, in average, than their pairs. For TR-RO $\mathbf{len_1}$ is higher than $\mathbf{len_2}$, so we expect more deletions for this pair of languages. The **edit** columns show how much words vary from one language to another based on their relationship (cognates or borrowings). For IT-RO both distances are small (0.26 and 0.29), as opposed to the other languages, where there is a more significant difference between the two (e.g., 0.26 and 0.52 for ES-RO). The small difference for IT-RO might make the discrimination between the two classes more difficult.

4.2 Baselines

Given the initial analysis presented above, we hypothesize that the distance between the words might be indicative of the type of relationship between them. Previous studies (Inkpen et al., 2005; Gomes and Lopes, 2011) show that related and non-related words can be distinguished based on the distance between them, but a finer-grained task, such as determining the type of relationship between the words, is probably more subtle. We compare our method with two baselines:

- A baseline which assigns a label based on the normalized edit distance between the words: given a test instance pair $word_1$ - $word_2$, we subtract the average normalized edit distance between $word_1$ and $word_2$ from the average normalized edit distance of the cognate pairs and from the average normalized edit distance between the borrowings and their etymons (computed on the training set; see Table 2), and assign the label which yields a smaller difference (in absolute value). In case of equality, the label is chosen randomly.

- A decision tree classifier, following the strategy proposed by Inkpen et al. (2005): we use the normalized edit distance as single feature, and we fit a decision tree classifier with the maximum tree depth set to 1. We perform 3-fold cross-validation in order to select the best threshold for discriminating between borrowings and cognates. Using the

[3] Sister languages are “languages which are related to one another by virtue of having descended from the same common ancestor (proto-language)” (Campbell, 1998).

[4] Romanian is always the recipient language in our dataset (i.e., the language that borrowed the words).

best threshold selected for each language, we further assign one of the two classes to the pairs of words in our test set.

4.3 Task Setup

We experiment with Naive Bayes and Support Vector Machines (SVMs) to learn orthographic changes. We put our system together using the Weka[5] workbench (Hall et al., 2009). For SVM, we employ the radial basis function kernel (RBF) and we use the wrapper provided by Weka for LibSVM (Chang and Lin, 2011). For each language pair, we split the dataset in two stratified subsets, for training and testing, with a 3:1 ratio. We experiment with different values for the n-gram size (n $\in$ $\{1, 2, 3\}$) and we perform grid search and 3-fold cross validation over the training set in order to optimize hyperparameters *c* and γ for SVM. We search over $\{1, 2, ..., 10\}$ for *c* and over $\{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$ for γ.

4.4 Results Analysis

Table 3 and Table 4 show the results of our experiment. The two baselines produce comparable results. For all pairs of languages, our method significantly improves over the baselines (99% confidence level)[6] with values between 7% and 29% for the F_1 score, suggesting that the n-grams extracted from the alignment of the words are better indicators of the type of relationship than the edit distance between them. The best results are obtained for TR-RO, with an F_1 score of 92.1, followed closely by PT-RO with 90.1 and ES-RO with 85.5. These results show that, for these pairs of languages, the orthographic cues are different with regard to the relationship between the words. For IT-RO we obtain the lowest F_1 score, 69.0.

In this experiment, we know beforehand that there is a relationship between the words, and our aim is to identify the type of relationship. However, in many situations this kind of a-priori information is not available. In a real scenario, we would have either to add an intermediary classifier for discriminating between related and unrelated words, or to discriminate between three classes: cognates, borrowings, and unrelated. We augment our dataset with unrelated words (determined based on their etymology), building a stratified dataset annotated with three classes, and we repeat the previous experiment. The performance decreases[7], but the results are still significantly better than chance (99% confidence level).

Lang.	Baseline #1			Baseline #2		
	P	R	F_1	P	R	F_1
IT-RO	50.7	50.7	50.7	64.4	54.5	45.0
PT-RO	79.3	79.0	79.2	80.1	80.0	80.0
ES-RO	78.6	78.4	78.5	78.6	78.5	78.4
TR-RO	61.1	61.0	61.1	62.5	59.8	57.6

Table 3: Weighted average precision (P), recall (R) and F_1 score (F_1) for automatic discrimination between cognates and borrowings.

4.5 Linguistic Factors

To gain insight into the factors with high predictive power, we perform several further experiments.

Part of speech. We investigate whether adding knowledge about the part of speech of the words leads to performance improvements. Verbs, nouns, adverbs and adjectives have language-specific endings, thus we assume that part of speech might be useful when learning orthographic patterns. We obtain POS tags from the DexOnline[8] machine-readable dictionary. We employ the POS feature as an additional categorical feature for the learning algorithm. It turns out that, except for PT-RO (F_1 score 92.3), the additional POS feature does not improve the performance of our method.

Syllabication. We analyze whether the system benefits from using the syllabified form of the words as input to the alignment algorithm. We are interested to see if marking the boundaries between the syllables improves the alignment (and, thus, the feature extraction). We obtain the syllabication for the words in our dataset from the RoSyllabiDict dictionary (Barbu, 2008) for Romanian words and several available Perl modules[9] for the other languages. For PT-RO and ES-RO the F_1 score increases by about 1%, reaching a value of 93.4 for the former and 86.7 for the latter.

[5] www.cs.waikato.ac.nz/ml/weka

[6] All the statistical significance tests reported in this paper are performed on 1,000 iterations of paired bootstrap resampling (Koehn, 2004).

[7] Weighted average F_1 score on the test set for SVM: IT-RO 63.8, PT-RO 77.6, ES-RO 74.0, TR-RO 86.1.

[8] www.dexonline.ro

[9] Lingua::ID::Hyphenate modules where ID $\in$ $\{$IT, PT, ES, TR$\}$, available on the Comprehensive Perl Archive Network: www.cpan.org.

Lang.	Naive Bayes				SVM					
	P	R	F_1	n	P	R	F_1	n	c	γ
IT-RO	68.6	68.2	68.3	3	69.2	69.1	69.0	3	10	0.10
PT-RO	92.6	91.7	92.1	3	90.1	90.0	90.0	3	3	0.10
ES-RO	85.3	84.5	84.9	3	85.7	85.5	85.5	2	2	0.10
TR-RO	89.7	89.4	89.5	3	90.3	90.2	90.1	3	6	0.01

Table 4: Weighted average precision (P), recall (R), F_1 score (F_1) and optimal n-gram size for automatic discrimination between cognates and borrowings. For SVM we also report the optimal values for c and γ.

Consonants. We examine the performance of our system when trained and tested only on the aligned *consonant skeletons* of the words (i.e., a version of the words where vowels are discarded). According to Ashby and Maidment (2005), consonants change at a slower pace than vowels across time; while the former are regarded as reference points, the latter are believed to carry less information useful for identifying the words (Gooskens et al., 2008). The performance of the system decreases when vowels are removed (95% confidence level). We also train and test the decision tree classifier on this version of the dataset, and its performance is lower in this case as well (95% confidence level), indicating that, for our task, the information carried by the vowels is helpful.

Stems. We repeat the first experiment using stems as input, instead of lemmas. What we seek to understand is whether the aligned affixes are indicative of the type of relationship between the words. We use the Snowball Stemmer[10] and we find that the performance decreases when stems are used instead of lemmas. Performing a χ^2 feature ranking on the features extracted from mismatches in the alignment of the related words reveals further insight into this matter: for all pairs of languages, at least one feature containing the $ character (indicating the beginning or the end of a word) is ranked among the 10 most relevant features, and over 50 are ranked among the 500 most relevant features. This suggests that prefixes and suffixes (usually removed by the stemmer) vary with the type of relationship between the words.

Diacritics. We explore whether removing diacritics influences the performance of the system. Many words have undergone transformations by the augmentation of language-specific diacritics when entering a new language (Ciobanu and Dinu, 2014a). For this reason, we expect diacritics to play a role in the classification task. We observe that, when diacritics are removed, the F_1 score on the test set is lower in almost all situations. Analyzing the ranking of the features extracted from mismatches in the alignment provides even stronger evidence in this direction: for all pairs of languages, more than a fifth of the top 500 features contain diacritics.

5 Conclusions

In this paper, we propose a computational method for discriminating between cognates and borrowings based on their orthography. Our results show that it is possible to identify the type of relationship with fairly good performance (over 85.0 F_1 score) for 3 out of the 4 pairs of languages we investigate. Our predictive analysis shows that the orthographic cues are different for cognates and borrowings, and that underlying linguistic factors captured by our model, such as affixes and diacritics, are indicative of the type of relationship between the words. Other insights, such as the syllabication or the part of speech of the words, are shown to have little or no predictive power. We intend to further account for finer-grained characteristics of the words and to extend our experiments to more languages. The method we propose is language-independent, but we believe that incorporating language-specific knowledge might improve the system's performance.

Acknowledgements

We thank the anonymous reviewers for their helpful and constructive comments. The contribution of the authors to this paper is equal. Liviu P. Dinu was supported by UEFISCDI, PNII-ID-PCE-2011-3-0959.

[10] http://snowball.tartarus.org

References

Michael Ashby and John Maidment. 2005. *Introducing Phonetic Science*. Cambridge University Press.

Quentin D. Atkinson, Russell D. Gray, Geoff K. Nicholls, and David J. Welch. 2005. From Words to Dates: Water into Wine, Mathemagic or Phylogenetic Inference? *Transactions of the Philological Society*, 103(2):193–219.

Ana-Maria Barbu. 2008. Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 1937–1941.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate Production using Character-based Machine Translation. In *Proceedings of the 6th International Joint Conference on Natural Language Processing, IJCNLP 2013*, pages 883–891.

Lyle Campbell. 1998. *Historical Linguistics. An Introduction*. MIT Press.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Alina Maria Ciobanu and Liviu P. Dinu. 2014a. An Etymological Approach to Cross-Language Orthographic Similarity. Application on Romanian. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1047–1058.

Alina Maria Ciobanu and Liviu P. Dinu. 2014b. Automatic Detection of Cognates Using Orthographic Alignment. In *Proceedings of the 52st Annual Meeting of the Association for Computational Linguistics, ACL 2014*, pages 99–105.

Alina Maria Ciobanu and Liviu P. Dinu. 2014c. Building a Dataset of Multilingual Cognates for the Romanian Lexicon. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*.

Antonella Delmestri and Nello Cristianini. 2010. String Similarity Measures and PAM-like Matrices for Cognate Identification. *Bucharest Working Papers in Linguistics*, 12(2):71–82.

Luís Gomes and José Gabriel Pereira Lopes. 2011. Measuring Spelling Similarity for Cognate Identification. In *Proceedings of the 15th Portugese Conference on Progress in Artificial Intelligence, EPIA 2011*, pages 624–633. Software available at `http://research.variancia.com/spsim`.

Charlotte Gooskens, Wilbert Heeringa, and Karin Beijering. 2008. Phonetic and Lexical Predictors of Intelligibility. *International Journal of Humanities and Arts Computing*, 2(1-2):63–81.

David Hall and Dan Klein. 2010. Finding Cognate Groups Using Phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010*, pages 1030–1039.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Robert Anderson Hall. 1960. *Linguistics and Your Language*. Doubleday New York.

Paul Heggarty. 2012. Beyond Lexicostatistics: How to Get More out of ”Word List” Comparisons. In *Quantitative Approaches to Linguistic Diversity: Commemorating the Centenary of the Birth of Morris Swadesh*, pages 113–137. Benjamins.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic Identification of Cognates and False Friends in French and English. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2005*, pages 251–257.

Brett Kessler. 2001. *The Significance of Word Lists*. Stanford: CSLI Publications.

Philipp Koehn and Kevin Knight. 2000. Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm. In *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pages 711–715.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004*, pages 388–395.

Grzegorz Kondrak, Daniel Marcu, and Keven Knight. 2003. Cognates Can Improve Statistical Translation Models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, HLT-NAACL 2003*, pages 46–48.

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

April McMahon, Paul Heggarty, Robert McMahon, and Natalia Slaska. 2005. Swadesh Sublists and the Benefits of Borrowing: an Andean Case Study. *Transactions of the Philological Society*, 103(2):147–170.

James W. Minett and William S.-Y. Wang. 2003. On Detecting Borrowing: Distance-based and Character-based Approaches. *Diachronica*, 20(2):289–331.

James W. Minett and William S.-Y. Wang. 2005. Vertical and Horizontal Transmission in Language Evolution. *Transactions of the Philological Society*, 103(2):121–146.

Andrea Mulloni and Viktor Pekar. 2006. Automatic Detection of Orthographic Cues for Cognate Recognition. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 2387–2390.

Andrea Mulloni. 2007. Automatic Prediction of Cognate Orthography Using Support Vector Machines. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop, ACL 2007*, pages 25–30.

Mirabela Navlea and Amalia Todirascu. 2011. Using Cognates in a French-Romanian Lexical Alignment System: A Comparative Study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2011*, pages 247–253.

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.

Ee-Lee Ng, Beatrice Chin, Alvin W. Yeo, and Bali Ranaivo-Malançon. 2010. Identification of Closely-Related Indigenous Languages: An Orthographic Approach. *International Journal of Asian Language Processing*, 20(2):43–62.