Transfer Learning for Cognate Identification in Low-Resource Languages

Eliel Soisalon-Soininen
Department of Computer Science
University of Helsinki
eliel.soisalon-soininen@helsinki.fi

Mark Granroth-Wilding
Department of Computer Science
University of Helsinki
mark.granroth-wilding@helsinki.fi

Introduction

In our on-going work, we are addressing the problem of identifying cognates across lexica of any pair of languages. In particular, we assume that the languages of interest are low-resource to the extent that no training data whatsoever, even in closely related languages, are available for the task. Instead, we investigate the performance of transfer learning approaches utilising training data from a completely unrelated language family.

Cognate identification is a core task in the *comparative method*, a collection of techniques used in historical linguistics, a field closely tied with linguistic typology (Shields, 2011). Cognate information is also useful for applications such as machine translation (Grönroos et al., 2018). In addition, knowledge of cognates is useful for second-language learning (Beinborn et al., 2014).

Cognate identification

In cognate identification, we are essentially given two string sets $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$. The task is to extract those pairs (x, y) in relation R:

$$R = \{(x, y) \in X \times Y|\ x \text{ is cognate with } y\ \}$$

Each element $x \in X$ and $y \in Y$ is a string over alphabets Σ_x and Σ_y respectively. The alphabet sets do not necessarily overlap.

Table 1 illustrates the difficulty of cognate identification. As can be seen, some cognates are straightforward with a similar form and meaning (e.g. *notte* - *noche*). On the other hand, there is large variation in the degree of similarity in terms of both form and meaning. However, common to all of these examples is that they exhibit *regular sound correspondences*, i.e. word segments regularly occurring in similar positions and contexts (Kondrak, 2012), such as *oa-e* and *th-d* in English-Swedish cognates. Therefore, cognate identification should rely on the identification of such pairs of symbols or, more generally, substrings.

Word A	**Word B**	**Meanings**
it: *notte*	es: *noche*	’night’
fi: *huvittava*	et: *huvitav*	’amusing’; ’interesting’
en: *attend*	fr: *attendre*	’attend’; ’wait’
en: *oath*	sv: *ed*	’oath’
fi: *pöytä*	sv: *bord*	’table’
en: *bite*	fr: *fendre*	’bite’; ’split’

Table 1: Examples of cognates with varying degree of similarity in form and meaning.

Most previous work attempts to design such a string similarity metric that would tend to assign a higher score to cognate than unrelated words. Common approaches include extensions of the traditional Levenshtein distance (Levenshtein, 1966) that either assign weights to pairs of symbols according to their phonetic properties (e.g. List, 2013; Kondrak, 2000), or that learn such weights from example cognates (Ciobanu and Dinu, 2014; Gomes and Pereira Lopes, 2011). McCoy and Frank (2018) use weights based on character embeddings.

In contrast to much of previous work, we make no strict assumptions about the degree of similarity in form or meaning that any two cognates should exhibit. Instead, following Rama (2016) and Jäger (2014), we treat regular correspondences as the main driving factor in the cognate relation and attempt to capture these in a completely data-driven manner. We aim to contribute to this line of research by considering the ability of our models to generalise across language families.

Models and experiments

In our experiments, we have trained our models with an etymological database of Indo-European languages (De Melo, 2014), and tested their performance on combinations of three lexica from Sami languages of the Uralic family. We have experimented with two similarity learning models, a Siamese convolutional neural network (S-CNN) based on Rama (2016) and a support vector machine (SVM) based on Hauer and Kondrak (2011), compared with a Levenshtein-distance (LD) baseline (Levenshtein, 1966). In addition, we have experimented with fine-tuning the S-CNN model in order to quantify the benefit of having small amounts of target-language training data.

The Levenshtein distance between two strings is the minimum number of insertions, deletions, and substitutions needed to transform one string to another. It is straightforward to turn this into a similarity metric. For the SVM, word pairs are encoded into vectors of the following features: Levenshtein distance, number of common bigrams, prefix length, lengths of both words, and the absolute difference between the lengths. The S-CNN is a two-input version of a convolutional neural network, where input words are encoded into matrices of concatenated one-hot vectors representing characters. As shown in Figure 1, the network creates a merged representation of a word pair, to be classified as cognate or unrelated.

Figure 2 shows the precision-recall curve for each model, including both a fine-tuned (with 500 target-language training pairs) and unadapted S-CNN. As expected, the fine-tuned S-CNN outperforms the other models. Interestingly, even the unadapted S-CNN simply relying on Indo-European training data outperforms the SVM and LD. This suggests that the S-CNN is able to more effectively capture such aspects of the cognateness relation that carry across language families.

Work in progress

We are currently investigating approaches to improve target-family performance with unsupervised methods of domain adaptation. One of our lines of work is to use an adversarial approach to making target-family word pair representations more similar to source-family representations, following (Tzeng et al., 2017). Another way to extend the S-CNN model is to use unsupervised multilingual character embeddings (Granroth-Wilding and Toivonen, 2019), trained with small corpora from the target languages. This could be a way to make characters across languages more comparable to each other, which is an obvious issue when dealing with orthographic forms of words.

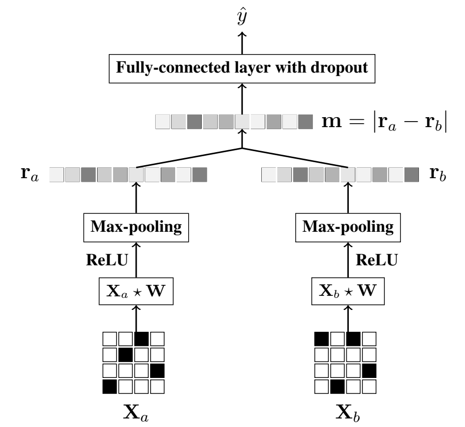
$\hat{y}$
Fully-connected layer with dropout
$\mathbf{m} = |\mathbf{r}_a - \mathbf{r}_b|$
$\mathbf{r}_a$
$\mathbf{r}_b$
Max-pooling
Max-pooling
ReLU
ReLU
$\mathbf{X}_a \star \mathbf{W}$
$\mathbf{X}_b \star \mathbf{W}$
$\mathbf{X}_a$
$\mathbf{X}_b$

Figure 1: Architecture of the S-CNN. Column vectors in input matrices represent one-hot-encoded characters. The same filter $\mathbf{W}$ is convolved with both inputs.

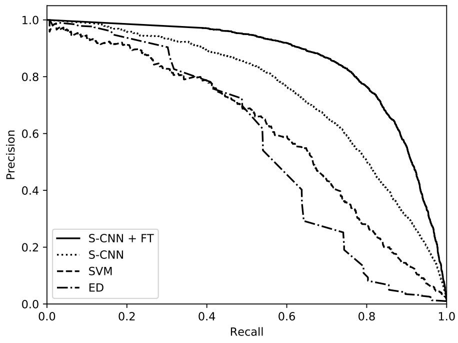

Figure 2: Precision-recall for Sami test set.

In addition to unsupervised methods, we also intend to compare our data-driven approaches with more linguistically-informed ones, in order to assess the benefit of such information.

Although we have thus far specifically focused on the problem of cognate identification, we believe that these methods could be extended to the study of other typological features of language.

References

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Readability for foreign language learning: The importance of cognates. *ITL-International Journal of Applied Linguistics*, 165(2):136–162.

Alina Maria Ciobanu and Liviu P Dinu. 2014. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 99–105.

Gerard De Melo. 2014. Etymological wordnet: Tracing the history of words. In *LREC*, pages 1148–1154. Citeseer.

Luís Gomes and José Pereira Lopes. 2011. Measuring spelling similarity for cognate identification. *Progress in Artificial Intelligence*, pages 624–633.

Mark Granroth-Wilding and Hannu Toivonen. 2019. Unsupervised learning of cross-lingual symbol embeddings without parallel data. *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 19–28.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2018. Cognate-aware morphological segmentation for multilingual neural translation. *arXiv preprint arXiv:1808.10791*.

Bradley Hauer and Grzegorz 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.

Gerhard Jäger. 2014. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. In *Quantifying Language Dynamics*, pages 155 – 204. Leiden, The Netherlands: Brill.

Grzegorz 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. Association for Computational Linguistics.

Grzegorz Kondrak. 2012. Similarity patterns in words. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 49–53. Association for Computational Linguistics.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Johann-Mattis List. 2013. *Sequence comparison in historical linguistics*. Ph.D. thesis, Heinrich-Heine-Universität Düsseldorf.

Richard T McCoy and Robert Frank. 2018. Phonologically informed edit distance algorithms for word alignment with low-resource languages. *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 102–112.

Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027.

Kenneth Shields. 2011. Linguistic typology and historical linguistics. In *The Oxford Handbook of Linguistic Typology*.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4.