A Classification-Based Approach to Cognate Detection Combining Orthographic and Semantic Similarity Information

Sofie Labat and Els Lefever
LT3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
sofie.labat@gmail.com, els.lefever@ugent.be

Abstract

This paper presents proof-of-concept experiments for combining orthographic and semantic information to distinguish cognates from non-cognates. To this end, a context-independent gold standard is developed by manually labelling English-Dutch pairs of cognates and false friends in bilingual term lists. These annotated cognate pairs are then used to train and evaluate a supervised binary classification system for the automatic detection of cognates. Two types of information sources are incorporated in the classifier: fifteen string similarity metrics capture form similarity between source and target words, while word embeddings model semantic similarity between the words. The experimental results show that even though the system already achieves good results by only incorporating orthographic information, the performance further improves by including semantic information in the form of embeddings.

1 Introduction

In general linguistics, the term *cognate* is defined as a "language or a linguistic form which is historically derived from the same source as another language/form" (Crystal, 2008, page 83). The assumption of common etymology is, however, often disregarded in the literature, because certain research areas such as psycho-linguistics or natural language processing (NLP) tend to shift their focus from diachronic to perceptual relatedness (Shlesinger and Malkiel, 2005; Mitkov et al., 2007; Schepens et al., 2013; Hansen-Schirra et al., 2017). We follow this second strand of research in that we define cognates as words with high formal and semantic cross-lingual similarity. Conversely, *false friends* are words which have similar forms, but which differ in their meaning.

The ability to distinguish cognates from non-cognates (and especially) false friends is an important skill for second language learners. Similarly, source language interference is a problem often experienced by translators that is partly caused by the influence of cognates and false friends. Research in natural language processing can address these bottlenecks by, for instance, developing computer tools that aid second language users.

Nevertheless, most studies have mainly focused on the detection of cognates (Bergsma and Kondrak, 2007; Hauer and Kondrak, 2011; Ciobanu and Dinu, 2014; Rama, 2016), while relatively little attention has been devoted to false friends (Frunza and Inkpen, 2007; Mitkov et al., 2007; Ljubešić and Fišer, 2013; Castro et al., 2018). Mitkov (2007) explains that the main goal of investigation is often the cross-lingual identification of equivalent lexical items, as such knowledge can be integrated in other applications. Automatic cognate detection has indeed proven very useful for NLP, e.g. to boost the performance of automatic alignment between related languages or to compile bilingual lexicons (Smith et al., 2017).

The aim of this research is twofold: (1) we introduce a context-independent gold standard which can be used to classify English-Dutch pairs of cognates and non-cognates (among which false friends); (2) we develop a supervised binary classifier able to identify cognates across the English-Dutch language pair on the basis of orthographic and semantic information. Since our focus lies on the detection of cognates for these proof-of-concept experiments, no distinction is made between false friends and non-equivalent words.

The remainder of this paper is organized as follows. In Section 2, we give a brief overview of

Proceedings of Recent Advances in Natural Language Processing, pages 602–610,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_071

the existing research and methodologies to cognate detection. Section 3 describes the data and annotation process used to create the context-independent gold standard for English-Dutch cognate pairs, while Section 4 gives an overview of the experimental setup and the two types of information sources, viz. orthographic and semantic similarity features, that were used. In section 5, we report on the results of our classifier (1) incorporating only orthographic features and (2) combining orthographic and semantic similarity features. Section 6 concludes this paper and gives directions for future research.

2 Related Research

Extensive lists of known cognates and false friends are hard to find and expensive to compose, since they require a considerable amount of time and effort from trained lexicographers (Schepens et al., 2012). Especially for low resource languages, this constitutes a serious issue. Therefore, most NLP research on cognates has mainly focused on the automatic detection of such cognate pairs. In the literature, there are three main methods to identify cognates: orthographic, phonetic and semantic approaches. The oldest approaches to tackle this task involve simple string similarity metrics as the longest common subsequence ratio (Melamed, 1999) or the normalized Levenshtein distance (Levenshtein, 1965). More recently, however, the attention has been drawn to machine learning techniques. For instance, Frunza et al. (2007) combine several orthographic similarity measures to train a machine classifier, while Gomes et al. (2011) design a new similarity metric that is able to learn spelling differences across languages.

Different types of approaches can also be combined to distinguish cognates, e.g. Kondrak et al. (2004) join orthographic and phonetic information to distinguish between similar drug names. In order to capture the phonetic similarity between words, Konrak (2000) further developed a software package, called ALINE, which portrays phonemes as vectors of phonetic features, thus creating a phonetic similarity measure. Nevertheless, Heeringa et al. (2010) find that simple phonetic transcriptions still seem to outperform phonetic similarity metrics that are based on phonetic features. Hence, Schepens et al. (2013) propose to calculate a substitution cost for each pair of phonetically transcribed words by taking the edit distance between them. For this research, we opted to only focus on the orthographic proximity, as sound metrics require an additional phonetic transcription, thus making them less-likely to be applied on large data sets. Moreover, Schepens et al. (2013) find that there is a high consistency between orthographic and phonetic similarity measures for Dutch-English cognate pairs.

Whereas orthographic and phonetic features have often been employed to model the similarity between candidate cognate pairs, semantic information has often been ignored. Mitkov (2007) beliefs that this is another result of the main focus of investigation, which is the identification of cognates rather than distinguishing cognates from false friends. Semantic evidence is, however, an important information source, as it can not only be used to represent the semantic (dis)similarity between word pairs, but it can also further increase the accuracy of cognate detection systems. In his own research, Mitkov (2007) distinguishes between two types of semantic approaches: taxonomic and distributional semantic similarity measures. Whereas the first group relies on the taxonomic structure of a resource such as WordNet (Miller, 1995), the second approach relies on large corpora. The latter methods are based on the Distributional Hypothesis (Harris, 1954), which states that words that appear in similar contexts tend to share similar meanings. The different approaches that leverage this principle are typically divided into two categories: count-based methods, such as Latent Semantic Analysis, and predictive methods, such as neural probabilistic language models, which have gained a lot of popularity in today's NLP community. On the one hand, count-based models count how often a given target word co-occurs with its neighbor words in a large text corpus, after which the resulting counts are mapped to a dense vector for each word. On the other hand, predictive models directly try to predict a word from its neighbors in terms of learned dense embedding vectors (Baroni et al., 2014). Word2vec (Mikolov et al., 2013) is a particularly computationally-efficient and popular example of predictive models for learning word embeddings from raw text. In this research, we will incorporate the more recent fastText word embeddings as implemented by Bojanowski et al. (2017).

3 Data Creation

To train and evaluate the cognate detection system, we created a novel context-independent gold standard by manually labelling English-Dutch pairs of cognates and false friends in bilingual term lists. In this section, we describe how the lists of candidate cognate pairs were compiled on the basis of the Dutch Parallel Corpus (Macken et al., 2011) and how a manual annotation was performed to create a gold standard for English-Dutch cognate pairs.

3.1 List of Candidate Cognate Pairs

To select a list of candidate cognate pairs, unsupervised statistical word alignment using GIZA++ (Och and Ney, 2003) was applied on the Dutch Parallel Corpus (DPC). This high-quality parallel corpus for Dutch, French and English consists of more than ten million words and is sentence-aligned. It contains five different text types and is balanced with respect to text type and translation direction. The automatic word alignment on the English-Dutch part of the DPC resulted in a list containing more than 500,000 translation equivalents. A first selection was performed by applying the Normalized Levenshtein Distance (NLD) (as implemented by Gries (2004)) on this list of translation equivalents and only considering equivalents with a distance smaller than or equal to 0.5. This resulted in a list with 28,503 Dutch-English candidate cognate pairs, which was manually labeled.

3.2 Creation of Gold Standard

To create the gold standard for cognate detection, an extensive set of guidelines was established (Labat et al., 2019). The guidelines propose a clearly defined method for the manual labeling of the following six categories:

1. **Cognate**: words which have a similar form and meaning in all contexts. Conform with our working definition for cognates, the source and target words do not need to be etymologically related.

2. **Partial cognate**: words which have a similar form, but only share the same meaning in some contexts.

3. **False friend**: words which have a similar form but a different meaning.

4. **Proper name**: proper nouns (e.g. persons, companies, cities, countries, etc.) and their derivations (e.g. *American*).

5. **Error**: word alignment errors and compound nouns of which one part is a cognate but the other part is missing in one of the languages (e.g. *peripherals - aansturingsperipherals*).

6. **No standard**: words that do not occur in the dictionary (e.g. *num_connectors*) and numbers (e.g. *adm12006e, VI*).

To decide on the correct label, we adopted a context-independent approach applying the following procedure: (1) for every candidate cognate pair, the dictionary Grote Van Dale[1] (henceforth: VD) was consulted; (2) the English word is looked up in the VD, e.g. *salon*, (3) the Dutch translation is inspected in the VD, e.g. *salon: "nice room"* and *salon: "(room for) gathering of people (e.g. from the literary world)"*.

Based on the previously obtained information, a decision is made: in case all meanings of the Dutch word correspond with the English word, we consider them "cognates", in case only part of the Dutch meanings correspond with the English word, we consider them "partial cognates", in case the words have different meanings, we consider them "false friends". An example of partial cognates is the pair *agent-agent*: the Dutch *agent* refers both to (1) a police man and to (2) a representative (e.g. business representative). As only the second meaning of the Dutch word is expressed by the English *agent*, these words are considered partial cognates.

Two important observations should be made. Firstly, we accorded more fine-grained labels in the gold standard that are described in great detail in the annotation guidelines (Labat et al., 2019). For cognates, a distinction was, for instance, made between cognates of which Part-of-Speech (PoS) and meaning are identical in both languages, cognates that differ in PoS (e.g. *organisatie-organizing*) and cognates that differ in agreement (e.g. *organisatie-organisations*). Secondly, it is important to note that a successful dictionary look-up never overruled the "proper name" annotation.

The resulting gold standard is context-independent. Hence, it can be used for both the development and the evaluation of machine

[1] https://www.vandale.be/

learning models that deal with cognate detection. Besides its applications in natural language processing, the gold standard can also form an important new resource for further research on cognates in linguistics, translation studies or psycho-linguistics.

4 Classification

This section describes the experimental setup and the two types of information sources, viz. orthographic similarity and semantic similarity, that were incorporated for the experiments.

4.1 Experimental Setup

In this paper, cognate detection was approached as a supervised classification task. To this end, we applied Support Vector Machines as implemented in sklearn (Pedregosa et al., 2011).

The data set used for the binary classification experiments consisted of the COGNATE pairs (labels "cognate" and "partial cognate") and NON-COGNATE pairs (labels "error" and "false friend"). The categories of "proper name" and "no standard" were removed from the data set as they are always identical translations and would boost the performance of the system in an artificial way. Table 1 gives an overview of the distribution of the two classes in the gold standard data set.

	Cognate	Non-cognate	Total pairs
GS	9,855	4,763	14,618

Table 1: Distribution of the "cognate" and "non-cognate" class labels in the gold standard (GS).

In order to train and test the system, we performed 5-fold cross-validation for which we fixed our 5 subsamples. Hyperparameter optimisation was performed by means of a 5-fold cross-validation grid search on the training folds, resulting in the following values: $kernel$ = RBF, C = 5, $class\ weight$ = None and $gamma$ = 5.

4.2 Orthographic Similarity Features

Fifteen different string similarity metrics were applied on the candidate cognates to measure the formal relatedness between source and target words. Eleven of these fifteen metrics were also used by Frunza et al. (2007). The following list briefly summarizes the orthographic features implemented:

- **Prefix** divides the length of the shared prefix by the length of the longest cognate in the pair.
- **Dice** (Brew and McKelvie, 1996) divides the number of common bigrams times two by the total number of bigrams in the cognate pair, as in $\frac{2 \times |bigrams(x)| \cap |bigrams(y)|}{|bigrams(x)| + |bigrams(y)|}$.
- **Dice (trigrams)** differs from Dice in that it uses trigrams instead of bigrams.
- **XDice** is a variant of Dice as it uses bigrams that are created out of trigrams by deleting the middle letter in them.
- **XXDice** incorporates the string positions of the bigrams into its metric. Therefore, the denominator is no longer multiplied by two, but by $\frac{2}{1 + (pos(x) - pos(y))^2}$.
- **LCSR** stands for the longest common subsequence ratio, which is two times the length of the longest subsequence over the summed length of both sequences.
- **NLS** or the Normalized Levenshtein Similarity equals one minus the minimum number of edits required to change one string sequence to another.
- **LCSR (bigrams), NLS (bigrams), LCSR (trigrams), and NLS (trigrams)** differ from their standard metrics in that they use, respectively, bigrams and trigrams to calculate their results.
- **Jaccard index** models the length of the intersection of both cognate strings over the length of the union of these strings.
- **Jaro-Winkler similarity** is the complement of the Jaro-Winkler distance. Word pairs that from their beginning correspond to a set prefix length will receive higher scores.
- **Spsim option 1 and Spsim option 2** are the only metrics which require supervised training, in order to learn grapheme mappings between language pairs (Gomes and Pereira Lopes, 2011). They are trained by performing 5-fold cross-validation on the positive instances (i.e. cognates) in the data set. Therefore, we created two different train

sets: option 1 includes cognate pairs which differ in agreement or PoS-tags, while option 2 only includes cognates and partial cognates.

4.3 Semantic Information

Besides features that model formal similarity between word pairs, we also included semantic information in our classifier. We opted for word embeddings, as these have shown to be very effective for various NLP tasks. In addition, word embeddings have not yet been used for the task of cognate identification. For the purpose of this research, we worked with fastText word embeddings that were pre-trained on the Wikipedia corpus with the skip-gram model proposed by Bojanowski et al. (2017). The model was trained with the default parameters and the length of the vector was set to 300. A disadvantage of using text-formatted pre-trained embeddings is that we could not generate embeddings for all words in the gold standard list. As a result, we only obtained word embeddings for 12,433 instances, while we have orthographic information for 14,618 instances. Table 2 gives an overview of the distribution of the two classes in the full and reduced gold standard data sets. The experimental results that we obtained for this subset are presented in Section 5.2.

	Cognate	Non-cognate	Total pairs
Ortho	9,855	4,763	14,618
Semantic	8,935	3,498	12,433

Table 2: Distribution of the "cognate" and "non-cognate" class labels in the full (*Ortho*) and reduced (*Semantic*) gold standard data sets.

We chose to work with fastText embeddings instead of regular Word2Vec embeddings because the former model uses n-grams to train its embeddings. In contrast to the Word2Vec models, fastText can create word embeddings for out-of-vocabulary words, which is especially important for low-frequent words. Although the current research only works with pre-trained word entries, in future research we plan to add out-of-vocabulary words by training word embeddings on domain-specific corpora more similar to the DPC corpus that was used to extract the list of candidate cognate pairs. This way, we hope to construct embeddings for all word pairs in the gold standard list.

Once the results for the Dutch and English monolingual embeddings were loaded, the Dutch embeddings were mapped to the English vector space by means of a pre-trained alignment matrix (Smith et al., 2017). Since the embeddings are then situated in the same vector space, one can easily compute the cosine similarity between the two words of a candidate cognate pair. Subsequently, this cosine similarity was used as a semantic feature for our machine learning system.

5 Experimental Results

This section describes the classification results for two sets of experiments, namely (1) a classifier incorporating fifteen orthographic similarity features and (2) a classifier combining the same set of orthographic similarity features with a semantic feature resulting from computing the cosine similarity between the word embeddings of the cognate pair.

5.1 Experiment 1: Orthographic Features

A first set of experiments was conducted to evaluate the performance of the orthographic similarity features for the task of cognate detection. Table 3 lists the averaged precision, recall and F1-score for all individual orthographic similarity features and their combination.

The results show a very good performance of the classifier combining all orthographic similarity information (average F-score of 84%). Especially precision improves considerably when combining the different orthographic similarity metrics. When looking into the results for the individual features, it is clear that some metrics perform very well in isolation, such as LCSR and NLS, which obtain F-scores of around 85% for the positive class ("Cognates") with good balance of precision and recall.

To get further insight in the informativeness of the various orthographic features, we also trained a conditional inference tree and random forest on the cognate data set. Figure 1 visualizes the model learned by the conditional inference tree at depth 3. The tree indicates which orthographic metric is the most important for that node in the tree. As can be observed in Figure 1, the longest common subsequence ratio is overall the most influential metric, followed by SpSim (option 1) and the Jaro-Winkler similarity.

In addition to the conditional inference tree, a

	Cognates			Non-cognates			Average score		
Metric	Prec	Rec	F-score	Prec	Rec	F-score	Prec	Rec	F-score
Prefix	77.43	87.84	82.31	65.17	47.03	54.62	71.30	67.44	68.46
Dice	73.38	91.99	81.63	65.04	30.91	41.84	69.21	61.45	61.73
Dice (3gr)	73.28	91.88	81.53	64.63	30.67	41.59	68.95	61.28	61.56
Jaccard	73.83	91.53	81.73	65.22	32.86	43.69	69.52	62.19	62.71
XDice	70.85	96.26	81.62	70.03	18.08	28.73	70.44	57.17	55.18
XXDice	76.10	92.54	83.52	72.15	39.88	51.35	74.12	66.21	67.43
LCSR	82.15	89.30	85.47	72.65	59.93	65.66	77.40	74.62	75.57
NLS	82.39	86.03	84.24	68.47	61.84	64.95	75.43	73.93	74.59
LCSR (2gr)	76.92	81.28	79.03	56.16	49.52	52.58	66.54	65.40	65.80
NLS (2gr)	76.80	81.02	78.84	55.74	49.31	52.26	66.27	65.17	65.55
LCSR (3gr)	73.28	91.88	81.53	64.63	30.67	41.59	68.95	61.28	61.56
NLS (3gr)	73.34	91.60	81.46	64.16	31.10	41.87	68.75	61.35	61.67
Jaro-Winkler	77.06	90.72	83.33	69.72	44.10	53.99	73.39	67.41	68.66
SpSim (opt.1)	86.01	79.01	82.35	62.83	73.38	67.68	74.42	76.19	75.02
SpSim (opt.2)	83.36	80.37	81.82	62.21	66.76	64.37	72.79	73.56	73.10
Combined	89.33	90.63	89.97	80.63	77.60	78.78	84.68	84.11	84.38

Table 3: Precision (Prec), Recall (Rec) and F1-score for the individual orthographic similarity features and for the classifier combining all features (%).

random forest was trained in order to further investigate the importance of each metric. Since a random forest uses lots of seeds (in our case: 123) in order to decide on the importance of each variable individually, it provides a somewhat more representative, validated picture of the influence of different metrics. An additional Somers' D value was computed for the random forest in order to check the goodness of fit. With a correlation score of 0.9528739, our random forest forms a good model for unseen data. Figure 2 shows that the model agrees with the conditional inference tree in that it also classifies LCSR, SpSim (option 1) and the Jaro-Winkler similarity as important metrics for the identification of cognates. It does, however, provide some additional information, as it shows that the normalized Levenshtein similarity is also very influential for this binary classification task.

5.2 Experiment 2: Orthographic and Semantic Features

In a second set of experiments, we combined all orthographic similarity features with a semantic feature expressing the cosine similarity between the two word embeddings. Table 4 shows the result of the classifiers incorporating (1) only semantic information and (2) a combination of orthographic and semantic similarity information. As this set of experiments is only conducted on that part of the data set for which word embeddings were retrieved, we also added the updated performance scores for all individual orthographic metrics on this reduced data set.

The classification results listed in Table 4 show some interesting findings. First of all, the embeddings in isolation already obtain good classification results for the "Cognates" class (F-score of 89.14%). Second, the classifier combining orthographic and semantic similarity features clearly outperforms the classifier only incorporating orthographic information.

An analysis of the output reveals that the semantic information indeed helps to detect cognate pairs showing less orthographic resemblance (e.g. *east–oost, older–ouderen, widespread–wijdverbreid, asleep–slaap, sweating–zweten, shame–schaamte, belief–geloof, whole–hele, swarm–zwerm, overheated–oververhitte*). In addition, the word embedding information also generates less false negatives. Examples of pairs that were wrongly labeled as cognates by the classifier relying on orthographic information and that are correctly labeled as non-cognates by the combined classifier are: *affects–effecten, unlocking–blokkering, investments–instrument, slit–gesplit, provide–profielen, brazier–branden, might–high, where–wateren*. On the other hand,

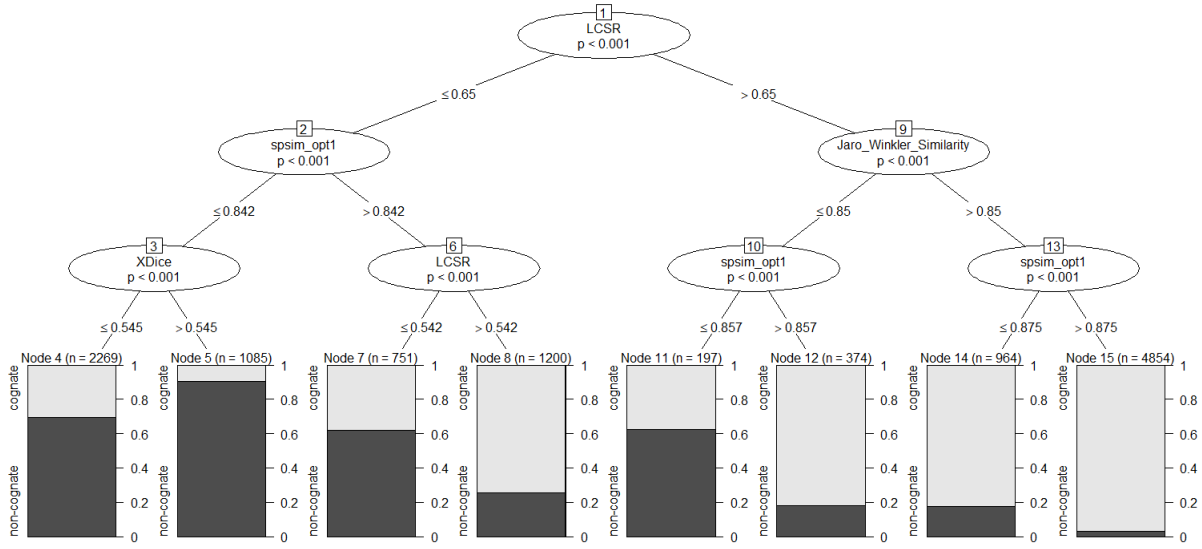

Figure 1: Conditional Inference Tree with depth 3 trained on the orthographic similarity features.

	Cognates			Non-cognates			Average score		
Metric	Prec	Rec	F-score	Prec	Rec	F-score	Prec	Rec	F-score
Prefix	82.30	87.99	85.05	62.74	51.66	56.65	72.52	69.82	70.85
Dice	80.40	91.23	85.47	65.84	43.19	52.15	73.12	67.21	68.81
Dice (3gr)	79.94	91.28	85.23	65.05	41.48	50.65	72.50	66.38	67.94
Jaccard	80.71	90.74	85.43	65.34	44.58	52.98	73.02	67.66	69.20
XDice	76.45	95.94	85.09	70.25	24.50	36.32	73.35	60.22	60.70
XXDice	79.89	94.15	86.43	72.56	39.45	51.09	76.22	66.80	68.76
LCSR	83.79	91.99	87.70	72.73	54.55	62.32	78.26	73.27	75.01
NLS	85.23	88.48	86.83	67.42	60.83	63.95	76.33	74.66	75.39
LCSR (2gr)	78.77	91.57	84.69	63.16	36.94	46.60	70.96	64.25	65.64
NLS (2gr)	79.09	90.57	84.44	61.69	38.82	47.64	70.39	64.69	66.04
LCSR (3gr)	79.94	91.28	85.23	65.05	41.48	50.65	72.49	66.38	67.94
NLS (3gr)	80.04	90.97	85.15	64.57	42.05	50.92	72.30	66.51	68.04
Jaro-Winkler	82.24	91.31	86.54	69.11	49.64	57.76	75.67	70.47	72.15
SpSim (opt.1)	85.58	81.05	83.23	57.40	65.01	60.89	71.49	73.03	72.06
SpSim (opt.2)	80.87	87.42	83.99	59.57	47.02	52.33	70.22	67.22	68.16
Sem	83.56	95.53	89.14	82.00	51.99	63.62	82.78	73.76	76.38
Ortho	89.46	91.23	90.33	76.42	72.54	74.42	82.94	81.88	82.38
Ortho + Sem	92.59	94.65	93.61	85.52	80.64	83.00	89.05	87.64	88.30

Table 4: Precision (Prec), Recall (Rec) and F1-score for the classifiers incorporating the fifteen individual orthographic features, the classifier incorporating only semantic information (*Sem*), the classifier incorporating the combined orthographic information (*Ortho*) and the classifier incorporating both orthographic and semantic similarity features (*Ortho + Sem*).

the combined classifier rarely introduces additional false negatives (seven instances in total, e.g. *lead–leiden, include–inhouden, docker–dokwerker*) or additional false positives (three instances in total: *told–toen, escapologist–escapist, because–bepaalde*).

6 Conclusion and Future Work

This paper presents preliminary experiments for combining orthographic and semantic similarity information for cognate detection. The experimental results already show promising scores for

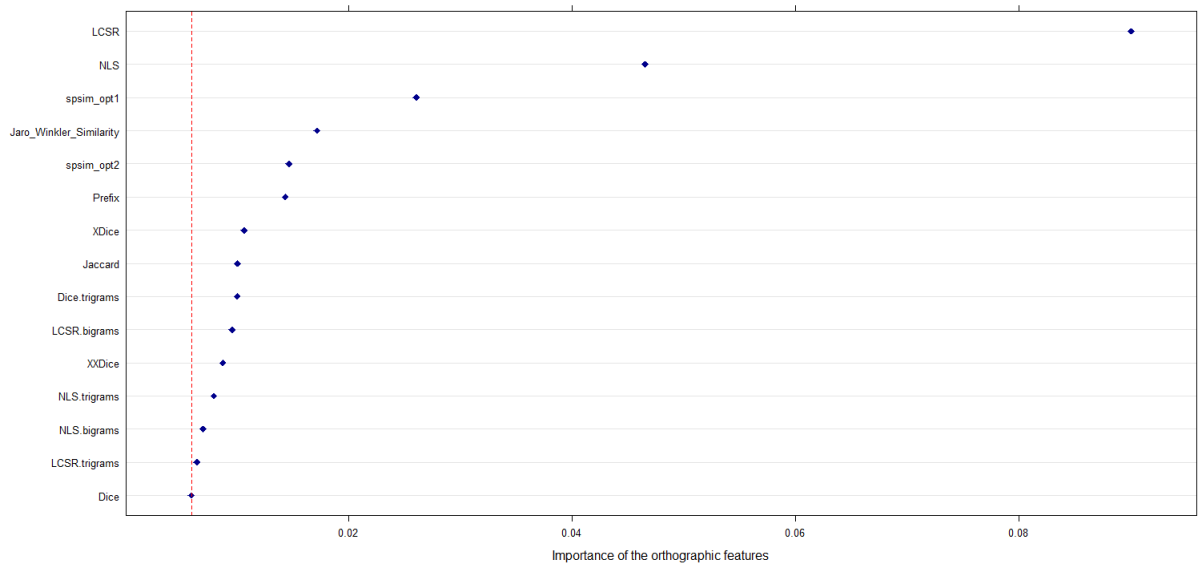

Figure 2: Random Forest indicating the importance the different orthographic similarity features for the current cognate identification task.

a classifier using merely orthographic similarity information. The results, however, revealed that adding semantic information capturing the cosine similarity between the word embeddings of the Dutch and English terms further improves the classification results considerably. As a result, we can conclude that combining orthographic and semantic similarity information is a viable approach to automatic cognate detection.

As we presented proof-of-concept results in this research, there is still a lot of room for future research. Firstly, the implementation of alternative word embeddings is an important direction for future work. We will perform additional experiments with (1) larger and different (e.g. domain-specific) corpora and (2) other embedding approaches to improve the semantic information based on embedding distance. We are confident this will result in high-level quality embeddings for all candidate cognate pairs.

Secondly, it would be interesting to perform multi-class experiments, where a distinction is made between cognates, false friends and non-related word pairs. To this end, a training and evaluation corpus containing cognate candidates in context will be built and manually annotated.

Finally, we plan to compile the corresponding gold standard set for French-Dutch, which is also part of the Dutch Parallel Corpus. This will allow an evaluation of our approach for a different language pair. In addition, this will enable us to perform trilingual machine learning experiments and to gain useful insights into cross-lingual cognate detection.

References

M. Baroni, G. Dinu, and G. Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 238–247.

S. Bergsma and G. Kondrak. 2007. Alignment-Based Discriminative String Similarity. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. pages 656–663.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5:135–146.

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

S. Castro, J. Bonanata, and A. Rosá. 2018. A High Coverage Method for Automatic False Friends Detection for Spanish and Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. pages 29–36.

A. Ciobanu and L. Dinu. 2014. Automatic Detection of Cognates Using Orthographic Alignment. In

52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference. volume 2, pages 99–105.

D. Crystal. 2008. *A Dictionary of Linguistics and Phonetics*. The language library. Blackwell, 6th edition.

O. Frunza and D. Inkpen. 2007. A tool for detecting French-English cognates and false friends. In *Actes de la 14me confrence sur le Traitement Automatique des Langues Naturelles*. Association pour le Traitement Automatique des Langues, pages 91–100.

L. Gomes and J. G. Pereira Lopes. 2011. Measuring Spelling Similarity for Cognate Identification. In L. Antunes and H. S. Pinto, editors, *Progress in Artificial Intelligence*. Springer, pages 624–633.

S. T. Gries. 2004. Shouldn't It Be Breakfunch? A Quantitative Analysis of Blend Structure in English. *Linguistics* 42(3):639–667.

S. Hansen-Schirra, J. Nitzke, and K. Oster. 2017. Predicting cognate translation. In S. Hansen-Schirra, O. Czulo, and S. Hofmann, editors, *Empirical modelling of translation and interpreting*, Language Science Press, chapter 1, pages 3–22.

Z. S. Harris. 1954. Distributional Structure. *WORD* 10(2-3):146–162.

B. Hauer and G. Kondrak. 2011. Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. pages 865–873.

W. Heeringa, J. Nerbonne, and P. Osenova. 2010. Detecting contact effects in pronunciation. In M. Norde, B. de Jonge, and C. Hasselblatt, editors, *Language Contact. New Perspectives.*, Benjamins, pages 131–154.

G. Kondrak. 2000. A New Algorithm for the Alignment of Phonetic Sequences. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*. pages 288–295.

G. Kondrak and B. Dorr. 2004. Identification of Confusable Drug Names: A New Approach and Evaluation Methodology. In *Proceedings of the 20th International Conference on Computational Linguistics*. pages 952–958.

S. Labat, L. Vandevoorde, and E. Lefever. 2019. Annotation Guidelines for Labeling English-Dutch Cognate Pairs, version 1.0. Technical report, Ghent University, LT3 15-01.

V. I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR* 163(4):845–848.

N. Ljubešić and D. Fišer. 2013. Identifying False Friends between Closely Related Languages. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*. pages 69–77.

L. Macken, O. De Clercq, and H. Paulussen. 2011. Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus. *Meta* 56(2):374–390.

I. D. Melamed. 1999. Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics* 25(1):107–130.

T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781.

G. A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11):39–41.

R. Mitkov, V. Pekar, D. Blagoev, and A. Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation* 21(1):29–53.

F.J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1):19–51.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Mller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

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

J. Schepens, T. Dijkstra, and F. Grootjen. 2012. Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition* 15(1):157–166.

J. Schepens, K. Paterson, T. Dijkstra, F. Grootjen, and W. J. B. van Heuven. 2013. Cross-Language Distributions of High Frequency and Phonetically Similar Cognates. *PLOS ONE* 8:1–15.

M. Shlesinger and B. Malkiel. 2005. Comparing Modalities: Cognates as a Case in Point. *Across Languages and Cultures* 6(2):173–193.

S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR* abs/1702.03859.