Simplified Access to Spanish Historical Newspapers: From Transcription to Visualization

Ana Garcia-Serrano
ETSI Informática
Universidad Nacional de Educación a Distancia - UNED
Madrid, Spain
agarcia@lsi.uned.es

ABSTRACT

The paper shows the first approaches taken to simplify historical texts using lexical and sentence simplification tasks. A simplification task is understood as the translation of a text into an expression with the same meaning as the original text and of any length. We are dealing with a corpus of historical newspapers written in Old Spanish and transcribed with different kinds of errors. We use ChatGPT as a support to obtain lexical and sentence simplification proposals for transcribed texts. The main goal of the project is to develop a new corpus to train a refined Language Model through experimentation and tackle the proposed problem of Old Spanish Automatic SentenceSsimplification.

KEYWORDS
Text Simplification, Natural Language Processing, Language Models, Transkribus, Artificial Intelligence Tools.

1. INTRODUCTION

Automatic Text Simplification (AST) is a natural language processing (NLP) task that aims to transform a text from one language to another or to a simpler language. Corpus-based linguistic research has been mainly dominated by statistical and neural models [9], although in the last five years, generative artificial intelligence (AGI) techniques have been used, driven by the emergence of Large Language Models (LLMs) [4, 5].

With this new paradigm, systems, now called models, learn from large amounts of text available mainly from large companies. Once a model is trained for a specific language and task, its use in other tasks or languages requires fine-tuning or transfer learning with much smaller datasets than the initial ones and with fewer computational resources [3, 8, 11, 14].

The textual information we consume daily is presented in a style and vocabulary that varies significantly depending on the source, the domain, and the intended communication goal [2,6]. For example, the language used on social media platforms is very different from that used in newspapers, academic journals, or medical insurance policies. Everyone has the right to access and understand information relevant to their life or that interests them.

The case study to be presented [12] comes from the work carried out in the CLARA-HD project 1 which addresses the understanding of historical texts from the newspaper El Diario de Madrid (DM), published from 1788 to 1825, digitized by the National Library of Spain (BNE)2.

The case study will show the steps and resources needed in an approach based on simplification, understood as the translation of a text to an expression with the same meaning as the original text (in Old Spanish) and of any length. This includes the use of AI-based tool, technology will be discussed, such as the Transkribus3 for the construction of transcribed corpus (see figure 1).

Figure 1: Visualization of the historical text (original) and the corresponding automatically transcribed text.

2. SIMPLIFICATION APPROACH: OUTLINE

We used ChaGPT both to lexical simplification and simplified sentences proposal. For the lexical simplification, the task to obtain a list of words candidates to simplify a given word, we use English and Spanish prompts as for example, (1) English Prompt: Give me ten simplified Spanish synonyms for the following word: “propiciado”, and (2) Spanish Prompt: Dame diez sinónimos mas sencillos en español para la palabra: “propiciado” [13]. In this case a previous step was needed to identify the words to be simplified. For the next step we selected experimentally the pre-trained language model necessary to obtain adequate and effective results in a lexical simplification task. We also performed complex comparisons of the tests carried out using

1 https://clara-nlp.uned.es/home/dh/
2 https://hemerotecadigital.bne.es/hd/card?oid=0001510462
3 https://www.transkribus.org/
available technology based on instructions (prompts) [1, 10, 13, 15]. From these tests we can conclude that the new linguistic models incorporated into GPT-3, especially the gpt-3.5-turbo model, significantly improve the performance of the previous ones. Our approximations manage to obtain 77% accuracy (ACC@1) compared to 65% of Baseline [1], despite not incorporating combinations of prompts. This result implies that the first substitute proposed by the model is found in the list of correct results in 77% of the cases, a significant result.

Mainly given the specificity of the DM corpus, the results were not accurate enough (see table 1) neither improve de baseline. So, we went into a new approach based on sentence simplification.

<table>
<thead>
<tr>
<th>Temp.</th>
<th>ACC@k</th>
<th>MAP@k</th>
<th>k=3</th>
<th>k=5</th>
<th>k=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>base</td>
<td>0.6521</td>
<td>0.4281</td>
<td>0.3239</td>
<td>0.1967</td>
<td></td>
</tr>
<tr>
<td>gpt-3.5-turbo</td>
<td>0.3</td>
<td>0.7771</td>
<td>0.5590</td>
<td>0.4187</td>
<td>0.2471</td>
</tr>
<tr>
<td>t-1s</td>
<td>0.5</td>
<td>0.7635</td>
<td>0.5394</td>
<td>0.4022</td>
<td>0.2371</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.7527</td>
<td>0.5357</td>
<td>0.4025</td>
<td>0.2358</td>
</tr>
</tbody>
</table>

Table 1: Experiments comparison. Extracted from [13].

The approach for the sentence simplification started with a corpus development for the further tuning of a LLM, consisted in a two steps process (see figure 2): first a ChatGPT simplified sentence was obtained from the automatically transcribed original sentence (with possible errors) and after, a human expert constructed the correct simplified sentence according to the original one by modifying the ChatGPT proposal [7].


Figure 2: Visualization of the original text, the corresponding automatically transcribed text, the ChatGPT proposal and the human expert proposal for the sentence simplification.

3. FINAL DISCUSSION

Lexical Simplification did not obtain good results to produce simplified sentences from historical newspapers advertisements (See figure 3). After the corresponding evaluation, we started a new approach using ChatGPT help for the development of a corpus that will allow us the design and evaluation of a new LLM refined to the corpus at the CLARA-DH project. Early results show the potential success of the new approach.

ACKNOWLEDGMENTS

This work has been supported by the Spanish Government and funded under the CLARA-HD project (PID2020-116001RB-C32). We would like to acknowledge the support of Andrés Rodríguez-Francés and Fernando Chuquimmarca members, for a while, of the UNED research group NL&IR > DH.

REFERENCES