Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)

Building an Extremely Low Resource Language to High Resource Language Machine Translation System from Scratch

Flammie A Pirinen
UiT Norgga árktalaš universitehta
Tromsø, Norway
firstname.lastname@uit.no

Linda Wiechetek
UiT Norgga árktalaš universitehta
Tromsø, Norway
firstname.lastname@uit.no

Abstract

Building a machine translation system for an extremely low-resource language is a problem in contemporary computational linguistics. In this article, we show how to use existing morpho-syntactic analysers and a modern rule-based machine translation system to rapidly build a baseline system for a language pair where a neural model approach is not feasible due to the total lack of high-quality parallel corpora. Our experiment produces a freely available open-source North Sámi to German machine translator, which provides us useful insights into rule-based machine translation of unrelated languages with varying levels of morphological complexity. As German is a language taught in Scandinavian schools this MT system would be of immediate relevance for Sámi school children learning German. In addition, there is a strong Finno-Ugric tradition in the German linguistics space that has in the past produced important publications on the Sámi language, so the system is immediately useful for researchers and enthusiasts as well as language users.

1 Introduction

1.1 Motivation

Machine translation is an important tool for language users. The most common contemporary method for implementing machine translation is to curate professionally translated texts and use machine learning methodology to learn the translations. This presupposes the availability of perhaps several millions of professionally translated sentences, which is unfeasible for under-resourced marginalised languages, where very little parallel corpora or even monolingual corpora are available. To put the low-resourcedness of North Sámi in context, the largest available monolingual corpus (SIKOR, 2018) is only 38 million tokens, and for the bilingual corpora at most 10,000s of aligned phrases, most of which are from Linux program GUI translations[1]. Given the circumstances, we do not find it reasonable to try to train a neural network for this task. The sensible solution is to use linguistic knowledge to build a rule-based machine translation system. What we are presenting in this article is a machine translation from North Sámi to German, a language pair that to our knowledge has not brought forth any system before, and that does not have enough resources for a neural machine translation system. Furthermore, our contribution consists in exploring a newly created module in a rule-based machine translation system, and we are looking at workflows for the rapid development of a baseline machine translator.

The rule-based system took us only some 100 hours to write and is the work of one programmer/linguist/advanced learner of German and native speaker of Finnish, an expert on Apertium - and one computational linguist, native speaker of German with high proficiency in North Sámi (but not a native speaker of it). The system described here is a work-in-progress, yet it is a proof-of-concept that rapid building of a machine translation system is plausible without big data corpus resources. Our motivation to build this system is two-fold: we are building a tool for users, as well as surveying the use of newly introduced techniques in a language pair that is not within the same language family and not English. This is also the *novel research* in our experiment: we provide further insights on the usage of the *new additions* to methodologies in a recently updated machine translation system in a typologically varied setting, that has not been tried before to our knowledge.

In the context of machine translation as a tool for supporting under-resourced language use, one must practice a certain level of carefulness in order

[1] https://opus.nlpl.eu/KDE4.php

to not cause more damage than good. For example, creating a system for generating large amounts of translations from the majority language to minority languages, for example, might sound like a lucrative offering to generate big data, but may result in creating larger bodies of automatically translated texts that overtake what there exists of naturally written texts which in the long run can be rather problematic. On the other hand, creating a system that translates well enough for *language understanding* (gisting) for majority language users will enable the minority language communities to wider use of their language in digital contexts. We stick to the ethics of not flooding the web with low-quality North Sámi text by building the system the other way around (German - North Sámi). Clean data is still of great value, and we do not want to put that in danger.

The machine translation system we created is freely available and open source in Apertium's GitHub repository[2]. The dependent North Sámi language model we developed earlier is also available at our github[3] and German model from Apertium's collection[4].[5]

1.2 Languages

North Sámi is a Finno-Ugric language belonging to the Uralic languages spoken in Norway, Sweden, and Finland by approximately 25,700 speakers (Eberhard et al., 2018). It is a synthetic language, where the open *parts-of-speech* (PoS) – e.g. nouns, adjectives – inflect for case, person, number, and more. The grammatical categories are expressed by a combination of suffixes and stem-internal processes affecting root vowels and consonants alike, making it perhaps the most fusional of all Uralic languages. In addition to compounding, inflection and derivation are common morphological processes in North Sámi. German, on the other hand, is an Indo-European language. In contrast to all previous work, there is neither language family similarity, nor geographical proximity or political relation. The latter would be the case for Sámi - Norwegian where despite language typological unrelatedness there are (even syntactic) loans due to coexistence and interaction of the languages.

As German was the previous century's language of science, a lot of scientific literature on the Sámi language was published in German. Newer publications include the North Sámi - German, German - North Sámi dictionary (Sammallahti and Nickel, 2006) of high quality (containing valencies, idiomatic phrases, examples of use). German has also been one of the languages that school children get to pick as a foreign language at school. For both these reasons, it makes sense to have MT systems between these two languages.

Morphologically, the languages have similar features: both are morphologically richer and suffixing, and mark case for nominals and some tense, aspect, and mood as well as person for verbs, however, North Sámi also marks other grammatical features such as possession and aktionsart in morphology. Both languages also have the productive compounding of nominals. The syntactic differences are notable, while the neutral word order for both is SVO, there are a number of mismatching features in the syntax: pro-drop for 1. and 2. person in Sámi, separable verbs in German, adverbial positioning, word order in sub-clauses, question clauses or after adverbial extensions, etc.

2 Background

Previous MT systems involving North Sámi are North Sámi - Lule Sámi (Tyers et al., 2009) (Wiechetek et al., 2010), North Sámi - Norwegian (Trosterud and Unhammer, 2012), North Sámi - South Sámi (Antonsen et al., 2016), North Sámi - Finnish (Pirinen et al., 2017). The systems were all based on previous versions of Apertium, the state-of-the-art in rule-based machine translation.

There is an Apertium-based system for translating North Sámi to Norwegian,[6] that has been in end-user use. As German and Norwegian (Bokmål) are related languages, we expect to be able to use them as a reference when implementing our system.

We chose to use Apertium (Khanna et al., 2021) as it is popular in the context of under-resourced languages. The system is based, roughly speaking, on doing a morpho-syntactic analysis of the source text, transferring the analysis to the target language morpho-syntactic description, and generating it into the target text. There is a diagrammatic presentation of the system pipeline in Figure 1. This means that the system consists of

[2] https://github.com/apertium/apertium-sme-deu

[3] https://github.com/giellalt/lang-sme

[4] https://github.com/apertium/apertium-deu

[5] For reproducibility purposes, the tag konvens2022 is available in the mentioned repos

[6] https://gtweb.uit.no/jorgal

morphological analyser-generators of target and source languages, based on finite-state morphology (Beesley and Karttunen, 2003), and a constraint grammar (Karlsson, 1990; Didriksen, 2010) for syntactic and semantic analysis suitable for transferring the source language structures to target language structures.

See examples (1) and (2) for a concrete example. In our experiment, we had pre-existing morphological analysers for North Sámi[7]and German[8], and we have written a bilingual translation dictionary as well as the grammatical rules.

(1) Boadát go dál?
come.V.2SG QST now.ADV?
'Are you coming now'

(2) Kommst du jetzt?
come.V.2SG you.PRN.2SG now.ADV?
'Are you coming now?'

From the example we see that there is some level of syntactic mapping to be done between the languages: North Sámi is generally pro-drop i.e. missing the subject pronoun morphologically encoded in the verb where German requires this. Furthermore, North Sámi indicates question with a question particle that is not easily glossed in English or German—perhaps an approximate gloss could be 'is it such that'—in German, the word order change indicates the question-format of the sentence.

We base our system on the tools developed within the *GiellaLT* infrastructure for North Sámi and tools developed within Apertium community for German, these include state-of-the-art FST-based morphological analyzers, with Constraint Grammar syntactic analysis and disambiguation. We have done a few slight adjustments to both monolingual systems, but our main work is in the bilingual part. In Figure 1, the part we work on concerns the part under *transfer*, specifically we have used the *recursive structural transfer* path in this experiment, which is a newly built part of Apertium in 2021 (Khanna et al., 2021).

To give an impression of concrete resources and rules, we show in Figure 2[9] what the dictionaries and the rules look like:

3 Development

We predominantly used pre-existing morphological analysers and morpho-syntactic disambiguation for the North Sámi morphological analysis and disambiguation and German morphological generation (and vice versa, but this direction was not the main objective of this article). Our contribution in terms of developed resources is a bilingual lexicon i.e. North Sámi to German translation dictionary, and the development of bilingual grammatical rules that determine for example word order changes and introduction of words that don't exist in the source language, such as articles.

The bilingual lexicon development was done by hand by a linguist, in the following three steps:

1. Translating words of initial reference bilingual corpus[10]
2. Translating high-frequency words (from SIKOR)[11]
3. Translating words from a random sample of large monolingual corpus (from SIKOR)

The final result has been verified by a linguist with near-native language skills. The first two steps ensure high coverage in general, whereas the third step is necessary to have high enough coverage in the genres of evaluation corpus for the human evaluation to even be possible.

The grammatical transfer was developed based on the reference bilingual corpus first. We ran the translation system through our reference corpus and located easy-to-fix syntactic differences, such as missing articles and pronouns, and local word order changes, and wrote the rules for those. We also needed to write transfer rules to account for purely morphological mismatches: for example, German only has grammatical cases: nominative, genitive, accusative, and dative, whereas North Sámi also has local cases and other cases that translate into prepositional phrases in German. The prepositions for each case do not translate one-to-one. Typically, one case will translate into several prepositions depending on the semantic/valency context.

The resulting lexicon and rules are summarised in Table 1.

[7] https://github.com/giellalt/lang-sme

[8] https://github.com/apertium/apertium-deu

[9] anonymised

[10] https://github.com/apertium/apertium-sme-deu/blob/master/sme-deu-corpus.txt

[11] https://gtsvn.uit.no/langtech/trunk/words/lists/sme/sme_lemma.freq

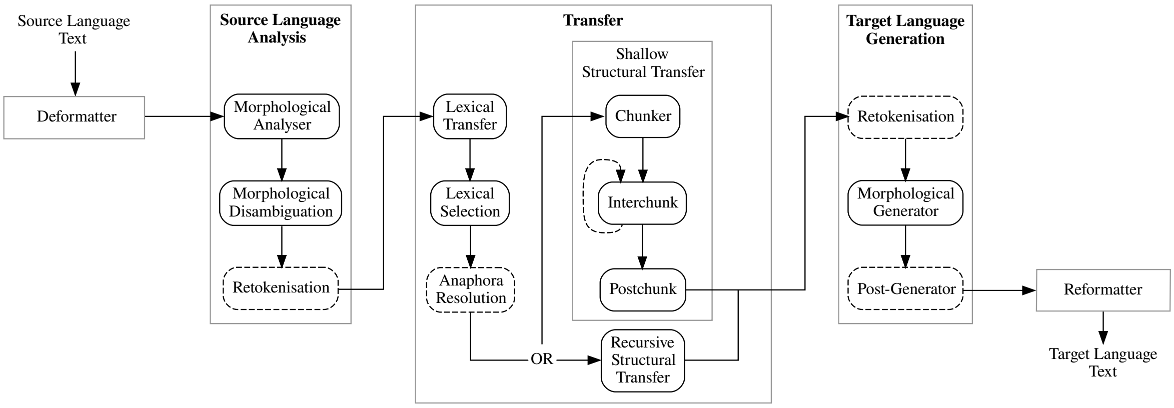

Figure 1: Apertium pipeline structure from (Khanna et al., 2021)

Bilingual dictionary

```
    <e><p><l>áddet</l><r>verstehen</r></p><par n=”vblex”/></e>
    <e><p><l>addit</l><r>geben</r></p><par n=”vblex”/></e>
    <e><p><l>addit</l><r>liefern</r></p><par n=”vblex”/></e>
    <e><p><l>álbmut</l><r>schaufeln</r></p><par n=”vblex”/></e>
    <e><p><l>álggahit</l><r>anfangen</r></p><par n=”vblex”/></e>
```

Syntactic rules

```
S -> VP NP { 1 _
    *(maybe_adp)[case=2.case]
    *(maybe_art)[number=2.number, case=2.case, gender=2.gender, def=ind]
    2 } ;
V -> %vblex {1[person = (if (1.tense = imp) ”” else 1.person),
               number = (if (1.number = du) pl else 1.number)] } ;
```

Figure 2: Bilingual dictionary format and syntactic rule format

4 Evaluation

As a corpus for evaluation of the translation quality, we randomly picked 300 paragraphs from *SIKOR*. This corpus is summarised in Table 1. We measured the naïve coverage of the monolingual analyser as well as our bilingual dictionary of the whole corpus to get an idea of how far we are in the process of building a translation dictionary suitable for any running texts.

4.1 Word Error Rate on Post-Edited text

We did a *Word Error Rate* (WER) test on our randomly selected corpus that was post-edited by a native speaker of German. Word error rate is a simple measure that calculates the proportion of the wrongly translated words, in this case when comparing the machine translation output to the translation that a human translator has post-edited. For example, if one word in a 10-word sentence is mistranslated, the word-error rate is 10 % and an exact match is 0 %. Notably, if the translation contains too many words, the word error rate can exceed 100 %. It is noteworthy that WER is also a rather naïve metric, for example, a wrong article or case is given the same weight as a completely wrong word. However, for understandability the latter is a much bigger obstacle than the wrong article. For the WER test, we used the `apertium-eval` tool available on their github[12]. The results of this evaluation are shown in Table 2.

5 Discussion and error analysis

One of the prevailing problems at this point of development is dictionary coverage. Creating the dictionary is one of the most time-consuming parts of the rule-based machine translation work. However, the resulting human-curated translation dictionary is a very valuable resource and therefore worth the effort. Once created, a translation dictionary can be included in any other future tool. Many of the errors we saw in the evaluation were due to low frequency, rather domain-specific words, such as *attorney general* or *vice candidate*, which had not been added to the bilingual dictionary yet.

[12] https://github.com/apertium/apertium-eval-translator

Data set	**Data size**	**Note**
Translation dictionary	4,340 LU pairs	newly built
Translation grammar	17 rules	newly built
German dictionary	100,390 LUs	extended
North Sámi dictionary	154,557 LUs	extended
Development corpus	1469 sentences	manually translated
SIKOR	38,94 Mtokens	monolingual corpus
Test set	7083 tokens	random sample

Table 1: *LU* is a lexical unit e.g. an entry in the dictionaries, *token* is a token in a running text e.g. word-form or punctuation, *Mtokens* is millions of tokens, and *sentences* in the text are based on our sentence boundary finding algorithm.

Corpus	**Naïve coverage**
Development corpus	99.8 %
Test set	88.2 %
SIKOR	84.6 %

Metric	**Test Corpus**
Post-Edit WER	77 %

Table 2: Evaluation of our North Sámi - German MT system

Some of the machine-translated sentences are intelligible despite grammatical errors. The translation of ex. (4) in ex. (3) requires lexical edits: *saamisch→Saamischsprachige*, *des Saamen→saamische*, *um→über*, *Lebensunterhalte→Gewerbe*, most of which are at least semantically related as can be seen in the correct translation of the sentence in ex. (4). In addition to the lexical edits, there are a number of word order issues, e.g. *treffen andere . . . →andere . . . treffen*. And also, e.g.*aufhören → hören . . . auf.*

(3) So können die Schüler treffen andere *saamisch, und lernen bißchen traditioneller *um *Lebensunterhalte *des Saamen.

(4) Nu besset oahppit deaivvadit eará
so können.3PL Schüler.PL treffen andere
sámegielagiiguin, ja oahppat
Saamischsprachig.KOM.PL, und lernen
veaháš árbevirolaš sámi
etwas traditionell saamisch
ealáhusaid birra.
Gewerbe.AKK.PL über;um
'So können die Schüler andere Saamischsprachige treffen, und ein bißchen über die traditionellen saamischen Gewerbe lernen.'

One of the interesting findings in this experiment is that, since the source and target languages are not related to each other[13] and the syntactic differences are notable, one focus of our work has been the tasks of word reordering and generation, which have typically been ignored in rule-based approaches to machine translation earlier. We found that the new recursive syntax-based approach in Apertium together with the high-quality Constraint Grammar-based syntactic analysis in the source language allows us to resolve reordering in an efficient way.

Looking at the edits we made in the post-edit, some errors are not as critical as the raw WER might suggest, for example, problems with the grammatical forms of the articles or compound splitting as well as separable verb processing may falsely increase the error rate more than it affects the readability. In the future, we will continue adding words as well as improve the description.

In a qualitative evaluation we found a lot of noise in the source text that affected the quality of our output. Noise in source texts is a much bigger problem in extremely low-resource languages like North Sámi and is due to newer or lacking language norms, lesser literacy and lesser use of the language in writing. (Wiechetek et al., 2022) We found the following types of noise: formatting errors and syllable splitting (potentially caused by corpus collection methods), spelling errors like accent errors and compound misspellings, grammatically doubtful sentences (potentially due to translations) and other grammatical errors like case errors.

6 Conclusion

We have developed the first North Sámi - German machine translation system in a short amount of

[13] Within Europe, the Finno-Ugric and Indo-European are as far apart as they can get.

time (100h) without any bilingual big data, based on the well-known Apertium system and the rule-based morpho-syntactic tools for North Sámi that are available in the *GiellaLT* infrastructure. The system is able to handle a number of syntactic transfer issues such as the generation of articles and longer distance reordering, such as the verb placement in a subordinate clause. We have evaluated our system and managed to develop a state-of-the-art system that is useful in terms of gisting, but still needs further development to serve as a post-editing tool. Our research contribution is not only an MT tool for a new language pair of completely unrelated languages but also, because of the unrelatedness, practical solutions to structural transfer problems that have been either ignored or marginalised in the past.

Acknowledgments

We thank Daniel Swanson from Apertium for his help and answers about the new system and Lene Antonsen for her help with the North Sámi Apertium usage.

References

Lene Antonsen, Trond Trosterud, and Francis M. Tyers. 2016. A North Saami to South Saami machine translation prototype. *Northern European Journal of Language Technology*, 4:11–27.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford.

Tino Didriksen. 2010. *Constraint Grammar Manual: 3rd version of the CG formalism variant*. GrammarSoft ApS, Denmark.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2018. *Ethnologue: Languages of the World*, twenty-fifth edition. SIL International, Dallas, Texas.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing unrestricted text. In *Proceedings of the 13th International Conference of Computational Linguistics*, volume 3, pages 168–173, Helsinki.

Tanmai Khanna, Jonathan North Washington, Francis Morton Tyers, Sevilay Bayatlı, Daniel Swanson, Flammie Pirinen, Irene Tang, and Héctor Alos i Font. 2021. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*.

Tommi Pirinen, Francis M. Tyers, Trond Trosterud, Ryan Johnson, Kevin Unhammer, and Tiina Puolakainen. 2017. North-sámi to Finnish rule-based machine translation system. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 115–122, Gothenburg, Sweden. Association for Computational Linguistics.

Pekka Sammallahti and Klaus Peter Nickel. 2006. *Sámi-duiskka sátnegirji=Saamisch-deutsches Wörterbuch*. Davvi Girji, Kárášjohka.

SIKOR. 2018. SIKOR uit norgga árktalaš universitehta ja norgga sámedikki sámi teakstačoakkáldat, veršuvdna 06.11.2018. `http://gtweb.uit.no/korp`. Accessed: 2018-11-06.

Trond Trosterud and Kevin Brubeck Unhammer. 2012. Evaluating North Sámi to Norwegian assimilation RBMT. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT 2012)*, 2013:03, pages 13–26, Gothenburg, Sweden. Chalmers University of Technology.

Francis M. Tyers, Linda Wiechetek, and Trond Trosterud. 2009. Developing Prototypes for Machine Translation between Two Sámi Languages. In *EAMT-2009*, pages 120–127, Barcelona, Spain. Universitat Politècnica de Catalunya.

Linda Wiechetek, Katri Hiovain-Asikainen, Inga Lill Sigga Mikkelsen, Sjur Moshagen, Flammie Pirinen, Trond Trosterud, and Børre Gaup. 2022. Unmasking the myth of effortless big data - making an open source multi-lingual infrastructure and building language resources from scratch. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1167–1177, Marseille, France. European Language Resources Association.

Linda Wiechetek, Francis M. Tyers, and Thomas Omma. 2010. Shooting at flies in the dark: Rule-based lexical selection for a minority language pair. In *Proceedings of the 7th International Conference on NLP (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 418–429, Berlin, Heidelberg. Springer.