Workflows for kickstarting RBMT in virtually No-Resource Situation

Tommi A Pirinen
Universität Hamburg,
Hamburger Zentrum für Sprachkorpora
Max-Brauer-Allee 60, Hamburg
tommi.antero.pirinen@uni-hamburg.de

Abstract

In this article we describe a work-in-progress best learnt practices on how to start working on rule-based machine translation when working with language that has virtually no pre-existing digital resources for NLP use. We use Karelian language as a case study, in the beginning of our project there were no publically available corpora, parallel or monolingual analysed, no analysers and no translation tools or language models. We show workflows that we have find useful to curate and develop necessary NLP resources for the language. Our workflow is aimed also for no-resources working in a sense of no funding and scarce access to native informants, we show that building core NLP resources in parallel can alleviate the problems therein.

1 Introduction

A lot of research goes into working with low-resource situation, however, in context of large international conferences today, loww-resources can mean anything from having millions and millions of lines of parallel corpus[1] to "anything except English". For this work we consider the lowest-resourced languages in the group of languages we work with, namely those having virtually no widely known publicly accessible or available resources at the start of our project, and for which we aim to search, curate and create the necessary resources. This is a work-in-progress, but we believe we have already gathered enough promising results to give some best recommended practices on how to start working on a language seemingly lacking all natural language processing (NLP) resources.

For the machine translation part we are working on doing a rule-based machine translation (RBMT) and specifically one between a minority language (Karelian) and a closely related more-resourced language (Finnish), in the first phase. The translator is bidirectional, i.e. we translate both Finnish to Karelian and vice versa. The work for majority language machine translations (e.g. English and Russian) is reserved for the future after some resources have been built. We have chosen this for a number of reasons, firstly the task is much easier when working with a related language than a typologically unrelated one, and secondly there is a body of good results using RBMT of closely related in further resource building, for example for Spanish-related languages in the Wikipedia content translation.

The article is organised as follows, first in Section 2 we describe the background and rationale for this project, in Section 3 we describe our approach and methodology for RBMT building, in Section 4 we describe our results so far and finally in the Section 5 we discuss findings and lay out future work.

2 Background

One of the problems, we have identified in the past in building NLP resources for minority languages, is that same or similar work ends up

[1] http://www.statmt.org/wmt19/parallel-corpus-filtering.html

doing multiple times between different scholars, or even within a single project of same minority language. This is not very ideal situation, when resources like native informants or skilled scholars are scarce. A typical example might be that a documentary linguistics effort builds a corpus of annotated texts, that includes hand annotated linguistic analysis, glosses and translations, while computational linguists build a morphological analyser, treebank and machine translator by hand from scratch as well. What we aim to achieve is synergy between these two different research practices.

The technological methodology used in this project is based on following:

- The rule-based machine translation is provided by apertium (Forcada et al., 2011).
- The morphological analyser-generator is based on the HFST engine (Lindén et al., 2009)
- The morphological disambiguation is based on Constraint Grammar (Karlsson, 1990)
- annotation format is based on Universal dependencies (Nivre et al., 2019)

This article describes what is still a work-in-progress, at this stage we are evaluating how the approach is and if we should make a project in building the supporting software for the methodology and language resource building. That is to say we have the workflows in place and the supporting software is built as we proceed with the project. Some of the workflows described here have been previously tested in building larger well-resourced machine translators, for example in (Pirinen, 2018). Based on experiences of this project, we could estimate that the effort needed is around 20,000 lines annotated and translated to get a comparable results as out of the box neural system (Pirinen, 2019), this is however a result achieved on two unrelated languages both of which aren't English, so results on related non-English languages may be different.

We have selected to use a rule-based approach to machine-translation for this project. Since rule-based approaches have somewhat fallen out of popularity in recent years, it needs strong arguments to select this approach in favour of others. For this purpose we have a check-list for which languages are to be used with which approaches first:

- Closely related languages: Finnish and Karelian are very closely related languages
- Lack of Parallel resources: Karelian has virtually zero digital resources
- Existence of written grammars: We have number of grammars to help (Zaikov, 2013)

One of the reasons we started to develop an approach to language resource creation that can produce multiple language resources fast, is that we have prior experience in 1. building computational linguistic resources like morphological analysers from the scratch without considering the corpus creation or documentary linguistics and 2. building language documentation corpora from the scratch without considering creation of dictionaries. The ideal result of this project is to develop a method that empowers computational linguists to work on their preferred form of language documentation and corpus creation and makes use of the expert work put in. This can always be achieved afterhands by scraping the produced corpora or data, but our plan is to introduce that as a part of workflow.

For other projects that have aimed to achieve similar goals, many are related to other rule-based machine translation efforts within the free/open source rule-based machine translation community, e.g.(?). On larger scale in the NLP community there have been several attempts to make computational linguists and documentary linguists work together towards common goal in this manner, for example (Maxwell and David, 2008; Blokland et al., 2015)

The basis of this RBMT system between Karelian and Finnish is that we also have a large coverage stable Finnish system already available (Pirinen, 2015). Karelian on the other hand has no resources, and is described by the ethnologuy as threatened[2]. We could have also tested an unrelated language

[2] https://www.ethnologue.com/language/krl

with large coverage dictionary, for example Russian-Karelian would be useful for the target audience, or build a machine translation between two under-resourced closely related languages, like Karelian and Livvi, which is a closely related language with slightly more resources than Karelian but much less than Finnish.

Finally, for social and political reasons, there is a growing interest in Karelian language and culture, and while there is a number of projects on the linguistic aspects and language learning, there is a lack of language technology-based projects in the field. Our aim is to fill that hole.

2.1 Languages

The language we use a case study is Karelian, a minority Uralic language spoken mainly in Republic of Karelia in Russia and in Finland. It is closely related to Finnish, Livvi and Ludic, but they are not mutually intelligible for an individual without at least some linguistic training. The naming of different languages and varieties related to Karelian is often confusing, what we aim to describe here is in line of ISO 639–3 language code krl; see the number 1 in the map in Figure 1[3] for the geographic distribution. For the machine translation task, in first phase we build a Karelian—Finnish translator.

3 Workflow

The workflow that we have reached at this point of the project is a synthesis of traditional workflow in documentary linguistics and workflows in building corpora and analyser writing, specifically in traditional rule-based systems. In documentational linguistics we have drawn experience and inspiration from SIL Fieldworks Explorer (FLeX) and the rule-based workflows are loosely based in tradition of Finite State Morphology.

The first part of the workflow is acquiring corpora, which for unresourced minority is relatively difficult task, at the beginning of our project we aimed to use web-as-corpus approach. During categorising the downloaded data into languages we found also a corpus

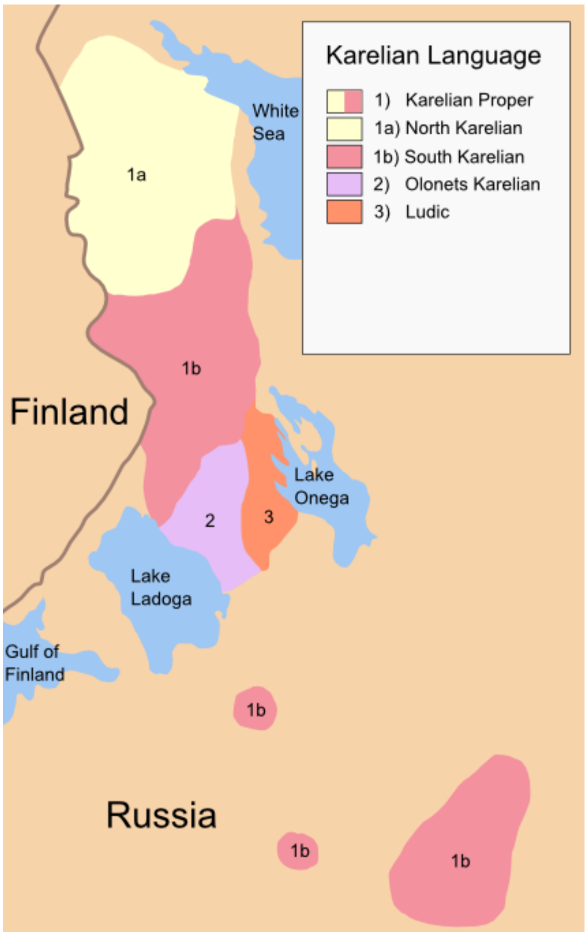

Figure 1: Map of Karelian languages, the number 1 is Karelian that we study in this article, numbers 2 and 3 are closely related languages that are in some literature refered to as Karelian as well, but are separate languages and do not belong under the krl language code in ISO standard.

[3] https://commons.wikimedia.org/wiki/File:Map_of_Karelian_dialects.png

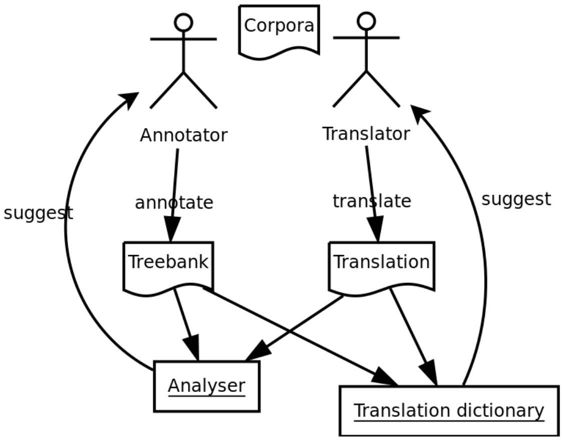

Figure 2: A UML-style chart of the annotation and translation process

repository with a free to use open source compatible licencing policy[4], which on top of expert made language classification has the advantage that we can keep full documents instead of shuffled sentences.

The actual corpus building workflow consists of two parts that can be alternated between, annotation and lexicon building. With annotation, we can work on any of the following tasks: lemmatising and pos tagging, morphological analysis, syntactic treebanking and machine translation. On the other side lexicon building we build the morphological lexicon for a finite-state analyser, and a bilingual lexicon for rule-based machine translation. A UML-style graph of the process is shown in figure 2.

The main contribution of this workflow is that both of the tasks feed into the other task, that is annotated corpora can be immediately used for entry generation of the lexicons, and the analysers and machine translators built from the lexicons are used to generate n-best lists from which annotators can choose the annotations.

We provide a real world example here: An annotator starts working on a new document that contains sentences: “Pelih ošallistu 13 henkie” (13 people participated in the play) the annotator annotates in UD format:

```
1 peliin peli  NOUN  Number=Sing|Case=Ill 2 obl _ _
2 ošallistu ošallistuo VERB  Number=Sing|Tense=Pres 0 root _ _
3 13 13 NUM Number=Sing|NumType=Card 4 nummod _ _
4 henkie henki NOUN  Number=Sing|Case=Par 2 nsubj _ _
```

and provides Finnish translation like “Peliin osallistui 13 henkilöä”. The annotation is used to generate entries for monolingual dictionary of Karelian, i.e. peli<n>, ošallistuo<vblex>, 13<num>, and henki<n>, the lexicon writer can simply fill in the necessary informations to inflect the words properly. The entries can likewise be generated to bilingual dictionary, if 1:1 translation match to existing target language analyses is trivial, we get peli<n>:peli<n> etc. among the suggested entries. Now, when the annotator gets back to annotating and translating the next sentences of the document and runs into: “Pelissä ”tapettih” šamoin Ilmarini” (Ilmarinen was also killed in the game), the first token “Pelissä” has suggested annotation peli NOUN Number=Sing|Case=Ine as well as suggested translation.

	Tokens	Sentences
Annotations	3094	228
Translations	1144	161

Table 1: The size of Karelian—Finnish corpus at the time of writing.

4 Results

In a short time we have managed to build a rule-based machine translation system. We detail the system in Table 1. The corpus built so far in this proto-typing phase of the project has been built by one expert annotator, working on spare time for three months in other words in only handful of work hours.

At the current moment we do not have enough bilingual corpora to measure the translation quality yet but we hope to include a BLEU and WER evaluations of the translation quality by the time we submit a camera-ready version of the paper.

The corpora will be released on github via the Apertium project for the translations and possibly also disambiguated corpora, and via Universal dependencies project for the annotated corpus. Both retain the CC BY licence of the original raw text data. The dictionaries and analysers are also released via the Apertium using the GNU General Public Licence.

5 Concluding remarks

We have found that we can rapidly build a solid base of natural language resourcses suitable for rule-based machine translation and we aim to extend the approach to more Uralic languages

[4] http://dictorpus.krc.karelia.ru/en/corpus/text

in near future. Furthermore the approach prototyped in this paper has been found very motivating and nice to work with in the future we will look at building a more approachable graphical user interface for it.

The approach we describe here is especially suitable in no-resources starting situation, even a limited amount of resources will open more workflows, more technical possibilities to aid in the initial part of the corpus building and resource building. However, we still think this approach may be useful as a part of balanced corpus building approach in a research project for any lesser researched language.

One of the things that we are looking forward to is to test the advances in neural methods in very low resource situation, (Neubig and Hu, 2018)[5]this would be particularly suitable for Karelian-to-Finnish direction as Finnish is well-resourced.

References

Blokland, Rogier, Marina Fedina, Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2015. Language documentation meets language technology. In Septentrio Conference Series, number 2, pages 8–18.

Forcada, Mikel L, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. Machine translation, 25(2):127–144.

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

Lindén, Krister, Miikka Silfverberg, and Tommi Pirinen. 2009. Hfst tools for morphology—an efficient open-source package for construction of morphological analyzers. In Mahlow, Cerstin and Michael Piotrowski, editors, sfcm 2009, volume 41 of Lecture Notes in Computer Science, pages 28—47. Springer.

Maxwell, Michael and Anne David. 2008. Joint grammar development by linguists and computer scientists. In Workshop on NLP for Less Privileged Languages, Third International Joint Conference on Natural Language Processing, pages 27–34, Hyderabad, India.

Neubig, Graham and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. arXiv preprint arXiv:1808.04189.

Nivre, Joakim, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Andre Kaasen, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman

[5] We thank the anonymous reviewers for bringing this line of work to our attention

Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Abigail Walsh Sarah McGuinness, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Pirinen, Tommi A. 2015. Development and use of computational morphology of finnish in the open source and open science era: Notes on experiences with omorfi development. SKY Journal of Linguistics, 28.

Pirinen, Tommi A. 2018. Rule-based machine-translation between finnish and german.

Pirinen, Tommi A. 2019. Neural and rule-based finnish nlp models—expectations, experiments and experiences. In Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages, pages 104–114.

Zaikov, Pekka. 2013. Vienankarjalan kielioppi.