Two Demonstrations of the Machine Translation Applications to Historical Documents

Miguel Domingo and Francisco Casacuberta
PRHLT Research Center - Universitat Politècnica de València
midobal@prhlt.upv.es, fcn@prhlt.upv.es

***Abstract*—We present our demonstration of two machine translation applications to historical documents. The first task consists in generating a new version of a historical document, written in the modern version of its original language. The second application is limited to a document's orthography. It adapts the document's spelling to modern standards in order to achieve an orthography consistency and accounting for the lack of spelling conventions. We followed an interactive, adaptive framework that allows the user to introduce corrections to the system's hypothesis. The system reacts to these corrections by generating a new hypothesis that takes them into account. Once the user is satisfied with the system's hypothesis and validates it, the system adapts its model following an online learning strategy. This system is implemented following a client–server architecture. We developed a website which communicates with the neural models. All code is open-source and publicly available. The demonstration is hosted at http://casmacat.prhlt.upv.es/mthd/.**

I. Introduction

Despite being an important part of our cultural heritage, historical documents are mostly accessible to scholars. This is due to both the linguistic properties of these documents—in the past, orthography changed depending on the time period and author due to a lack of an spelling convention—and the nature of language—which evolves with the passage of time. This creates a language barrier which increases the difficulty of comprehending historical documents.

Different research topics on historical documents focus on tackling this language barrier as well as different aspects related to the document's language richness, which is also part of our cultural heritage. For example, historical manuscripts are automatically digitized and transcribed [1]. Documents are automatically translated into a modern version of their original language to make them accessible to a broader audience [2]. And orthography is normalized to account for the lack of a spelling convention [3].

In this work, we focus on two of these topics: language modernization and spelling normalization. The first one aims to generate a modern version of a historical document to make its content available to a broader audience. The second one aims to achieve an orthography consistency—without altering the document's content—to reduce the variability derived from the lack of spelling conventions and facilitate the work of scholars. We present a demonstration that showcases neural machine translation (NMT) applications to these two topics, and provides an interactive, adaptive framework for scholars to increase their productivity when working on these tasks.

II. Applications

Our demonstration showcases NMT applications to two different historical documents research topics: language modernization and spelling normalization.

A. Language modernization

This research topic aims to tackle the language barrier inherent in historical documents in order to make them available to a broader audience. To do so, it proposes to automatically generate a new version of a historical document, written in the modern version of the document's original language. A common approach to this problem is to tackle modernization as a conventional MT task [4], [5]. Results showed that, while there is still room for improvement, modernization techniques successfully decrease the comprehension difficulty of historical documents—indicated by both automatic metrics and human evaluation [2].

Additionally, while modernization's goal is limited to bringing a better understanding of historical documents to a general audience—language-related losses may appear during the process—scholars are in charge of different tasks that require them to generate modernizations of the highest quality (e.g., producing a comprehensive contents document for non-experts [6]). Thus, we provided our system of an interactive, adaptive framework to increase the scholars productivity when working on these tasks. The modernization system was developed following Domingo and Casacuberta [2]. For the interactive, adaptive framework, we followed Peris and Casacuberta [7].

B. Spelling normalization

In order to account for the lack of spelling conventions, which were not created until recent years, spelling normalization aims to achieve an orthography consistency by adapting a document's spelling to modern standards. This linguistics problem suppose an additional challenge for the effective natural language processing for these documents. Through the years, different approaches to this problem have been researched (e.g., [3], [8]). Some of them tackle spelling normalization as a conventional MT problem (e.g., [9], [10]). We developed our normalization system following Domingo and Casacuberta [11], which approached spelling normalization using character-based NMT and obtained significant improvements according to several automatic metrics. Like in the previous task, we followed Peris and Casacuberta [7] for the interactive, adaptive framework.

III. System description

Our system is composed of two main elements: the client and the server. The client is an HTML website, which interacts with the user through javascript and communicates with the server via the HTTP protocol, using the PHP curl tool. The server is the core element. It contains the NMT systems, which were developed with *NMT-Keras* [12], and it is deployed as a Python HTTP server that handles the client's requests. All code is open-source an publicly available at https://github.com/midobal/mthd.

Initially, an old sentence is presented to the user in the client website. When the user requests an automatic modernization/normalization (see Section II), the client communicates the server via PHP. Then, the server queries the NMT system, which generates an initial hypothesis. After that, the hypothesis is sent back to the client website.

At this point, the interactive-predictive process starts: The user searches the hypothesis for the first error and introduces a correction with the keyboard (writing one or more characters). Once the user finishes typing, the client reacts to this feedback by sending a request to the server. This request contains the old sentence and the user feedback (the sequence of characters that conform the prefix). Then, the NMT system produces an alternative hypothesis coherent with the user's feedback and sends it back to the client website. This process is repeated until the user finds the system's hypothesis satisfactory. Fig. 1 illustrates one step of this process.

Once the user is satisfied with the system's hypothesis, they can validate it. Then, the system is incrementally updated with this new sample following an online learning setup [7]. Hence, in future interactions, the system will be progressively updated and able to generate better hypothesis.

Fig. 2 illustrates an example of how to performed a task using the client server. After having selected the task to perform, a list of old sentences will appear. When you click on "Modernize/Normalize", the system will generate an initial hypothesis. If you desire to improve this hypothesis, you can click on the left box and type a correction. The system will, then, generate a new hypothesis to take that correction into account. You can repeat this process for as many corrections as you desire to make. Finally, you can click "Validate" to tell the system that you are happy with the modernization/normalization and, if the *Learn from sample* option is activated (in blue), the system will use the sample to improve its model.

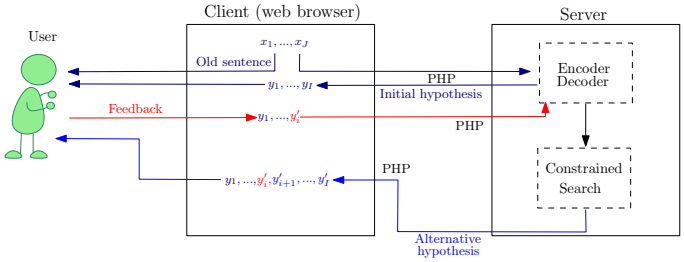

Fig. 1: System architecture. The client presents the user an old sentence and a prediction. Then, the user introduces a feedback signal for correcting this prediction (in this example, they are validating the prefix $y_1, .., y_{i-1}$ and correcting the word y'_i). After that, the old sentence and the user's feedback is sent to the server, which generates an alternative hypothesis that takes into account the user corrections (in this example, a new suffix $y'_{i+1}, .., y'_I$ that completes the user's feedback).

IV. Conclusions and future work

We presented a demonstration of two MT applications to historical documents. We described its client–server architecture and developed a website to facilitate the use of the system.

As a future work, we would like to improve our website's front end. This improvement could allow us to create a better visualization of the hypothesis, which is specially relevant for the spelling normalization task. Additionally, relevant attributes of the neural system could be visualize in order to help to understand better the model predictions and behavior.

Acknowledgments

The research leading to these results has received funding from Generalitat Valenciana (GVA) under project PROMETEO/2019/121. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for part of this research, and Andrés Trapiello and Ediciones Destino for granting us permission to use their book in our research.

Fig. 2: Frontend of the client website. As the button "Modernize" is clicked (or "Normalize", depending on the task your are performing), an initial hypothesis for the old sentence appears in the right area. Then, the user can introduce corrections of this text. The system will react to each correction, producing alternative hypotheses coherent with the user feedback. Once the user is satisfied with the modernization hypothesis, they can click in the "Validate" button to accept the hypothesis.

REFERENCES

[1] A. H. Toselli, L. A. Leiva, I. Bordes-Cabrera, C. Hernández-Tornero, V. Bosch, and E. Vidal, "Transcribing a 17th-century botanical manuscript: Longitudinal evaluation of document layout detection and interactive transcription," *Digital Scholarship in the Humanities*, vol. 33, no. 1, pp. 173–202, 2017.

[2] M. Domingo and F. Casacuberta, "Modernizing historical documents: A user study," *Pattern Recognition Letters*, vol. 133, pp. 151–157, 2020.

[3] J. Porta, J.-L. Sancho, and J. Gómez, "Edit transducers for spelling variation in old spanish," in *Proceedings of the workshop on computational historical linguistics*, 2013, pp. 70–79.

[4] M. Domingo, M. Chinea-Rios, and F. Casacuberta, "Historical documents modernization," *The Prague Bulletin of Mathematical Linguistics*, vol. 108, pp. 295–306, 2017.

[5] S. Sen, M. Hasanuzzaman, A. Ekbal, P. Bhattacharyya, and A. Way, "Take help from elder brother: Old to modern english nmt with phrase pair feedback," in *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, 2019, in press.

[6] C. Monk, "Custumale Roffense: An overview of the thirteenth-century custumal of Rochester Cathedral priory," https://www.themedievalmonk.com/the-rochester-customs-book.html, 2018.

[7] Á. Peris and F. Casacuberta, "Online learning for effort reduction in interactive neural machine translation," *Computer Speech & Language*, vol. 58, pp. 98–126, 2019.

[8] A. Baron and P. Rayson, "VARD2: A tool for dealing with spelling variation in historical corpora," *Postgraduate conference in corpus linguistics*, 2008.

[9] Y. Scherrer and T. Erjavec, "Modernizing historical slovene words with character-based smt," in *Proceedings of the Workshop on Balto-Slavic Natural Language Processing*, 2013, pp. 58–62.

[10] M. Bollmann, "Normalization of historical texts with neural network models," Ph.D. dissertation, Sprachwissenschaftliches Institut, Ruhr-Universität, 2018.

[11] M. Domingo and F. Casacuberta, "Enriching character-based neural machine translation with modern documents for achieving an orthography consistency in historical documents," in *Proceedings of the International Workshop on Pattern Recognition for Cultural Heritage*, 2019, pp. 59–69.

[12] A. Peris and F. Casacuberta, "NMT-Keras: a Very Flexible Toolkit with a Focus on Interactive NMT and Online Learning," *The Prague Bulletin of Mathematical Linguistics*, vol. 111, pp. 113–124, 2018.