Challenges of Using Character Level Statistical Machine Translation for Normalizing Old Estonian Texts

Gerth Jaanimäe[1]

[1] *University of Tartu, Institute of Estonian and General Linguistics, Jakobi 2, Tartu, 51005, Estonia*

Abstract

This paper reports on experiments of normalizing the 19th century Estonian parish court records. Converting the historical texts from old to contemporary spelling system, also known as normalizing, can be challenging in itself due to the fact that there was no single orthographic standard or if there even was, often the rules were not strictly followed, so there was a lot of variation in the texts. This paper also concentrates on the more specific issues related to Estonian as a morphologically rich language and presents the initial results of applying the character level statistical machine translation normalization on the parish court records from the 19th century. Morphological richness and the peculiarities of the old orthography can create the problem of ambiguity, which we attempted to solve using word bigrams instead of single words for training. Also, as the annotated training data is scarce and we assumed that more of it helps us obtain better results, we tested the idea to create the artificial additional training data, the so-called silver standard. The old texts which's spellings were closest to modern Estonian were converted to the old spelling system, which is much simpler than the reverse process, and after that added to the training set.

Keywords

natural language processing, historical texts, corpus linguistics, text normalization

1. Introduction

Historical texts are invaluable resource for linguists, historians, genealogists and other people who use digital archives in their work. In the linguistic point of view these writings are interesting for the reason that they can provide an insight to the dialects, vocabulary and grammar used in the time period they were written in. These writings can be difficult to analyze automatically due to the differences between modern and old orthographies. Thus, the tools designed for contemporary language usually perform worse on them and they have to be converted to modern form, or in other words normalized [1]. Another approach would be adapting the tools to older orthography, however it would be very time consuming.

Estonian, which belongs to Finno-Ugric language family and on which this research is based, is a morphologically rich language, meaning that many different word forms can be created, and thus more material is needed to cover the vocabulary. Another issue that can occur is that some of the words normalized can create forms which are homonymous with forms of another word, which may cause falsely recognized lemmas for a given word. Automatic detection of these errors can be complicated, mainly because these words are often morphologically correct, and the sentences formed by them can also be in accordance with the rules of syntax.

The dataset that is used in this research consists of parish court records written in the 19th century. These texts were written mostly in Estonian and provide a valuable insight into the way of life, relationships and the language that was used colloquially during this time period. Some of these texts

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), March 15–18, 2022, Uppsala, Sweden.
EMAIL: gerth.jaanimae@ut.ee
ORCID: 0000-0002-9588-1642


were written in old Estonian orthography, some in modern and a little portion in the so-called transitional spelling system. Also, the texts contain a sizeable amount of dialectal variation.

These varieties make them especially interesting from the linguistic point of view, however at the same time make them more difficult to normalize.

In this paper we discuss the issues described above and present the initial results of applying the statistical machine translation method for normalizing the Estonian texts written in the 19th century.

The paper consists of the following sections. Section 2 gives an overview of the data used in this research and describes the issues related to it. In Section 3, the normalization method and related work is described. Section 4 provides an overview of the preprocessing, and the normalization experiments themselves. Section 5 gives the summary of the results of the experiments and attempts to give the reasons behind them. In Section 6, the reasons are elaborated further and future plans are briefly discussed.

2. Description of the dataset

The dataset analyzed in this research consists of parish court records written in the 19th century. Automatic analysis of these texts would make it possible to perform keyword searches and use different NLP applications that are designed for standard language. While there exist NLP tools for standard Estonian, such as a Python library called ESTNLTK [2], the researched material have some features that make it impossible or extremely difficult to apply them off the shelf. Also, as Estonian morphology contains fusional elements, searching different keywords using regular expressions would be impossible or at least a lot of hard work. For example the genitive and partitive forms for South Estonian word susi ‘wolf’ is soe and sutt.

Not only is the material written in older spelling system and non-standard Estonian, they were also hand written and due to a big variation in the handwriting styles, it would be difficult and error-prone to use optical character recognition on them. Thus the texts were first manually transcribed by volunteers in the crowdsourcing project launched by the National Archives of Estonia.[2] After that further processing and analysis could be performed.

Many of these writings are written in old spelling system which was introduced around the end of the 17th century and was heavily influenced by German orthography at the time. The main rules were as follows: 1. The long vowel of a stressed open syllable is marked by a single letter. 2. The long vowel of a stressed closed syllable is marked by a digraph. 3. The short vowel of a stressed open syllable is marked by a double consonant [3].

The old spelling system was also ambiguous as the Table 1 shows [4]. Although for a human it is quite easy to make the correct decision based on the context, it would be incredibly difficult for the normalization algorithm to know, which of the modern equivalents is the correct one.

To make matters more complicated there were two written languages in parallel use until the end of the 19th century representing North vs. South Estonian. Eventually the North Estonian language and spelling standard became the single standard for the whole country. The spelling standard Estonians know and use today was introduced in 1843 and started gaining popularity in the 1870s. This means that although there is some material in the dataset written in Modern Estonian orthography, most of it is written in older spelling and some of it during a transitional period, where people still wrote some words in the earlier spelling out of habit [3].

[2] https://www.ra.ee/vallakohtud/

Table 1
Differences between old and modern Estonian orthographies

Old spelling	Modern spelling	Meaning
ma	*maa*	land
ma	*ma*	I
ramat	*raamat*	book
maalt	*maalt*	from the country
munna	*muna*	egg
teggi	*tegi*	did
kolli	*koli*	stuff
kolli	*kolli*	monster (genitive form)

South Estonian used to be considered a dialect of Estonian, but nowadays many linguists classify it as a separate language due to numerous grammatical and phonological differences suggesting that the South Estonian language branched off the Proto-Finnic language earlier on [5]. As the main goal of this research is to normalize the texts to standard Estonian, North and South Estonian are still treated as dialects. The data can be divided into nine different dialectal areas which in turn can be grouped into North and South Estonian dialects.

North Estonian: central, insular, coastal, western, eastern and northeastern dialects.

South Estonian: Mulgi, Tartu and Võru dialects.

Mulgi dialect was an interesting case as the official language in this area was North Estonian, although colloquially South Estonian was spoken instead.

In addition to the sizable amount of dialectal variation, there are more challenges in normalizing these texts. Morphological richness, meaning that cases and derivations are used instead of prepositions and postpositions, poses some extra challenges in normalization. Main one being that there are inevitably many more different wordforms to normalize and thus probability of mistakes will be significantly increased. Also, there would be much smaller amount of frequently occurring prepositions that would automatically increase the scores reflecting the quality of normalization.

Another problem is the small amount of manually annotated data for training the machine learning algorithm as the annotation process is time consuming, human resources are limited and due to dialectal variations, data from one region often does not work for normalizing texts from another region.

3. Method

The method of normalizing older texts by converting them to standard modern spelling can be achieved using many different methods, such as dictionaries, rule-based approach, edit distances, machine translation etc.

3.1. Method

The method used in the current investigation is often referred to as character level statistical machine translation, where the old and modern spelling systems are treated as two separate languages.

Also, as the "languages" are similar enough, the words are processed as sentences and characters as words. This makes it possible to translate the patterns of letters instead of just individual words, thus making it more flexible, compared to, for example, the dictionary-based method [6]. In order to overcome the challenges described in the previous section, the following processes were implemented.

In order to mitigate the problem of ambiguity, the bigrams, or in other words word pairs were used instead of giving a single word at a time for the algorithm to process. Therefore, the problem of ambiguity could possibly be solved thanks to the collocations providing the translated words the context.

The issue of scarcity of data could be solved by creating more artificial data for the algorithm to learn from, a so-called silver standard. The conversion from the contemporary spelling to old spelling

can be achieved with a small amount of rules and thus could be done more easily than the reverse process. The texts were converted to the old spelling system and the pairs of texts were given for the machine translation algorithm to learn.

3.2. Related work

Using character level statistical machine translation for normalizing historical texts is nothing new. One of the first experiments with this methods was to normalize old Slovene texts written in the 18th and 19th century [6]. It has been also extensively tested in order to compare its performance on English, Swedish, German, Icelandic and Hungarian [7]. Although there are more state of the art methods today, such as ones based on neural networks, which usually have better performance, they require large amounts of data. Some researchers have also found out that the method even performs worse on the smaller dataset [8] [9].

4. Description of experiments and setup

In order to evaluate the statistical machine translation for normalizing the text material the following preprocessing and experiments were performed.

A small set of parish court records, 153 in total, was randomly chosen for manual annotation and normalization. The annotation consists of morphological information, such as lemma and inflectional information. It also contains the normalized form for every given word, which is the main interest in our research.

Before training, the tokens were separated by newline, the letters by whitespace and the punctuation was removed. The corpus was then divided into nine smaller datasets according to the dialectal variations. After that these smaller datasets were randomly divided into training set, development set and test set in size of 75%, 5% and 20% respectively. The software used for the translation process was Moses[3].

Training the models consisted from two steps. First the target language model is trained and after that the translation model. For the former the parallel corpus is needed and for the latter the corpus in the target language, or in our case, normalized words are required.

The scripts and related files are uploaded to Github.[4]

The following subsections describe different types of experiments.

4.1. Baseline translation

The manually annotated corpus was used to train both the language model and translation model without any additions or modifications. As the target language or in our case normalized forms are in the same language for every dialect, the training sets were merged into single file for the language model.

The training set was used to train the translation models and the development set to tune them using minimal error rate training (mert). The accuracy on the test sets were calculated by comparing the translation to the normalized form found in the test set.

For cross validation purposes the corpus was shuffled in ten iterations into train, development and test sets and the macro-average was taken. Table 2 describes how many tokens the datasets contain.

[3] https://www.statmt.org/moses/

[4] https://github.com/gerthjaanimae/csmt-parish-court-records

Table 2
Sizes of the datasets in tokens in the annotated corpus

Dialect	Records	Training set	Development set	Test set	Total
Eastern	3	880	58	236	1174
Central	79	17842	1189	4759	23790
Northeastern	2	375	25	100	500
Western	23	5063	337	1351	6751
Mulgi	21	5735	382	1530	7647
Coastal	7	1543	102	413	2058
Insular	40	9391	626	2505	12522
Tartu	26	6413	427	1711	8551
Võru	40	9269	617	2473	12359

4.2. Translation using the silver standard

In order to improve the quality of the translation and give the training algorithm more data to learn, artificial data, the so called silver standard was introduced.

As converting texts from contemporary Estonian to old spelling system is much simpler than the reverse process, the old parish court records that had the spelling closest to modern Estonian were transformed into older orthography. In order to determine the texts to be converted, they were morphologically analysed using Vabamorf tagger, which is a tool for extracting morphological information from a given word and to determine if a word belongs to modern Estonian or not. It is contained in ESTNLTK library [2]. The texts that got the highest percentage of words in accordance to modern Estonian (about 1100 texts) were transformed to old system using the automatic syllabifier from the ESTNLTK library and some hand-crafted rules. The main ones being: the single letter refering to the consonant in the first syllable was doubled if the vowel was short. For example koli > kolli ‘stuff’. The double letters refering to a long vowel in a first syllable were singled. For example kooli > koli ‘to school’.

As a result the train and development sets got significantly larger. The test set remained the same as described in the experiment above. After appending the silver standard to the portion for the train and dev-sets, 90% of it went for former and 10% for latter.

Afterwards the process was identical to the one described above.

Table 3 describes how many words the datasets contain within the silver standard corpus.

Table 3
Sizes of the datasets in words in the silver standard corpus

Dialect	Training set	Development set	Test set	Total
Eastern	2158	240	234	2632
Central	99065	11008	4758	114831
Northeastern	439	49	100	588
Western	19545	2172	1350	23067
Mulgi	5506	612	1529	7647
Coastal	1482	165	411	2058
Insular	9016	1002	2504	12522
Tartu	6156	685	1710	8551
Võru	8899	989	2471	12359

4.3. Translation using larger language model

The process was identical to the baseline experiment, except the contemporary Estonian part of the silver standard corpus was added to train the target language model.

For comparison the language model in the baseline translation contained about 57000 tokens and the larger language model about 164000 tokens.

4.4. Translation using bigrams

As the older spelling of Estonian could be ambiguous with one written form possibly corresponding to two different forms in the contemporary standard spelling (see section 2), the use of bigrams was tested to mitigate this problem.

As the word pairs containing punctuation were removed, the datasets became smaller.

Otherwise the experiment was identical to the baseline translation.

5. Results

5.1. Results of the text normalization

The following table describes the macro-average accuracies across 10 iterations on the test sets.

Table 4

Results of the character level statistical machine translation experiments on the test sets

Dialect	Baseline translation	Silver standard	Large language model	Word bigrams
Central	88.05%	82.07%	89.68%	86.26%
Coastal	86.90%	89.59%	81.53%	83.39%
Eastern	88.39%	88.25%	88.43%	82.66%
Insular	86.39%	86.00%	87.34%	81.31%
Northeastern	72.4%	76.6%	80%	72.2%
Western	84.37%	81.2%	85.31%	81.32%
Mulgi	84.01%	85.88%	86.88%	81.59%
Tartu	78.94%	82.27%	83.89%	71.76%
Võru	83.02%	83.52%	86.41%	75.75%
Average accuracy for all dialects	83.61%	83.93%	85.5%	79.58%

As can be observed from Table 4, the best results were obtained by using the baseline translation model together with large target language model. The explanation could be that the larger language model helps the algorithm to learn the patterns existing in the target language. The scores were the lowest using word bigrams, which could have a simple reason that within the longer strings the n-grams create, the probability of mistakes increases significantly. Across the dialects, the scores were the highest when normalizing the texts written in central dialect. That can be easily explained by the fact that modern standard Estonian is largely based on that dialect [3].

The scores were the lowest when normalizing the texts written in the northeastern dialect due to the small amount of data for training. It also has to be reminded that Mulgi, Tartu and Võru dialects belong into South Estonian and the rest into North Estonian dialects.

5.2. Results of the morphological analysis

In order to measure the performance of the normalization on the bigger corpus, the "translated" texts were analyzed using Vabamorf tagger, which outputs the inflectional information for a given word and if it cannot be retrieved, we can deduce that it is not in accordance to the modern Estonian orthography [2].

The morphological analysis was performed first on the unnormalized texts and after that different translation methods were compared. The corpus consisted of around 25000 records.

Although it is a very rough and error-prone estimate, as some of the words can easily get incorrect analyses due to the fact that some of the old and dialectal word forms are homonymous with modern ones. For example, pesnud in standard modern Estonian means 'washed', in South-Estonian it means 'beaten'. Regardless of the issues it still gives a general overview of the performance of the method used on the larger data that has not been annotated.

Table 5
Scores of the morphological analysis

Dialect	Not normalized	Baseline translation	Silver standard	Large language model	Word bigrams
Central	73.03%	85.27%	83.79%	87.82%	81.76%
Coastal	78.00%	83.35%	83.09%	80.53%	81.03%
Eastern	63.22%	74.66%	74.76%	79.62%	71.67%
Insular	69.27%	82.43%	83.38%	86.73%	80.31%
Northeastern	76.73%	83.10%	85.64%	76.00%	84.50%
Western	72.41%	79.59%	80.41%	85.81%	78.84%
Mulgi	67.23%	83.40%	82.02%	85.94%	80.00%
Tartu	74.02%	77.55%	80.31%	84.40%	73.85%
Võru	57.71%	82.13%	82.95%	86.13%	73.35%
Whole corpus	69.81%	80.79%	81.06%	85.19%	77.22%

As is evident from Table 5, the scores were also the highest when using baseline translation together with large target language model and the lowest using word bigrams. The results are very likely the same as described in section 5.1. The results across the dialects were not so clear cut as in previous section, but the same tendencies also apply here, except the northeastern dialect, that ranked surprisingly high. The reasons for that could be that the texts were already relatively close to modern Estonian and due to the small amount of texts there is also lower amount of variation in vocabulary and thus also lower probability of mistakes.

6. Discussion

As it can be seen from the previous sections, the accuracy was the highest when performing the experiments using the larger language and baseline translation model and the lowest using word bigrams. The results remained almost the same when comparing the accuracies of baseline and silver standard experiments. The scores of the morphological analysis reflect similar results.

Although we expected much better results from the silver standard experiments, it is still too early to draw a definite conclusion and the silver standard might simply need further development and tuning. For example, the unstressed syllables are occasionally still incorrectly converted.

Also, the scores seem to be in accordance with the related work in character level machine translation. As can be seen from the Pettersson et al. experiments [7], the method performed better on English, Swedish and German (over 90% accuracy) and worse on Hungarian and Icelandic texts (around 80% and 70% accuracy respectively). One of the reasons was that both, the Hungarian and Icelandic texts, came from earlier time period compared to others, the other was most probably due to the fact

that Hungarian is morphologically very rich and Icelandic richer compared to English [7]. As the same can be said about Estonian language, the lower accuracy can be expected.

Also, as mentioned in section 5.2, the scores reflecting the amount of words being in accordance with modern Estonian spelling system are rough estimates. There are some examples were even a human, who usually has more knowledge about word meanings and context than the machine, can normalize a word in a wrong way, let alone an algorithm. For example, töisel päeval means 'on the second day' in South Estonia. However, it can be easy to mistakenly give it a meaning 'on the day when people were hard at work', which is the meaning of the phrase in contemporary Estonian.

Also, as the distribution of data across different dialects is uneven, it can contribute to the occasionally inconsistent results. It would be interesting to test the combinations of different dialects that have some features in common.

7. Conclusion

Character level statistical machine translation showed promising results in normalizing old Estonian texts written in the 19th century. However, there is still a lot of work to be done in order to improve the quality and mitigate various issues that cropped up during the process. Mainly the silver standard has yet to be improved. Also combining the machine translation with some hand-crafted rules is something that might improve the quality of the normalization. It would be also important to gather the statistics about words that are already in their contemporary form, but still get erroneously normalized.

8. Acknowledgements

The author wishes to thank his supervisors Kadri Muischnek, Siim Orasmaa and Külli Prillop for their help and support and the National Archives of Estonia for the cooperation. This work has been supported by the national programme "Estonian language and culture in the digital age" grant EKKD29.

9. References

[1] M. Piotrowski, Natural language processing for historical texts, Morgan & Claypool, 2012.
[2] S. Laur, S. Orasmaa, D. Särg, P. Tammo, EstNLTK 1.6: Remastered Estonian NLP Pipeline, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 7152−7160.
[3] M. Erelt, Estonian language, volume 1 of Linguistica Uralica Supplementary Series, Estonian Academy Publishers, 2007.
[4] R. Raag, Talurahvakeelest riigikeeleks, Atlex, Tartu, 2008.
[5] P. Kallio, The diversification of Proto-Finnic, in: volume 18 of Studia Fennica, Fibula, fabula, fact: The Viking Age in Finland, Suomalaisen Kirjallisuuden Seura, 2014, pp. 155–170.
[6] Y. Scherrer, T. Erjavec, Modernizing historical Slovene words with character-based SMT, in: 4th Biennial Workshop on Balto-Slavic Natural Language Processing, 2013.
[7] E. Pettersson, J. Tiedemann, B. Megyesi, An SMT approach to automatic annotation of historical text, in: Proceedings of the workshop on computational historical linguistics, 2013.
[8] G. Tang, F. Cap, E. Pettersson, J. Nivre, An Evaluation of Neural Machine Translation Models on Historical Spelling Normalization, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1320–1331.
[9] N. Korchagina, Normalizing medieval German texts: From rules to deep learning, in: Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Linköping University Electronic Press, 2017, pp. 12–17.

10. Appendix

Table 6

Detailed results of the morphological analysis

Dialect	Not normalized		Baseline translation		Silver standard	
	Number of words analyzed	Percentage	Number of words analyzed	Percentage	Number of words analyzed	Percentage
Central	781533 / 1070212	73.03%	920342 / 1079288	85.27%	903427 / 1078193	83.79%
Coastal	19924 / 25542	78.00%	21374 / 25643	83.35%	21312 / 25649	83.09%
Eastern	335425 / 530571	63.22%	397637 / 532616	74.66%	397211 / 531299	74.76%
Insular	127510 / 184077	69.27%	153608 / 186354	82.43%	155333 / 186293	83.38%
Northeastern	709 / 924	76.73%	772 / 929	83.10%	793 / 926	85.64%
Western	221394 / 305732	72.41%	245921 / 308990	79.59%	248385 / 308882	80.41%
Mulgi	51587 / 76737	67.23%	64264 / 77052	83.40%	63210 / 77071	82.02%
Tartu	488030 / 659357	74.02%	516597 / 666120	77.55%	534237 / 665201	80.31%
Võru	163521 / 283357	57.71%	234506 / 285522	82.13%	236412 / 285013	82.95%
Whole corpus	2189633 / 3136509	69.81%	2555021 / 3162514	80.79%	2560320 / 3158527	81.06%

Dialect	Large language model		Word bigrams	
	Number of words analyzed	Percentage	Number of words analyzed	Percentage
Central	941064 / 1071540	87.82%	879963 / 1076215	81.76%
Coastal	20655 / 25649	80.53%	20783 / 25649	81.03%
Eastern	423828 / 532306	79.62%	381669 / 532566	71.67%
Insular	161607 / 186324	86.73%	149390 / 186012	80.31%
Northeastern	703 / 925	76.00%	785 / 929	84.50%
Western	262656 / 306104	85.81%	243366 / 308677	78.84%
Mulgi	66187 / 77016	85.94%	61425 / 76780	80.00%
Tartu	560173 / 663746	84.40%	482008 / 652671	73.85%
Võru	245785 / 285349	86.13%	211507 / 288358	73.35%
Whole corpus	2682658 / 3148959	85.19%	2430896 / 3147857	77.22%