Modernizing historical Slovene words with character-based SMT

Yves Scherrer, Tomaž Erjavec

Modernizing historical Slovene words with character-based SMT

Yves Scherrer
ALPAGE
Université Paris 7 Diderot & INRIA
5 Rue Thomas Mann, Paris, France
yves.scherrer@inria.fr

Tomaž Erjavec
Dept. of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

We propose a language-independent word normalization method exemplified on modernizing historical Slovene words. Our method relies on character-based statistical machine translation and uses only shallow knowledge. We present the relevant lexicons and two experiments. In one, we use a lexicon of historical word–contemporary word pairs and a list of contemporary words; in the other, we only use a list of historical words and one of contemporary ones. We show that both methods produce significantly better results than the baseline.

1 Introduction

A lot of recent work deals with detecting and matching cognate words in corpora of closely related language varieties. This approach is also useful for processing historical language (Piotrowski, 2012), where historical word forms are matched against contemporary forms, thus normalizing the varied and changing spelling of words over time. Such normalization has a number of applications: it enables better full-text search in cultural heritage digital libraries, makes old texts more understandable to today's readers and significantly improves further text processing by allowing PoS tagging, lemmatization and parsing models trained on contemporary language to be used on historical texts.

In this paper, we try to match word pairs of different historical stages of the Slovene language. In one experiment we use character-based machine translation to learn the character correspondences from pairs of words. In the second experiment, we start by extracting noisy word pairs from monolingual[1] lexicons; this experiment simulates a situation where bilingual data is not available.

The rest of this paper is structured as follows: Section 2 presents related work, Section 3 details the dataset used, Section 4 shows the experiments and results, and Section 5 concludes.

2 Related work

The most common approach to modernizing historical words uses (semi-) hand-constructed transcription rules, which are then applied to historical words, and the results filtered against a contemporary lexicon (Baron and Rayson, 2008; Scheible et al., 2010; Scheible et al., 2011); such rules are often encoded and used as (extended) finite state automata (Reffle, 2011). An alternative to such deductive approaches is the automatic induction of mappings. For example, Kestemont et al. (2010) use machine learning to convert 12th century Middle Dutch word forms to contemporary lemmas.

Word modernization can be viewed as a special case of transforming cognate words from one language to a closely related one. This task has traditionally been performed with stochastic transducers or HMMs trained on a set of cognate word pairs (Mann and Yarowsky, 2001). More recently, character-based statistical machine translation (C-SMT) (Vilar et al., 2007; Tiedemann, 2009) has been proposed as an alternative approach to translating words between closely related languages and has been shown to outperform stochastic transducers on the task of name transliteration (Tiedemann and Nabende, 2009).

For the related task of matching cognate pairs in bilingual non-parallel corpora, various language-independent similarity measures have been proposed on the basis of string edit distance (Kondrak and Dorr, 2004). Cognate word matching has been shown to facilitate the extraction of translation lexicons from comparable corpora (Koehn and Knight, 2002; Kondrak et al., 2003; Fišer and Ljubešić, 2011).

[1] For lack of a better term, we use "monolingual" to refer to a single diachronic state of the language, and "bilingual" to refer to two diachronic states of the language.

For using SMT for modernizing historical words, the only work so far is, to the best of our knowledge, Sánchez-Martínez et al. (2013).

3 The dataset

In this section we detail the dataset that was used in the subsequent experiments, which consists of a frequency lexicon of contemporary Slovene and training and testing lexicons of historical Slovene.[2]

3.1 The lexicon of contemporary Slovene

Sloleks is a large inflectional lexicon of contemporary Slovene.[3] The lexicon contains lemmas with their full inflectional paradigms and with the word forms annotated with frequency of occurrence in a large reference corpus of Slovene. For the purposes of this experiment, we extracted from Sloleks the list of its lower-cased word forms (930,000) together with their frequency.

3.2 Corpora of historical Slovene

The lexicons used in the experiments are constructed from two corpora of historical Slovene.[4] The texts in the corpora are, *inter alia* marked up with the year of publication and their IANA language subtag (`sl` for contemporary Slovene alphabet and `sl-bohoric` for the old, pre-1850 Bohorič alphabet). The word tokens are annotated with the attributes *nform*, *mform*, *lemma*, *tag*, *gloss*, where only the first two are used in the presented experiments.

The *nform* attribute contains the result of a simple normalization step, consisting of lower-casing, removal of vowel diacritics (which are not used in contemporary Slovene), and conversion of the Bohorič alphabet to the contemporary one. Thus, we do not rely on the C-SMT model presented below to perform these pervasive, yet deterministic and fairly trivial transformations.

The modernized form of the word, *mform* is the word as it is (or would be, for extinct words) written today: the task of the experiments is to predict the correct *mform* given an *nform*.

[2]The dataset used in this paper is available under the CC-BY-NC-SA license from `http://nl.ijs.si/imp/experiments/bsnlp-2013/`.

[3]Sloleks is encoded in LMF and available under the CC-BY-NC-SA license from `http://www.slovenscina.eu/`.

[4]The data for historical Slovene comes from the IMP resources, see `http://nl.ijs.si/imp/`.

Period	Texts	Words	Verified
18B	8	21,129	21,129
19A	9	83,270	83,270
19B	59	146,100	146,100
Σ	75	250,499	250,499

Table 1: Size of goo300k corpus.

Period	Texts	Words	Verified
18B	11	139,649	15,466
19A	13	457,291	17,616
19B	270	2,273,959	65,769
Σ	293	2,870,899	98,851

Table 2: Size of foo3M corpus.

The two corpora were constructed by sampling individual pages from a collection of books and editions of one newspaper, where the pages (but not necessarily the publications) of the two corpora are disjoint:[5]

- **goo300k** is the smaller, but fully manually annotated corpus, in which the annotations of each word have been verified;[6]
- **foo3M** is the larger, and only partially manually annotated corpus, in which only the more frequent word forms that do not already appear in goo300k have verified annotations.

The texts have been marked up with the time period in which they were published, e.g., 18B meaning the second half of the 18th century. This allows us to observe the changes to the vocabulary in 50-year time slices. The sizes of the corpora are given in Table 1 and Table 2.

3.3 Lexicons of historical Slovene

From the two corpora we have extracted the training and testing lexicons, keeping only words (e.g., discarding digits) that have been manually verified. The training lexicon, L_{goo} is derived from the goo300k corpus, while the test lexicon, L_{foo} is derived from the foo3M corpus and, as

[5]The corpora used in our experiments are slightly smaller than the originals: the text from two books and one newspaper issue has been removed, as the former contain highly idiosyncratic ways of spelling words, not seen elsewhere, and the latter contains a mixture of the Bohorič and contemporary alphabet, causing problems for word form normalization. The texts older than 1750 have also been removed from goo300k, as such texts do not occur in foo3M, which is used for testing our approach.

[6]A previous version of this corpus is described in (Erjavec, 2012).

Period	Pairs	Ident	Diff	OOV
18B	6,305	2,635	3,670	703
19A	18,733	12,223	6,510	2,117
19B	30,874	24,597	6,277	4,759
Σ	45,810	31,160	14,650	7,369

Table 3: Size of L_{goo} lexicon.

Period	OOV	Pairs	Ident	Diff
18B	660	3,199	493	2,706
19A	886	3,638	1,708	1,930
19B	1,983	10,033	8,281	1,752
Σ	3,480	16,029	9,834	6,195

Table 4: Size of L_{foo} lexicon.

mentioned, contains no ⟨*nform*, *mform*⟩ pairs already appearing in L_{goo}. This setting simulates the task of an existing system receiving a new text to modernize.

The lexicons used in the experiment contain entries with *nform*, *mform*, and the per-slice frequencies of the pair in the corpus from which the lexicon was derived, as illustrated in the example below:

```
benetkah     benetkah   19A:1 19B:1
aposteljnov apostolov 19A:1 19B:1
aržati       aržetu⋆    18B:2
```

The first example is a word that has not changed its spelling (and was observed twice in the 19th century texts), while the second and third have changed their spelling. The asterisk on the third example indicates that the *mform* is not present in Sloleks. We exclude such pairs from the test lexicon (but not from the training lexicon) since they will most likely not be correctly modernized by our model, which relies on Sloleks. The sizes of the two lexicons are given in Table 3 and Table 4. For L_{goo} we give the number of pairs including the OOV words, while for L_{foo} we exclude them; the tables also show the numbers of pairs with identical and different words. Note that the summary row has smaller numbers than the sum of the individual rows, as different slices can contain the same pairs.

4 Experiments and results

We conducted two experiments with the data described above. In both cases, the goal is to create C-SMT models for automatically modernizing historical Slovene words. In each experiment, we create three different models for the three time periods of old Slovene (18B, 19A, 19B).

The first experiment follows a supervised setup: we train a C-SMT model on ⟨*historical word, contemporary word*⟩ pairs from L_{goo} and test the model on the word pairs of L_{foo}. The second experiment is unsupervised and relies on monolingual data only: we match the old Slovene words from L_{goo} with modern Slovene word candidates from Sloleks; this noisy list of word pairs then serves to train the C-SMT model. We test again on L_{foo}.

4.1 Supervised learning

SMT models consist of two main components: the translation model, which is trained on bilingual data, and the language model, which is trained on monolingual data of the target language. We use the word pairs from L_{goo} to train the translation model, and the modern Slovene words from L_{goo} to train the language model.[7] As said above, we test the model on the word pairs of L_{foo}. The experiments have been carried out with the tools of the standard SMT pipeline: GIZA++ (Och and Ney, 2003) for alignment, Moses (Koehn et al., 2007) for phrase extraction and decoding, and IRSTLM (Federico et al., 2008) for language modelling. After preliminary experimentation, we settled on the following parameter settings:

- We have obtained the best results with a 5-gram language model. The beginning and the end of each word were marked by special symbols.
- The alignments produced by GIZA++ are combined with the *grow-diag-final* method.
- We chose to disable distortion, which accounts for the possibility of swapping elements; there is not much evidence of this phenomenon in the evolution of Slovene.
- We use *Good Turing discounting* to adjust the weights of rare alignments.
- We set 20% of L_{goo} aside for *Minimum Error Rate Training*.

The candidates proposed by the C-SMT system are not necessarily existing modern Slovene words. Following Vilar et al. (2007), we added a

[7] It is customary to use a larger dataset for the language model than for the translation model. However, adding the Sloleks data to the language model did not improve performances.

			Supervised		Unsupervised	
Period	Total	Baseline	No lex filter	With lex filter	No lex filter	With lex filter
18B	3199	493 (15.4%)	2024 (63.3%)	2316 (72.4%)	1289 (40.3%)	1563 (48.9%)
19A	3638	1708 (46.9%)	2611 (71.8%)	2941 (80.0%)	2327 (64.0%)	2644 (72.7%)
19B	10033	8281 (82.5%)	8707 (86.8%)	9298 (92.7%)	8384 (83.6%)	8766 (87.4%)

Table 5: Results of the supervised and the unsupervised experiments on L_{foo}.

lexicon filter, which selects the first candidate proposed by the C-SMT that also occurs in Sloleks.[8]

The results of these experiments, with and without lexicon filter, are shown in Table 5. As a baseline, we consider the words that are identical in both language varieties. Without lexicon filter, we obtain significant improvements over the baseline for the first two time spans, but as the language varieties become closer and the proportion of identical words increases, the SMT model becomes less efficient. In contrast to Vilar et al. (2007), we have found the lexicon filter to be very useful: it improves the results by nearly 10% absolute in 18B and 19A, and by 5% in 19B.

4.2 Unsupervised learning

The supervised approach requires a bilingual training lexicon which associates old words with modern words. Such lexicons may not be available for a given language variety. In the second experiment we investigate what can be achieved with purely monolingual data. Concretely, we propose a bootstrapping step to collect potential cognate pairs from two monolingual word lists (the historical words of L_{goo}, and Sloleks). We then train the C-SMT system on these hypothesized pairs.

The bootstrapping step consists of searching, for each historical word of L_{goo}, its most similar modern words in Sloleks.[9] The similarity between two words is computed with the BI-SIM measure (Kondrak and Dorr, 2004). BI-SIM is a measure of graphemic similarity which uses character bigrams as basic units. It does not allow crossing alignments, and it is normalized by the length of the longer string. As a result, this measure captures a certain degree of context sensitivity, avoids counterintuitive alignments and favours associations between words of similar lengths. BI-SIM is a language-independent measure and therefore well-suited for this bootstrapping step.

For each old Slovene word, we keep the correspondences that maximize the BI-SIM value, but only if this value is greater than 0.8.[10] For the 18B slice, this means that 812 out of 1333 historical words (60.9%) have been matched with at least one modern word; 565 of the matches (69.6%, or 42.4% of the total) were correct.

These word correspondences are then used to train a C-SMT model, analogously to the supervised approach. As for the language model, it is trained on Sloleks, since the modernized forms of L_{goo} are not supposed to be known. Due to the smaller training set size, MERT yielded unsatisfactory results; we used the default weights of Moses instead. The other settings are the same as reported in Section 4.1. Again, we conducted experiments for the three time slices. We tested the system on the word pairs of the L_{foo} lexicon, as above. Results are shown in Table 5.

While the unsupervised approach performs significantly less well on the 18B period, the differences gradually diminish for the subsequent time slices; the model always performs better than the baseline. Again, the lexicon filter proves useful in all cases.

5 Conclusion

We have successfully applied the C-SMT approach to modernize historical words, obtaining up to 57.0% (absolute) accuracy improvements with the supervised approach and up to 33.5% (absolute) with the unsupervised approach. In the future, we plan to extend our model to modernize entire texts in order to take into account possible tokenization changes.

[8] In practice, we generated 50-best candidate lists with Moses, and applied the lexicon filter on that lists. In case none of the 50 candidates occurs in Sloleks, the filter returns the candidate with the best Moses score.

[9] In order to speed up the process and remove some noise, we excluded hapaxes from L_{goo} and all but the 20,000 most frequent words from Sloleks. We also excluded words that contain less than four characters from both corpora, since the similarity measures proved unreliable on them.

[10] This threshold has been chosen empirically on the basis of earlier experiments, and allows us to eliminate correspondences that are likely to be wrong. If several modern words correspond to the same old word, we keep all of them.

Acknowledgements

The authors thank the anonymous reviewers for their comments — all errors, of course, remain our own. This work has been partially funded by the LabEx EFL (ANR/CGI), operation LR2.2, by the EU IMPACT project "Improving Access to Text" and the Google Digital Humanities Research Award "Language models for historical Slovenian".

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, UK. Aston University.

Tomaž Erjavec. 2012. The goo300k corpus of historical Slovene. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC'12*, Paris. ELRA.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of Interspeech 2008*, Brisbane.

Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'11)*, pages 125–131.

Mike Kestemont, Walter Daelemans, and Guy De Pauw. 2010. Weigh your words – memory-based lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25:287–301.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition (SIGLEX 2002)*, pages 9–16, Philadelphia.

Philipp Koehn, Hieu Hoang, Alexandra Birch, 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 *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), demonstration session*, Prague.

Grzegorz Kondrak and Bonnie Dorr. 2004. Identification of confusable drug names: A new approach and evaluation methodology. In *In Proceedings of COLING 2004*, pages 952–958.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of NAACL-HLT 2003*.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, pages 151–158, Pittsburgh.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Ulrich Reffle. 2011. Efficiently generating correction suggestions for garbled tokens of historical language. *Natural Language Engineering*, 17:265–282.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2010. Annotating a Historical Corpus of German: A Case Study. In *Proceedings of the LREC 2010 Workshop on Language Resources and Language Technology Standards*, Paris. ELRA.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. A Gold Standard Corpus of Early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 124–128, Portland, Oregon, USA, June. Association for Computational Linguistics.

Felipe Sánchez-Martínez, Isabel Martínez-Sempere, Xavier Ivars-Ribes, and Rafael C. Carrasco. 2013. An open diachronic corpus of historical Spanish: annotation criteria and automatic modernisation of spelling. Research report, Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Alicante. `http://arxiv.org/abs/1306.3692`.

Jörg Tiedemann and Peter Nabende. 2009. Translating transliterations. *International Journal of Computing and ICT Research*, 3(1):33–41. Special Issue of Selected Papers from the fifth international conference on computing and ICT Research (ICCIR 09), Kampala, Uganda.

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

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, Prague.