HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
FACULTY OF SCIENCE

TRANSFER LEARNING FOR COGNATE DETECTION IN LOW-RESOURCE LANGUAGES

Eliel Soisalon-Soininen
Mark Granroth-Wilding

Department of Computer Science,
firstname.lastname@helsinki.fi

INTRODUCTION

In our on-going work, we are addressing the problem of identifying **cognates** across **unannotated** vocabularies of **any pair of languages**. We assume that the languages of interest are **low-resource** to the extent that no training data whatsoever, even in closely related languages, is available for the task.

Instead, we investigate the performance of language-independent **transfer learning** approaches, utilising training data from a completely **unrelated, higher-resource** language family.

COGNATE DETECTION

Cognates are words in different languages that share an etymological root in a common proto-language. Cognate detection is central to the **comparative method**, a collection of techniques used in historical linguistics, closely tied with linguistic typology [1]. Cognate information is also useful for applications such as machine translation [2] and knowledge of cognates is useful for second-language learning [3].

We are given two sets X and Y whose elements are strings over alphabets Σ_x and Σ_y. The task is to extract pairs in relation R:

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

The **alphabets do not necessarily overlap**, since the orthographies of different languages vary. This issue is often circumvented by using phonetic transcriptions of words, which we lack for our low-resource case.

Word x	Word y	Meaning of x	Meaning of y
it: *notte*	es: *noche*	'night'	'night'
en: *attend*	fr: *attendre*	'attend'	'wait'
fi: *huvittava*	et: *huvitav*	'amusing'	'interesting'
en: *oath*	sv: *ed*	'oath'	'oath'
fi: *pöytä*	sv: *bord*	'table'	'table'
en: *bite*	fr: *fendre*	'bite'	'split'

Table 1: Examples of cognates, i.e. etymologically related words. The degree of similarity in form and meaning may vary quite substantially.

Table 1 illustrates the difficulty. All of these examples exhibit **regular sound correspondences**, i.e. word segments regularly occurring in similar positions and contexts [4], such as *oa–e* and *th–d* in English–Swedish cognates. Therefore, cognate detection should rely on detecting such correspondences, between pairs of single characters or short substrings, at the level of **orthography** or **phonology**.

In contrast to previous work, we make no strict assumptions about the degree of **similarity in form or meaning** that cognates should exhibit. Instead, following [5] and [6], we treat **regular correspondences** as the main driving factor in the cognate relation and attempt to capture these in a completely data-driven manner. Our main contribution is to consider the ability of models to generalise **across language families**.

MODELS

In our experiments, we have examined the performance of two similarity learning models:

- Support vector machine (SVM), based on [7]. Word pairs are encoded into vectors of the following features: edit distance; number of common bigrams; prefix length; lengths of both words; absolute difference between lengths.
- Siamese convolutional neural network (S-CNN), based on [5]. The network takes pairs of words (represented by concatenated character vectors) as input and creates a merged representation, to be classified as cognate or unrelated. Figure 1 shows the network.

We use the string edit distance (Levenshtein distance, ED) [8] as a **baseline** in our experiments.

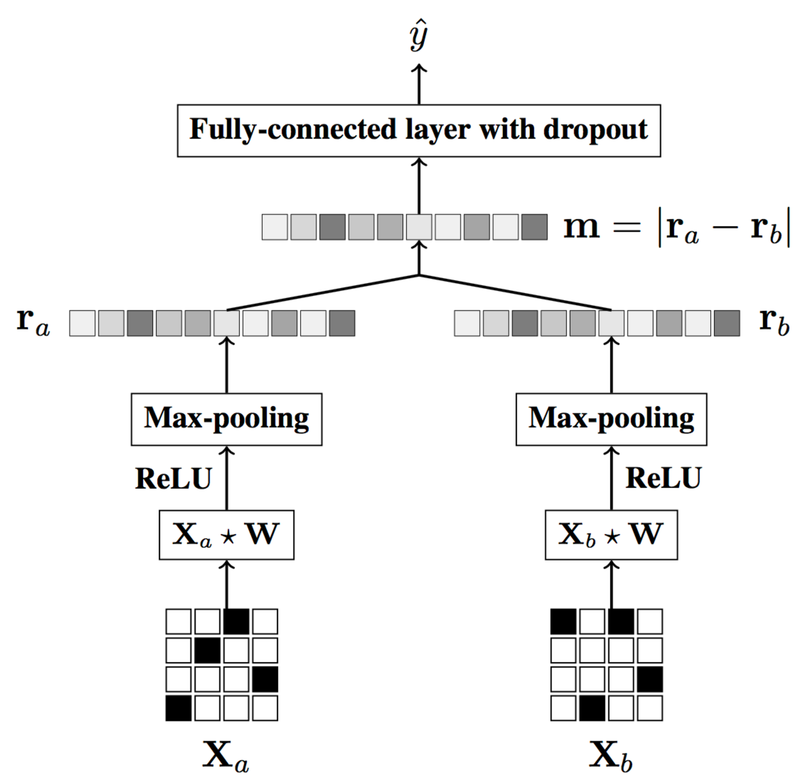

Figure 1: The S-CNN architecture. Column vectors in input matrices represent one-hot-encoded characters. The filter **W** is convolved over character sequences.

DATASETS

We obtained our **training dataset** IE-TRAIN from the Etymological WordNet [9], a database specifying cognateness and other etymological word relationships. It has been mined from Wiktionary and its entries are mostly from widely-spoken **Indo-European** languages. As our **low-resource test data**, we use unannotated word lists from three **Sami languages** of the **Uralic** language family. We have retrieved these from dictionaries compiled by Giellatekno [10]. We sampled a small set of known cognates to fine-tune the S-CNN model (see below). For **evaluation**, we obtained gold-standard cognate sets from Álgu [11], an etymological database for Sami languages.

Dataset	# cognate	# all pairs
IE-TRAIN	73,238	732,380
sma–sme	1,460	11,234 × 47,312
sma–sms	838	11,234 × 29,401
sme–sms	2,188	47,312 × 29,401

Table 2: Summary of datasets. Languages: South Sami (sma), North Sami (sme), Skolt Sami (sms).

INDO-EUROPEAN MODELS FOR SAMI COGNATES

Figure 2 compares the two **similarity learning models** with the edit distance baseline. The models are trained on Indo-European cognate pairs and applied without modification to cognate identification on Sami languages.

Since our gold-standard database is not complete, we cannot know whether a given word pair is *not* a cognate pair. Therefore, we evaluate the **recall** of known cognate pairs: proportion of annotated pairs in the set ranked as most likely cognates.

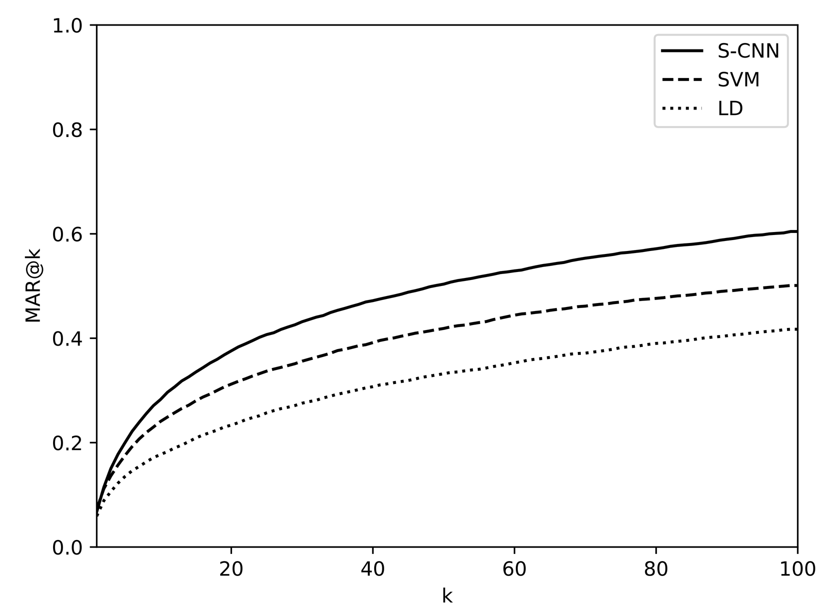

Figure 2: Models trained only on Indo-European data, tested on Sami vocabularies. MAR@k refers to recall@k, averaged over pairs of Sami vocabularies and query words, for $k = 1 \dots 100$.

Since the S-CNN outperforms other models in Figure 2, we try **fine-tuning** it with a small set of positive and negative **examples of Sami cognates**. Figure 3 shows precision-recall curves of fine-tuned and unadapted S-CNN, SVM, and ED (baseline).

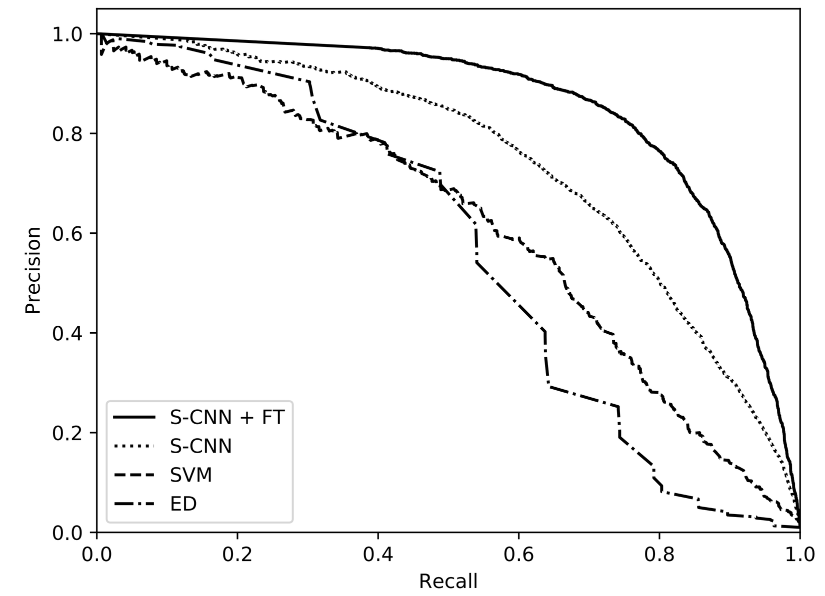

Figure 3: Precision-recall curves for models tested on Sami vocabularies. S-CNN + FT was pre-trained on IE-TRAIN and fine-tuned on a set of 500 cognate pairs from Sami. S-CNN and SVM were trained only on IE-TRAIN.

RESULTS

Unsurprisingly, the **fine-tuned S-CNN** outperforms the **unadapted models**. The unadapted S-CNN simply relying on Indo-European training data outperforms SVM and LD. This suggests that the **S-CNN** may be better able to capture aspects of cognateness that **carry over 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, similarly to the method of [12] for domain adaptation of images. Another way to extend the S-CNN model is to use **unsupervised multilingual character embeddings** [13], trained with small corpora from the target languages. This could be a way to make characters across languages more comparable to each other, thus tackling the issue that orthographies are often not directly comparable.

ACKNOWLEDGEMENTS

This work has been funded by the Academy of Finland *Digital Language Typology* project (no. 12933481).

REFERENCES

1. Kenneth Shields. 2011. Linguistic typology and historical linguistics. In *The Oxford Handbook of Linguistic Typology*.
2. Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2018. Cognate-aware morphological segmentation for multilingual neural translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 386–393.
3. Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate production using character-based machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 883–891.
4. Johann-Mattis List. 2013. *Sequence comparison in historical linguistics*. Ph.D. thesis, Heinrich-Heine-Universität Düsseldorf.
5. 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.
6. Gerhard Jäger. 2014. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. In *Quantifying Language Dynamics*, Brill, Leiden, the Netherlands, pages 155–204.
7. 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.
8. Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8):707–710.
9. Gerard de Melo. 2014. Etymological WordNet: tracing the history of words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.
10. The research group of Sami language technology at the University of Tromssa. `http://giellatekno.uit.no/index.eng.html`.
11. Álgu, Etymological database of Sami languages, `http://kaino.kotus.fi/algu/`.
12. Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recog- nition (CVPR)*, volume 1, page 4.
13. 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.