5th International Conference on Corpus Linguistics (CILC2013)

Splitting Complex Sentences for Natural Language Processing Applications: Building a Simplified Spanish Corpus

José Camacho Collados*

Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, 08290, Spain

Abstract

This paper presents a new Spanish parallel corpus of original and syntactically simplified texts. The simplification carried out basically consists of opportunistically splitting a complex original sentence into several simple ones. This parallel corpus is envisioned as a first step in order to create an automatic syntactic simplification system to be used as a preprocessing tool for other Natural Language Processing tasks such as Text Summarization, Information Extraction, parsing or Machine Translation. The corpus has been evaluated by human annotators regarding its grammaticality and preservation of meaning. The results suggest that the meaning of simplified and original sentences is almost identical.

© 2013 The Authors. Published by Elsevier Ltd. Open access under CC BY-NC-ND license. Selection and peer-review under responsibility of CILC2013.

Keywords: text simplification; syntactic simplification; parallel corpus; spanish; natural language processing.

1. Introduction

Text simplification (TS) consists of reducing the complexity of a text while preserving its meaning (Chandrasekar et al., 1996; Siddharthan, 2002 and 2004). It is mainly divided in two groups regarding the type of simplification considered: syntactic and lexical. This research presents a Spanish parallel corpus of original and syntactically simplified sentences which is aimed to be used as a tool for other Natural Language Processing (NLP) applications such as Text Summarization, Information Extraction, parsing or Machine Translation (MT). Simple sentences are usually much easier to process by NLP tasks and the simplification carried out along this research follows this idea.

* Corresponding author. Tel.: +34 637204734
E-mail address: josecamachocollados@gmail.com
It is mainly based on splitting complex sentences without damaging the original meaning and keeping its grammaticality.

This paper is organized as follows: Section 2 briefly introduces previous researches about the subject; Section 3 explains the simplification rules applied on this research; Section 4 presents the parallel corpus and its validation; Section 5 introduces several applications of syntactic simplification (SS); and Section 6 addresses some conclusions and ideas for future research.

2. Related work

2.1. Text simplification systems

Chandrasekar et al. (1996) was the first serious attempt of creating a general architecture for a TS system. The system is aimed to be used as a preprocessing tool to improve a parser and consequently, to improve other NLP applications such as Machine Translation and Text Summarization. He also introduces for the first time the possibility of using TS for Information Retrieval tasks. Apart from this aim, the system could be also used to make a text easier to understand for people with problems at reading. As far as improving a full parser is concerned, the main issue of this TS system was to reduce the complexity of syntactic structures. This was achieved by opportunistically splitting the complex sentences into two or more simple ones.

After this first attempt, Siddharthan (2002, 2004 and 2011) developed a new architecture for a TS system following the first steps given by Chandrasekar et al. (1996). The applications for this new TS system are quite similar to the ones proposed in previous works but he introduces quite remarkable improvements for the automatic TS system, such as the introduction of a new stage: regeneration.

2.2. Corpus-based text simplification

There are also other works that addresses the need of a parallel corpus of original and simplified sentences (Petersen & Ostendorf, 2007; Aluisio et al. 2008; Specia et al., 2009; Specia, 2010). Obtaining such corpus could be quite useful for new approaches in TS. Firstly, it would be useful to carry out a deeper analysis of the task, which could lead to some new ideas or improvements of rule-based TS systems.

Specia et al. (2009) followed this direction and worked on building a Brazilian Portuguese parallel corpus of original and simplified sentences with both lexical and syntactic simplifications. The main goal of the corpus is to help people with low level of literacy or some other cognitive disabilities. The corpus consists of texts from a corpus of news which were simplified by a linguist expert in TS. This work also includes Simplification Annotation Editor, which could be used in future similar researches for other corpus on different languages.

Specia (2010) experimented afterwards with a quite simple corpus-based TS approach. The goal is the same as Specia et al. (2009), for people with difficulties at reading. She basically used a Statistical Machine Translation (SMT) method to deal with the TS problem and check the results. The SMT was carried out without making many changes, so that the approach could be easily improved by adapting the framework to the particular TS problem. The corpus used was taken from two Brazilian newspapers and was manually annotated by native Brazilian speakers depending on the type of simplification within the sentences.

2.3. Spanish Text Simplification

A few attempts have been made about text simplification for the Spanish language. However, all these previous attempts have been focused on making the text more accessible for people with difficulties at reading such as foreigners learning Spanish or aphasic readers. Therefore all the simplifications considered followed this direction and in this case, lexical simplification has been proved to be easier and more reliable than syntactic.

Bott & Saggion (2011a) studies the problem of TS for the Spanish language. The goal of Bott et al (2011a) is to create a text easy to read for people with learning disabilities, so it differs in this point from the main goal of this research. It is a preliminary study where they analyze a corpus of news and the respective simplified one. They
address the need of getting a parallel corpus in order to be able to use it for the creation of a reliable TS system. Bott & Saggion (2011b) followed the work from their last paper and explain in detail an algorithm to align a parallel corpus of news and their simplified ones at a sentence level. As there was not training data for task (there was no available manually aligned parallel corpora to use), they relied on unsupervised learning.

Drndarević et al. (2013) presents a two-component (syntactic and lexical) automatic text simplification system for the Spanish language in order to make the text easier to read for people with cognitive disabilities. The system managed to get simpler sentences without seriously damaging their grammaticality and preservation of meaning with the original sentences.

3. Methodology

The simplification considered in the corpus is basically reduced in splitting long and complex sentences in simple ones. Since this research is aimed to be used to build an automatic simplification system, all the rules have been carefully selected in order to achieve this goal in the future. Subsection 3.1 presents the split rules and subsection 3.2 handles the regeneration stage, which is a more complex stage of the syntactic simplification process.

3.1. Syntactic simplification split rules

To begin with, the first condition to split a sentence is the number of conjugated verbs. The sentences which are to be simplified must have at least two conjugated verbs. However, not all this kind of sentences will be simplified, as there are some kind of structures which do not allow splitting and preservation of meaning simultaneously. It is also important to notice that in most sentences there are different types of structures to be simplified simultaneously, which makes the task harder. Subsection 3.1.1 handles the coordinate sentences and subsection 3.1.2 focuses on the subordinate sentences.

3.1.1. Coordination:

The original sentence is split on the position of the coordination nexus or articulation point as called by Chandrasekar et al. (1996). It is usually suggested repeating some noun phrases in order to improve the understandability and a best processing by other NLP applications. Example (4) is an illustrative example of how the simplification is handled on coordinate sentences.

(1)  
  
\begin{enumerate}
  \item a. *Fishes swim in the sea and butterflies fly in the sky.*
  \item b. – *Fishes swim in the sea.*
  \item – *Butterflies fly in the sky.*
\end{enumerate}

3.1.2. Subordination:

The simplification is done by taking into consideration the subordination type. It is also worth noticing that there are certain kinds of subordination which do not allow a syntactic simplification without change of meaning. Therefore every case has been studied separately and some cases of subordination and its simplification are summarized below:

- **Non-restrictive relative clauses:** This kind of structure is split without adding any other element which might cause a mistreatment by other NLP applications. The relative clause begins with a comma followed by a connector such as *que* or *quien*. Example (2) is a simple and representative example for the Spanish language of this type of simplification.

(2)  
\begin{enumerate}
  \item a. *Juan, que es aún muy joven, consiguió el premio.*
  \item [Juan, who is still very young, got the prize.]
\end{enumerate}
b. – Juan es aún muy joven.
   [Juan is still very young.]
   – Juan consiguió el premio.
   [Juan got the prize.]

- **Effect:** This type of structure contains the cause-effect relation. They are connected by a conjunction which indicates the end of the cause and the beginning of the effect. Therefore, the splitting of the sentence will be done at the conjunction’s position. Another effect connector such as *therefore* in English is introduced at the beginning of the second sentence, as we can appreciate in Example (3). In Spanish the connector introduced will be “*Por lo tanto,*”.

(3)  

a. The cat ate poisoned food, so it died.

b. – The cat ate poisoned food.
   – Therefore, the cat died.

- **Causal:** The same kind of relation cause-effect appears in this structure. However, the cause is placed after the effect in this case. They are connected by a different causal connector such as *because* on the English language (*porque* on the Spanish language) - Example (4). The output sentences include the effect in the first position and the cause or reason in the second sentence.

(4)  

a. Dogs can’t fly because they don’t have wings.

b. – Dogs can’t fly.
   – Reason: They don’t have wings.

- **Indirect speech to direct speech:** This structure is currently simplified just on very specific cases. More precisely, in the cases where a communication verb such as *decir, comunicar or explicar* are followed by the relative *que* and a sentence concerning the communication of the speaker. Most of the communication verbs included for this task were extracted from *Diccionario combinatorio del español contemporáneo* (Bosque, 2004). The original sentence is split into two sentences, the second one introduced by “:”, as we can observe in Example (5):

(5)  

a. El jugador dijo que el presidente estuvo con el equipo antes del partido.
   [The player said that the president was with the team before the match.]

b. – El jugador dijo:
   [The player said:]
   – El presidente estuvo con el equipo antes del partido.
   [The president was with the team before the match.]

3.2. Regeneration

This is the stage concerning the last part of the simplification process. It was firstly introduced by Siddharthan (2002) in the creation of an automatic TS system. Once the original sentence has been split into several ones, there are still some unresolved issues such as the simplified sentences order and anaphoric references which need to be tackled in this last stage.
3.2.1. Anaphoric references and repetition of noun phrases

Since preserving the meaning of the original sentence is one of the main premises in TS, some noun phrases which should appear on an anaphoric way in the simplified sentences will be repeated instead and the use of pronouns will be considerably reduced. We can appreciate in Example (2) how the noun phrase “the cat” is repeated in both sentences. Although the readability of the sentence could be damaged, this will lead to a better understanding of the simplified sentences with no ambiguity. To improve the readability, these anaphoric references, albeit repeated, will appear as shortened as possible.

3.2.2. Sentence reordering

The split phase within the simplification process gives a set of new sentences with no order. Giving them a right order is in some cases essential to fully understand the message of the original sentence. Although most of the sentences keep the order as found in the original sentences, there are some cases where a sentence reordering is needed. This reordering is not a difficult task for a native Spanish but an automatic system might find it extremely difficult. Therefore further researches should include a suitable algorithm (maybe coming out using Machine Learning techniques) for this task.

4. Corpus

4.1. Creation of the parallel corpus

The corpus chosen as a reference was the AnCora Corpus (Taulé et al., 2008), which consists of Spanish newspaper texts annotated at syntactic and morphological level. Newspaper texts are quite representative of the natural language and complex enough for the simplification task. Once the corpus was transferred from the original XML format to text format, the sentences from the AnCora Corpus were manually simplified as explained in section 3 in separate text files. On this way a parallel corpus of original and simplified texts was built. As we can see in Fig. 1, the original sentences and the simplified ones are easily aligned, as each original sentence is separated from each other on a new line, which is respected on the simplified part (right part on Fig. 1) no matter how many new sentences have been created.

The corpus currently counts with 3000 original sentences and the respective simplified ones. It is already available for research purposes if required.

![Fig.1: Creation of the parallel corpus in text format.](image)

4.2. Parallel corpus validation

The evaluation of the corpus has been done regarding two issues (grammaticality and preservation of meaning), similar to the one used in Drndarević et al. (2013) for the evaluation of their SS system. To carry out this, six native
Spanish speakers (three of them linguists and three of them holding a non-related university degree) were required to fill three excel sheets. The first one was to evaluate the grammaticality of the original sentences from AnCora corpus; the second one about the grammaticality of the simplified sentences from AnCora Corpus; and the third one concerning the preservation of meaning of original and simplified sentences. To do so, thirty complex sentences were randomly selected from thirty different texts of the corpus.

Table 1: Corpus validation (grammaticality and preservation of meaning)

<table>
<thead>
<tr>
<th>Preservation of meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammaticality</td>
</tr>
<tr>
<td>Average</td>
</tr>
<tr>
<td>Positive (4-5)</td>
</tr>
<tr>
<td>Neutral (3)</td>
</tr>
<tr>
<td>Negative (1-2)</td>
</tr>
</tbody>
</table>

All the evaluations were done on a 1-5 scale. For the grammaticality measure, 1 means that sentence is completely ungrammatical and the 5 that the sentence is completely grammatical. For the preservation of meaning, 1 means that there is no preservation of meaning at all and 5 that the meaning of the simplified sentence and the original is identical. Table 1 shows the total average for each evaluation and from the results we can appreciate how grammaticality is not damaged on the simplified sentences (4.74 without simplification - 4.66 with simplification). A paired two-sample t-test at a 0.05 level suggests that the difference is not statistically significant (t(179)=1.513; p-value=0.132). There are even a few cases where the simplification actually improves the grammaticality of the original sentences.

Considering 4-5 as a positive attitude of the annotator towards the grammaticality/preservation of meaning, 3 a neutral attitude, and 1-2 a negative attitude, table 2 also shows the general results of the annotator’s attitude. The grammaticality and preservation of meaning from simplified sentences get an incredibly positive attitude by the annotator (over 97% in both cases).

Regarding the inter-rater reliability for the preservation of meaning task, a statistical test at a 0.05 level was carried out among the six annotators. The results (F(5,174)=1.577, p-value=0.169) show an inter-annotator agreement quite high.

5. Applications

Text Simplification applications are basically divided in two main groups. The first one uses TS in order to make a text accessible for certain people with problems in reading complex structures. In this group we can differentiate, for instance, language learners and people suffering some kind of language disability such as aphasia. People suffering this disorder become unable to read and understand some structures, so TS attempts to modify the text so that the complex structures disappear.

The second main group includes the TS systems aiming to improve other NLP applications. The simplification rules of this research were established following this direction. The applications which could benefit from this approach are Text Summarization, Information Extraction (Evans 2011), parsing or Machine Translation (Temnikova, 2012).

5.1. Text Summarization

The idea of TS used in Text Summarization is to reduce the information extracted by each sentence, keeping just the relevant one. Since Text Summarization is sometimes based on the information extracted from different sentences, having the information better divided in sentences might be extremely useful for the task (Chandrasekar et al., 1996; Siddharthan, 2002).
5.2. Information Extraction

Automatic Information Extraction systems usually work better when the complexity of a sentence is low. They are usually based on some kind of automatic parsers in order to extract the relevant information. Example (6) shows how SS could help store different types of information inside a complex sentence into several simpler sentences.

(6) a. Obama, who was born in Kenya, is currently the US president. (One sentence, two statements)

b. – Obama was born in Kenya. (One sentence, one statement)
– Obama is currently the US president. (One sentence, one statement)

5.3. Parsing

The goal of TS in parsing is due to the improvement of performance if given a short sentence as input. Longer sentences give more cases of ambiguity, so that TS aims to reduce the length of the sentences while preserving the meaning as much as possible (Chandrasekar et al. 1996; Siddharthan, 2002).

5.4. Machine Translation

SS is applied on the source language prior to the translation, as we can appreciate in Example (7). This example uses Spanish-Korean Google Translate as MT system. In this example, the MT system clearly works better on the simplified sentence.

(7) a. La policía, que últimamente ha sido duramente criticada, ha detenido a dos secuestradores.

Google Translate: . [Somebody (unknown) has been heavily criticized lately police have arrested two kidnappers.]

b. – La policía últimamente ha sido duramente criticada.

Google Translate: . [Police have been heavily criticized lately.]
– La policía ha detenido a dos secuestradores.

Google Translate: 경찰은 두 남치범을 잡기 했다. [Police have arrested two kidnappers.]

There are several reasons why SS may be beneficial for a MT system. Firstly, as we have seen in the previous subsection, parsing could be benefited by this SS system. Since parsing is usually an essential tool to preprocess an input sentence, especially for rule-based MT approaches, the MT system would be directly improved. Secondly, Rule-based MT systems tend to backfire when the complexity of a given sentence is quite high and work better when this complexity is low, as previous researches on controlled languages have proved (Kaji, 1999; Mitamura, 1999; Cardey et al., 2004). Therefore a preprocessing tool to reduce the complexity of a sentence such as a SS system could be quite appropriate. The split of the given sentence into several simple ones would lead to a better performance of these MT systems.

As far as Statistical Machine Translation (SMT) is concerned, the length of a sentence is inversely proportional to the accuracy of the system (Koehn, 2010). This is not only due to the increasing number of possible alignments in long sentences, but also to the reliability of the language model, which mainly relies on n-grams found on a monolingual text. The combination of these two factors makes a SMT system be quite more reliable in short and medium size sentences.
Last but not least is the fact that SS is carried out just on the source language no matter the target language. These kinds of monolingual approaches are quite requested nowadays because of a purely mathematical reason. The total number of language pairs (and consequently the MT systems to be developed) is much larger than the total number of languages in the world. The difference is quite significant, as the number of language pairs is from order $n^2$, being $n$ the total number of languages. That is the reason why the MT tasks tend to be divided as much as possible and focus on source and target language separately. Monolingual tasks can be used in many different MT systems in contrast to working on specific MT systems directly.

6. Conclusions and future work

This paper presents a manually created parallel corpus of original and simplified sentences which is envisioned as a tool to develop an automatic syntactic simplification system. This corpus may be used to carry out some extrinsic evaluations regarding the other NLP applications mentioned along the paper. Since the preservation of meaning between original and simplified sentences have been proved in the previous section, the corpus is already available to be used as a preprocessing tool for other NLP tasks such as Text Summarization, Information Extraction, parsing and Machine Translation.

Further researches should focus on the creation of this system, by using this parallel corpus either as a dataset for supervised Machine Learning or as a reference corpus for rule-based approaches. Especially important will be how to tackle the regeneration stage. It may be useful to have a human evaluation in order to decide how to handle the anaphoric references (whether keeping them or not), noun phrase repetitions and any other elements affecting the readability and preservation of meaning between sentences.

Acknowledgements

Thank you to Obra Social “La Caixa” for supporting my Eramus Mundus International Masters in Natural Language Processing and Human Language Processing.

References


