Automatic Recovery of Punctuation Marks and Capitalization Information for Iberian Languages

Fernando Batista[1,2], *Isabel Trancoso*[1,3], *Nuno Mamede*[1,3]

[1]L^2F - Spoken Language Systems Laboratory - INESC ID Lisboa
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
http://www.l2f.inesc-id.pt/
[2]DCTI – ISCTE - Institute of Science, Technology and Management, Portugal
[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: jingle 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[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.

[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*; *exclamation mark*; *question mark*; *colon*; *semicolon;* and *quotation marks*. However, most of these marks rarely occur and are quite difficult to automatically insert or evaluate. Hence, most studies focus either on *full stop* or in *comma*, which have much higher corpus frequencies.

Comma is the most frequent punctuation mark, but it is also the most problematic because it serves many different purposes. It can be used to: introduce a word, phrase or construction; separate long independent constructions; separate words within a sentence; separate elements in a series; separate thousands in a number; and also to prevent misreading. [5] describes a method for inserting *commas* into text, and presents a qualitative evaluation based on the user satisfaction, concluding that the system performance is qualitatively higher than the sentence accuracy rate would indicate.

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 positional 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].

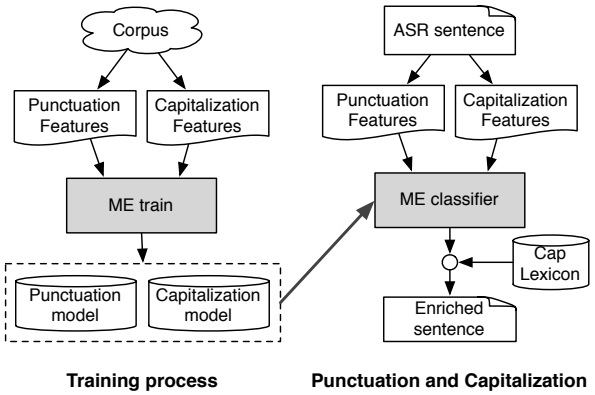

Figure 1: Rich transcription tasks block diagram.

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 specially 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 multiclass 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

	#Words	Duration	Planned	Spont.	WER
Train	477k	52h	54.6%	32.1%	11.3%
Devel	66k	7h	51.2%	37.6%	20.8%
Eval	135k	15h	55.6%	35.5%	20.3%

Table 1: Portuguese BN corpus properties.

	#Words	Duration	Planned	Spont.	WER
Train	152k	15h	71.6%	10.6%	11.0%
Devel	25k	3h	72.5%	11.2%	17.2%
Eval	16k	2h	67.9%	14.7%	18.9%

Table 2: Spanish BN corpus properties.

metrics: Precision, Recall and SER (Slot Error Rate) [22]. Only capitalized words and punctuation marks are considered as slots and used by these metrics. Hence, for example, the SER for the capitalization task is computed by dividing the number of capitalization errors by the number of capitalized words in the reference data.

Tables 1 and 2 show details of Portuguese and Spanish BN corpora subsets, respectively, which were used for training and evaluation. The Portuguese corpus is a subset of the BN European Portuguese Corpus, collected during 2000 and 2001. The Spanish BN corpus is a recent corpus, collected during 2008 and 2009. The manual orthographic transcription of these corpora provides the reference data, and includes punctuation marks and capitalization information. For each corpus we had access not only to the manual transcription, but also to the automatic transcription. Whereas the manual transcriptions already contain reference punctuation and capitalization, this is not the case of the automatic transcriptions. The required reference was produced by means of word alignments between the manual and automatic transcription. The higher WER (Word Error Rate) of the Portuguese corpus may be attributed to the larger proportion of spontaneous speech, as well the higher complexity of the Portuguese phonological system.

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 evaluation data was annotated by different people, using possibly different criteria, which explains the higher SER values for the *comma*. Results are strongly affected by the presence of recognition errors, as shown in the performance difference between manual and automatic transcripts. The Spanish corpus contains only a small portion of spontaneous speech, which causes less impact on the overall results.

The following features are used for a given word w in the position i of the corpus: w_i, w_{i+1}, $2w_{i-2}$, $2w_{i-1}$, $2w_i$, $2w_{i+1}$, $3w_{i-2}$, $3w_{i-1}$, p_i, p_{i+1}, $2p_{i-2}$, $2p_{i-1}$, $2p_i$, $2p_{i+1}$, $3p_{i-2}$, $3p_{i-1}$ (lexical), $GenderChgs_1$, $SpeakerChgs_1$, and $TimeGap_1$ (acoustic), where: w_i is the current word, w_{i+1} is the word that follows and $nw_{i\pm x}$ is the n-gram of words that starts x positions after or before the position i; p_i is part-of-speech of the current word, and $np_{i\pm x}$ is the part-of-speech n-gram for words starting x positions after or before the position i.

Corpus	Manual Transc.			ASR output			Newspapers		
	Prec	Rec	SER	Prec	Rec	SER	Prec	Rec	SER
Portuguese	84.4	86.7	29.1	73.2	77.7	50.4	93.8	86.5	19.0
Spanish	94.7	85.6	19.0	77.6	74.1	47.1	95.1	83.2	20.8

Table 5: Capitalization results for the BN corpora.

$GenderChgs_1$, and $SpeakerChgs_1$ correspond to changes in speaker gender, and speaker clusters; $TimeGap_1$ corresponds to the time period between the current and following word. For the moment, only lexical and acoustic features are being used in this task. Nevertheless, prosodic features, which already proved useful for this task, will be included in future experiments. All the existing punctuation marks were replaced by a *full stop* or a *comma*: ".": ";", "!", "?", "..." => *full stop*; ",", "-" => *comma*.

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"). The following features were used for a given word w in the position i of the corpus: w_i, $2w_{i-1}$, $2w_i$, $3w_{i-2}$, $3w_{i-1}$.

The Portuguese capitalization model was trained with a newspaper corpus collected from 1999 to 2004 and containing about 148M words. The Spanish capitalization model was trained with the content of online text, daily collected since 2003, and containing about 79M words. The original texts were normalized and all the punctuation marks removed, making them close to speech transcriptions. Only data previous to the evaluation data period was used for training.

The retraining approach described in Section 3 was followed, and the most recent capitalization model was used for processing each evaluation subset. Table 5 shows the corresponding results for both languages. While the performance is similar for written corpora in both languages, there is a significant difference for the speech data, where the performance is better for Spanish. One important explanation is related with the small portion of the Spanish spontaneous speech, which causes little impact on the overall results. Nevertheless, the worse performance for the Portuguese data is also due to the unusual topic covered in the news by that time (War on Terrorism). Many foreign names, which can be rarely found in the news, were used by that time.

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 daily process BN shows on-the-fly, for automatic subtitling.

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-

Focus	Manual Transcripts									Automatic Transcripts								
	Full stop			Comma			ALL			Full stop			Comma			ALL		
	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER
All	81.1	65.9	49.5	41.6	29.5	111.9	60.3	45.5	69.2	68.7	60.5	67.1	29.8	21.4	128.9	48.9	38.5	88.7
Planned	85.7	68.7	42.8	34.8	25.7	122.5	64.1	49.6	60.0	75.6	66.4	55.1	26.5	23.4	141.4	53.7	47.2	78.6
Spontaneous	71.5	59.7	64.1	46.7	32.6	104.6	55.2	40.9	79.8	53.2	47.3	94.3	33.0	20.2	120.7	40.8	28.3	101.6

Table 3: Punctuation results for the Portuguese BN corpus.

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

Table 4: Punctuation results for the Spanish BN corpus.

rent lexical and acoustic features.

6. Acknowledgements

The authors would like to thank to Hugo Meinedo and Helena Moniz for their useful hints and Raquel Martinez for her 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

[1] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman, "Measuring the readability of automatic speech-to-text transcripts," in *Proc. of Eurospeech*, pp. 1585–1588, 2003.

[2] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast news subtitling system in portuguese," in *Proc. of ICASSP 2008*, pp. 1561–1564, 2008.

[3] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, ch. 10 - Speech Recognition: Advanced Topics. Prentice Hall, 2008.

[4] J.-H. Kim and P. Woodland, "Automatic capitalisation generation for speech input," *Computer Speech & Language*, vol. 18, no. 1, pp. 67–90, 2004.

[5] D. Beeferman, A. Berger, and J. Lafferty, "Cyberpunc: a lightweight punctuation annotation system for speech," *Proc. of the ICASSP-98*, pp. 689–692, 1998.

[6] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," in *Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding*, pp. 35–40, 2001.

[7] J. Kim and P. C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," in *Proc. Eurospeech*, pp. 2757–2760, 2001.

[8] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," in *Proc. of the ISCA Workshop: ASR-2000*, pp. 228–235, 2000.

[9] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communications*, vol. 32, no. 1-2, pp. 127–154, 2000.

[10] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[11] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *ICASSP 2009*, (Taipei, Taiwan), 2009.

[12] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. of the ICSLP*, pp. 917 – 920, 2002.

[13] L. V. Lita, A. Ittycheriah, S. Roukos, and N. Kambhatla, "tRuEcasIng," in *Proc. of ACL-03*, pp. 152–159, 2003.

[14] E. Brown and A. Coden, "Capitalization recovery for text," *Information Retrieval Techniques for Speech Applications*, pp. 11–22, 2002.

[15] E. Brill, "Some advances in transformation-based part of speech tagging," in *AAAI '94: Proc. of the 12th national conference on Artificial intelligence*, vol. 1, pp. 722–727, 1994.

[16] A. Mikheev, "A knowledge-free method for capitalized word disambiguation," in *Proc. of the ACL-99*, pp. 159–166, 1999.

[17] C. Chelba and A. Acero, "Adaptation of maximum entropy capitalizer: Little data can help a lot," *Proc. of the EMNLP '04*, 2004.

[18] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, "Recovering capitalization and punctuation marks for automatic speech recognition: Case study for portuguese broadcast news," *Speech Communication*, vol. 50, no. 10, pp. 847–862, 2008.

[19] F. Batista, N. Mamede, and I. Trancoso, "Language dynamics and capitalization using maximum entropy," in *Proc. of ACL-08: HTL - Short Papers*, pp. 1–4, 2008.

[20] F. Batista, N. Mamede, and I. Trancoso, "The impact of language dynamics on the capitalization of broadcast news," in *Proc. of Interspeech 2008*, Sep. 2008.

[21] H. Daumé III, "Notes on CG and LM-BFGS optimization of logistic regression." http://hal3.name/megam/, 2004.

[22] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proc. of the DARPA BN Workshop*, 1999.