Modernizing historical Slovene words with character-based SMT
Yves Scherrer, Tomaž Erjavec

To cite this version:
Yves Scherrer, Tomaž Erjavec. Modernizing historical Slovene words with character-based SMT. BSNLP 2013 - 4th Biennial Workshop on Balto-Slavic Natural Language Processing, Aug 2013, Sofia, Bulgaria. 2013. <hal-00838575>

HAL Id: hal-00838575
https://hal.inria.fr/hal-00838575
Submitted on 26 Jun 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 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).
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.

<table>
<thead>
<tr>
<th>Period</th>
<th>Texts</th>
<th>Words</th>
<th>Verified</th>
</tr>
</thead>
<tbody>
<tr>
<td>18B</td>
<td>8</td>
<td>21,129</td>
<td>21,129</td>
</tr>
<tr>
<td>19A</td>
<td>9</td>
<td>83,270</td>
<td>83,270</td>
</tr>
<tr>
<td>19B</td>
<td>59</td>
<td>146,100</td>
<td>146,100</td>
</tr>
<tr>
<td>Σ</td>
<td>75</td>
<td>250,499</td>
<td>250,499</td>
</tr>
</tbody>
</table>

Table 1: Size of goo300k corpus.

<table>
<thead>
<tr>
<th>Period</th>
<th>Texts</th>
<th>Words</th>
<th>Verified</th>
</tr>
</thead>
<tbody>
<tr>
<td>18B</td>
<td>11</td>
<td>139,649</td>
<td>15,466</td>
</tr>
<tr>
<td>19A</td>
<td>13</td>
<td>457,291</td>
<td>17,616</td>
</tr>
<tr>
<td>19B</td>
<td>270</td>
<td>2,273,959</td>
<td>65,769</td>
</tr>
<tr>
<td>Σ</td>
<td>293</td>
<td>2,870,899</td>
<td>98,851</td>
</tr>
</tbody>
</table>

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

²The dataset used in this paper is available under the CC-BY-NC-SA license from http://nl.ijs.si/imp/experiments/bsnlp-2013/.
³Sloleks is encoded in LMF and available under the CC-BY-NC-SA license from http://www.slovenscina.eu/.
⁴The data for historical Slovene comes from the IMP resources, see http://nl.ijs.si/imp/.
⁵The 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.
⁶A previous version of this corpus is described in (Erjavec, 2012).
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:

\[
\begin{array}{llll}
\text{benetkah} & \text{benetkah} & 19A:1 & 19B:1 \\
\text{aposteljnov} & \text{apostolov} & 19A:1 & 19B:1 \\
\text{aržati} & \text{aržetu} & * & 18B:2 \\
\end{array}
\]

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.

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

Table 3: Size of Lgoo lexicon.

<table>
<thead>
<tr>
<th>Period</th>
<th>OOV</th>
<th>Pairs</th>
<th>Ident</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>18B</td>
<td>660</td>
<td>3,199</td>
<td>493</td>
<td>2,706</td>
</tr>
<tr>
<td>19A</td>
<td>886</td>
<td>3,638</td>
<td>1,708</td>
<td>1,930</td>
</tr>
<tr>
<td>19B</td>
<td>1,983</td>
<td>10,033</td>
<td>8,281</td>
<td>1,752</td>
</tr>
<tr>
<td>Σ</td>
<td>3,480</td>
<td>16,029</td>
<td>9,834</td>
<td>6,195</td>
</tr>
</tbody>
</table>

Table 4: Size of Lfoo lexicon.

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 \((\text{historical word, contemporary word})\) pairs from Lgoo and test the model on the word pairs of Lfoo. The second experiment is unsupervised and relies on monolingual data only: we match the old Slovene words from Lgoo 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 Lfoo.

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 Lgoo, to train the translation model, and the modern Slovene words from Lgoo to train the language model.7 As said above, we test the model on the word pairs of Lfoo. 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 Lgoo 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

7It 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.
Table 5: Results of the supervised and the unsupervised experiments on \( L_{\text{foo}} \).

<table>
<thead>
<tr>
<th>Period</th>
<th>Total</th>
<th>Baseline</th>
<th>Supervised</th>
<th>Unsupervised</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>No lex filter</td>
<td>With lex filter</td>
</tr>
<tr>
<td>18B</td>
<td>3199</td>
<td>493 (15.4%)</td>
<td>2024 (63.3%)</td>
<td>2316 (72.4%)</td>
</tr>
<tr>
<td>19A</td>
<td>3638</td>
<td>1708 (46.9%)</td>
<td>2611 (71.8%)</td>
<td>2941 (80.0%)</td>
</tr>
<tr>
<td>19B</td>
<td>10033</td>
<td>8281 (82.5%)</td>
<td>8707 (86.8%)</td>
<td>9298 (92.7%)</td>
</tr>
</tbody>
</table>

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 \( L_{\text{goo}} \), and Sloleks). We then train the C-SMT system on these hypothesized pairs.

The bootstrapping step consists of searching, for each historical word of \( L_{\text{goo}} \), its most similar modern words in Sloleks. 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. 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 \( L_{\text{goo}} \) 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 \( L_{\text{foo}} \) 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.
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


