Automatic Recovery of Punctuation Marks and Capitalization Information for Iberian Languages

Fernando Batista\textsuperscript{1,2}, Isabel Trancoso\textsuperscript{1,3}, Nuno Mamede\textsuperscript{1,3}

\textsuperscript{1}L2F - Spoken Language Systems Laboratory - INESC ID Lisboa
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
http://www.l2f.inesc-id.pt/
\textsuperscript{2}DCTI – ISCTE - Institute of Science, Technology and Management, Portugal
\textsuperscript{3}IST – Technical University of Lisbon, Portugal
{fmmb,imt,njm}@l2f.inesc-id.pt

Abstract

This paper shows experimental results concerning automatic enrichment of the speech recognition output with punctuation marks and capitalization information. The two tasks are treated as two classification problems, using a maximum entropy modeling approach. The approach is language independent as reinforced by experiments performed on Portuguese and Spanish Broadcast News corpora. The discriminative models are trained for a language using spoken and written corpora from that language. This paper provides the first results on Spanish Broadcast News data and the first comparative study between Portuguese and Spanish, on this subject.

Index Terms: Rich Transcription, Capitalization, Punctuation marks, Speech processing

1. Introduction

The text produced by a standard speech recognition system consists of raw single-case words, without punctuation marks, with numbers written as text, and with many different types of disfluencies. The missing information makes this representation format hard to read and understand [1], and pose problems to further automatic processing. Capitalization is important for improving human readability, parsing, and NER (Named Entity Recognition). Punctuation marks, or at least sentence boundaries, are important for parsing, information extraction, machine translation, extractive summarization and NER.

These tasks are important modules of the Broadcast News (BN) processing system developed at our lab, which integrates several other core technologies, in a pipeline architecture: single detection, audio segmentation, automatic speech recognition (ASR), topic segmentation and indexation, and summarization. The first modules of this system, including punctuation and capitalization, were optimized for on-line performance, given their deployment in the fully automatic subtitling system that is running on the main news shows of the public TV channel in Portugal, since 2008 [2]. This BN processing chain was originally developed for European Portuguese, but was already ported to other varieties of Portuguese (Brazilian and African). The goal of the current work was to port the punctuation and capitalization modules to Spanish, a language for which we recently developed our ASR system, thereby supporting the language independence of our approaches. This paper provides the first results on Spanish BN data and the first comparative study between Portuguese and Spanish, concerning this subject.

This paper is organized as follows: Sections 2 and 3 describe the related work and the adopted approach. Section 4 presents experimental results concerning automatic punctuation and capitalization of Portuguese and Spanish. Section 5 presents the final remarks and the future work.

2. Related work

Whereas speech-to-text core technologies have been developed for more than 30 years, the metadata extraction/annotation technologies are receiving significant importance only in the recent years. For example, [3] contains an entire section dedicated to this subject, while this topic is only briefly mentioned in the first version of this book, published in 2000. Producing rich transcripts usually involves the process of recovering structural information and the creation of metadata from that information. Recovering punctuation marks and capitalization are two relevant MDA (Metadata Annotation) tasks, which contribute to enriching the final recognition output.

The first joint initiatives concerning automatic rich transcription of speech started around 2002. The five year project DARPA-sponsored EARS program supported the goal of advancing the state-of-the-art in automatic rich transcription of speech. The NIST RT evaluation series\textsuperscript{1} is another important initiative that supports some of the goals of the EARS program, providing means to investigate and evaluate STT (speech-to-text) and MDE (Metadata Extraction) technologies, and promote their integration. Nevertheless, despite the emerging rich transcription efforts, only a few of the most important MDE tasks are covered by these evaluation plans.

Two different rich transcription methods are proposed and evaluated by [4]. The first method consists of adapting the ASR system for dealing with both punctuation and capitalization. This is done by duplicating each vocabulary entry with the possible capitalized forms, modeling the full-stop with silence, and training with capitalized and punctuated text. The second method consists of using a ruled-based NE tagger and punctuation generation. The paper shows that the first method produces worse results, due to the distorted and sparser language model (LM), suggesting a separation between the recognition process and the enriching tasks. The rest of this section describes in more detail the previous work related to each one of the tasks.

\textsuperscript{1}http://www.nist.gov/speech/tests/rt/
2.1. Punctuation

Different punctuation marks can be used in spoken texts, including: comma; period or full stop ; ... marks and capitalization for the Portuguese and Spanish languages. The evaluation is performed using the performance model \( T \) using huge amounts of data can be found in [11]. Approaches can be found in [18]. A recent study on the impact of (CRF). A study comparing generative and discriminative approaches consists of a cascade of different simple position heuristics. Other approaches include Maximum Entropy period, which consists of a cascade of different simple position heuristics. A similar approach was described in [11], showing that using larger training data sets lead to improvements in performance. [12] describes a maximum entropy (ME) based method for inserting punctuation marks into spontaneous conversational speech, which covers comma, full stop, and question mark. Bigram-based features, combining lexical and prosodic features, achieve the best results on the ASR output.

The work conducted by [6] and [7] uses a general HMM framework that allows the combination of lexical and prosodic clues for recovering punctuation marks. A similar approach was also used for detecting sentence boundaries by [8, 9, 10]. A study, using purely text-based n-gram language models, can be found in [11], showing that using larger training data sets lead to improvements in performance. [12] describes a maximum entropy (ME) based method for inserting punctuation marks into spontaneous conversational speech, which covers comma, full stop, and question mark. Bigram-based features, combining lexical and prosodic features, achieve the best results on the ASR output.

2.2. Capitalization

The capitalization task, also known as truecasing [13], consists of assigning the proper case information to each input word, which may depend on the context. Proper capitalization can be found in many information sources, such as newspaper articles, books, and most of the web pages. Besides improving the readability of texts, capitalization provides important semantic clues for further text processing tasks. The capitalization is not usually considered as a topic by itself. A typical approach, when dealing with processes where capitalization is expected, consists of modifying the process that usually relies on case information in order to suppress the need of that information [14]. An alternate approach is to previously recover the capitalization information, which can also benefit other processes that use case information.

A common approach for capitalization relies on n-gram LMs estimated from a corpus with case information [13, 4]. Another approach consists of using a rule-based tagger, as described in [15], which was shown to be robust to speech recognition errors, while producing better results than case sensitive language modeling approaches. [16] describes an approach to the disambiguation of capitalized words where capitalization is expected, such as the first word of the sentence or after a period, which consists of a cascade of different simple position heuristics. Other approaches include Maximum Entropy Markov Models (MEMM) [17] and Conditional Random Fields (CRF). A study comparing generative and discriminative approaches can be found in [18]. A recent study on the impact of using huge amounts of data can be found in [11].

3. Approach description

The same approach is used for the punctuation and capitalization tasks, which can be treated as two classification tasks. Our experiments use a discriminative approach, based on maximum entropy (ME) models, which provide a clean way of expressing and combining different aspects of the information. This is especially useful for the punctuation task, given the broad set of lexical, acoustic and prosodic features that can be used. This approach requires all information to be expressed in terms of features causing the resultant data file to become several times larger than the original one. On the other hand, the memory required for training with this approach increases with the size of the corpus (number of observations). This constitutes a problem, making it difficult to use large corpora for training. However, the classification is straightforward, making it interesting for on-the-fly usage.

Capitalization models are usually trained using large written corpora, which contain the required capitalization information. The consequent memory problem is solved by splitting the corpus into several subsets, and then iteratively retraining with each one separately. The first subset is used for training the first ME model, which is then used to provide initial weights for the next iteration over the next subset. This process goes on until all subsets are used. Although the final ME model contains information from all corpora subsets, events occurring in the latest training sets gain more importance in the final model. As the training is performed with the new data, the old models are iteratively adjusted to the new data. This approach provides a clean framework for language dynamics adaptation: (1) new events are automatically considered in the new models; and (2) with time, unused events slowly decrease in weight [19, 20].

Figure 1 illustrates the classification approach for both tasks. An updated capitalization lexicon containing the capitalization of new words and mixed-case words can be used as a complement for capitalization.

The experiments described in this paper use the MegaM tool [21] for training the ME models, which is open source and efficiently implements limited memory BFGS for multi-class problems (usually outperforms Iterative Scaling methods).

4. Experimental results

This section describes some experiments recovering punctuation marks and capitalization for the Portuguese and Spanish languages. The evaluation is performed using the performance
Table 1: Portuguese BN corpus properties.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>#Words</th>
<th>Duration</th>
<th>Planned</th>
<th>Spont.</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>477k</td>
<td>52h</td>
<td>54.6%</td>
<td>32.1%</td>
<td>11.3%</td>
</tr>
<tr>
<td>Devel</td>
<td>66k</td>
<td>7h</td>
<td>51.2%</td>
<td>37.6%</td>
<td>20.8%</td>
</tr>
<tr>
<td>Eval</td>
<td>135k</td>
<td>15h</td>
<td>55.6%</td>
<td>35.5%</td>
<td>20.3%</td>
</tr>
</tbody>
</table>

Table 2: Spanish BN corpus properties.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>#Words</th>
<th>Duration</th>
<th>Planned</th>
<th>Spont.</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>152k</td>
<td>15h</td>
<td>71.6%</td>
<td>10.6%</td>
<td>11.0%</td>
</tr>
<tr>
<td>Devel</td>
<td>25k</td>
<td>3h</td>
<td>72.5%</td>
<td>11.2%</td>
<td>17.2%</td>
</tr>
<tr>
<td>Eval</td>
<td>16k</td>
<td>2h</td>
<td>67.9%</td>
<td>14.7%</td>
<td>18.9%</td>
</tr>
</tbody>
</table>

Table 5: Capitalization results for the BN corpora.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Manual Transc.</th>
<th>ASR output</th>
<th>Newspapers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>SER</td>
</tr>
<tr>
<td>Portuguese</td>
<td>84.4</td>
<td>86.7</td>
<td>29.1</td>
</tr>
<tr>
<td>Spanish</td>
<td>94.7</td>
<td>85.6</td>
<td>19.0</td>
</tr>
</tbody>
</table>

4.1. Punctuation

The punctuation experiments use only the BN data, collected from broadcasted TV shows. Tables 3 and 4 show the results achieved for the Portuguese and Spanish data, respectively. The overall results are affected by the comma detection performance, which mostly achieves above 100% SER. The Portuguese punctuation experiments were conducted over BN data collected from 2000 to 2001, and containing about 148M words. The Spanish punctuation experiments were conducted with the content of online text, daily collected since 2003, and containing about 79M words. The original texts were preprocessed by the capitalization module, which has been integrated in a speech recognition system, currently being used to cover the Portuguese and Spanish spontaneous speech.

4.2. Capitalization

The capitalization experiments assume that the capitalization of the first word of each sentence is performed in a separated processing stage (e.g. after punctuation), since its correct graphical form depends on its position in the sentence. Our experiments consider four ways of writing a word: lower-case, first-capitalized, all-upper, and mixed-case (e.g. “McGyver”).

5. Conclusions

This paper presents a language independent approach for recovering punctuation marks and capitalization over speech data. Experiments were conducted over Portuguese and Spanish BN corpora. The described approach is now implemented by two punctuation and capitalization modules, which have been integrated in a speech recognition system, currently being used to cover the Portuguese and Spanish spontaneous speech.

We plan to port the punctuation and capitalization modules to other languages, for which we recently developed our ASR system, such as English and Brazilian Portuguese. We are currently trying to further improve the performance of the punctuation module by introducing prosodic features, besides the cur-
Table 3: Punctuation results for the Portuguese BN corpus.

<table>
<thead>
<tr>
<th>Focus</th>
<th>Manual Transcripts</th>
<th>Automatic Transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec Rec SER Comma ALL</td>
<td>Prec Rec SER Comma ALL</td>
</tr>
<tr>
<td>All</td>
<td>87.0 67.4 42.7 51.2 31.3 98.5 71.4 49.5 61.7</td>
<td>76.9 58.9 58.9 43.5 23.5 107.0 63.1 41.2 73.9</td>
</tr>
<tr>
<td>Planned</td>
<td>87.9 67.3 41.9 52.8 31.4 96.6 72.8 49.3 59.5</td>
<td>82.3 58.1 54.3 48.0 23.5 102.0 68.3 40.9 69.0</td>
</tr>
<tr>
<td>Spontaneous</td>
<td>85.3 71.8 40.5 49.4 28.9 100.7 69.9 49.5 66.2</td>
<td>68.0 62.2 67.1 32.7 20.2 121.3 53.0 40.3 86.9</td>
</tr>
</tbody>
</table>

Table 4: Punctuation results for the Spanish BN corpus.

6. Acknowledgements

The authors would like to thank to Hugo Meinedo and Helena Moniz for their useful hints and Raquel Martinez for their support with the Spanish corpus. This work was funded by FCT projects PTDC/PLP/72404/2006 and CMU-PT/HuMach/0039/2008. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

7. References


