Finding Sami Cognates with a Character-Based NMT Approach

Mika Hämäläinen
University of Helsinki, mika.hamalainen@helsinki.fi

Jack Rueter
jack.rueter@helsinki.fi, jack.rueter@helsinki.fi

Follow this and additional works at: https://scholar.colorado.edu/scil-cmel

Part of the Computational Linguistics Commons

Recommended Citation
Available at: https://scholar.colorado.edu/scil-cmel/vol1/iss1/6
Finding Sami Cognates with a Character-Based NMT Approach

Mika Hämäläinen
Department of Digital Humanities
University of Helsinki
mika.hamalainen@helsinki.fi

Jack Rueter
Department of Digital Humanities
University of Helsinki
jack.rueter@helsinki.fi

Abstract

We approach the problem of expanding the set of cognate relations with a sequence-to-sequence NMT model. The language pair of interest, Skolt Sami and North Sami, has too limited a set of parallel data for an NMT model as such. We solve this problem on the one hand, by training the model with North Sami cognates with other Uralic languages and, on the other, by generating more synthetic training data with an SMT model. The cognates found using our method are made publicly available in the Online Dictionary of Uralic Languages.

1 Introduction

Sami languages have received a fair share of interest in purely linguistic study of cognate relations. Although various schools of Finno-Ugric studies have postulated contrastive interpretations of where the Sami languages should be located within the language family, there is strong evidence demonstrating regular sound correspondence between Samic and Balto-Finnic, on the one hand, and Samic and Mordvin, on the other. The importance of this correspondence is accentuated by the fact that the Samic might provide insight for second syllable vowel quality, as not all Samic-Mordvin vocabulary is attested in Balto-Finnic (cf. Korhonen, 1981). The Sami languages themselves (there are seven written languages) also exhibit regular sound correspondence, even though cognates, at times, may be opaque to the layman. One token of cognate relation studies is the Álgu database (Kotus, 2006), which contains a set of inter-Sami cognates. Cognates have applicability in NLP research for low-resource languages as they can, for instance, be used to induce the predicate-argument structures from bilingual vector spaces (Peirsman and Padó, 2010).

The main motivation for this work is to extend the known cognate information available in the Online Dictionary of Uralic Languages (Hämäläinen and Rueter, 2018). This dictionary, at its current stage, only has cognate relations recorded in the Álgu database.

Dealing with true cognates in a non-attested hypothetical proto-language presupposes adherence to a set of sound correlations posited by a given school of thought. Since Proto-Samic is one such language, we have taken liberties to interpret the term cognate in the context of this paper more broadly, i.e. not only words that share the same hypothesized origin in Proto-Samic are considered cognates (hence forth: true cognates), but also items that might be deemed loan words acquired from another language at separate points in the temporal-spatial dimensions. This more permissive definition makes it possible to tackle the problem computationally easier given the limitation imposed by the scarcity of linguistic resources.

Our approach does not presuppose a semantic similarity of the meaning of the cognate candidates, but rather explores cognate possibilities based on grapheme changes. The key idea is that the system can learn what kinds of changes are possible and typical for North Sami cognates with other Uralic languages in general. Taking leverage from this more general level knowledge, the model can learn the cognate features between North Sami and Skolt Sami more specifically.

We assimilate this problem with that of normalization of historical spelling variants. On a higher level, historical variation within one language can be seen as discovering cognates of different temporal forms of the language. Therefore, we want to take the work done in that vein for the first time in the context of cognate detection. Using NMT (neural machine translation) on a character level has been shown to be the single most accurate...
method in normalization by a recent study with historical English (Hämäläinen et al., 2018).

In this paper, we use NMT in a similar character level fashion for finding cognates. Furthermore, due to the limited availability of training data, we present an SMT (statistical machine translation) method for generating more data to boost the performance of the NMT model.

2 Related Work

Automatic identification of cognates has received a fair share of interest in the past from different methodological stand points. In this section, we will go through some of these approaches.

Ciobanu and Dinu (2014) propose a method based on orthographic alignment. This means a character level alignment of cognate pairs. After the alignment, the mismatches around the aligned pairs are used as features for the machine learning algorithm.

Another take on cognate detection is that of Rama (2016). This approach employs Siamese convolutional networks to learn phoneme level representation and language relatedness of words. They based the study on Swadesh lists and used hand-written phonetic features and 1-hot encoding for the phonetic representation.

Cognate detection has also been done by looking at features such as semantics, phonetics and regular sound correspondences (St. Arnaud et al., 2017). Their approach implements a general model and language specific models using support vector machine (SVM).

Rama et al. (2017) present an unsupervised method for cognate identification. The method consists of extracting suitable cognate pairs with normalized Levenshtein distance, aligning the pairs and counting a point-wise mutual information score for the aligned segments. New sets of alignments are generated and the process of aligning and scoring is repeated until there are no changes in the average similarity score.

3 Finding Cognates

In this section, we describe our proposed approach in finding cognates between North Sami and Skolt Sami. We present the dataset used for the training and an SMT approach in generating more training data.

3.1 The Data

Our training data consists of Álgu (Kotus, 2006), which is an etymological database of the Sami languages. From this database, we use all the cognate relations recorded for North Sami to all the other Finno-Ugric languages in the database. This produces a parallel dataset of North Sami words and their cognates in other languages.

The North Sami to other languages parallel dataset consists of 32905 parallel words, of which 2633 items represent the correlations between North Sami and Skolt Sami.

We find cognates for nouns, adjectives, verbs and adverbs recorded in the Giellatekno dictionaries (Moshagen et al., 2013) for North Sami and Skolt Sami. These dictionaries serve as an input for the trained NMT model and for filtering the output produced by the model.

3.2 The NMT Model

For the purpose of our research we use OpenNMT (Klein et al., 2017) to train a character based NMT model that will take a Skolt Sami word as its input and produce a potential North Sami cognate as its output. We use the default settings for OpenNMT.

We train a sequence to sequence model with the list of known cognates in other languages as the source data and their North Sami counterparts as the target data. In this way, the system learns a good representation of the target language, North Sami, and can learn what kind of changes are possible between cognates in general. Thus, the model can learn additional information about cognates that would not be present in the North Sami-Skolt Sami parallel data.

In order to make the model adapt more to the North Sami-Skolt Sami pair in particular, we continue training the model with only the North Sami-Skolt Sami parallel data for an additional 10 epochs. The idea behind this is to bring the model closer to the language pair of interest in this research, while still maintaining the additional knowledge it has learned about cognates in general from the larger dataset.

3.3 Using SMT to Generate More Data

Research in machine translation has shown that generating more synthetic parallel data that can be
noisy in the source language end but is not noisy in the target end, can improve the overall translations of an NMT model (Sennrich et al., 2015). In light of this finding, we will try a similar idea in our cognate detection task as well.

Due to the limited amount of North Sami-Skolt Sami training data available, we use SMT instead of NMT to train a model that will produce plausible but slightly irregular Skolt Sami cognates for the word list of North Sami words obtained from the Giellatekno dictionaries.

We use Moses (Koehn et al., 2007) baseline\(^2\) to train a translation model to the opposite direction of the NMT model with the same parallel data. This means translating from North Sami to Skolt Sami. We use the same parallel data as for the NMT model, meaning that on the source side, we have North Sami and on the target side we have all the possible cognates in other languages. The parallel data is aligned with GIZA++ (Och and Ney, 2003).

Since we are training an SMT model, there are two ways we can make the noisy target of all the other languages resemble more Skolt Sami. One is by using a language model. For this, we build a 10-gram language model with KenLM (Heafield et al., 2013) from Skolt Sami words recorded in the Giellatekno dictionaries.

The other way of making the model more aware of Skolt Sami in particular is to tune the SMT model after the initial training. For the tuning, we use the Skolt Sami-North Sami parallel data exclusively so that the SMT model will go more towards Skolt Sami when producing cognates.

We use the SMT model to translate all of the words extracted from the North Sami dictionary into Skolt Sami. This results in a parallel dataset of real, existing North Sami words and words that resemble Skolt Sami. We then use this data to continue the training of the previously explained NMT model for 10 additional epochs.

### 3.4 Using the NMT Models

We use both of the NMT models, i.e. the one without SMT generated additional data and the one with the data separately to assess the difference in their performance. We feed in the extracted Skolt Sami words from the dictionary and translate each word to a North Sami word as it would look like if there were a cognate for that word in North Sami.

The approach produces many non-words which we filter out with the North Sami dictionary. The resulting list of translated words that are actually found in the North Sami dictionary are considered to be potential cognates found by the method.

### 4 Results and Evaluation

In this section, we present the results of both of the NMT models, the one without SMT generated data and the one with generated data. The results shown in Table 1 indicate that the model with the additional SMT generated data outperformed the other model. The evaluation is based on a 200 randomly selected cognate pairs output by the models. These pairs have then been checked by an expert linguist according to principles outlined in (4.1).

<table>
<thead>
<tr>
<th></th>
<th>NMT</th>
<th>NMT + SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>67.5%</td>
<td>83%</td>
</tr>
</tbody>
</table>

Table 1: Percentage of correctly found cognates

Table 2 gives more insight on the number of cognates found and how they are represented in the original Álgu database. The results show that while the models have poor performance in finding the cognates in the training data, they work well in extending the cognates outside of the known cognate list.

<table>
<thead>
<tr>
<th></th>
<th>NMT</th>
<th>NMT + SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same as in Álgu</td>
<td>75</td>
<td>61</td>
</tr>
<tr>
<td>North Sami word in Álgu but no cognates with Skolt Sami</td>
<td>211</td>
<td>226</td>
</tr>
<tr>
<td>North Sami word in Álgu with other Skolt Sami cognates</td>
<td>646</td>
<td>577</td>
</tr>
<tr>
<td>North Sami word not in Álgu</td>
<td>848</td>
<td>936</td>
</tr>
<tr>
<td><strong>Cognates found in total</strong></td>
<td><strong>1780</strong></td>
<td><strong>1800</strong></td>
</tr>
</tbody>
</table>

Table 2: Distribution of cognates in relation to Álgu

As one of the purposes of our work is to help evaluate and develop etymological research, we will conduct a more qualitative analysis of the correctly and incorrectly identified cognates for the better working model, i.e. the one with SMT generated data. This means that the developers should

---

\(^2\)As described in http://www.statmt.org/moses/?n=moses.baseline
be aware not only of the comparative-linguistic rules etymologists use when assessing the regularity of cognate candidates but semantics as well.

In an introduction to Samic language history (Korhonen, 1981: 110–114), the proto-language is divided into 4 separate phases. The first phase involves vowel changes in the first and second syllables (ä–e » e–ẹ, u–e » ö–ẹ, e–ä » e–ẹ, i–e » e–ẹ, etc.) followed by vowel rotation in the first syllable dependent on the quality of the second-syllable vowel (e–ẹ » e–ẹ but e–ä » ẹ–ä). The second phase entails the loss of quantitative distinction for first-syllable vowels such that high vowels are normalized as short, and non-high first-syllable vowels are normally long (e–ẹ » ë–ẹ, ẹ–ä » ë–ë, etc.).

In Table 3, below, we provide an approximation of a few hypothesized sound changes for three words that are attested in Balto-Finnic and Mordvin, alike. *kité ‘hand; arm’ has true cognates in Finnish käsi, Northern Sami giehta, Skolt Sami kiot, Erzya ked’ and Moksha käd’, *tule ‘fire’ is represented by Finnish tuli, Northern Sami dolla, Skolt Sami toll, and Mordvin tol, while *pesä ‘nest’ is attested in Finnish pesä, Northern Sami beassi, Skolt Sami pie’ss, Erzya pize, and Moksha piza. The Roman numerals in the table correspond to four separate phases in Proto-Samic.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>hand/arm</td>
<td>käte</td>
<td>ketę</td>
<td>kęte</td>
<td>kietę</td>
</tr>
<tr>
<td>fire</td>
<td>tule</td>
<td>tole</td>
<td>tole</td>
<td>tole</td>
</tr>
<tr>
<td>nest</td>
<td>pesä</td>
<td>pęsę</td>
<td>pęsę</td>
<td>peası</td>
</tr>
</tbody>
</table>

Table 3: Illustration of some vowel correlations in 4 phases of Proto-Samic

In the evaluation, our attention was drawn to the adherence of (143) items to accepted sound correlations while there were (57) candidates that failed in this respect (cf. Korhonen, 1981; Lehtiranta, 2001; Aikio, 2009). Irregular sound correlation can be exemplified in the North Sami word bierdna ‘bear’ and its counterpart the Skolt Sami word peđ’rn (the prime indicates palatalization in Skolt Sami orthography) ‘bear cub’. The former appears to represent the word type found in ‘hand’ North Sami giehta and Skolt Sami kiot, whereas the latter represents the word type found in ‘nest’ North Sami beassi and Skolt Sami pie’ss and ‘swamp’ North Sami jeaggi and Skolt Sami peđ’gğ. Hence, on the basis of the North Sami word bierdna ‘bear’, one would posit a Skolt Sami form *piormn, whereas the Skolt Sami word peđ’rn ‘bear cub’ would presuppose a North Sami form *beardnî. Both types have firm representation in both languages, so it would seem that these borrowings have entered the languages at separate points in the spatio-temporal dimensions.

4.1 Analysis of the Correct Cognates

Correct cognates were selected according to two simples principles of similarity. On the one hand, there was the principle of conceptual similarity in their referential denotations (i.e., this refers to future work in semantic relations). On the other hand, a feasible cognate pair candidate should demonstrate adherence or near adherence to accepted sound law theory. The question of adherence versus near adherence indicated here can be directed to concepts of sound law theory, where conceivable irregularities may further be attributed to points in spatio-temporal dimensions (i.e., when and where a particular word was introduced into the lexica of the two languages involved in the investigation).

In the investigation of 200 random cognate pair candidates, 166 cognate pair candidates exhibited conceptual similarity which in some instances surpassed what might have been discovered using a bilingual dictionary. Of the 166 acceptable cognate pairs 131 candidate pairs demonstrated regular correlation to received sound law theory.

Adherence to concepts of sound law theory can be observed in the alignment of the North Sami words čuõka ‘mosquito’ and ada ‘marrow’ with their Skolt Sami counterparts čuõskk and õ̂õ̂t, respectively. Although these words may appear opaque to the layman, and thus this alignment might be deemed dubious at first, awareness of cognate candidates in the Erzya Mordvin šeskke ‘mosquito’ and uďem ‘marrow’ helps to alleviate initial misgivings.

As may be observed above, North Sami frequently has two-syllable words where Skolt Sami attests to single-syllable words. This relative length correlation between North Sami and Skolt Sami is described through measurement in

42
There was one attested correlation for a 5-syllable word *engelasgiella* 'English language' in North Sami and its 3-syllable counterpart in Skolt Sami *engglâskiöll*. Since we are looking at a compound word with 3-to-2 and 2-to-1 correlations, we can assume that our model is recognizing individual adjacent segments within a larger unit.

Correct cognates do not necessarily require etymologically identical source forms or structure. The recognized cognate pairs represent both recent loan words or possible irregularities in sound law theory (35) and presumably older mutual lexicon (131) (see 4.2, below). They also attest to differed structure and length (i.e., this may also include derivation and compounding). While a majority of the cognate candidate pairs linked words sharing the same derivational level, 11 represented instances of additional derivation in either the North Sami or Skolt Sami word, and 3 recognized instances where one of the languages was represented by a compound word.

### 4.2 Analysis of the Incorrect Cognates

Incorrect cognates often offer vital input for cognate detection development. There are, of course, words pairs that diverge in regard to both accepted sound law theory and semantic cohesion. These pairs have not yet been applied to development. In contrast, word pairs that appear to adhere to sound law theory yet are not matched semantically might be regarded as false friends. These pairs can be potentially useful in further development.

Of 34 semantically non-feasible candidates, 12 stood out as false friends. One such example pair is observed in the North Sami *älgg* 'beginning' and the Skolt Sami *älgg* 'piece of firewood'. These two words, it should be noted, can be associated with the Finnish cognates *alku* 'beginning' and *halko* 'piece of split firewood [NB! there is a loss of the word initial *h*]', respectively. Since the theoretically expected vowel equivalent of the first syllable *a* in Finnish is *uo* and *ue* in North Sami and Skolt Sami, respectively, we might assume that neither word comes from a mutual Samic-Finnic proto-language.

We do not know to what extent random selection has affected our results. Had the first North Sami noun *älg* been replaced with its paradigmatic verb *älgit* 'begin', the Skolt Sami *dłğged*, also translated as 'begin', would have shown direct correlation for *ä* and *ä* in North Sami and...
Skolt Sami, respectively. The second noun, meaning ‘piece of split firewood’, is actually hálgu in North Sami, which simply demonstrates h re- 
tention and the possible recognition problems faced in the absence of semantic knowledge.

4.3 Summary of the Analyses
The cognate candidates were evaluated according to two criteria: One query checked for concep-
tual similarity (correct vs incorrect), and the other checked for regularity according to received sound 
law theory. While the majority (65%) of the word pairs evaluated were both conceptually similar and 
correlated to received sound law theory, an addi-
tional 17% of the candidates represented irregular 
sound correlation, as indicated by the figures 131 
and 35 below, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Similar</th>
<th>Dissimilar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular</td>
<td>131</td>
<td>12</td>
</tr>
<tr>
<td>Irregular</td>
<td>35</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 4: Cognate candidate evaluation

The presence of an 11% negative score for both 
sound law regularity and conceptual similarity in-
dicates an improvement requirement of at least 
6% before the machine can be considered relevant (95%). The 6% attestation of false friend discov-
ery, however, displays an already existing accu-
ricacy in our algorithm.

5 Conclusions and Future Work
In this paper, we have shown that using a character-based NMT is a feasible way of expand-
ing a list of cognates by training the model mostly 
on the cognate pairs for North Sami words in lan-
guages other than Skolt Sami. Furthermore, we 
have shown that an SMT model can be used to 
generate synthetic parallel data by pushing the 
model more towards the direction of Skolt Sami 
by introducing a Skolt Sami language model and 
tuning the model with Skolt Sami - North Sami 
parallel data.

In our evaluation, we have only considered the 
best cognate produced by the NMT model with the 
idea of one-to-one mapping. However, it is pos-
sible to make the NMT model output more than one 
possible translation. In the future, we can conduct 
more evaluation for a list of top candidates to see 
whether the model is able to find more than one 
cognate for a given word and whether the overall 
recall can be improved for the words where the 
top candidate has been rejected by the dictionary 
check as a non-word.

We have currently limited our research in cog-
nates between Skolt Sami and North Sami where 
the translation direction of the NMT model has 
been towards North Sami. An interesting future 
direction would be to change the translation direc-
tion. In addition to that, we are also interested in 
trying this method out on other languages recorded 
in the Álgu database.

We are also interested in conducting research 
that is more linguistic in its nature based on the 
cognate list produced in this paper. This will shed 
more light in the current linguistic knowledge of 
cognates in the Sami languages. The current re-
sults of the better working NMT model are re-
leased in the Online Dictionary for Uralic Lan-
guages3.

References
Ante Aikio. 2009. The Saami Loan Words in Finnish 
and Karelian, first edition. Faculty of Humanities of 
the University of Oulu, Oulu. Dissertation.

Alina Maria Ciobanu and Liviu P Dinu. 2014. Au-
tomatic detection of cognates using orthographic 
alignment. In Proceedings of the 52nd Annual Meet-
ing of the Association for Computational Linguistics 

Mika Hämäläinen and Jack Rueter. 2018. Advances 
in Synchronized XML-MediaWiki Dictionary De-
velopment in the Context of Endangered Uralic Lan-
guages. In Proceedings of the Eighteenth EURALEX 
International Congress, pages 967–978.

Mika Hämäläinen, Tanja Säily, Jack Rueter, Jörg 
Tiedemann, and Eetu Mäkelä. 2018. Normalizing 
early english letters to present-day english spelling. 
In Proceedings of the Second Joint SIGHUM Work-
shop on Computational Linguistics for Cultural 
Heritage, Social Sciences, Humanities and Litera-
ture, pages 87–96.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H 
Clark, and Philipp Koehn. 2013. Scalable modified 
Kneser-Ney language model estimation. In Pro-
cedings of the 51st Annual Meeting of the Associa-
tion for Computational Linguistics (Volume 2: Short 
Papers), volume 2, pages 690–696.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean 
Senellart, and Alexander M. Rush. 2017. Open-
NMT: Open-Source Toolkit for Neural Machine 
Translation. In Proc. ACL.

3http://akusanat.com/


