Comparing Rule-based and SMT-based Spelling Normalisation for English Historical Texts

Gerold Schneider
Institute of Computational Linguistics and Department of English
University of Zurich
gschneid@ifi.uzh.ch

Eva Pettersson
Department of Linguistics and Philology
Uppsala University
eva.pettersson@lingfil.uu.se

Michael Percillier
Department of English
University of Mannheim
percillier@uni-mannheim.de

Abstract

To be able to use existing natural language processing tools for analysing historical text, an important preprocessing step is spelling normalisation, converting the original spelling to present-day spelling, before applying tools such as taggers and parsers. In this paper, we compare a probablistic, language-independent approach to spelling normalisation based on statistical machine translation (SMT) techniques, to a rule-based system combining dictionary lookup with rules and non-probabilistic weights. The rule-based system reaches the best accuracy, up to 94% precision at 74% recall, while the SMT system improves each tested period.

1 Introduction

Language technology for historical texts poses several challenges, as earlier stages of languages are under-resourced. But language technology is helpful both to researchers in Digital Humanities and Diachronic Linguistics. Natural Language Processing (NLP) tools are needed at all levels of processing, but spelling is a particularly obvious candidate, for at least two reasons. First, historical variants not only differ from present-day spellings. They also often lack normalisation within their period – the same word often appears with several different spellings inside the same document. Thus, even simple lexicon-based research is hampered by complex corpus queries and low recall. Second, spelling variants can affect all other subsequent processing levels – tokenisation, part-of-speech tagging and parsing. For example, frequent variants like *call'd* for *called* lead to a tokenisation error, which in turn results in wrong tagging (*call_NN d_MD*), and as a consequence parsing quality is also affected. Rayson et al. (2007), Scheible et al. (2011) and Schneider et al. (2014) report that about half of the changes induced by automatic spelling normalisation lead to improved tagging and parsing, which makes it a vital contributor to improved tagging and parsing of historical texts.

Several approaches for mapping historical variants to present-day standard spelling have been proposed. For English, on which we are going to focus in this article, VARiant Detector 2 (VARD) (Baron and Rayson, 2008) is a popular spelling normalisation tool, but there are other possible approaches. Pettersson et al. (2014) compared three statistical approaches: 1) a filtering approach, 2) a Levenshtein-distance approach, and 3) a character-based statistical machine translation (SMT) approach. These approaches were applied to five languages, and for four of these (including English), the SMT-based approach yielded the best results.

In this paper, we compare the results of applying the SMT-based spelling normalisation approach to the ARCHER corpus of historical English and American texts, to the results achieved for VARD2 on the same corpus. The comparison is interesting as the approaches are significantly different: SMT is a probablistic, language-independent approach, whereas VARD2 combines lexicon-lookup with rules and non-probabilistic weights.

2 Data and Methods

2.1 The ARCHER Corpus

As corpus of application, we use ARCHER (Biber et al., 1994), a historical corpus sampled from British and American texts from 1600-1999 and across several registers. Its current version (V 3.2) contains 3.2 million words. Since there are increasingly fewer non-standard spelling variants in later texts, we have only used texts until 1850.

The increasing scarcity of non-standard spelling also gives rise to a new research question: from which point on does spelling normalisation introduce more errors than correcting the few remaining non-standard spelling variants?

For the training phase, we have manually annotated 109 documents (about 200,000 words), stratified by 4 periods (1650-99, 1700-49, 1750-99, 1800-49), with a total of 6,975 manual normalisations. For evaluation, we have manually annotated a further 30 documents, containing 1,467 normalisations. The ARCHER corpus has been carefully sampled and aims to be genre-balanced, which provides us with a realistic real-world scenario.

A first observation that we have made is that while the amount of non-standard spelling decreases (from a mean of 315 per document in the period 1600-1649 to 24 in the period from 1800-49), the variance is very large (the standard deviation in the period 1600-1649 is 266, in the period from 1800-49 it is 52), indicating that individual styles vary considerably.

2.2 SMT

In the SMT-based approach, spelling normalisation is treated as a translation task, which could be solved using statistical machine translation (SMT) techniques. To address changes in spelling rather than the full translation of words and phrases, the translation system is trained on sequences of characters instead of word sequences.

In our experiments, we use the same settings for SMT-based spelling normalisation as presented in Pettersson et al. (2013), that is a phrase-based translation model, using Moses with all its standard components (Koehn et al., 2007), and IRSTLM for language modelling (Federico et al., 2008). For aligning the characters in the historical part of the corpus to the corresponding characters in the modern version of the corpus, the word alignment toolkit GIZA++ (Och and Ney, 2003) is applied, implementing the IBM models commonly used in SMT (Brown, 1993). The same default settings as for standard machine translation are used, with the following exceptions:

1. The system is trained on sequences of characters instead of word sequences.
2. Reordering is switched off during training, since it is unlikely that the characters are to be reordered across the whole word.
3. The maximum size of a phrase (sequence of characters) is set to 10, a setting previously shown to be successful for character-based machine translation between closely related languages (Tiedemann , 2009).

2.3 VARD2

The automatic normalisation tool VARD2 (Baron and Rayson, 2008) is a rule-based system, which can be customized to learn more rules from annotated corpora and adapt weights to them. The first version of VARD was a pure dictionary-based system. VARD2 extends this approach as follows.

First, every word that is not found in the tool's present-day English (PDE) spelling lexicon is marked as a candidate. Second, PDE variants for candidates are found and ranked, according to the following three methods:

1. the original VARD replacement dictionary
2. a variant of SoundEx, which maps phonetically similar words onto each other
3. letter replacement rules, which represent commmon patterns of spelling variation, for example interchanging *v* and *u* or dropping word-final *e*.

These rules are given a non-probabilistic confidence score, and each replacement candidate is also weighted by edit distance. When further annotated corpora are added, the replacement dictionary is extended and the weights of the three methods are optimised.

As VARD2 is a rule-based and non-probabilistic system, the question arises how it performs in comparison to state-of-the-art statistical approaches. It has been shown, for example in the domain of part-of-speech tagging (Samuelsson and Voutilainen, 1997; Loftsson, 2008), that carefully written rule-based systems can perform at the same level or better than statistical systems.

3 Results

3.1 Annotation, Inter-Annotator Agreement

For evaluating the SMT method, we used the manual annotation of ARCHER (split into 90% training and 10% evaluation) as the first evaluation method. For the evaluation of VARD2, and for comparing VARD2 to SMT, we used the manually annotated 30 documents described in Section 2.1.

When annotating the evaluation set, we noticed that while in most cases normalisation is

clear, there are several reasons why inter-annotator agreement is considerably lower than 100%. Four important reasons are: first, there are cases where it is unclear if a variant is PDE or not. A good example is *thou hast* where VARD2 by default changes *hast* to *have*, although this is, in the opinion of one annotator, rather a change of morphological inflection than of spelling. Second, if dictionaries list alternative readings (e.g. British and American), should one normalise? Third, it is unclear how strict to be with hyphenation: should *sun-shine* or *bridle-way* be corrected? Fourth, particularly in the recent texts, where only every 100th or 200th word has a non-standard spelling, it is very easy to overlook variants.

A subset of our evaluation corpus, comprising 7 documents, was annotated by two of the authors. On the possible 529 normalisations, they agreed on 439, which corresponds to an inter-annotator agreement of 83%. We corrected obvious oversights and otherwise took the annotations of the author who had annotated the training set.

3.2 SMT

For the SMT-based experiments, we need to train a *translation model* and a *language model*. For the translation model, we use pairs of historical word forms mapped to their corresponding normalised spelling, to calculate the likelihood that certain sequences in the target language (i.e. the modern spelling) are translations of the sequences in the source language (i.e. the historical word forms). Such word pairs were extracted from the training part of the ARCHER corpus (as described in Section 2.1) and split into a pure training part and a tuning part (as required by the Moses system) by extracting every 10th word form to the tuning part, and the rest of the word forms to the training part. For language modeling, a monolingual target language corpus is used for modeling the probabilities that any candidate translation string would occur in the target language. For this purpose, we use the British National Corpus (BNC) of approximately 100 million words sampled to represent a wide cross-section of British English from the late 20th century (BNC, 2007). We filter hapax legomena, i.e. take all word forms that appear at least twice in the BNC. In addition, the manually normalised part of the training corpus is added to the language model, to include archaic word forms that are unlikely to occur in the BNC corpus.

Historical texts are marked by a high degree of spelling variance and spelling inconsistencies, leading to data sparseness when applying different kinds of NLP tools to the data. It is therefore interesting to explore whether adding historical data in general could improve normalisation accuracy, or if the data need to be representative of the specific time period targeted. We therefore split both the training and the evaluation parts of the ARCHER corpus into three subcorpora, containing texts from the 17th, 18th, and 19th century respectively. This way, we can evaluate normalisation accuracy for each subcorpus, when trained on data from all three centuries, and when trained on data from the specific time period only.

For the SMT-based approach, we then ran experiments by 1) training on the full corpus of manually normalised historical text, 2) training on the correct century only (17th, 18th or 19th), and 3) adding dictionaries in two ways:

(a) Historical word forms that are found in the manually normalised part of the training corpus are left unchanged.

(b) A normalisation candidate suggested by the SMT system is only accepted if it occurs in the BNC corpus.

Full Test Corpus	
Unnormalised	97.21
Training corpus	98.00
Training corpus + Dict (a)	98.14
Training corpus + Dict (b)	98.01
Training corpus + Dict (a) & (b)	98.14
17th Century Part of the Test Corpus	
Unnormalised	93.88
Full training corpus	96.60
17th century part of the training corpus	96.89
18th Century Part of the Test Corpus	
Unnormalised	98.65
Full training corpus	98.75
18th century part of the training corpus	98.69
19th Century Part of the Test Corpus	
Unnormalised	98.95
Full training corpus	99.10
19th century part of the training corpus	99.15

Table 1: Normalisation accuracy, per word, for different parts of the corpus. dict = adding dictionaries for lexical filtering.

As shown in Table 1, normalisation accuracy improves for all parts of the corpus, using the SMT-based approach to spelling normalisation. There are 421 cases where the SMT-based system has modified the original spelling to a spelling identical to the manually defined gold standard spelling, e.g.:

happinesse → *happiness*
onely → *only*
relligious → *religious*
iustices → *justices*
loue → *love*

In contrast, there are 44 cases where the SMT-based system has suggested a modification that is different from the gold standard spelling, i.e. precision errors. In most of these cases, the normalisation system has failed, but there are also instances that seem to be due to mistakes in the manually defined gold standard.

In 762 cases, the normalisation system has left the original word form unchanged, even though the manually defined gold standard suggest a normalisation, i.e.we have a recall error. A manual error analysis shows that one of the major cause of recall errors involves apostrophes, e.g.:

mans ↛ *man's*
o'er ↛ *over*
redeem'd ↛ *redeemed*
y'are ↛ *you're*

Other common causes of recall error are connected to endings like *-ie*, *-y*, *e* and *eth*, e.g.:

flie ↛ *fly*
easie ↛ *easy*
disdaine ↛ *disdain*
gipsey ↛ *gypsy*
captaine ↛ *captain*
seemeth ↛ *seems*

Furthermore, using the manually normalised part of the training data as a filter, leaving word forms that occur in this data set unnormalised, has a positive effect on normalisation accuracy. The main reason is the otherwise incorrect normalisation of frequently occurring function words, such as *thy* and *thee*. In the manual normalisation process, these word forms have been left as they are. The SMT-based system would however, without lexical filtering, normalise these word forms into *they* and *the* respectively, due to the strong preference for these word forms in the language model. Even though the manually normalised version of the ARCHER training data, including word forms such as *thy* and *thee*, have been added to the language model, these occurrences are outnumbered by the occurrences of the much more frequent English word forms *the* and *they* in the BNC part of the language model.

The second lexical filtering, where normalisation candidates suggested by the SMT system are only accepted if occurring in the BNC corpus, also leads to a small (non-significant) improvement of the normalisation accuracy. The results presented for the time-specific subcorpora are thus based on lexical filtering using both methods.

It is interesting to note that for both 17th century data and 19th century data, the best normalisation accuracy is achieved if a smaller data set containing time-specific data only is used, rather than adding training data from all three centuries.

3.3 VARD2 Performance on Evaluation Set

The results of applying VARD2 are given in Table 2, in terms of precison, recall, and per-word rates. The results using the default rules provided with the VARD2 distribution are in the second column, and using the training from the manually annotated 109 ARCHER documents (in addition to the default rules provided in the VARD2 distribution) in the third column, and best SMT in the fourth column.

Table 2 shows five points. First, VARD2 improves spelling (in the sense of mapping it to PDE variants) in most settings, except when applying the defaults settings to the latest period, 19th century texts.

Second, the training with ARCHER has considerably improved results.

Third, we have tested the effect of training on the entire ARCHER or only the appropriate century and show the results in the second last column. The effect of training VARD2 on different periods could be relatively small, as the default rules are not deleted, the new rules are just added and the the weights adapted. Using less training data leads to results with higher precision and lower recall.

Fourth, the task gets increasingly difficult in later periods, which is related to the fact that only very few tokens need normalisation, as we have already observed in the discussion of inter-annotator agreement. The performance in the 19th century is partly so low because there are only very few words that require correction, thus absurd cor-

	VARD Default	+ trained on ALL ARCHER	+ trained on ARCHER ct.	best SMT
Full Evaluation Corpus (N=838,W=29167)				
Precision	89.54	94.36	–	80.48
Recall	76.61	73.89	–	64.43
Unnorm. words	97.13	97.13	–	97.13
Correct words	99.07	99.11	–	98.53
17th Century Part of the Evaluation Corpus (N=507,W=9682)				
Precision	88.31	94.81	99.43	78.93
Recall	74.56	72.00	69.42	67.26
Unnorm. words	94.76	94.76	94.76	94.76
Correct words	98.15	98.33	98.38	97.34
18th Century Part of the Evaluation Corpus (N=92,W=11478)				
Precision	83.75	92.42	100.00	78.31
Recall	72.82	66.30	65.22	70.65
Unnorm. words	99.20	99.20	99.20	99.20
Correct words	99.67	99.69	99.72	99.61
19th Century Part of the Evaluation Corpus (N=61,W=9617)				
Precision	90.63	87.23	100.00	60.71
Recall	47.54	67.21	24.59	27.87
Unnorm. words	99.36	99.36	99.36	99.36
Correct words	99.64	99.73	99.52	99.42

Table 2: Normalisation accuracy of VARD, in percent, for the evaluation corpus, and split by century, comparing the VARD default rules, and the effect of training on 109 manually annotated ARCHER documents, and a comparison to SMT. N=number of manual changes, W=number of words

rections such as changing *idiotism* to *idiocy* affect precision. Recall is strongly affected by rare words and rare but correct variants, such as *silicious* which is not corrected to *siliceous*. It might be advisable to stop using historical spelling correction already at 1800 instead of 1850.

Fifth, the SMT system performs slightly below the highly costumized VARD tool. We elaborate on this point in the following section.

4 VARD2 and SMT in comparison

Among the items that VARD2 failed to detect, hyphenation stood out in particular (e.g. *sun-shine* which should be changed to *sunshine*). On the other hand, it overgeneralizes from 2nd person singular verb forms to plural forms (e.g. *hast* and *darest* are changed to *have* and *dare*). As these are frequent forms, they have a substantial numerical impact. VARD2 also overnormalises proper names (e.g. *ALONZO* to *ALONSO*), which often keep histoical spellings in PDE. The detection of proper names in historical texts is far from trivial, however, as also common nouns and verbs are often capitalised.

When inspecting the errors made by the SMT system, we have observed the following types of errors:

- Overgeneralisation, e.g.: *whether* has incorrectly been suggested to be normalised to *wheather*.
- Undergeneralisations, e.g.: *complements* is not normalised to *compliments*, because the word *complements* also exists, with a different meaning.
- Foreign words: for example, the Latin word *mater* is incorrectly normalised to *matter*
- Inter-annotator questions, e.g.: *hath* is normalised to *have*, *insomuch* to *inasmuch*, *emphatical* to *emphatic*
- Oversights, spurious errors: Some of the suggested normalisations are correct, even though classified as incorrect when compared to the gold standard.

5 Related Work

Apart from the SMT-based approach to spelling normalisation originally described in Pettersson et al. (2013), and applied to the ARCHER corpus in this study, character-based SMT-techniques have also been implemented by Scherrer and Erjavec (2013), for the task of normalising historical Slovene. They tried both a supervised and an unsupervised learning approach. In the supervised setting, the translation model was trained on a set of 45,810 historical-modern Slovene word pairs, whereas the language model was trained on the same data set but only including the modern word forms. In addition, a lexicon filter was used, in which normalisation candidates proposed by the translation model were only accepted if they were also found in the Modern Slovene Sloleks dictionary. In the unsupervised setting, the historical-to-modern training data was created inbased on separate lists of historical word forms and modern word forms, where the historical word forms were mapped to modern word forms based on string similarity comparisons between the word forms occurring in the two lists. Their evaluation showed an increase in normalisation accuracy from 15.4% to 48.9% for 18th century test data using the unsupervised setting. In the supervised setting, accuracy improved further to 72.4%.

6 Conclusions and Outlook

We have compared a probablistic, language-independent approach to spelling normalisation based on SMT, to a carefully crafted and highly adapted rule-based system. The latter has slightly higher performance (up to 94% precision at 74% recall) while the former is more general and fully language-independent. We have tested various settings, and shown that training with smaller century-specific data sets performs better, and that statistical SMT can be improved in several ways, e.g. by constraining the dictionary to forms seen in present-day spelling.

As future work, we would like to assess the results of succeeding NLP tasks, such as tagging and parsing, based on normalised data. We will also try to improve normalisation results further by combining the two approaches in various ways. One way would be to add automatically normalised word forms using VARD to the training data for the SMT-based system. This would be considered a semi-supervised method, in which both manually revised and automatically annotated data are used for training the SMT-based system. Another way of combining the two systems would be to use the normalisations suggested by the SMT-based system to guide the VARD system in the ranking process, in cases where several normalisation candidates are given in VARD.

Many of the remaining errors are hard to correct with purely word-based approaches. We would like to investigate if using limited context can improve results.

References

Alistair Baron and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Birmingham. Aston University.

Douglas Biber, Edward Finegan, and Dwight Atkinson. 1994. Archer and its challenges: Compiling and exploring a representative corpus of historical english registers. In Udo Fries, Peter Schneider, and Gunnel Tottie, editors, *Creating and using English language corpora, Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993*, pages 1–13. Rodopi, Amsterdam.

BNC Consortium. 2007. The British National Corpus, Version 3. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. `http://www.natcorp.ox.ac.uk/`.

Peter Brown, Vincent Della Pietra, Stephen Della Pietra, and Robert Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2), pages 263–311.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. in *Proceedings of Interspeech 2008*, pages 1618–1621.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst, 2007. Moses: Open Source Toolkit for Statistical Machine Translation. in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Hrafn Loftsson. 2008. Tagging icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1).

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment

Models. *Computational Linguistics*, 1(29), pages 19–51.

Eva Pettersson, Beáta Megyesi, and Jörg Tiedemann. 2013. An SMT approach to automatic annotation of historical text. In *Proceedings of the NoDaLiDa 2013 workshop on Computational Historical Linguistics*.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2014. A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014*, pages 32–41, Gothenburg, Sweden.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics 2007*. University of Birmingham, UK.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. In *Proceedings of the ACL-HLT 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*, Portland, Oregon.

Christer Samuelsson and Atro Voutilainen. 1997. Comparing a linguistic and a stochastic tagger. In *Proceedings of of ACL/EACL Joint Conference*, Madrid.

Yves Scherrer and Tomaž Erjavec. 2013. Modernizing historical Slovene words with character-based SMT. In *Proceedings of the 4th Biennial Workshop on Balto-Slavic Natural Language Processing*, pages 58–62.

Gerold Schneider, Hans Martin Lehmann, and Peter Schneider. 2014. Parsing Early Modern English corpora. *Literary and Linguistic Computing*, first published online February 6, 2014 `doi:10.1093/llc/fqu001`.

Jörg Tiedemann. 2009. Character-based PSMT for closely related languages. *Proceedings of 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, pages 12–19.