Automatic conversion of colloquial Finnish to standard Finnish

Inari Listenmaa
Chalmers Institute of Technology
Sweden
inari@chalmers.se

Francis M. Tyers
HSL-fakultehta,
UiT Norgga árktalaš universitehta,
N-9015 Norway
francis.tyers@uit.no

Abstract

This paper presents a rule-based method for converting between colloquial Finnish and standard Finnish. The method relies upon a small number of orthographical rules combined with a large language model of standard Finnish for ranking the possible conversions. Aside from this contribution, the paper also presents an evaluation corpus consisting of aligned sentences in colloquial Finnish, orthographically-standardised colloquial Finnish and standard Finnish. The method we present outperforms the baseline of simply treating colloquial Finnish as standard Finnish, but is outperformed by a phrase-based MT system trained by the evaluation corpus. The paper also presents preliminary results which show promise for using normalisation in the machine translation task.

1 Introduction

Most language technology tools are designed or trained based on standard language forms, where they exist. The application of these tools to non-standard language can cause a substantial decrease in quality for example in machine translation, parsing and part-of-speech tagging (Eisenstein, 2013). Non-standard language can have different orthographic conventions, along with different morphology, syntax and stylistics.

For language-technology researchers working on non-standard forms of language, there are two clear options: either create new tools to process non-standard text, or create tools to preprocess non-standard text, standardising it to be subsequently processed by existing tools.

This paper evaluates a number of methods for converting colloquial Finnish to standard Finnish and describes a parallel corpus for evaluation.

2 Related work

There are a number of areas of research related to the task of text normalisation. Text proofing tools, such as spelling and grammar checkers (Kukich, 1992) can be used to encourage adherence to particular orthographic or grammatical norms. Accent and diacritic restoration — for example in Scannell (2011) — is similar in that it aims to bring text closer to standard orthography in order to facilitate treatment by automatic tools. Another related area is machine translation between different written norms of the same language, for example between Norwegian Bokmål and Norwegian Nynorsk (Unhammer and Trosterud, 2009).

Scannell (2014) presents a method for normalising pre-standardised text in Irish to the modern standard. The method relies on a translation model consisting of word-to-word correspondences in addition to spelling rules. Each word-to-word mapping has the same conditional probability and a penalty is assigned to each spelling rule application. Decoding works by processing the source sentence word-for-word left-to-right, keeping track of the possible 'hypothesis' translations and their probabilities, and when the end of sentence is reached, the most probable is output.

2.1 Colloquial Finnish

Viinikka and Voutilainen (2013) describe the common meaning of the terms colloquial (*puhekieli*) and standard (*yleiskieli* or *kirjakieli*) Finnish: standard language is unified in morphology and vocabulary, following the regulations of a language board; colloquial language shows local and idiolectal variation, and has structures that are characteristic to spoken variety, such as discourse particles and incomplete clauses.

We illustrate the differences with the following example from our data set. Sentence 1 is the original colloquial version. The gloss shows the ac-

Colloquial	Normalised	Standardised
tai emmä tiiä olikse erikseen joku nuorisoalennus	tai en#minä tiedä oliko#se erikseen jokin nuorisoalennus	tai en minä tiedä oliko erikseen nuorisoalennus
toistaseks tullu kaks kysymystä	toistaiseksi tullut kaksi kysymystä	toistaiseksi on tullut kaksi kysymystä
ja sit 2009 just ennenku menin Japaniin	ja sitten 2009 juuri ennen#kuin menin Japaniin	ja sitten 2009 juuri ennen kuin menin Japaniin

Table 1: Example sentences from the parallel corpus. The # mark represents a missing word boundary.

tual word-by-word translation, and the translation shows similar style and register in English.

(1) *seiskakin* *oli* *vaan* *silleen* *et*
seven-ALSO was just like.that that
fonotaksista *päättelin*
phonotactics.ELA I.deduced

'also the seventh, it was like, I just deduced it from phonotactics'

For the normalised version,[1] we changed only morphology and vocabulary. On the lexical level, the word *seiska* 'number 7' is colloquial style, and in the standard translation it is replaced by the ordinal *seitsemäs* 'seventh'. Other changes in the normalised version target common morphological or phonological phenomena, such as restoring the reduced diphthong in *vaan* → *vain*. The original sentence and the normalised translation are shown below, aligned word by word.

(2) *seiskakin* *oli* *vaan* *silleen* *et*
seitsemäskin oli vain sillä#lailla että
fonotaksista *päättelin*
fonotaksista päättelin

The syntactic structure of the original sentence is markedly spoken; the word *seiska* is topicalised, and the main information "deduced from phonotactics" is in a subordinate clause. The translation into standard Finnish is shorter and more precise, leaving just the main information.

(3) *päättelin* *seitsemännenkin*
I.deduced seventh-ALSO
fonotaksista
phonotactics.ELA

'I deduced the seventh also from phonotactics'

[1]The normalised version is converted orthographically and lexically, but not syntactically or stylistically.

Section	Tokens		
	dev	test	train
Colloquial	1,003	1,012	5,103
Normalised	1,003	1,012	5,103
Standardised	1,000	991	4,982

Table 2: Statistics on sentences from the parallel corpus.

3 Corpus

Our evaluation corpus was created by manually translating texts in colloquial Finnish to standard Finnish. The corpus is freely available and published under the CreativeCommons CC-BY-SA 3.0 licence[2]. The texts were extracts from internet relay chat (IRC) conversations. We performed the conversion process in two steps, the first step involved simple orthographic normalisation, for example *oon* → *olen* 'I am'. Syntactic and stylistic conversions were not applied at this stage. The second conversion step normalised the text both orthographically and syntactically/stylistically. Table 1 presents an excerpt from each of the three parts of the corpus.

The corpus was split into three parts, development, testing and training. The development and testing portions contain approximately 1,000 words each, with the remaining approximately 5,000 words for training phrase-based and character-based models.[3] Table 2 gives statistics on the number of words in each section.

[2]https://svn.code.sf.net/p/apertium/svn/languages/apertium-fin/texts/normalisation/

[3]The corpus is split into 14 files of 500 words each. Files 01–02 were used for development; 03–04 for testing and 05–14 for training.

Input:	Mä oon Tomminkaa ‘I am with Tommi’.		
Step 1	Mä oon Tomminkaa Minä oon Tomminkaa		apply rule 1: mä → minä
Step 2	Mä oon Tomminkaa Minä oon Tomminkaa Mä olen Tomminkaa Minä olen Tomminkaa		apply rule 2: oon → olen
Step 3	Mä oon Tomminkaa Minä oon Tomminkaa Mä olen Tomminkaa Minä olen Tomminkaa Mä oon Tommin kanssa Minä oon Tommin kanssa Mä olen Tommin kanssa Minä olen Tommin kanssa		apply rule 3: (?+)nkaa → \1n kanssa
Step 4	Minä olen Tommin kanssa Minä oon Tommin kanssa Mä olen Tommin kanssa Mä oon Tommin kanssa Minä olen Tomminkaa Minä oon Tomminkaa Mä olen Tomminkaa Mä oon Tomminkaa	-4.5811 -7.8174 -8.0941 -8.8651 -9.2045 -12.4408 -12.7176 -13.4885	rank candidates
Output:	Minä olen Tommin kanssa		

Table 3: Example trace of the normalisation method. Rules are applied in order to each of the possible candidate translations in turn. The candidates are then ranked using an *n*-gram language model of standard Finnish and either an *n*-best list or the best candidate is output.

4 Experiments

4.1 Rule-based normalisation

For the rule-based normalisation we applied a set of regular-expression based replace rules to the input text to produce all the possible candidate sentences in standard Finnish and then used a target-language model to rank the possible candidates. The candidate with the highest rank was selected as the normalised sentence. For the target-language model we used the Finnish side of the English–Finnish EuroParl parallel corpus (Koehn, 2005).

We developed two sets of rules:

- `rules-1`: 273 rules from Karlsson (2008)’s grammar of Finnish (§95–97). The rules took around one hour to implement.

- `rules-2`: 98 rules written by examining the development corpus, these rules also took approximately one hour to implement.

The rules included both simple one-to-one (‘mä’ → ‘minä’) and one-to-many (‘emmä’ → ‘en minä’) word correspondences, and also regular expression substitutions which could match a prefix or a suffix (‘(?+)nkaa’ → ‘\1n kanssa’).

Table 3 gives an example trace of the system on a simple sentence using three replace rules.

4.2 Statistical machine translation

The statistical-machine translation approaches were implemented using the Moses toolkit (Koehn et al., 2007). The training set up was that used for the baseline system in the WMT shared tasks on machine translation.[4]

The target-language model corpus, trained using KenLM (Heafield, 2011), used was the same as in the rule-based experiments.

We trained models based on two approaches, the first being phrase-based machine translation (PBMT, Zens et al. (2002)) and the second on character-based machine translation

[4] `http://www.statmt.org/wmt11/baseline.html`

(CBMT, Nakov and Tiedemann (2012); Tiedemann (2009)).

For both approaches we trained two systems, the first used the *normalised* part of the corpus as the target language; the second used the *standardised* part of the corpus as the target language.

The idea behind this was that the *normalised* part of the corpus would be closer to the original colloquial text than the *standardised* part, making it easier to learn the alignment model.

Character-based

Nakov and Tiedemann (2012) present a method of statistical machine translation on the character level between related languages that takes advantage of phrase-based machine translation architecture. The method relies on preprocessing the input and output by inserting spaces in between the characters of words, for example the string 'mä meen Helsinkiin' would become 'm ä $ m e e n $ H e l s i n k i i n' with a unigram model, or 'mä ä$ $m me ee en n$ $H He el ls si in nk ki ii in' with a bigram model.

After preprocessing, the corpora are processed as with the phrase-based system, with the difference that the language model order is increased from 5 to 10-grams.

4.3 End-to-end translation

In order to evaluate how well the different normalisation strategies worked in combination with another language technology tool, we performed an end-to-end experiment involving machine translation. To evaluate this, we took the colloquial portion of the test corpus and manually translated it to English. For each of the best-performing systems we first passed the colloquial text through, and then translated the output to English using a widely-used online machine translation engine with Finnish to English. We compared the output to translating the text to English without the standardiser.

5 Results

Tables 4 and 5 present the results of the experiments.

The baseline was made by calculating the metric scores between the standardised 'reference' and the colloquial input. The results show that all conversion methods outperform the baseline.

Our rule-based method performs similarly to the character-based machine translation systems.

System	PER	WER	BLEU
Baseline	46.12	48.04	26.31
rules-1	38.27	41.19	32.65
rules-2	**38.17**	**35.25**	**36.41**
rules-c	36.56	39.68	34.68
CBMT-cn	43.09	48.34	33.55
CBMT-cs	46.22	52.27	29.21
PBMT-cn	**28.05**	**35.42**	**48.37**
PBMT-cs	27.95	36.13	46.76

Table 4: Results for the normalisation task. The system `rules-c` is the combination of the rules in `rules-1` and `rules-2`. The figures in bold are the best results for rule-based and SMT methods.

System	PER	WER	BLEU
Colloquial	41.02	69.57	12.11
Normalised	30.73	59.73	22.56
Standardised	33.88	61.61	19.49
rules-2	37.94	69.35	12.92
CBMT-cn	65.59	86.25	13.06
PBMT-cn	**35.69**	**65.14**	**17.75**

Table 5: Results for the Finnish to English translation task. The first three rows are the results from translating the sections of the corpus.

Both the character-based systems and the rule-based system achieve around half of the performance of the phrase-based system.

Out of the rule-based systems, the set of rules which was created by examining the development corpus outperforms both the set of rules from the grammar and the combined rules. The rules from the grammar capture more general tendencies, whereas the rules from the development corpus are more lexicalised. Since the testing corpus is small and only contains text from a single author, the higher performance of the second rule set could also be an due to overfitting.

It is interesting to note that MT systems trained on the normalised section of the corpus outperform those trained on the standardised corpus. One explanation for this could be that the corpus size is small, so that the word alignments are not as reliable on the standardised corpus which is by nature not a word-for-word conversion.

The systems for normalisation are able to improve out-of-vocabulary rates in many cases, most likely as the online statistical system that we used is trained on more formal texts. Frequent contrac-

tions such as *onks?* 'is there?', *oo* 'be.CONNEG'[5] and *vaa* 'only' are found untranslated in the output, but are easily converted by the systems.

6 Future work

Although a reasonable size for a test corpus, the corpus is still too small for wide-coverage experiments. We intend to expand the corpus size to at least 10,000 words. Another weakness of the corpus is that it contains text from a single author. We would ideally like to add texts from other authors and other colloquial genres—the challenge here is finding text that is both free of privacy issues and available to release under a free/open-source sence.

As for the methods, we would like to follow Scannell (2014) in incorporating 'translation' probabilities into our rule-based normalisation model. Our current model relies exclusively on the target-language model probability, however some rules may be more reliable or probable than others.

7 Conclusions

We have presented a parallel corpus of colloquial Finnish and standard Finnish – to our knowledge the first of its kind – and an evaluation of methods for converting colloquial Finnish to standard Finnish.

We have shown that converting from colloquial Finnish to standard Finnish substantially helps with the Finnish to English machine translation task.

Acknowledgments

We thank Joonas Kylmälä for translating the colloquial sentences into normalised and standardised versions.

References

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

[5]The word *oo* is the negative form of the verb *olla* 'to be' in Finnish.

Fred Karlsson. 2008. *Finnish: An Essential Grammar*. Routledge, Abingdon, Oxon.

Philipp Koehn, Hieu Hoang, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Demonstration session at the Annual Meeting of the Association for Computational Linguistics (ACL2007)*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24(4):377–439, December.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 301–305.

Kevin Scannell. 2011. Statistical unicodification of african languages. *Language Resources and Evaluation*, 45(3):375–386.

Kevin Scannell. 2014. Statistical models for text normalization and machine translation. In *Proceedings of the Celtic Language Technology Workshop at COLING 2014*.

Jörg Tiedemann. 2009. Character-based PSMT for Closely Related Languages. In *Proceedings of 13th Annual Conference of the European Association for Machine Translation (EAMT09)*, pages 12–19.

Kevin Unhammer and Trond Trosterud. 2009. Reuse of free resources in machine translation between nynorsk and bokml. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 35–42.

Jenni Viinikka and Eero Voutilainen. 2013. Ääniä ilmassa, merkkejä paperilla – puhutun ja kirjoitetun kielen suhteesta. *Kielikello*.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *KI - 2002: Advances in Artificial Intelligence. 25. Annual German Conference on AI, KI 2002*, volume 2479, pages 18–32. Springer Verlag.