An Interactive Machine Translation Framework for Modernizing Historical Documents

Abstract

Due to the nature of human language, historical documents are hard to comprehend by contemporary people. This limits their accessibility to scholars specialized in the time period in which the documents were written. Modernization aims at breaking this language barrier by generating a new version of a historical document, written in the modern version of the document's original language. However, while it is able to increase the document's comprehension, modernization is still far from producing an error-free version. In this work, we propose a collaborative framework in which a scholar can work together with the machine to generate the new version. We tested our approach on a simulated environment, achieving significant reductions of the human effort needed to produce the modernized version of the document.

Introduction

In recent years, awareness of the importance of preserving our cultural heritage has increased. Historical documents are an important part of that heritage. In order to preserve them, there is an increased need in creating digital text versions which can be search and automatically processed BIBREF0. However, their linguistic properties create additional difficulties: due to the lack of a spelling convention, orthography changes depending on the time period and author. Furthermore, human language evolves with the passage of time, increasing the difficulty of the document's comprehension. Thus, historical documents are mostly accessible to scholars specialized in the time period in which each document was written.

Modernization tackles the language barrier in order to increase the accessibility of historical documents. To achieve this, it generates a new version of a historical document in the modern version of the language in which the document was originally written (fi:Shakespeare shows an example of modernizing a document). However, while modernization has been successful in order to increase the comprehension of historical documents BIBREF1, BIBREF2, it is still far from creating error-free modern versions. Therefore, this task still needs to be carried out by scholars.

Interactive machine translation (IMT) fosters human–computer collaborations to generate error-free translations in a productive way BIBREF4, BIBREF5. In this work, we proposed to apply one of these protocols to historical documents modernization. We strive for creating an error-free modern version of a historical document, decreasing the human effort needed to achieve this goal.

The rest of this document is structured as follows: se:work introduces the related work. Then, in se:IMT we present our protocol. se:exp describes the experiments conducted in order to assess our proposal. The results of those experiments are presented and discussed in se:res. Finally, in se:conc, conclusions are drawn.

Related Work

While the lack of a spelling convention has been extensively researched for years BIBREF6, BIBREF7, BIBREF8, modernization of historical documents is a younger field. BIBREF1 organized a shared task in order to translate historical text to contemporary language. The main goal of this shared task was to tackle the spelling problem. However, they also approached document modernization using a set of rules. BIBREF9 proposed a modernization approach based on statistical machine translation (SMT). A neural machine translation (NMT) approach was proposed by BIBREF2. Finally, BIBREF10 extracted parallel phrases from an original parallel corpus and used them as an additional training data for their NMT approach.

Despise the promising results achieved in last years, machine translation (MT) is still far from producing high-quality translations BIBREF11. Therefore, a human agent has to supervise these translation in a post-editing stage. IMT was introduced with the goal of combining the knowledge of a human translator and the efficiency of an MT system. Although many protocols have been proposed in recent years BIBREF12, BIBREF13, BIBREF14, BIBREF15, the prefix-based remains as one of the most successful approaches BIBREF5, BIBREF16, BIBREF17. In this approach, the user corrects the leftmost wrong word from the translation hypothesis, inherently validating a correct prefix. With each new correction, the system generates a suffix that completes the prefix to produce a new translation.

Interactive Machine Translation

Classical IMT approaches relay on the statistical formalization of the MT problem. Given a source sentence $\mathbf {x}$, SMT aims at finding its most likely translation $\hat{\mathbf {y}}$ BIBREF18:

For years, the prevailing approach to compute this expression have been phrase-based models BIBREF19. These models rely on a log-linear combination of different models BIBREF20: namely, phrase-based alignment models, reordering models and language models; among others BIBREF21, BIBREF22. However, more recently, this approach has shifted into neural models (see se:NMT).

Interactive Machine Translation ::: Prefix-based Interactive Machine Translation

Prefix-based IMT proposed a user–computer collaboration that starts with the system proposing an initial translation $\mathbf {y}$ of length I. Then, the user corrects the leftmost wrong word y_i, inherently validating all preceding words. These words form a validated prefix $\tilde{\mathbf {y}}_p$, that includes the corrected word $\tilde{y}_i$. The system reacts to this user feedback, generating a suffix $\hat{\mathbf {y}}_s$ that completes $\tilde{\mathbf {y}}_p$ to obtain a new translation of $\mathbf {x}:\hat{\mathbf {y}}~=~\tilde{\mathbf {y}}_p\,\hat{\mathbf {y}}_s$. This process is repeated until the user accepts the complete system suggestion. fi:IMT illustrates this protocol.

BIBREF5 formalized the suffix generation as follows:

which can be straightforwardly rewritten as:

This equation is very similar to eq:SMT: at each iteration, the process consists in a regular search in the translations space but constrained by the prefix $\tilde{\mathbf {y}}_p$.

Interactive Machine Translation ::: Neural Machine Translation

In NMT, eq:SMT is modeled by a neural network with parameters $\mathbf {\Theta }$:

This neural network usually follows an encoder-decoder architecture, featuring recurrent networks BIBREF23, BIBREF24, convolutional networks BIBREF25 or attention mechanisms BIBREF26. Model parameters are jointly estimated on large parallel corpora, using stochastic gradient descent BIBREF27, BIBREF28. At decoding time, the system obtains the most likely translation using a beam search method.

Interactive Machine Translation ::: Prefix-based Interactive Neural Machine Translation

The prefix-based IMT protocol (see se:PBIMT) can be naturally included into NMT systems since sentences are generated from left to right. In order to take into account the user's feedback and generate compatible hypothesis, the search space must be constraint. Given a prefix $\tilde{\mathbf {y}}_p$, only a single path accounts for it. The branching of the search process starts once this path has been covered. Introducing the validated prefix $\tilde{\mathbf {y}}_p$, eq:NMT becomes:

which implies a search over the space of translations, but constrained by the validated prefix $\tilde{\mathbf {y}}_p$ BIBREF15.

Experiments

In this section, we present our experimental conditions, including translation systems, corpora and evaluation metrics.

Experiments ::: MT Systems

SMT systems were trained with Moses BIBREF29, following the standard procedure: we estimated a 5-gram language model—smoothed with the improved KneserNey method—using SRILM BIBREF30, and optimized the weights of the log-linear model with MERT BIBREF31.

We built our NMT systems using NMT-Keras BIBREF32. We used long short-term memory units BIBREF33, with all model dimensions set to 512. We trained the system using Adam BIBREF34 with a fixed learning rate of 0.0002 and a batch size of 60. We applied label smoothing of 0.1 BIBREF35. At inference time, we used beam search with a beam size of 6. We applied joint byte pair encoding to all corpora BIBREF36, using $32,000$ merge operations.

Statistical IMT systems were implemented following the procedure of word graph exploration and generation of a best suffix for a given prefix described by BIBREF5. Neural IMT systems were built using the interactive branch of NMT-Keras.

Experiments ::: Corpora

The first corpus used in our experimental session was the Dutch Bible BIBREF1. This corpus consists in a collection of different versions of the Dutch Bible: a version from 1637, another from 1657, another from 1888 and another from 2010. Except for the 2010 version, which is missing the last books, all versions contain the same texts. Moreover, since the authors mentioned that the translation from this last version is not very reliable and, considering that Dutch has not evolved significantly between 1637 and 1657, we decided to only use the 1637 version—considering this as the original document—and the 1888 version—considering 19$^{\mathrm {th}}$ century Dutch as modern Dutch.

We selected El Quijote BIBREF2 as our second corpus. This corpus contains the famous 17$^{\mathrm {th}}$ century Spanish novel by Miguel de Cervantes, and its correspondent 21$^{\mathrm {st}}$ century version. Finally, we used El Conde Lucanor BIBREF2 as a third corpus. This data set contains the original 14$^{\mathrm {th}}$ century Spanish novel by Don Juan Manuel, and its correspondent 21$^{\mathrm {st}}$ century version. Due to the small size of the corpus, we decided to use it only as a test. Additionally, unable to find a suitable training corpus, we used the systems built for El Quijote—despite the original documents belonging to different time periods—in order to modernize El Conde Lucanor.

ta:corp presents the corpora statistics.

Experiments ::: Metrics

In order to measure the gains in human effort reduction, we made use of the following metrics:

BIBREF37: measures the number of words edited by the user, normalized by the number of words in the final translation.

BIBREF5: measures the number of mouse actions made by the user, normalized by the number of characters in the final translation.

Additionally, to evaluate the quality of the modernization and the difficulty of each task, we made use of the following well-known metrics:

BiLingual Evaluation Understudy (BLEU) BIBREF38: computes the geometric average of the modified n-gram precision, multiplied by a brevity factor that penalizes short sentences.

Translation Error Rate (TER) BIBREF39: computes the number of word edit operations (insertion, substitution, deletion and swapping), normalized by the number of words in the final translation.

We used sacreBLEU BIBREF40 for ensuring consistent BLEU scores. For determining whether two systems presented statistically significant differences, we applied approximate randomization tests BIBREF41, with $10,000$ repetitions and using a p-value of 0.05.

Experiments ::: User Simulation

Due to the high costs of an evaluation involving human agents, we carried out an automatic evaluation with simulated users whose desired modernizations correspond to the reference sentences.

At each iteration, the user corrects the leftmost wrong word from the system's hypothesis. With this correction, a new prefix is validated. The associated cost of this correction is of one mouse action and one word stroke. The system, then, reacts to this feedback, generating a new suffix that completes the prefix to conform a new hypothesis. This process is repeated until hypothesis and reference are the same.

Results

ta:quality presents the quality of the modernization. Both SMT and NMT approaches were able to significantly improved the baseline. That is, the modernized documents are easier to comprehend by a contemporary reader than the original documents. An exception to this is El Conde Lucanor. The SMT approach yielded significant improvements in terms of TER, but was worse in terms of BLEU. Moreover, the NMT approach yielded worst results in terms of both BLEU and TER. Most likely, this results are due to having used the systems trained with El Quijote for modernizing El Conde Lucanor (see se:corp).

When comparing the SMT and NMT approaches, we observe that SMT yielded the best results in all cases. This behavior was already perceived by BIBREF2 and is, most likely, due to the small size of the training corpora—a well-known problem in NMT. However, while the goal of modernization is making historical documents as easier to comprehend by contemporary people as possible, our goal is different. In this work, our goal is to obtain an error-free modern copy of a historical document. To achieve this, we proposed an interactive collaboration between a human expert and our modernizing system, in order to reduce the effort needed to generate such copy. ta:effort presents the experimental results.

Both SMT and NMT approaches yielded significant reductions of the human effort needed to modernize the Dutch Bible (up to 48 points in terms of WSR and 8 in terms of MAR) and El Quijote (up to 7 points in terms of WSR and 1 of MAR). For El Conde Lucanor, however, both approaches resulted in an increased of the effort need to generate an error-free modern version. This behavior was to be expected since the modernization quality for El Conde Lucanor was very low. Therefore, the system consistently generated wrong suffixes, resulting in the user having to make more corrections.

Regarding the performance of both approaches, SMT achieved the highest effort reduction. This was reasonably expected since its modernization quality was better. However, in past neural IMT works

BIBREF15, the neural IMT approach was able to yield further improvements despite having a lower translation quality than its SMT counterpart. Most likely, the reason of this is that, due to the small training corpora, the neural model was not able to reach its best performance, Nonetheless, we should address this in a future work.

Results ::: Qualitative Analysis

fi:exIMT shows an example of modernizing a sentence from El Quijote with the interactive SMT approach. While the system's initial suggestion contains five errors, with the IMT protocol, the user only needs to make three corrections. With each correction, the system is able to improve its suggestions, reducing the total effort needed to achieve an error-free modernization. Note that this example has been chosen for illustrative purposes of a correct functioning of the system. The average sentences from El Quijote are longer, and there are times in which the system fails to take the human knowledge into account, resulting in an increase of the number of corrections. Nonetheless, as seen in se:res, overall the system is able to significantly decrease the human effort.

fi:exINMT contains an example of modernizing the same sentence as in fi:exIMT, using the interactive NMT approach. This is an example in which the system fails to take into account the user's corrections, resulting in an increase of the human effort. It is specially worth noting the introduction of non-existing words such as durdos and duradas. This problem was probably caused by an incorrect segmentation of a word, via the byte pair encoding process, and should be address in a future work. Nonetheless, as seen in se:res, overall the system is able to significantly decrease the human effort.

Conclusions and Future Work

In this work, we proposed a collaborative user–computer approach to create an error-free modern version of a historical document. We tested this proposal on a simulated environment, achieving significant reductions of the human effort. We built our modernization protocol based on both SMT and NMT approaches to prefix-based IMT. Although both systems yielded significant improvements for two data sets out of three, the SMT approach yielded the best results—both in terms of the human reduction and in the modernization quality of the initial system.

As a future work, we want to further research the behavior of the neural systems. For that, we would like to explore techniques for enriching the training corpus with additional data, and the incorrect generation of words due to subwords. We would also like to develop new protocols based on successful IMT approaches. Finally, we should test our proposal with real users to obtain actual measures of the effort reduction.

Acknowledgments

The research leading to these results has received funding from the European Union through Programa Operativo del Fondo Europeo de Desarrollo Regional (FEDER) from Comunitat Valencia (2014–2020) under project Sistemas de frabricación inteligentes para la indústria 4.0 (grant agreement IDIFEDER/2018/025); and from Ministerio de Economía y Competitividad (MINECO) under project MISMIS-FAKEnHATE (grant agreement PGC2018-096212-B-C31). We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for part of this research.