Abstract

Historical documents are mostly accessible to scholars specialized in the period in which the document originated. In order to increase their accessibility to a broader audience and help in the preservation of the cultural heritage, we propose a method to modernized these documents. This method is based in statistical machine translation, and aims at translating historical documents into a modern version of their original language. We tested this method in two different scenarios, obtaining very encouraging results.

1. Introduction

An inherent problem in historical documents is the language in which they are written. Human language evolves with the passage of time, increasing its comprehension for contemporary people. This problem limits the accessibility of historical documents to scholars specialized in the time period in which the document was originated. To break the language barrier, these documents could be translated into a modern version of the language in which they were written.

Most scholars consider a modern version of a historical document to be that version in which words have been updated to match contemporary spelling. This way, the document preserves its original meaning and is easier to read. Fig. 1 shows an example of a historical document with modern spelling. Despite that the new version of the document is easier to read for a person who speaks Spanish, its content is still difficult to comprehend if that person is not specialized in the period in which the document was written.

For this reason, the concept of modernization that we propose does not consist only on updating the spelling. We also propose to update the lexicon and grammar to
En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos los sábados, lantejas los viernes, algún palomino de añadidura los domingos, consumían las tres partes de su hacienda. El resto de ella concluían sayo de velarte, calzas de velludo para las fiestas, con sus pantuflos de lo mismo, y los días de entre semana se honraba con su vellorí de lo más fino.

Additional problems arise with historical manuscripts. Besides the language barrier, these kind of documents have extra difficulties particular to their author. For instance, they contain a lot of abbreviated words. These abbreviations do not follow any known standard and are usually particular to the time period and writer of the document, with the same writer changing her style during the years. Moreover, in
many occasions, the same word inconsistently appears abbreviated or fully written throughout the same document. Fig. 3 shows an example of a historical manuscript in which this problem is present. The transcription of the manuscript is known as a transliteration, and the version in which abbreviations have been expanded to their corresponding words is known as paleographic version.

Figure 3. Example of a historical manuscript with abbreviations. The left text is a transliteration of the manuscripts, and the right text is known as a paleographic version of the document. Words in **bold** represent abbreviations and their corresponding expansions. Words in *italic* denote words which inconsistently appear abbreviated and fully written throughout the text. Additionally, beginning of sentences have been truecased. The texts from the example belong to the Alcaraz corpus (Villegas et al., 2016).

In this work, we propose a method to translate historical documents to a contemporary version of the language in which they were written. With this modernized version of a document, we aim at increasing the accessibility of historical documents to a broader audience, as well as helping in the preservation of the cultural heritage: e.g., given a transliteration of a manuscript, this method could be applied to obtain the corresponding paleographic version.

The rest of this paper is structured as follows: **Section 2** presents our modernization approach. Then, in **Section 3**, we describe the experiments conducted in order to assess our proposal. After that, in **Section 4**, we present the results of those experiments. Finally, conclusions are drawn in **Section 5**.

2. Modernization

In this section, we present a method to translate a historical document into a contemporary version of its language. We also describe two additional techniques to enhance translation quality.
2.1. Statistical Machine Translation

In order to achieve the modernization of historical documents, we propose an approach based on Statistical Machine Translation (SMT). SMT has as a goal to find the best translation $\hat{y}$ of a given source sentence $x$ (Brown et al., 1993):

$$\hat{y} = \arg \max_y \Pr(y \mid x)$$

(1)

During years, phrase-based models (Koehn, 2010) have been the prevailing approach to compute this expression. These models rely on a log-linear combination of different models (Och and Ney, 2002): namely, phrase-based alignment models, re-ordering models and language models; among others (Zens et al., 2002; Koehn et al., 2003). However, in the last few years, neural machine translation (Sutskever et al., 2014; Bahdanau et al., 2015) has had a great impact. This novel approach is based on the use of neural networks for carrying out the translation process.

Therefore, considering the document’s original language as a source and the modern version of that language as the target, we propose to use phrase-based SMT to obtain a modernized version of the document.

2.2. Data Selection

In order to successfully apply SMT for modernizing a historical document, we need training data as similar as possible as the document to modernize. However, this is not always feasible. To cope with this problem, we propose to use a data selection technique which has been successfully used in SMT to increase the training data with sentences from corpora of different domains than the text to translate, which are as similar as possible to this text.

Infrequent n-grams recovery strategy (Gascó et al., 2012) increases the training corpus by selecting from other corpora the sentences closest to the test set. These sentences contain those n-grams that have been seldom observed in the test set. i.e., the infrequent n-grams. An n-gram is considered infrequent when it appears less times than a given infrequency threshold $t$. Therefore, the idea is to construct a training corpus by selecting from the available corpora those sentences which contain the most infrequent n-grams.

Let $X$ be the set of n-grams that appear in the sentences to be translated; $m$ one of these n-grams; $R(m)$ the counts of $m$ in a given source sentence $x$ from the available corpora; and $t$ a given infrequency threshold. Then, the infrequency score $i(x)$ is defined as:

$$i(x) = \sum_{m \in X} \min(1, R(m))t$$

(2)

Therefore, the sentences from the available corpora are scored using Eq. (2). Then, at each iteration, the sentence $x^*$ with the highest score $i(x^*)$ is selected and added
to the training corpus. After that, $x^*$ is removed from the available corpora and the counts of the n-grams $R(m)$ are updated within $x^*$. Consequently, the scores of the corpora are updated. This process is repeated until all the n-grams within $X$ reach frequency $t$. Once the process is finished, the resulting corpus will be the one used for training the systems.

### 2.3. Byte Pair Encoding

A common problem in SMT are those rare and unknown words which the system has never seen. This could be a bigger problem when modernizing historical documents due to the constants evolution of the language as well as, in the case of manuscripts, the aforementioned problem with abbreviations (see Section 1). An innovative solution to tackle this problem is Byte Pair Encoding (BPE) (Sennrich et al., 2016).

Based on the intuition that various word classes are translatable via smaller units than words, this technique aims at encoding rare and unknown words as sequences of subwords units. To achieve this, the symbol vocabulary is initialized with the character vocabulary, and each word is represented as a sequence of characters—plus a special end-of-word symbol. After that, all symbol pairs are iteratively counted. Then, each occurrence of the most frequent pair $(A, B)$ is replaced with a new symbol $AB$. This process is repeated as many times as new symbols to create. Once the encoding is learned, BPE is applied to the training corpora to obtain a representation as sequences of subwords units. Then, the SMT system is trained using the encoded corpora. At the end of the process, the generated text—which has been translated into an encoded version of the target language—is decoded.

### 3. Experiments

In this section, we describe the experiments conducted in order to assess our proposal. We also present the corpora and metrics, and describe the set up of our framework.

#### 3.1. Corpora

To test our proposal, we selected the corpora distributed at the CLIN2017 Shared Task on Translating Historical Text:\[1\]

**Bible:** A collection of books from different version of the Dutch bible. Mainly, a version from 1637, another from 1657, another from 1888 and another from 2010. All versions are composed by the same books, except from the 2010’s version, which is missing the last part of the content.

[^1]: [https://ifarm.nl/clin2017st/](https://ifarm.nl/clin2017st/)
Dutch Literature: A collection of texts from Dutch literary classics from the 17th century. It contains a small development partition and a test partition. The test partition is composed by a collection of texts from a different decade of the 17th century.

The goal of the shared task was to translate historical documents from 17th to 21st century Dutch. However, the translation they were looking for consisted in replacing all the words that did not occur in a standard lexicon. Therefore, the aim of the shared task was to update the spelling to 21st century standards, and not to obtain a version of the documents that matches nowadays Dutch.

While the Dutch literature corpus was created with the aim of updating the spelling, the Bible corpus contains the same books in different versions of Dutch (i.e., the Dutch spoken in the moment they were written). This last corpus was given as a training material for the shared task, and contains a test partition for translating a document from 17th to 19th century Dutch. Therefore, we decided to use this corpus to assess our proposal—considering 19th century Dutch as modern Dutch. Additionally, we make use of the Dutch literature corpus to evaluate our method in the context of only updating the spelling. Table 1 shows the corpora statistics.

<table>
<thead>
<tr>
<th></th>
<th>Bible 1637–1888</th>
<th>Bible 1637–2010</th>
<th>Dutch literature 17th–21st century</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>T</td>
<td>V</td>
</tr>
<tr>
<td>Train</td>
<td>37K</td>
<td>927/917K</td>
<td>57/45K</td>
</tr>
<tr>
<td>Development</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Test</td>
<td>5000</td>
<td>148/141K</td>
<td>11/9K</td>
</tr>
</tbody>
</table>

Table 1. Corpora statistics. |S| stands for number of sentences, |T| for number of tokens and |V| for size of the vocabulary. K denotes thousand. The bible corpus is extracted from different versions of the Dutch bible. The Dutch literature corpus is composed by a collection of texts extracted from various Dutch literary classics.

For the task of modernizing historical documents, we limited the training corpora to the 1637–1888 partition of the Bible corpus (since we are considering 19th century Dutch to be the contemporary version of Dutch). Additionally, to enrich the language model, we collected all 19th century works available at the Digitale Bibliotheek voor de Nederlandse letteren2 and added them to the training data.

2http://dbnl.nl/
The 1637–2010 and 1657–2010 partitions of the Bible corpus were proportionated with a warning about the quality of the 2010’s version. Therefore, for the task of updating the spelling to 21st century Dutch, instead of limiting the training data to these two partitions we made use of all the available partitions. More precisely, we selected those sentences from the training corpora which were better suited for the task (see Section 2.2). Additionally, in a similar way as in the previous task, we collected all 21st century works from the Digitale Bibliotheek voor de Nederlandse letteren to enrich the language model.

3.2. Metrics

In order to assess our proposal, we made use of the following well known metrics: BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002): computes the geometric average of the modified n-gram precision, multiplied by a brevity factor that penalizes short sentences. Translation Error Rate (TER) (Snover et al., 2006): computes the number of word edit operations (insertion, substitution, deletion and swapping), normalized by the number of words in the final translation.

3.3. SMT Systems

SMT systems were trained with the Moses toolkit (Koehn et al., 2007), following the standard procedure: optimizing the weights of the log-lineal model with MERT (Och, 2003), and estimating a 5-gram language model, smoothed with the improved Kneser-Ney method (Chen and Goodman, 1996), with SRILM (Stolcke, 2002). Moreover, since source and target have similar linguistic structures—the target language is an evolution of the source language—we used monotonous reordering. The corpora were lowercased and tokenized using the standard scripts, and the translated text was truecased with Moses’ truecaser.

The systems in which BPE was used (see Section 2.3) were trained in the same way. The only difference is that the corpora were previously encoded using BPE, and the translated text was decoded afterwards. BPE encoding was learned and applied using the scripts kindly provided by Sennrich et al. (2016). In learning the encoding, the default values for the number of symbols to create and the minimum frequency to create a new symbol were used.

4. Results

This section presents the results of the experiments conducted in order to assess our proposal. We first evaluate our method for modernizing a historical document using the Bible corpus (see Section 3.1) and, then, we additionally test our method in a context in which only the spelling needs to be updated, using the Dutch literature.
corpus. Confidence intervals (p = 0.05) were computed for all metrics by means of bootstrap resampling (Koehn, 2004).

4.1. Document Modernization

The first task consisted in applying our proposed method for obtaining a version of a historical document in modern language. Table 2 shows the results obtained in this task. As a baseline, we compare the quality of the original document with respect to its modern version. Additionally, the shared task from which the corpus was obtained (see Section 3.1) provided an extra baseline. This second baseline was generated by applying some unspecified translation rules to the original document.

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>13.5 ± 0.3</td>
<td>57.0 ± 0.3</td>
</tr>
<tr>
<td>Baseline₂</td>
<td>50.8 ± 0.4</td>
<td>26.5 ± 0.3</td>
</tr>
<tr>
<td>SMT</td>
<td>64.8 ± 0.4</td>
<td>17.0 ± 0.3</td>
</tr>
<tr>
<td>SMT + LM₂</td>
<td>65.1 ± 0.4</td>
<td>17.3 ± 0.3</td>
</tr>
<tr>
<td>SMT_{BPE}</td>
<td>64.8 ± 0.4</td>
<td>17.4 ± 0.3</td>
</tr>
<tr>
<td>SMT_{BPE} + LM₂</td>
<td>66.7 ± 0.4</td>
<td>16.2 ± 0.3</td>
</tr>
</tbody>
</table>

Table 2. Experimental results for the document modernization task using the Bible corpus. Baseline system corresponds to considering the original document as the modernized document. Baseline₂ was proportionated as part of the shared task and was obtained by applying certain translation rules to the original document. SMT is the standard SMT system. SMT + LM₂ is the SMT system trained with an additional language model. SMT_{BPE} is the standard system in which the training corpus has been encoded using BPE. SMT_{BPE} + LM₂ is the system in which the training corpus has been encoded using BPE and an additional language model is used during the training process. Best results are denoted in **bold**.

The proposed standard SMT system greatly improves this first baseline, both in terms of BLEU (around 51 points of improvement) and TER (around 40 points of improvement). Moreover, it also improves significantly the second baseline (around 14 points of BLEU and 9 points of TER). Finally, enriching the system by adding an additional language model does not significantly improve the results of the standard system. Most likely, this is due to the training data being very similar to the document we are modernizing (they all belong to the same version of the Bible). For this reason, the language model obtained from the training data is robust enough to do the modernization without additional help.

Encoding the training corpora with BPE (see Section 2.3) to reduce the number of unknown words brings similar results to just using the standard system. Once more,
the similarity between training and test reduces the vocabulary problem. Nonetheless, combining the use of BPE with the additional language model obtains a significant improve over the standard system (around 2 points of BLEU and 1 points of TER). Most likely, this is due to BPE taking profit from the additional language model to better learn how to generate subword units.

4.2. Standard Spelling

The second task consisted in updating the spelling of a historical document to match current standards. Although our proposed method aims at obtaining a version of the document with modern language, we wanted to assess how the method would work in this context. Similarly as in the previous task, we considered as baseline the quality of the original document in comparison to the document with the updated spelling.

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>TER</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>29.9 ± 1.8</td>
<td>32.4 ± 1.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SMT</td>
<td>48.1 ± 1.8</td>
<td>22.0 ± 0.8</td>
<td>49.9 ± 1.8</td>
<td>20.2 ± 0.8</td>
</tr>
<tr>
<td>+ LM₂</td>
<td>49.4 ± 1.8</td>
<td>21.2 ± 0.8</td>
<td>49.8 ± 1.8</td>
<td>20.9 ± 0.8</td>
</tr>
<tr>
<td>SMT獒BPE</td>
<td>48.6 ± 1.6</td>
<td>24.2 ± 0.9</td>
<td>49.2 ± 1.6</td>
<td>23.7 ± 0.8</td>
</tr>
<tr>
<td>+ LM₂</td>
<td>47.9 ± 1.7</td>
<td>25.5 ± 0.9</td>
<td>49.9 ± 1.7</td>
<td>23.7 ± 0.8</td>
</tr>
</tbody>
</table>

Table 3. Experimental results for the standard spelling task using the Dutch literature corpus. Baseline system correspond to considering the original document as the document with the updated spelling. SMT is the standard SMT system. SMT + LM₂ is the SMT system trained with an additional language model. SMT獒BPE is the standard system in which the training corpus has been encoded using BPE. SMT獒BPE + LM₂ is the system in which the training corpus has been encoded using BPE and an additional language model is used during the training process. Best results are denoted in bold.

Our standard SMT system greatly improves the baseline, obtaining increases of around 18 points of BLEU and 10 points of TER. Similarly as in the previous task, enriching the system with an additional language model does not obtain significant improvements. This is probably due to the nature of the task: only non-standard words should be change, independently of semantic correctness. The language model, however, has only been trained with sentences which are semantically correct.

In this case, encoding the training corpus with BPE (see Section 2.3) to mitigate the number of unknown words does not improve results. Not even when using an additional language model. Most likely, the nature of the task makes more difficult for BPE to learn to create subword units.
When using data selection to create a new training corpus formed only by those sentences which are more similar to the document (see Section 2.2), we obtain a significant improve in terms of TER. Results for BLEU, however, are not significantly different to training with all the available corpora. Similarly to what happened before, enriching the system with an additional language model does not obtain significant improvements.

As in the previous case, encoding the training corpus with BPE to reduce the number of unknown words does not improve results. BLEU values are more or less within the same confidence interval, while TER significantly increases around 3 points. Enriching the system with an additional language model also obtains similar results.

Finally, in comparison to the results of the shared task from which this corpus was obtained (see Section 3.1), our approach would have placed 6th out of 9. It is worth noting, however, that while the aim of the shared task was to update the spelling to modern standards without aiming for semantic correctness, our method aimed at obtaining modern semantic, lexicon and grammar.

5. Conclusions and Future Work

In this work, we have presented a method, based on SMT, to translate a historical document to a modern version of its original language. With this method, we aim at increasing the accessibility of historical documents to a broader audience as well as helping in the preservation of the cultural heritage.

Experimental results show that the proposed method significantly increases the quality of the document—with respect to the modern language. However, due to the lack of available corpora, we tested our proposal on a corpus in which the training data is very similar to the document to modernize. This is not often the case with historical documents. Therefore, we should test our method in a framework in which the document to translate has few similarities with the training data.

We also proposed two alternatives for solving two common problems in SMT which also affect to the modernization task. The first of these alternatives, to find training data as similar to the document as possible, was not tested due to the training data already being similar to the document. The second alternative, to tackle rare and unseen words, significantly improves the results achieved by the basic method.

Additionally to the modernization of historical documents, we have tested our method for updating the spelling of a historical document according to modern standards. Experimental results show that our proposal succeeds at standardizing the spelling. However, when comparing to other approaches to this problem, our method still needs some improvements. Nonetheless, this task searches for updating the spelling without aiming for semantic correctness, while our proposal aims at obtaining a modern version of the language—including its spelling and semantic.

We also tested the previously mentioned alternatives. Using data selection techniques to find training data as similar as possible to the document significantly im-
proves results. However, due to the nature of the task, the second alternative does not improve results.

As a future work, besides obtaining more corpora to being able to work in a more common framework, we want to assess our proposal with historical manuscripts to see how it behaves with the additional difficulties inherent in the manuscripts. Additionally, it would be interesting to use our method to generate the paleographic version of a transliterated transcript.

Acknowledgements

The research leading to these results has received funding from the Ministerio de Economía y Competitividad (MINECO) under project CoMUN-HaT (grant agreement TIN2015-70924-C2-1-R), and Generalitat Valenciana under project ALMAMATER (grant agreement PROMETEOII/2014/030).

Bibliography


Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks, 2014.


Address for correspondence:
Miguel Domingo
midobal@prhlt.upv.es
Universitat Politècnica de València
PRHLT Research Center
Camino de Vera s/n, 46022 Valencia, Spain

306