Dialect Text Normalization to Normative Standard Finnish

Niko Partanen
Department of Finnish, Finno-Ugrian and Scandinavian Studies
University of Helsinki

Mika Hämäläinen
Department of Digital Humanities
University of Helsinki

Khalid Alnajjar
Department of Computer Science
University of Helsinki

firstname.lastname@helsinki.fi

Abstract

We compare different LSTMs and transformer models in terms of their effectiveness in normalizing dialectal Finnish into the normative standard Finnish. As dialect is the common way of communication for people online in Finnish, such a normalization is a necessary step to improve the accuracy of the existing Finnish NLP tools that are tailored for normative Finnish text. We work on a corpus consisting of dialectal data from 23 distinct Finnish dialect varieties. The best functioning BRNN approach lowers the initial word error rate of the corpus from 52.89 to 5.73.

1 Introduction

Normalization is one of the possible preprocessing steps that can be applied to various text types in order to increase their compatibility with tools designed for the standard language. This approach can be taken in an essentially similar manner with dialectal texts, historical texts or colloquial written genres, and can be beneficial also as one processing step with many types of spoken language materials.

Our study focuses to the normalization of dialect texts, especially within the format of transcribed dialectal audio recordings, published primarily for linguistic research use. However, the dialectal correspondences in this kind of material are comparable to phenomena in other texts where dialectal features occur, the results are expected to be generally applicable.

This paper introduces a method for dialect transcript normalization, which enables the possibility to use existing NLP tools targeted for normative Finnish on these materials. Previous work conducted in English data indicates that normalization is a viable way of improving the accuracy of NLP methods such as POS tagging (van der Goot et al., 2017). This is an important motivation as the non-standard colloquial Finnish is the de facto language of communication on a multitude of internet platforms ranging from social media to forums and blogs. In its linguistic form, the colloquial dialectal Finnish deviates greatly from the standard normative Finnish, a fact that lowers the performance of the existing NLP tools for processing Finnish on such text.

2 Related work

Automated normalization has been tackled in the past many times especially in the case of historical text normalization. A recent meta-analysis on the topic (Bollmann, 2019) divides the contemporary approaches into five categories: substitution lists like VARD (Rayson et al., 2005) and Norma (Bollmann, 2012), rule-based methods (Baron and Rayson, 2008; Porta et al., 2013), edit distance based approaches (Hauser and Schulz, 2007; Amoia and Martinez, 2013), statistical methods and most recently neural methods.

For statistical methods, the most prominent recent ones have been different statistical machine translation (SMT) based methods. These methods often assimilate the normalization process with a regular translation process by training an SMT model on a character level. Such methods have been used for historical text (Pettersson et al., 2013; Hämäläinen et al., 2018) and contemporary dialect normalization (Samardzic et al., 2015).

Recently, many normalization methods utilized neural machine translation (NMT) analogously to the previous SMT based approaches on a character level due to its considerable ability in addressing the task. Bollmann and Søgaard (2016) have used a bidirectional long short-term memory (bi-LSTM) deep neural network to normalize historical German on a character level. The authors have also tested the efficiency of the model

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 141–146
Hong Kong, Nov 4, 2019.

when additional auxiliary data is used during the training phase (i.e. multi-task learning). Based on their benchmarks, normalizations using the neural network approach outperformed the ones by conditional random fields (CRF) and Norma, where models trained with the auxiliary data generally had the best accuracy.

Tursun and Cakici (2017) test out LSTM and noisy channel model (NCM), a method commonly used for spell-checking text, to normalize Uyghur text. In addition to the base dataset ($\approx$ 200 sentences obtained from social networks, automatically and manually normalized), the authors have generated synthetic data by crawling news websites and introducing noise in it by substituting characters with their corresponding corrupted characters at random. Both of the methods have normalized the text with high accuracy which illustrates the their effectiveness. Similarly, Mandal and Nanmaran (2018) had employed an LSTM network and successfully normalized code-mixed data with an accuracy of 90.27%.

A recent study on historical English letters (Hämäläinen et al., 2019) compares different LSTM architectures finding that bi-directional recurrent neural networks (BRNN) work better than one-directional RNNs, however different attention models or deeper architecture do not have a positive effect on the results. Also providing additional data such as social metadata or century information makes the accuracy worse. Their findings suggest that post-processing is the most effective way of improving a character level NMT normalization model. The same method has been successfully applied in OCR post-correction as well (Hämäläinen and Hengchen, 2019).

3 Data

Finnish dialect materials have been collected systematically since late 1950s. These materials are currently stored in the Finnish Dialect Archive within Institute for the Languages of Finland, and they amount all in all 24,000 hours. The initial goal was to record 30 hours of speech from each pre-war Finnish municipality. This goal was reached in the 70s, and the work evolved toward making parts of the materials available as published text collections. Another approach that was initiated in the 80s was to start follow-up recordings in the same municipalities that were the targets of earlier recording activity.

Later the work on these published materials has resulted in multiple electronic corpora that are currently available. Although they represent only a tiny fraction of the entire recorded material, they reach remarkable coverage of different dialects and varieties of spoken Finnish. Some of these corpora contain various levels of manual annotation, while others are mainly plain text with associated metadata. Materials of this type can be characterized by an explicit attempt to represent dialects in linguistically accurate manner, having been created primarily by linguists with formal training in the field. These transcriptions are usually written with a transcription systems specific for each research tradition. The result of this type of work is not simply a text containing some dialectal features, but a systematic and scientific transcription of the dialectal speech.

The corpus we have used in training and testing is the Samples of Spoken Finnish corpus (Institute for the Languages of Finland, 2014). It is one of the primary traditional Finnish dialect collections, and one that is accompanied with hand-annotated normalization into standard Finnish. The size of corpus is 696,376 transcribed words, of which 684,977 have been normalized. The corpus covers 50 municipalities, and each municipality has two dialect samples. The materials were originally published in a series between 1978-2000. The goal was to include various dialects systematically and equally into the collection. The modern digital corpus is released under CC-BY license, and is available with its accompanying materials and documentation in the Language Bank of Finland.[1]

The data has been tokenized and the normative spellings have been aligned with the dialectal transcriptions on a token level. This makes our task with normalization model easier as no preprocessing is required. We randomly sort the sentences in the data and split them into a training (70% of the sentences), validation (15% of the sentences) and test (15% of the sentences) sets.

4 Dialect normalization

Our approach consists of a character level NMT model that learns to translate the dialectal Finnish to normative spelling. We experiment with two different model types, one being an LSTM based BRNN (bi-directional recurrent neural network) approach as taken by many in the past, and the

[1] http://urn.fi/urn:nbn:fi:lb-201407141

other is a transformer model as it has been reported to outperform LSTMs in many other sequence-to-sequence tasks.

For the BRNN model, we use mainly the OpenNMT (Klein et al., 2017) defaults. This means that there are two layers both in the encoder and the decoder and the attention model is the general global attention presented by Luong et al. (2015). The transformer model is that of Vaswani et al. (2017). Both models are trained for the default 100,000 training steps.

We experiment with three different ways of training the models. We train a set of models on a word level normalization, which means that the source and target consist of single words split into characters by white spaces. In order to make the models more aware of the context, we also train a set of models on chunked data. This means that we train the models by feeding in 3 words at a time; the words are split into characters and the word boundaries are indicated with an underscore character (_). Lastly we train one set of models on a sentence level. In this case the models are trained to normalize full sentences of words split into characters and separated by underscores.

In terms of the size of the training data, the word level data consists of 590k, the chunk level of 208k and the sentence level of 35k parallel rows. All of the models use the same split of training, testing and validation datasets as described earlier. The only difference is in how the data is fed into the models.

5 Results & Evaluation

We evaluate the methods by counting the word error rate[2] (WER) of their output in comparison with the test dataset. WER is a commonly used metric to assess the accuracy of text normalization.

Table 1 shows the WERs of the different methods. The initial WER of the non normalized dialectal text in comparison with the normalized text is shown in the column *No normalization*. As we can see from this number, the dialectal text is very different from the standardized spelling. Both the word level and chunk level normalization methods reach to a very high drop in the WER meaning that they manage to normalize the text rather well. Out of these, the chunk level BRNN achieves the best results. The performance is the worst in the sentence level models, even to a degree that the transformer model manages to make the WER higher than the original.

5.1 Error analysis

Table 2 illustrates the general performance of the model, with errors marked in bold. The example sentence fragments are chosen by individual features they exhibit, as well as by how well they represent the corpus data.

Since the model accuracy is rather high, the errors are not very common in the output. We can also see clearly that the chunk model is able to predict the right form even when form is reduced to one character, as on line 5.

Since the dialectal variants often match the standard Finnish, over half of the forms need no changes. The model learns this well. Vast majority of needed changes are individual insertions, replacements or deletions in the word end, as illustrated in Table 2 at lines 2, 4, 6, 7, 15, 16, 17 and 18. However, also word-internal changes are common, as shown at lines 11 and 12. Some distinct types of common errors can be detected, and they are discussed below.

In some cases the errors are clearly connected to the vowel lengthening that does not mark ordinary phonological contrast. Line 3 shows how the dialectal pronoun variant of *he* 'he / she', *het*, is occasionally present in dialect material as *heet*, possibly being simply emphasized in a way that surfaces with an unexpected long vowel. This kind of sporadic vowel lengthening is rare, but seems to lead regularly to a wrong prediction, as these processes are highly irregular. This example also illustrates that when the model is presented a rare or unusual form, it seems to have a tendency to return prediction that has overgone no changes at all.

The model seems to learn relatively well the phonotactics of literary Finnish words. However, especially with compounds it shows a trait to classify word boundaries incorrectly. A good example of this is *ratapölökynterv̀aaų̆""kon* 'railroad tie treatment machine', for which the correct analysis would be 'rata#pölkyn#tervaus#kone'[3], but the model proposes 'rata#pölkyn#terva#uskoinen' which roughly translates as 'railroad tie creosote believer'. The latter variant is semantically quite awkward, but morphologically possible. This

[2]We use the implementation provided in https://github.com/nsmartinez/WERpp

[3]Here # is used for the illustrative purpose to indicate word boundaries within the compound

	No normalization	Words		Chunks of 3		Sentences	
		BRNN	Transformer	BRNN	Transformer	BRNN	Transformer
WER	52.89	6.44	6.34	**5.73**	6.1	46.52	53.23

Table 1: The word error rates of the different models in relation to the test set

	source	correct target	prediction
1	joo	joo	joo
2	ettɛ	että	että
3	heet	he	**heet**
4	uskovah	uskovat	uskovat
5	n	niin	niin
6	e͡ttɛ	että	että
7	sinn	sinne	sinne
8	e͡i	ei	ei
9	ole	ole	ole
10	,	,	,
11	kukhaan	kukaan	kukaan
12	ymmärtänny	ymmärtänyt	ymmärtänyt
13	mennä	mennä	mennä
14	.	.	.
15	Artjärveŋ	Artjärven	Artjärven
16	kirkolt	kirkolta	kirkolta
17	mennäh	mennään	mennään
18	sinneh	sinne	sinne
19	Hiiteläh	Hiitelään	**Hiitelässä**

Table 2: Examples from input, output and prediction

phonotactic accuracy makes selection of correct analysis from multiple predicted variants more difficult, as it is not possible to easily detect morphologically valid and invalid forms. The longer words such as this also have more environments where normalization related changes have to be done, which likely makes their correct prediction increasingly difficult.

In word level model there are various errors related to morphology that has eroded from the dialectal realizations of the words, or correspond to a more complicated sequences. Long vowel sequences in standard Finnish often correspond to diphthongs or word internal -h- characters, and these multiple correspondence patterns may be challenging for the model to learn. Chunk model performs few percentages better than word model in predictions where long vowel sequences are present, which could hint that the model benefits from wider syntactic window the neighbouring words can provide. On line 19 a case of wrongly selected spatial case is illustrated.

There are cases where dialectal wordforms are ambiguous without context, i.e. standard Finnish cases adessive (-lla) and allative (-lle) are both marked with single character (-l). Various sandhi-phenomena at the word boundary also blurren the picture by introducing even more possible interpretations, such as *vuoristol laitaa*, where the correct underlying form of the first element would be *vuoriston* 'mountain-GEN'. The decision about correct form cannot be done with information provided only by single forms in isolation. The chunk level model shows small but consistent improvements in these cases. This is expected, as the word level model simply has no context to make the correct prediction.

It is important to note that since the model is trained on linguistic transcriptions, its performance is also limited to this context. For example, in the transcriptions all numbers, such as years and dates, are always written out as words. Thereby the model has never seen a number, and is doesn't process them either. Improving the model with additional training data that accounts this phenomena should, on the other hand, be relatively straightforward. Similarly the model has had only very limited exposure to upper case characters and some of the punctuation characters used in ordinary literary language, which should all be taken into account when attempting to use the model with novel datasets.

6 Conclusion & Future work

The normalization method we have proposed reaches remarkable accuracy with this dialectal transcription dataset of spoken Finnish. The error rate is so low that even if manual normalization would be the ultimate target, doing this in combination with our approach would make the work manifold faster. We have tested the results with large enough material that we assume similar method would work in other conditions where same preliminary conditions are met. These are sufficiently large amount of training data and systematic transcription system used to represent the dialectal speech.

Future work needs to be carried out to evaluate the results on different dialectal Finnish datasets, many of which have been created largely within the activities described earlier, but which are also continuously increasing as research on Finnish is a very vibrant topic in Finland and elsewhere. This method could also be a very efficient in increasing the possibilities for natural language processing of other contemporary spoken Finnish texts. Our method could also be easily used within OCR correction workflows, for example, as a step after automatic error correction.

Situation is essentially similar, to our knowledge, also in other countries with comparable history of dialectal text collection. Already within Finnish archives there are large collections of dialectal transcriptions in Swedish, as well as in the endangered Karelian and Sami languages. Applying our method into these resources would also directly improve their usability. However, it has to be kept in mind that our work has been carried out in a situation where the manually annotated training data is exceptionally large. In order to understand how widely applicable our method is for an endangered language setting, it would be important to test further how well the model performs with less data.

The performance with less data is especially crucial with low-resource languages. Many endangered languages around the world have text collections published in the last centuries, which, however, customarily use a linguistic transcription system that deviates systematically from the current standard orthography. Such a legacy data can be highly useful in language documentation work and enrich modern corpora, but there are challenges in normalization and further processing of this data (Blokland et al., 2019). The approach presented in our paper could be applicable into such data in various language documentation situations, and the recent interest the field has displayed toward language technology creates good conditions for further integration of these methods (Gerstenberger et al., 2016).

We have released the chunk-level BRNN normalization model openly on GitHub as a part of an open-source library called Murre[4]. We hope that the normalization models developed in this paper are useful for other researchers dealing with a variety of downstream Finnish NLP tasks.

[4]https://github.com/mikahama/murre

7 Acknowledgements

Niko Partanen's work has been conducted within the project Language Documentation meets Language Technology: The Next Step in the Description of Komi, funded by the Kone Foundation.

References

Marilisa Amoia and Jose Manuel Martinez. 2013. Using comparable collections of historical texts for building a diachronic dictionary for spelling normalization. In *Proceedings of the 7th workshop on language technology for cultural heritage, social sciences, and humanities*, pages 84–89.

Alistair Baron and Paul Rayson. 2008. VARD2: A tool for dealing with spelling variation in historical corpora. In *Postgraduate conference in corpus linguistics*.

Rogier Blokland, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2019. Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. In *Workshop on Computational Methods for Endangered Languages, Honolulu, Hawai'i, USA, February 26–27, 2019*, volume 2, pages 24–30. University of Colorado.

Marcel Bollmann. 2012. (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, Lisbon, Portugal.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcel Bollmann and Anders Søgaard. 2016. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 131–139, Osaka, Japan. The COLING 2016 Organizing Committee.

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology*, 4:29–47.

Rob van der Goot, Barbara Plank, and Malvina Nissim. 2017. To normalize, or not to normalize: The impact of normalization on Part-of-Speech tagging. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 31–39.

Mika Hämäläinen and Simon Hengchen. 2019. From the Paft to the Fiiture: a fully automatic NMT and Word Embeddings Method for OCR Post-Correction. In *Recent Advances in Natural Language Processing*, pages 432–437. INCOMA.

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 Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 87–96.

Mika Hämäläinen, Tanja Säily, Jack Rueter, Jörg Tiedemann, and Eetu Mäkelä. 2019. Revisiting NMT for normalization of early English letters. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 71–75, Minneapolis, USA. Association for Computational Linguistics.

Andreas W Hauser and Klaus U Schulz. 2007. Unsupervised learning of edit distance weights for retrieving historical spelling variations. In *Proceedings of the First Workshop on Finite-State Techniques and Approximate Search*, pages 1–6.

Institute for the Languages of Finland. 2014. Suomen kielen näytteitä - Samples of Spoken Finnish [online-corpus], version 1.0. http://urn.fi/urn:nbn:fi:lb-201407141.

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

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Soumil Mandal and Karthick Nanmaran. 2018. Normalization of transliterated words in code-mixed data using Seq2Seq model & Levenshtein distance. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 49–53, Brussels, Belgium. Association for Computational Linguistics.

Eva Pettersson, Beáta Megyesi, and Jörg Tiedemann. 2013. An SMT approach to automatic annotation of historical text. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*, 087, pages 54–69. Linköping University Electronic Press.

Jordi Porta, José-Luis Sancho, and Javier Gómez. 2013. Edit transducers for spelling variation in Old Spanish. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*, 087, pages 70–79. Linköping University Electronic Press.

Paul Rayson, Dawn Archer, and Nicholas Smith. 2005. VARD versus WORD: a comparison of the UCREL variant detector and modern spellcheckers on english historical corpora. *Corpus Linguistics 2005*.

Tanja Samardzic, Yves Scherrer, and Elvira Glaser. 2015. Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of the 7th Language and Technology Conference*. ID: unige:82397.

Osman Tursun and Ruket Cakici. 2017. Noisy Uyghur text normalization. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 85–93, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.