Recovering Capitalization and Punctuation Marks on Speech Transcriptions

Fernando Batista[1,2] and Nuno Mamede[1,3]

[1] L2F – INESC-ID, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
http://www.l2f.inesc-id.pt
[2] ISCTE-IUL, Av. das Forças Armadas, 1649-026 Lisboa, Portugal
[3] IST, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
Fernando.Batista@inesc-id.pt, Nuno.Mamede@inesc-id.pt

Thesis defended in May 2011
http://www.inesc-id.pt/pt/indicadores/Ficheiros/4467.pdf

Abstract. This work addresses two metadata annotation tasks, involved in the production of rich transcripts: automatic capitalization, and punctuation marks recovery. The main focus concerns broadcast news, using both manual and automatic speech transcripts. Different capitalization models were analysed and compared, and results support the ideia that generative approaches capture the structure of written corpora better, while the discriminative approaches are robust to ASR errors and suitable for dealing with speech transcripts. The so-called language dynamics has been addressed, and results indicate that the capitalization performance is affected by the temporal distance between the training and testing data. In what concerns the punctuation task, this study covers the three most frequent marks: *full stop*, *comma*, and *question mark*, combining lexical, acoustic, and prosodic information. Much of the research described here is language independent, but a special focus is given to the Portuguese language. This work provides the first evaluation results of these two tasks over European Portuguese broadcast news data.

Keywords: Rich transcription, capitalization, punctuation, Generative and discriminative methods, Language dynamics

1 Introduction

TV stations, radio broadcasters, and other media organizations are now producing large quantities of digital audio and video data on a daily basis. Automatic Speech Recognition (ASR) systems can now be applied to such sources in order to enrich them with additional information for applications, such as indexing, cataloging, subtitling, translation and multimedia content production. The ASR output consists of raw text, often in lowercase format, without any punctuation information, numbers are represented with words instead of symbols, and possibly containing different types of disfluencies. Even if useful for many applications, like indexing and cataloging, for other tasks, such as subtitling and

multimedia content production, the ASR output would benefit from other information. In general, enriching the speech output aims at enhancing information for a better human and machine processing.

Speech units do not always correspond to sentences, as established in the written sense. They may, in fact, be quite flexible, elliptic, restructured, and even incomplete. Taking into account this idiosyncratic behavior, the notion of *utterance* in [11] or *sentence-like unit* (SU) in [12] is often used instead of *sentence*. Detecting positions where a punctuation mark is missing, roughly[4] corresponds to the task of detecting a SU, or finding the SU boundaries. SU boundaries provide a basis for further natural language processing, and their impact on subsequent tasks has been analyzed in many speech processing studies [10, 16, 17].

The capitalization task, also known as truecasing, consists of assigning to each word of an input text its corresponding case information, which sometimes depends on its context. Proper capitalization can be found in many information sources, such as newspaper articles, books, and most of the web pages. Many computer applications, such as word processing and e-mail clients, perform automatic capitalization along with spell correction and grammar check. One important aspect related with capitalization concerns neologisms that are frequently introduced, and also archaisms. This so-called language dynamics are relevant and must be taken into consideration in what concerns capitalization.

This work addresses the tasks of recovering punctuation marks and capitalization, which are two metadata extraction tasks (MDE) that take part in the production of rich transcripts. Both tasks are critical for the legibility of speech transcripts, and are now gaining increasing attention from the scientific community. Consider for example the following extract "*the crisis was expected last may a man died...*". The extract does not provide any clue concerning the speaker pauses and intonation, making it very difficult to decide what has been said: "*The crisis was expected. Last May a man died*" or "*The crisis was expected last May. A man died*". These last two versions of the extract are easier to read and to process. Besides improving human readability, punctuation marks and capitalization provide important information for parsing, machine translation (MT), information extraction, summarization, Named Entity Recognition (NER), and other downstream tasks that are usually also applied to written corpora. Apart from the insertion of punctuation marks and capitalization, enriching speech recognition covers other important activities, such as speaker identification, and the detection and filtering of disfluencies, which are not covered by the scope of this work.

2 Motivation and Goals

While the speech-to-text core technologies have been developed for more than 30 years [9], metadata extraction/annotation technologies only have gained sig-

[4] *Roughly* because, for instance, units delimited by *commas* often do not correspond to sentences.

nificant importance during the latest years. The recent advances in the ASR systems, together with the increase of computational resources, make it now possible to process broadcast speech signals, continuously being produced by a number of broadcasters. The Portuguese recognition system, developed at the Spoken Language Laboratory (L^2F) [1], is a state-of-the-art ASR system, now being applied to different domains of the Portuguese language. It has been applied since the beginning of 2003 to the main TV news show, produced by the National Portuguese Broadcaster RTP. Nowadays, it is being used for processing the most important news shows, produced by all Portuguese TV broadcasters. The system performs two different tasks: live close-captioning, and multimedia content production for offline usage. The original content produced by this system was still difficult to read and to process, mainly because of the incorrect segmentation, and also because a number of basic information was missing, such as capital letters. Enriching the speech recognition output is an important asset for a speech recognition system that performs tasks like online captioning or produces multimedia content. Hence, the main motivation behind this thesis consists of producing enhanced transcripts, that can be applied to real life speech recognition systems. More specifically, the main target corresponds to correctly address the tasks of recovering punctuation marks and capitalization information, when dealing with speech transcripts, produced by an automatic speech recognition system. Accordingly, one of the expected outcomes is a prototype module, incorporating Rich Transcription tasks, for integration in the L^2F recognition system.

The output of a speech recognition system includes a broad set of lexical, acoustic and prosodic information, such as: time gaps between words, speaker clusters and speaker gender, which can be combined to produce the best results. On the other hand, the speech signal is also an important source of information when certain features, such as the pitch and energy, are not available on the recognition output. An important initial goal addressed by this thesis consisted of investigating and evaluating different punctuation and capitalization methods. An important requirement for a given method is its ability in combining all the available and relevant information. An additional outcome expected from this work is to better understand the individual contribution of each feature to the final performance, in both tasks.

Current computational resources allow the manipulation of large-sized data, and the application of complex learning methods on such data. On the other hand, we are now witnessing the mass production of online written content in the web. Different Portuguese newspaper companies are now publishing their news and last minute news content freely available on the web. This written corpus constitutes an important resource for processing the Portuguese language, and it also provides important basic information for speech processing. Given that only a limited set of manually labelled speech training data is now available for Portuguese, one of the main goals of this thesis consisted of using additional sources of information whenever possible, including large written corpora.

3 Strategy

This study considers both punctuation and capitalization tasks as two classification tasks, thus sharing the same approach. The approach is based on logistic regression classification models, a discriminative approach, corresponding to maximum entropy (ME) classification for independent events, firstly applied to natural language problems in [7]. Hence, both tasks were performed using ME models, a state-of-the-art approach that is suitable for dealing with speech transcripts, which includes both read and spontaneous speech, the latter being characterized by more flexible linguistic structures and by adjustments to the communicative situation [8]. The use of a discriminative approach facilitates the combination of different data sources and different features for modeling the data. It also provides a framework for learning with new data, while slowly discarding unused data, making it interesting for problems that comprise language variations in time, such as capitalization. With this approach, the classification of an event is straightforward, making it interesting for on-the-fly integration, with strict latency requirements.

The capitalization of a word depends mostly on the context where that word appears, and can be regarded as a sequence labeling or a lexical ambiguity resolution problem. The Hidden Markov Model (HMM) framework is a typical approach, used since the early studies, that can be easily applied to such problems. That is because computational models for sequence labeling or lexical ambiguity resolution usually involve language models built from n-grams, which can also be regarded as Markov models. For that reason, capitalization experiments reported include comparative results achieved using an HMM-based approach. Rather than comparing with other approaches, punctuation experiments focus on the usage of additional information sources, and the wide range of features provided by the speech data.

4 Experiments

Most of the experiments performed in the scope of this study aim at processing broadcast news data, but other information sources, like written newspaper corpora, have been used to complement the relatively small size of the speech corpora. The automatic transcripts for all the speech corpora were produced by the L^2F recognition system. The reference punctuation and capitalization for the automatic transcripts were provided by means of alignments between the manual and the automatic transcripts. This is not a trivial task because of the recognition errors. The data contains information coming from the speech recognition system, as well as other reference information coming from the manual transcripts. The word boundaries that have been previously identified automatically by the speech recognition system were adjusted by means of post-processing rules and prosodic features (pitch, energy and duration). The final content is also available as an XML file and contains not only pitch and energy, extracted directly from the speech signal, but also information concerning phones, syllable boundaries and syllable stress. Written corpora contains information that

is specially important for capitalization. The Portuguese corpora used in these experiments consist of online editions of Portuguese newspapers, collected from the web (at L^2F). Some of the data has been collected during this work time spam, allowing to perform experiments with the most recent data. The English written corpus is the North American News Text Supplement, available from the LDC. All the written corpora was normalized in order to be closer to speech transcripts, and therefore to be used for training speech-like models.

All the evaluation use standard performance metrics: Precision, Recall, F-measure and SER (Slot Error Rate) [13]. Concerning the capitalization task, only capitalized words (not lowercase) are considered as slots and used by these metrics. For the punctuation task, slots correspond to punctuation marks. Hence, for example, the punctuation SER is computed by dividing the number of punctuation errors by the number of punctuation marks in the reference, and corresponds to the NIST error rate for sentence boundary detection.

Our initial work on *capitalization recovery* revealed that generative methods produce better results for written corpora, while the ME-based approach works better with speech transcripts, also suggesting that the impact of the recognition errors is stronger for these generative approaches [3]. The following step in this study consisted of analysing the impact of language variation in the capitalization task. This was partly motivated by the daily BN Subtitling, which led the L^2F speech group to use a baseline vocabulary combined with a daily modification of the vocabulary [14] and a re-estimation of the language model. This dynamic language modeling provided an interesting scenario for our capitalization experiments. Maximum entropy models proved to be suitable to perform the capitalization task, specially when dealing with language dynamics. This approach provided a clean framework for learning with new data, while slowly discarding unused data. Most of the experiments performed compare the capitalization performance when performed both in written corpora and in speech transcripts. Individual results concerning manual and automatic transcriptions were also considered, revealing the impact of the recognition errors on this task. For both types of transcription, results show evidence that the performance is affected by the temporal distance between training and testing sets [5, 4], which is in agreement with other related work for NER [15]. Such conclusions led us to the proposal and evaluation of three different approaches for updating the capitalization module. The most promising approach consisted of iteratively retraining a baseline model with the new available data, using small corpora subsets, causing the performance to increase about 1.6% (absolute SER) when dealing with manual transcripts. Results reveal that capitalization models must be updated on a periodic basis [2]. The small improvements gained in terms of capitalization suggested that dynamically updated models may play a small role, but the updating does not need to be done daily. A number of recent experiments on automatic capitalization, reflecting the most recent training and testing conditions, with more accurate results, confirmed that an HMM-based approach is suitable for dealing with written corpora, but ME and CRFs achieved a better performance when applied to speech data. The effect of the language variation over

time was again studied for the English and Spanish data, confirming that the interval between the training and testing periods is relevant for the capitalization performance.

Concerning the experiments on the automatic recovery of punctuation marks, an exploratory work analysing the occurrence of the different punctuation marks for different languages has been performed. Such an analysis, considering both written corpora and speech transcripts, contributed to better understand the usage of each punctuation mark across languages. Results show that Portuguese broadcast news transcripts have a higher number of *commas* when compared with English and Spanish. The BN data contains a greater number of sentences and more intra-sentence punctuation marks when comparing to newspaper written corpora, confirming that speech sentences are shorter. Initial experiments were performed using lexical and acoustic features, firstly for basic sentence boundary detection, and then for discriminating the two most frequent punctuation marks: *full stop* and *comma* [3]. The initial results were improved by adding prosodic features, besides the existing lexical, time-based and speaker-based features; and by making use of punctuation information that can be found in large written corpora. Independent results were achieved for manual and automatic transcripts, allowing to assess the impact of the speech recognition errors on this task. Independent results were also achieved for spontaneous and planned speech. The contribution of each feature was analysed separately, making it possible to measure its influence on the automatic punctuation recovery. The punctuation module was then extended to obtain a better detection of the basic punctuation marks, *full stop* and *comma*, and also to deal with the *question mark*. Reported experiments were performed both on manual transcripts and directly over the automatic speech recognition output, using lexical, acoustic and prosodic features. Results pointed out that combining all the features usually conducts to the best performance.

5 Conclusions

This study addresses the tasks of recovering capitalization and punctuation marks from spoken transcripts, produced by ASR systems. These two practical RT tasks were performed using the same discriminative approach, based on maximum entropy, adequate for combining different data sources and features for characterizing the data and for on-the-fly integration, which is of great importance for tasks such as online subtitling, characterized by strict latency requirements. Reported experiments were conducted both over Portuguese and English BN data, allowing to compare the performance on the two languages [6]. Both force aligned and automatic transcripts experiments were used, allowing to measure the impact of the recognition errors.

Capitalized words and named entities are intrinsically related, and are influenced by time variation effects. For that reason, the so-called language dynamics have been analyzed for the capitalization task. The ME modeling approach provides a clean framework for learning with new data, while slowly discard-

ing unused data, making it interesting for addressing problems that comprise language variations in time. Language adaptation results clearly indicate, for both languages, that the capitalization performance is affected by the temporal distance between the training and testing data. Hence, our proposal states that different capitalization models should be used for different time periods. Capitalization experiments were also performed with an HMM-based tagger, a common approach in this type of problem. While the HMM-based approach captured the structure of written corpora better, the ME-based approach proved to be better suited for speech transcripts, which includes portions of spontaneous speech, characterized by a more flexible linguistic structure when compared to written corpora, and also more robust to ASR errors.

In what regards the punctuation task, this paper covers the three most frequent punctuation marks: *full stop*, *comma*, and *question mark*. Detecting *full stops* and *commas* is performed firstly, and corresponds to segmenting the speech recognizer output stream. *Question marks* are detected afterwards, making use of the previously identified segmentation boundaries. Rather than comparing with other approaches, reported punctuation experiments focused on the usage of additional information sources and diverse linguistic structures that can be found on the speech data. Two different scenarios were explored for improving the baseline results for *full stop* and *comma*. The first made use of the punctuation information that can be found in large written corpora. The second consisted of introducing prosodic features, besides the initial lexical, time-based and speaker-based features. The first scenario yielded improved results for all force aligned data, and for all the Portuguese data. The *comma* detection improved significantly, specially for Portuguese aligned data (7.8%). These findings support two basic ideas: results are better for Portuguese, because our English data is quite heterogeneous and has a higher WER; the most significant gains concerning *comma* derive from the fact that it depends more on lexical features. The second scenario provided even better results, for both languages and both punctuation marks, with improvements ranging from 3% to 8% (absolute). The best results were again achieved for Portuguese, but this time they are mainly related with *full stop*. We have concluded that, in both languages, the linguistic structure related with punctuation marks is being captured in different ways regarding the distinct punctuation marks: *commas* are being identified mostly by lexical features, while *full stops* depend more on prosodic ones. The most significant gains come from combining all the available features. As for *question marks*, there is a gain for the recognized Portuguese and for the aligned English data, but differences are not significant, due to the relatively small number of *question marks* in the corpora.

Acknowledgements

This work was funded by the FCT projects PTDC/PLP/72404/2006 and CMU-PT/HuMach/0039/2008, and partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and by DCTI, ISCTE-IUL.

References

1. R. Amaral, H. Meinedo, D. Caseiro, I. Trancoso, J. P. Neto. A prototype system for selective dissemination of broadcast news in European Portuguese. *EURASIP Journal on Advances in Signal Processing*, 2007(37507), May 2007.
2. F. Batista, R. Amaral, I. Trancoso, N. Mamede. Impact of dynamic model adaptation beyond speech recognition. *Proc. of the IEEE Workshop on Spoken Language Technology (SLT 2008)*, Goa, India, December 2008.
3. F. Batista, D. Caseiro, N. Mamede, I. Trancoso. Recovering capitalization and punctuation marks for automatic speech recognition: Case study for Portuguese broadcast news. *Speech Communication*, 50(10):847–862, 2008.
4. F. Batista, N. Mamede, I. Trancoso. The impact of language dynamics on the capitalization of broadcast news. *Proc. of the 9^{th} Annual Conference of the International Speech Communication Association (Interspeech 2008)*, Sep. 2008.
5. F. Batista, N. Mamede, I. Trancoso. Language dynamics and capitalization using maximum entropy. *Proc. of 46^{th} Annual Meeting of the Association of Computational Linguistics: Human Language Technologies (ACL-08): HLT, Short Papers*, strony 1–4. ACL, 2008.
6. F. Batista, H. Moniz, I. Trancoso, N. J. Mamede. Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. *IEEE Transaction on Audio, Speech and Language Processing*, Special Issue on New Frontiers in Rich Transcription, 2012.
7. A. L. Berger, S. A. D. Pietra, V. J. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.
8. E. Blaauw. *On the Perceptual Classification of Spontaneous and Read Speech*. Research Institute for Language and Speech, 1995.
9. S. Furui. 50 years of progress in speech and speaker recognition. *Proc. SPECOM 2005*, strony 1 – 9, Patras, Greece, October 2005.
10. M. Harper, B. Dorr, J. Hale, B. Roark, I. Shafran, M. Lease, Y. Liu, M. Snover, L. Yung, A. Krasnyanskaya, R. Stewart. Parsing and spoken structural event detection. *2005 Johns Hopkins Summer Workshop Final Report*, 2005.
11. D. Jurafsky, J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, wydanie second, 2009.
12. Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, M. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transaction on Audio, Speech and Language Processing*, 14(5):1526–1540, 2006.
13. J. Makhoul, F. Kubala, R. Schwartz, R. Weischedel. Performance measures for information extraction. *Proc. of the DARPA Broadcast News Workshop*, Herndon, VA, Feb. 1999.
14. C. Martins, A. Teixeira, J. P. Neto. Dynamic language modeling for a daily broadcast news transcription system. *Proc. of ASRU 2007*, 2007.
15. C. Mota. *How to keep up with language dynamics? A case study on Named Entity Recognition*. Praca doktorska, IST, Universidade Técnica de Lisboa, 2008.
16. J. Mrozinsk, E. W. Whittaker, P. Chatain, S. Furui. Automatic sentence segmentation of speech for automatic summarization. *Proc. of ICASSP'06*, 2006.
17. M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tür, M. Harper, D. Hillard, J. Hirschberg, H. Ji, J. G. Kahn, Y. Liu, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wang, C. Wooters. Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine*, 25(3):59–69, 2008.