Bilingual Experiments on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts

Fernando Batista, *Member, IEEE,* Helena Moniz, Isabel Trancoso *Fellow, IEEE,* and Nuno Mamede *Member, IEEE*

***Abstract*—This paper focuses on the tasks of recovering capitalization and punctuation marks from texts without that information, such as spoken transcripts, produced by automatic speech recognition systems. These two practical rich transcription tasks were performed using the same discriminative approach, based on maximum entropy, suitable for on-the-fly usage. Reported experiments were conducted both over Portuguese and English broadcast news data. Both force aligned and automatic transcripts were used, allowing to measure the impact of the speech 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 addressed for the capitalization task. Language adaptation results indicate, for both languages, that the capitalization performance is affected by the temporal distance between the training and testing data. In what regards the punctuation task, this paper covers the three most frequent punctuation marks: *full stop*, *comma*, and *question marks*. Different methods were explored for improving the baseline results for *full stop* and *comma*. The first uses punctuation information extracted from large written corpora. The second applies different levels of linguistic structure, including lexical, prosodic, and speaker related features. The *comma* detection improved significantly in the first method, thus indicating that it depends more on lexical features. The second method provided even better results, for both languages and both punctuation marks, best results being achieved mainly for *full stop*. As for *question marks*, there is a small gain, but differences are not very significant, due to the relatively small number of *question marks* in the corpora.**

***Index Terms*—Automatic Speech Processing, Rich Transcription; Capitalization, Punctuation Marks, Language Dynamics, Natural Language Processing**

I. Introduction

LARGE quantities of multimedia data are now being disseminated by TV stations, radio stations, newspapers, books, the Internet, and other communication means. The digital support makes it possible to use computers to analyze, learn and automatically process such data. Automatic Speech Recognition (ASR) systems are now being used daily to process radio and TV shows, in order to produce information for automatic indexing, cataloging, searching, and for on-line subtitling. Nonetheless, the text produced by a standard ASR system consists of raw single-case words without punctuation, which makes this representation format hard to read [1], and poses problems to further automatic processing.

Speech units do not always correspond to sentences, as established in the written sense. They may be quite flexible, elliptic, restructured, and even incomplete. Taking into account this idiosyncratic behavior, the notion of *utterance* [2] or *sentence-like unit* (SU) [3], [4] is often used instead of *sentence*. Detecting positions where a punctuation mark is missing, roughly[1] 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 [5], [6], [7].

The capitalization task, also known as truecasing [8], consists of assigning to each word of an input text its corresponding case information, which sometimes depends on its context. One important aspect related with capitalization concerns neologisms that are frequently introduced, and also archaisms. These so-called language dynamics are relevant and must be taken into consideration.

This paper addresses two rich transcription (RT) tasks: automatic capitalization, and punctuation recovery. Besides improving human readability, punctuation marks and capitalization provide important information for parsing, machine translation (MT), information extraction, summarization, Named Entity Recognition (NER), and further text processing tasks. Both tasks are performed using the same maximum entropy (ME) modeling approach, a discriminative approach, 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 [9]. 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

All the authors are with the Spoken Language Laboratory – L^2F, INESC-ID, Lisbon, Portugal (e-mail: {Fernando.Batista, Helena.Moniz, Isabel.Trancoso, Nuno.Mamede}@inesc-id.pt). The first author is also with Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal. The second one is also with FLUL/CLUL, University of Lisbon, Portugal. The third and fourth authors are also with IST, Lisbon, Portugal.

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

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 that can be easily applied to such problems. That is because computational models for sequence labeling or lexical ambiguity resolution usually involve language models (LM) built from n-grams, which can also be regarded as Markov models. For that reason, some capitalization experiments reported here include comparative results achieved using an HMM-based tagger. Rather than comparing with other approaches, reported punctuation experiments focus on the usage of additional information sources, and the wide range of features provided by the speech data.

The Broadcast News (BN) processing system, developed at the Spoken Language System Lab (L^2F) [10], integrates several core technologies, in a pipeline architecture: jingle detection, audio pre-processing, ASR, punctuation and capitalization, on-the-fly subtitling generation, topic segmentation and indexing, and summarization. The first modules of this system, including punctuation and capitalization, are 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 [11]. Most of the research reported here uses this processing chain, and is being used to improve the corresponding punctuation and capitalization modules.

This document is structured as follows: Section II covers the related work and the different approaches that have been used to tackle punctuation and capitalization. Section III describes the proposed approach. Section IV presents the data, and analyzes the vocabulary usage along time. Sections V and VI report experiments concerning capitalization and punctuation, respectively. Section VII concludes and presents future work.

II. Related Work

Recovering punctuation marks and capitalization are two relevant Metadata Annotation (MDA) tasks that involve the process of recovering structural information and the creation of metadata from that information. While the speech-to-text core technologies have been developed over more than 30 years, the metadata extraction/annotation technologies are receiving significant importance only during the latest years. For example, [2], published in 2009, contains an entire section dedicated to this subject (*Chapter 10 - Speech Recognition: Advanced Topics*), while this topic was only briefly mentioned in the first version of that book, published in 2000.

The first joint initiatives concerning automatic rich transcription of speech started around 2002, concomitantly with the DARPA-sponsored EARS program and the NIST RT evaluation series. One of the targets of the five year project EARS program was to advance the state-of-the-art in automatic RT. The NIST RT evaluation series[2] is another important initiative that supports some of the goals of the EARS program, providing means to investigate and evaluate speech-to-text (STT) and Metadata Extraction (MDE) technologies, and promote their integration.

A fair question for punctuation and capitalization is whether the ASR system can be adapted for dealing with both tasks, instead of creating additional modules. The work reported in [12] answers this question by proposing and evaluating two methods: i) adapting the ASR system for dealing with both punctuation and capitalization, by duplicating each vocabulary entry with the possible capitalized forms, modeling the full stop with silence, and training with capitalized and punctuated text, and ii) using a rule-based named entity tagger and punctuation generation. The paper shows that the first method produces worse results, due to the distorted and sparser language model, thus suggesting the separation of the punctuation and capitalization tasks from the speech recognition system.

Capitalization is not usually considered as a topic by itself. A typical approach consists of modifying the procedure that usually relies on case information in order to suppress the need for that information [13]. An alternate approach is to previously recover the capitalization information, thus benefiting other processes that use case information. A common approach for this problem relies on n-gram LMs estimated from a corpus with case information [12], [8], [14]. Another approach consists of using a rule-based tagger, as described in [15], which was shown to be robust to ASR errors, while producing better results than case sensitive language modeling approaches. [16] uses an approach based on Maximum Entropy Markov Models (MEMM), and studies the impact of using increasing amounts of training data as well as a small amount of adaptation data. A study comparing generative and discriminative approaches can be found in [17]. Experiments on huge corpora sets using different n-gram orders are performed in [14], concluding that using larger training data sets leads to increasing improvements in performance, but the same tendency is not achieved by using higher n-gram order LMs. Other related work includes bilingual capitalization models for capitalizing MT outputs, using Conditional Random Fields [18].

Most of the words and structures of a language are not subject to substantial diachronic changes. However, the usage of neologisms and archaisms introduces dynamics in the lexicon. This problem is addressed for Portuguese BN in the work of [19], which proposes a daily adaptation of the vocabulary and LM to the topic of current news, based on texts daily available on the Web. Also concerning this subject, [20] and [21] analyze the relation between corpora variation over time and the NER performance, proving that, as the time gap between training and test data increases, the performance of a named entity tagger based on co-training [22], [23] also decreases. These studies have shown that, as the time gap between corpora increases, the similarity between the corpora and the names shared between those corpora also decreases. The language adaptation problem concerning capitalization has been addressed by [24], [25], concluding that the capitalization performance is influenced by the training data period. All these studies emphasize the relation between named entities and capitalized words, showing that they are influenced by time variation effects.

Different punctuation marks can be used in spoken texts,

[2]http://www.nist.gov/speech/tests/rt/

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 *comma*, which have much higher corpora frequencies. *Comma* is the most frequent punctuation mark, but it is also the most problematic because of its multifunctionality, e.g., used to separate thousands in numbers and also used in different syntactic contexts. Punctuation marks are closely related with syntactic, and semantic properties. Thus, the presence/absence of a *comma* in specific locations may influence the grammatical judgments of the SUs. As synthesized by [26], *commas* should not be placed between: i) the subject and the predicate; ii) the verb and the arguments; iii) the antecedent and the restrictive relative clause; and iv) before the copulative conjunction *e/and*. Then again, *commas* should separate: i) adverbial subordinate clauses; ii) appositive modifiers; iii) parenthetical constituents; iv) anteposed constituents; v) asyndetically coordinated constituents; and vi) vocatives.

European Portuguese (EP), as other languages, has different interrogative types [27]: yes/no questions, alternative[3] questions, wh- and tag questions. A yes/no question requests a yes or no answer (*Estão a ver a diferença?/Can you see the difference?*). In EP they generally present the same syntactic order as a statement, contrarily to English that may encode the *yes/no* interrogative with an auxiliary verb and subject inversion. An alternative question presents two or more hypotheses (*Acha que vai facilitar ou vai ainda tornar mais difícil?/Do you think that it will make it easier or will it make it even harder?*) expressed by the disjunctive conjunction *ou/or*. A wh-question has a wh interrogative pronoun or adverb, such as *o que/what*, *quem/who*, *quando/when*, etc., corresponding to what is being asked about (*Qual é a ideia?/What is the idea?*). In a tag question, an interrogative clause is added to the end of a statement (*Isto é fácil, não é?/This is easy, isn't it?*).

In a previous study [28], we analyzed different corpora in order to see if the weight of the features was dependent on the nature of the corpus and on the most characteristic types of interrogatives in each. We concluded that the percentage of interrogatives was highly dependent on the nature of the corpus. For our map-task corpus, interrogatives represent 22.0% of all the punctuation marks, and similar values (20.4%) are found in a university lectures corpus – a proportion ten times more than in BN (2.1%). This difference is related not only to the percentage of interrogatives across different corpora, but also to their subtypes. In BN *yes/no* questions account for 47.0% of all interrogatives, *wh-questions* for 40.4%, while *tags* and *alternative* questions only 10.0% and 2.6%, respectively. These percentages compare well with the ones for newspapers, but not with the ones of the other corpora analyzed. The highest percentage of *tag* questions is found in the university lecture corpus (40.4%). Whereas the highest percentage of *yes/no* questions is reported in the map-task corpus (73.6%).

[29] 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. Concerning punctuation recovery, [30] and [31] report a general HMM framework that allows the combination of lexical and prosodic clues for recovering *full stop*, *comma*, and *question marks*. A similar approach was also used for detecting sentence boundaries by [32], [33], [4]. [31] also combines 4-gram LMs with a CART (Classification and Regression Tree) and concludes that prosodic information highly improves the results. [34] describes a ME-based method for inserting punctuation marks into spontaneous conversational speech, where the punctuation task is considered as a tagging task and words are tagged with the appropriate punctuation. It covers three punctuation marks: *comma*, *full stop*, and *question mark*; and the best results on the ASR output are achieved using bigram-based features and combining lexical and prosodic features. [35] proposes a multi-pass linear fold algorithm for sentence boundary detection in spontaneous speech, which uses prosodic features The paper focuses on the relation between sentence boundaries and their correlates, pitch breaks and pitch duration. Other recent studies have shown that the best performance for the punctuation task is achieved when prosodic, morphologic and syntactic information are combined [4], [7], [36].

III. Approach Description

Punctuation and capitalization tasks are treated here as two classification tasks, sharing the same approach. The approach is based on logistic regression classification models, which corresponds to the maximum entropy classification for independent events, firstly applied to natural language problems in [37]. This approach provides a clean way of expressing and combining different aspects of the information. A ME model estimates the conditional probability of the events given the corresponding features. Let us consider the random variable $y \in C$ that can take k different values, corresponding to the classes c_1, c_2, ... ,c_k. The ME model is given by the following equation:

$$P(c|d) = \frac{1}{Z_\lambda(F)} \times exp\left(\sum_i \lambda_{ci} f_i(c,d)\right)$$

determined by the requirement that $\sum_{c \in C} P(c|d)$=1. $Z_\lambda(F)$ is a normalizing term, used just to make the exponential a true probability, and is given by:

$$Z_\lambda(F) = \sum_{c' \in C} exp\left(\sum_i \lambda_{c'i} f_i(c',d)\right)$$

f_i are feature functions corresponding to features defined over events, and $f_i(c,d)$ is the feature defined for a class c and a given observation d. The index i indicates different features, each of which has associated weights λ_{ci}, one for each class. The ME model is estimated by finding the parameters λ_{ci} with the constraint that the expected values of the various feature functions match the averages in the training data. These parameters ensure the maximum entropy of the distribution

[3]In the literature, alternative questions may not be considered as a type of interrogative, rather as a subtype. For the sake of distinguishing alternative questions from disjunctive declarative clauses, we included the alternatives as well.

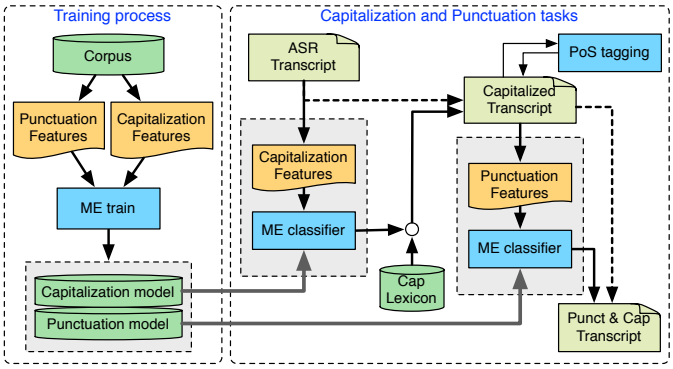

Figure 1. Block diagram of the capitalization and punctuation tasks.

and also maximize the conditional likelihood $\prod_i P(y^{(i)}|d^{(i)})$ of the training samples. Decoding is conducted for each sample individually and the classification is straightforward, making it interesting for on-the-fly usage. ME is a probabilistic classifier, a generalization of Boolean classification, that provides probability distributions over the classes. The single-best class corresponds to the class with the highest probability, and is given by:

$$\hat{c} = \underset{c \in C}{argmax}\, P(c|d)$$

The ME models used in this study are trained using the `MegaM` tool [38], which uses an efficient implementation of conjugate gradient (for binary problems) and limited memory BFGS (for multiclass problems).

A. Application to Rich Transcription

Figure 1 illustrates the classification approach for both tasks, where the left side of the picture represents the training process using a set of predefined features, and the right side corresponds to classification using previously trained models. An updated lexicon containing the capitalization of new and mixed-case words (e.g., "McGyver" is an example of a mixed-case word) can be used as a complement for producing the final capitalization form. Notice, however, that our evaluation results involve the classification only. As shown in the figure, capitalization comes first in the classification pipeline, thus producing suitable information for feeding a part-of-speech tagger. Subsequently, part-of-speech information is used to aid detecting the punctuation marks, corresponding to SU boundaries. The capitalization of the first word of each sentence is assigned in a post-processing step, based on the previously detected SU boundaries.

B. Training with Large Corpora

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. The memory required for training with this approach increases with the size of the corpus (number of observations). The `MegaM` tool, used in our experiments, requires the training to be performed in a single machine and using all the training data in a single step. That constitutes a problem when large corpora is used, which is the case of capitalization models that use large written corpora. The memory problem can be 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 values for the weights of the next iteration over the next subset. This process goes on, comprising several epochs, 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: i) new events are automatically considered in the new models; ii) the final model collects information from all corpora subsets; and iii) with time, unused events slowly decrease in weight [24], [25].

Table I
USER ANNOTATION AGREEMENT FOR THE PUNCTUATION MARKS IN THE PORTUGUESE BN CORPUS, IN TERMS OF COHEN'S KAPPA VALUES.

full stop	*comma*	*question marks*	*all punctuation marks*
0.890	0.557	0.870	0.705

IV. CORPORA AND EVALUATION METRICS

Experiments here reported comprise the Portuguese and English languages, and use data available for these languages. Speech transcripts and written corpora differ in many aspects, but they also share important information concerning punctuation marks and capitalization. Hence, large written newspaper corpora are used as a way of improving the transcript models.

A. Broadcast News Corpora

The Portuguese corpus used in these experiments is the speech recognition subset of the BN European Portuguese Corpus, collected during 2000 and 2001 [39]. The manual orthographic transcriptions of this corpus were recently revised by an expert linguist, thereby removing many inconsistencies in terms of punctuation marks that affected our previous results. The previous version of this corpus was manually transcribed by different annotators, who did not follow consistent criteria in terms of punctuation marks. The revision process focused mostly on correcting punctuation marks and on adding disfluency annotation [40]. In order to assess the user agreement between the original and the revised versions, in terms of punctuation marks, we have calculated Cohen's kappa values [41] for each punctuation mark. The corresponding results are shown in Table I, revealing that the most consistent punctuation marks are the *full stop* and the *question mark*. Most of the differences concern *comma*, and they are often due to different criteria when marking disfluencies. Since our previous data had no disfluency identification, the annotators often delimited the disfluency sequences with commas. Moreover, they also applied a naive criterion of corresponding a *comma* to a silent pause, even if that did not respect the syntactic structure.

The English BN corpus used in our experiments combines different corpora subsets, available from the Linguistic

Table II
BROADCAST NEWS CORPORA PROPERTIES.

Subset		#words	Dur	*full stop*	*comma*	*qmark*	WER
PT	Train	477k	46h	4.6%	6.8%	0.2%	14.1%
	Devel	66k	6h	5.0%	6.5%	0.3%	18.9%
	Eval	135k	18h	4.4%	7.0%	0.1%	18.5%
EN	Train	711k	81h	5.0%	3.4%	0.3%	15.5%
	Devel	66k	6h	5.4%	4.8%	0.3%	13.1%
	Eval	99k	9h	5.1%	4.7%	0.2%	21.6%

Data Consortium (LDC). The first 94% of the LDC1998T28 corpus (HUB4 1997 BN training data) was used for training and the rest was used for development. The first 80% of the LDC2005T24 corpus (RT-04 MDE Training Data Text/Annotations) was used for training, 10% for development, and the last 10% for evaluation. The evaluation data also includes the LDC corpora LDC2000S86/88 (HUB4 1998/99 BN evaluation) and LDC2007S10 (NIST RT03 evaluation data). Each subset has been produced in a different time period, built for different purposes, encoded with different annotation criteria, and available in different formats as well. Combining these heterogeneous corpora demanded normalization strategies specifically adapted for each subset.

Properties of the Portuguese and English BN corpora used in our experiments are shown in Table II. *Dur* values represent the duration of all speech sequences (silences not included). It is interesting to observe that the *comma* is the most frequent punctuation mark in the Portuguese corpora, while the *full stop* is the most frequent one in English. This is consistent with the widespread notion that sentences are longer in written Portuguese. The *question mark* is the third most frequent punctuation mark, still its frequency is quite residual.

The manual orthographic transcription of these corpora provides the reference data, and includes punctuation marks and capitalization information. The speech recognition system [11], a state-of-the-art system, was used to produce two transcription versions: force aligned and fully automatic transcripts. The words in the former are constrained to reference words and are therefore more accurate (e.g., the ASR system was unable to align in certain regions of overlapping speech, or dramatic cases of insufficient energy). The latter is more prone to contain misrecognized words, quantified by a recognition Word Error Rate (WER), as shown in Table II. Both versions of the transcripts include time marks for each unit of analysis. Whereas the manual transcripts already contain reference punctuation marks and capitalization, this is not the case in the automatic ones. The required reference was produced by means of word alignments between the manual and automatic transcripts. The alignment was performed using the NIST SCLite tool[4], followed by an automatic post-processing stage, for correcting possible SCLite errors and aligning special words which can be written/recognized differently (e.g. U.S.A. and USA). Both corpora were automatically annotated with part-of-speech information: MARv [42] was used for Portuguese, while TreeTagger [43] was used for the English data.

[4]available from http://www.nist.gov/speech.

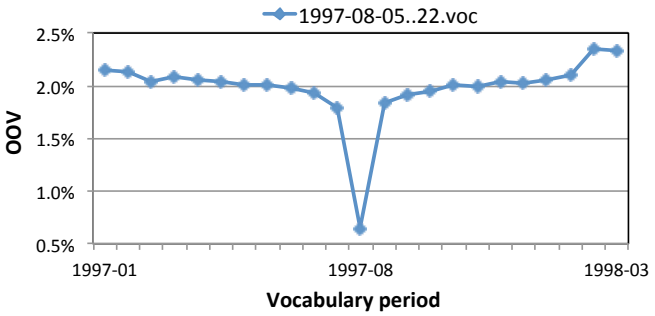

Figure 2. Vocabulary coverage on written newspaper corpora.

B. Newspaper Corpora

Written newspaper corpora are especially useful for training and improving the capitalization models. Nevertheless, they also provide important lexical information for punctuation detection in general, being of particular interest for some types of *question marks* that depend mostly on lexical information.

The Portuguese written corpus corresponds to on-line editions of the Portuguese Newspaper "Publico", collected from 1999 to 2004 and containing about 148 million words. The last two subsets of 2M words were used for development and evaluation, respectively. The English written corpus corresponds to the LDC corpus LDC1998T30 (North American News Text Supplement). For these experiments, only the NYT (New York Times) portion of the corpus was used. The data was collected from January 1997 to April 1998 and contains about 213M words, after cleaning the corpus and removing problematic text (unknown chars, etc). About 211M words were used for training, 574K for development, and 1.2M for evaluation.

The original texts were normalized, making them more appropriate for training models that can be used with speech transcripts. For the experiments here described, only data previous to the evaluation data period was used for training.

C. Analysis of the Language Variations over Time

Although the relation of time effects and punctuation conventions may be considered interesting, we conducted our time effect analysis exclusively for the capitalization task, since named entities are more prone to be influenced by short-time effects than punctuation conventions. This has to do with several reasons. Firstly, time effects in punctuation usually take into account texts from several decades (or even centuries), instead of short periods of time like the ones reported in our data. For instance, in 1838, Alexandre Herculano, a famous Portuguese writer[5], described punctuation conventions used in his time that are considered ungrammatical in contemporary Portuguese (e.g., a long subject is separated from the predicate by a comma) [26]. Secondly, changes in the conventional usages of punctuation marks, reported in recent years, are mainly associated with semicolon usage – a punctuation mark with residual frequencies across corpora (BN 0.2%; newspapers 0.7%; and university lectures 0.1%). Thirdly, punctuation is

[5]Alexandre Herculano, Opúsculo V, edição crítica de [critical edition by] J. Custódio and J. M. Garcia. Lisboa, Presença. 1986.

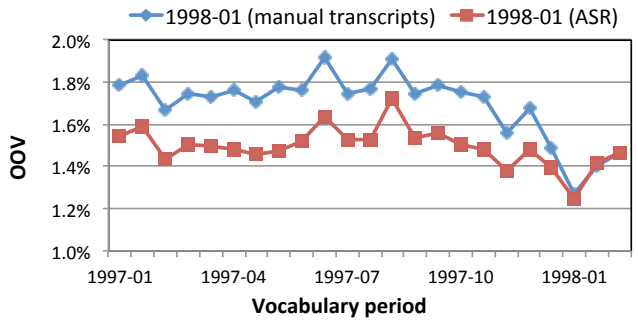

Figure 3. Vocabulary coverage for BN speech transcripts.

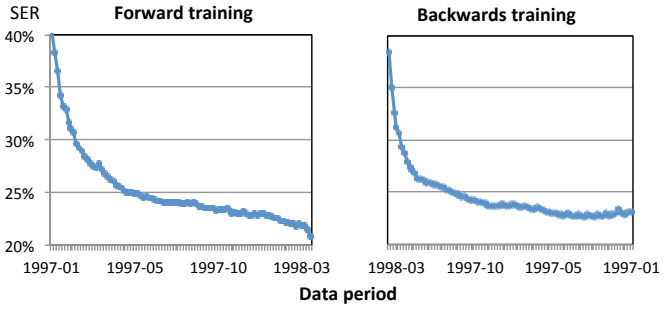

Figure 4. Forward and Backwards training results over written corpora.

diverse across corpora from the same period of time. However, that was not a major issue, since only BN is being analyzed.

In order to better understand the way we should train our capitalization models, we have started by analyzing the newspaper corpus for establishing a relation between the vocabulary usage and the time-line. The English corpus was split into several subsets, each containing about 8 million words. Each subset, comprising about 88K unique words, was named with the month corresponding to the first data in that subset. In order to assess the relation between the word usage and the language period, several vocabularies were created with the 50K most frequent words appearing in each set (roughly corresponds to a frequency greater than two). Then, the coverage of each vocabulary was checked against one of the subsets. Figure 2 shows the results for the chosen subset, containing data from August 1997, and located in the middle of the corpus time span. The best coverage is, as expected, achieved with the vocabulary built from the testing subset, but a more important result is that the number of OOVs (Out of Vocabulary Words) decreases as the time gap between the vocabulary and the testing period gets smaller.

The previous experiment was also performed on speech transcripts, by selecting a piece of speech data from the English BN corpus. Most of the English BN data does not have a reference to the corresponding collection time period, especially the evaluation subsets. Therefore, the coverage of each one of the previous 23 vocabularies was tested against a subset from the LDC1998T28 corpus, corresponding to January 1998. Figure 3 shows the corresponding results, highlighting that the coverage is better for vocabularies built from data collected nearby the testing data period. These results point to vocabulary changes across time, mainly because of the named entities. This subject is further addressed in the next section, where several capitalization experiments show how this affects the capitalization task.

Due to space limitations, the corresponding analysis for Portuguese is not performed here. However, the same relation between the vocabulary usage and the time-line was previously established for Portuguese written corpora [24], and for BN speech transcripts [25]. It would be interesting to compare the same period of time in both languages to measure the impact of new named entities, e.g., at the beginning of the Iraq war, or during U.S.A. presidential elections. That would depict the timeline effects of the same event on both languages. Unfortunately, we do not have data suitable to perform such experiments.

D. Evaluation Metrics

All the evaluation presented in this paper uses the performance metrics: Precision, Recall, F-measure and SER (Slot Error Rate) [44]. 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 capitalization SER is computed by dividing the number of capitalization errors (misses and false alarms) by the number of capitalized words in the reference, and the punctuation SER is computed by dividing the number of punctuation errors by the number of punctuation marks in the reference.

V. Capitalization Task

This paper assumes that the first word of each sentence is processed in a separated processing stage (e.g., after punctuation), since its correct graphical form depends on its position in the sentence. Thus, evaluation results do not consider them. Capitalization experiments here described discriminate between four capitalization classes, corresponding to four ways of writing a word: lower-case, first-capitalized, all-upper, and mixed-case (e.g., “McGyver”).

The capitalization models were trained with the previously described newspaper corpora, after removing all the punctuation marks. The retraining approach described in Section III-B was followed, with subsets of two million words each. Each epoch was retrained three times, using 500 iterations. For performance reasons, each capitalization model was also limited to 1.5 million weights. The following features were used for a given word w in the position i of the corpus: w_i, $2w_{i-1}$, $2w_i$, $3w_{i-2}$, $3w_{i-1}$, $3w_i$, 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.

In order to assess the impact of the language variations in time, we have used two different strategies for training, based on the data period. The first capitalization models were trained by starting with the oldest data available and by retraining each epoch with more recent data. The second set of capitalization models were trained backwards, using the newest data first and retraining each epoch with data older than the one used in the previous epoch. Figure 4 shows

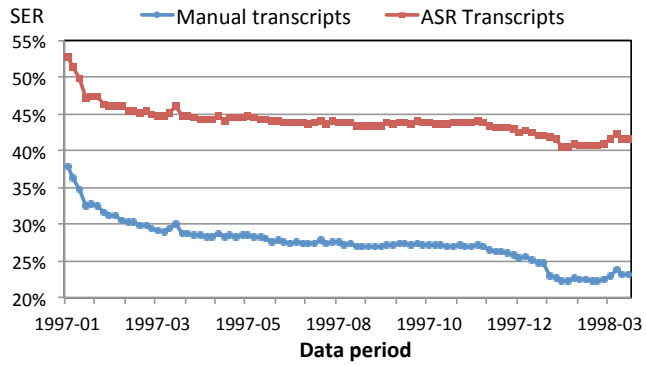

Figure 5. Results for the 1998-01 transcripts, using forward training.

results of applying each capitalization model to the newspaper corpora evaluation subset. While the performance of models trained with the forward strategy consistently increases, the performance of models produced with the backwards strategy does not increase after a certain number of epochs and even decreases. Although both experiments use the same training data, the best result is achieved with the last model created using the forward strategy, because the latest training data time period was closest to the evaluation time period. The small performance difference between the forward and backwards strategy is related to the relatively short period of time, less than one and a half year of data, covered by the English written corpus. During such a small period of time, the vocabulary changes are quite limited. Notice, however, that both results were achieved using exactly the same data, justifying the preference for the forward strategy.

Each one of the previous capitalization models, created using the forward strategy, has also been applied directly over speech transcripts. The evaluation was conducted over data collected during January 1998, extracted from the LDC1998T28 corpus, and corresponding to about 100k words. Figure 5 shows the results for manual and automatic transcripts, again revealing that the best models are the ones closest to the evaluation data period. In fact, the best model is the one built from data of the same period, despite the training and evaluation data being from different sources. The performance differences between manual and automatic transcripts reflect the WER impact. Another important conclusion arising from the two previous charts is that the amount of data is an important performance factor. In fact, our results show that the performance increases consistently as more data is provided. [24], [25] present the corresponding experiments for Portuguese. Due to space limitations, and because similar conclusions were achieved, this subject will not be further extended here for Portuguese.

The performance of the ME models was compared against the performance an HMM-based approach, commonly used for this task. An HMM-based tagger, implemented by the `disambig` tool from the SRILM toolkit [45], was used to perform the capitalization. This is a generative approach, that makes use of n-gram LMs. Our experiments use a trigram LM, using backoff estimates, as implemented by the `ngram-count` tool from the same toolkit, without *n*-gram

Table III
CAPITALIZATION RESULTS FOR THE ME AND HMM APPROACHES.

Evaluation data		ME				HMM			
		Prec	Rec	F	SER	Prec	Rec	F	SER
PT	Written	95.1	85.3	89.9	18.8	94.4	90.6	92.5	14.4
	Manual	94.8	88.0	91.3	16.5	84.8	91.4	87.9	24.7
	ASR	82.7	81.7	82.2	34.9	69.3	85.9	76.7	51.5
EN	Written	96.2	81.6	88.3	20.8	94.9	88.5	91.6	15.3
	Manual	94.3	82.4	88.0	22.2	91.9	84.9	88.2	22.2
	ASR	83.9	73.1	78.1	40.4	77.8	75.3	76.5	45.5

discounts. The HMM-based tagger uses a hidden-event n-gram LM [46], and can be used to perform capitalization directly from the LM. This implementation of the HMM-based tagger can use different algorithms for decoding. Results in this paper use the Viterbi decoding algorithm, where the output is the sequence with the higher joint posterior probability. This is a straightforward method, producing fast results, and often used by the scientific community for this task[6].

Table III shows results achieved with both methods, using the same training and evaluation data. As a first result, the ME approach achieves a better precision, while the HMM-based approach achieves a better recall. Results indicate that the HMM-based approach is better for written corpora, while the ME approach is significantly better for speech transcripts. Several reasons may explain this fact: i) the information expressivity is not the same in both methods, e.g., ME experiments here described do not use the information concerning the previous word (w_{i-1}) as an isolated feature, while that information is available in the 3-gram LM used by the HMM-based approach; ii) the ME-based approach is not much influenced by the context, which is quite important when dealing with speech units that may be flexible, elliptic, and even incomplete; and iii) the restricted training conditions used for limiting computational resources. The WER impact is bigger for the HMM-based approach, because different words may cause completely different search paths. A possible explanation for the bigger difference between the methods in the Portuguese speech data may be related with the proportion of spontaneous/prepared speech in both corpora. We know that Portuguese transcripts contain a high percentage of spontaneous speech (35%), much higher than our data for the Spanish BN (11%), but, unfortunately, this information is not available for the English data.

These results are difficult to compare to other related work, mainly because of the different evaluation sets, but also because of the different evaluation metrics and applied criteria. For example, sometimes it is not clear whether the evaluation takes into consideration the first word of each sentence. However, these results are consistent with the work reported by [14], which achieves 88.5% F-measure (89% prec., 88% recall) on written corpora (Wall Street Journal), and 83% F-measure (83% prec., 83% recall) on manual transcripts.

VI. PUNCTUATION TASK

This section addresses the punctuation task, covering the three most frequent punctuation marks: *full stop*, *comma*,

[6]For example, it was part of the baseline suggested in the IWSLT2006 workshop competition

Table IV
PUNCTUATION RESULTS FOR BN TRANSCRIPTS.

		Force aligned transcripts				Fully automatic transcripts			
		Prec	Rec	F	SER	Prec	Rec	F	SER
PT	Full stop	78.7	73.7	76.1	46.3	68.4	63.2	65.7	66.0
	Comma	67.8	41.4	51.4	78.2	63.3	30.3	41.0	87.2
	ALL	73.3	54.3	62.4	56.3	66.2	43.4	52.4	68.2
EN	Full stop	79.2	70.8	74.7	47.8	71.1	64.6	67.7	61.7
	Comma	66.2	16.1	25.9	92.1	65.1	16.1	25.8	92.6
	ALL	76.7	45.1	56.8	60.5	69.9	41.7	52.3	68.1

Table V
PUNCTUATION RESULTS, BOOTSTRAPPING FROM WRITTEN CORPORA.

		Force aligned transcripts				Fully automatic transcripts			
		Prec	Rec	F	SER	Prec	Rec	F	SER
PT	Full stop	77.8	76.9	77.3	45.1	66.8	66.8	66.8	66.4
	Comma	69.3	52.9	60.0	70.5	62.9	39.9	48.8	83.7
	ALL	73.2	62.5	67.4	50.1	64.9	50.5	56.8	65.0
EN	Full stop	77.1	74.8	75.9	47.5	69.7	61.2	65.2	65.4
	Comma	64.2	22.6	33.5	90.0	56.5	16.4	25.4	96.2
	ALL	73.9	50.3	59.9	57.5	66.7	40.1	50.1	70.5

and *question mark*. Detecting *full stops* and *commas* depends mostly on a local context, usually two or three words. However, most *interrogatives,* specially wh-questions, depend on the words that are used at the beginning or/and at the end of the sentence (e.g., *quem disse o quê?/who said what?*), which means that the SU boundaries must be previously known. For that reason, we distinguish two separate sub-tasks: the first, using a local context, for detecting *full stops* and *commas*; the second, for detecting *question marks*, using properties of the whole sentence as features.

A. *Recovering* Full stop *and* Comma

The following experiments recover only *full stops* and *comma*s, where all the other marks were converted into one of these two, according to the following rules: ".": ";", "!", "?", "..." => *full stop*; ",", "-" => *comma*. The following features were 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}, $2p_{i-2}$, $2p_{i-1}$, $2p_i$, $2p_{i+1}$, $3p_{i-2}$, $3p_{i-1}$, $GenderChgs_1$, $SpeakerChgs_1$, and $TimeGap_1$, where: w_i is the current word, w_{i+1} is the word that follows, and nw_k is the n-gram of words that starts at position $k = i \pm x$; np_k is the part-of-speech n-gram, of words starting at position k. $GenderChgs_1$, and $SpeakerChgs_1$ correspond to changes in speaker gender, and speaker clusters; $TimeGap_1$ corresponds to the time period between the current and the following word.

Table IV shows punctuation results for both languages, revealing that the *full stop* is easier to detect. Precision is consistently better than recall, suggesting that the system usually prefers to avoid mistakes than to add incorrect slots. The WER impact, in terms of SER, is about 12% for Portuguese and about 8% for English. [14] considers different LMs, ranging from 58 million to 55 billion of training words. The smallest LM achieves an F-score of 37% for *comma* and 46% for *full stop*, while the largest LM achieves 52% for *comma* and 62% for *full stop*. The significant performance increase suggests that our results (obtained with less than one million words of speech transcripts) can be much improved by using larger training sets. Their best F-score concerning the *full stop* (62%) is lower than results presented here for English, but they do not make use of any acoustic information.

1) Retrain from a Written Corpora Model: An initial idea for improving our punctuation results, consisted of using punctuation information extracted from written corpora. For that purpose, we have firstly trained a punctuation model using written corpora, and then trained a new punctuation model with transcripts, bootstrapping from the written corpora model. Table V presents the obtained results, that can be directly compared with results from Table IV. From the comparison, regular trends are found: i) Portuguese performance increased considerably; ii) English fully automatic transcripts is the only condition where bootstrapping does not promote better results; iii) the performance for the force aligned data is consistently improved; iv) comma detection always improves, and significantly for Portuguese force aligned data (about 8%). These findings support two basic ideas: results are better for Portuguese, because English data is quite heterogeneous and has a higher WER; the most significant gains concerning *comma* derive from the fact that this specific punctuation mark depends more on lexical features (e.g., *..., por exemplo/for instance, ...*), similar to observations from [36].

Additionally to the above bootstrapping method for improving the transcripts model, an alternative was also tested. The idea consisted of using the prediction of the written corpora model as a complement to the transcripts data. Three different features (COMMA, FULLSTOP, SPACE) were appended to the feature vector of each event in the transcripts data, with the corresponding probabilities, provided by the written corpora model. Models trained with the improved data achieve better performances than using solely information coming from the transcripts. Nevertheless, in general, this method is still worse than the first method tested, based on boostrapping.

2) Introducing Prosodic Information: The other strategy for improving our initial results consisted of adding prosodic features, besides the existing lexical, time-based and speaker-based features. We do know that there is no one-to-one mapping between prosody and punctuation [47]. Silent pauses, for instance, can not be directly transformed into punctuation marks for different reasons, e.g., prosodic constraints regarding the weight of a constituent, speech rate, style, different pragmatic functions, on-line planning. However, prosodic information can be used to improve the punctuation detection. For example, [31] concludes that F-measure can be improved by 19% relative.

The feature extraction stage involved several steps. The first step consisted of extracting pitch and energy from the speech signal, which was achieved using the Snack Sound Toolkit[7]. Durations of phones, words, and interword-pauses were extracted from the recognizer output. We normalized f_0 values in order to remove micro-prosodic and octave jump effects from the pitch track. Another important step consisted of marking the syllable boundaries as well as the syllable stress. A set of syllabification rules was designed for Portuguese and applied to the lexicon. The rules account fairly well for the canonical

[7]http://www.speech.kth.se/snack/

Table VI
PUNCTUATION RESULTS FOR BN TRANSCRIPTS, ADDING PROSODY.

		Force aligned transcripts				Fully automatic transcripts			
		Prec	Rec	F	SER	Prec	Rec	F	SER
PT	Full stop	80.4	79.9	80.1	39.6	69.8	71.2	70.5	59.6
	Comma	70.5	54.5	61.5	68.3	65.3	39.7	49.4	81.4
	ALL	75.1	64.6	69.5	47.7	67.6	52.2	58.9	62.2
EN	Full stop	79.3	75.7	77.5	44.0	71.3	64.0	67.4	61.8
	Comma	63.0	24.8	35.6	89.8	56.4	16.6	25.7	96.2
	ALL	74.9	51.9	61.3	55.8	67.9	41.7	51.6	68.6

Table VII
IMPACT OF PROSODIC FEATURES RECOVERING *full stop* AND *comma*.

Corpus		Force aligned transcripts			Fully automatic transcripts		
		WB	SL	WB+SL	WB	SL	WB+SL
PT	Full stop	40.4	41.6	39.6	57.2	59.8	59.6
	Comma	68.6	69.1	68.3	81.8	82.9	81.4
	ALL	48.1	48.8	47.7	61.9	63.0	62.2
EN	Full stop	44.1	46.2	44.0	62.1	62.0	61.8
	Comma	89.3	89.5	89.8	95.1	95.8	96.2
	ALL	55.8	56.7	55.8	68.4	69.0	68.6

pronunciation of native words, but still need improvements for words of foreign origin. As for English, we used tsylb2[8], an automatic phonological-based syllabication algorithm. Finally, we have calculated the maximum, minimum, median, and slope values for pitch and energy in each word, syllable, and phone. Duration was also calculated for each one of the previous units.

Underlying the prosodic feature extraction process is the linguistic evidence that pitch contour, boundary tones, energy slopes, and pauses are crucial to delimit SUs across languages [48]. First, we have tested if the features would perform better on different units of analysis: phones, syllables and/or words. Supported by linguistic findings for EP [49], [50], [51], [52], [53], we hypothesized that the stressed and post-stressed syllables would be relevant units of analysis to automatically identify punctuation marks. When considering the word as a window of analysis, we are also accounting for the information in the pre-stressed syllables as well.

Features were calculated for each *word transition*, with or without a pause, using: the last word, last stressed syllable and last voiced phone from the *current word*, and the first word, and first voiced phone from the *following word*. The following set of features has been used: f_0 and energy slopes in the words before and after a silent pause, f_0 and energy differences between these units, and also duration of the last syllable and the last phone. Table VI shows the results, achieving significant gains relatively to the previous results, for both languages, both types of transcripts, and both punctuation marks, ranging from 3% to 8% SER. Better results are again achieved for Portuguese, but in contrast to the ones from Section VI-A1, they are mainly related to the *full stop*.

Table VII outlines the contribution of each prosodic feature *per se*. The word-based (WB) features turned out to be the most reliable ones, whereas syllable-based (SL) features achieved only small gains relatively to previous results. The best results were always achieved either combining all the features or using the word-based features alone. These results partially agree with the ones reported in [54], regarding the contribution of each prosodic parameter, and also the set of features used, where the most influential feature turned out to be f_0 slope in the words and between word transitions for Portuguese.

[8]B. Fisher. The tsylb2 program, Aug. 1996. National Institute of Standards and Technology Speech.

Table VIII
QUESTION MARKS RESULTS, USING LEXICAL FEATURES ONLY.

Corpus		Cor	Ins	Del	Prec	Rec	F	SER
PT	Written	1100	236	1740	82.3	38.7	52.7	69.6
	Align	143	24	272	85.6	34.5	49.1	71.3
	Recog	74	27	315	73.3	19.0	30.2	87.9
EN	Written	993	81	668	92.5	59.8	72.6	45.1
	Align	155	22	109	87.6	58.7	70.3	49.6
	Recog	100	27	152	78.7	39.7	52.8	71.0

B. Detection of Question Marks

This section concerns the automatic detection of *question mark*s, which corresponds to detecting which sentences are interrogatives. This is an extension to the punctuation module, which was initially designed to deal with *full stop* and *comma* only. We will follow the previous ME approach, but now this is a binary problem, and each event corresponds to an entire sentence, instead of being a word. This section assumes that sentence boundaries are given by manual annotations.

1) Experiments with lexical features only: The initial set of experiments was performed using lexical information only. For each sentence, the following features were used, covering all the words in the sentence: w_i, w_{i+1}, $2w_{i-2}$, $2w_{i-1}$, $2w_i$, $2w_{i+1}$, $3w_{i-2}$, $3w_{i-1}$, $start\_x$, x_end, len, where: $start\_y$ and y_end features were used for identifying n-grams of words occurring either at the beginning or at the end of the sentence, and len corresponds to the number of words in the sentence. We have started by creating a model, for each language, from the written corpora described in section IV-B. Then, its prediction was used, as a complement, for training the other models, from the transcripts. Only two features were added, with the corresponding probabilities provided by the initial model. The performance of the resultant models is better than: i) using only the information coming from the transcripts; ii) using the bootstrapping method, previously applied in Section VI-A1, because it is an easier problem (binary), and the reduced number of question marks found in the BN corpora cause the method to converge too fast, losing most of the information given by the initial model.

Table VIII shows the performance of applying the previous models to: written corpora, force aligned, and automatic transcripts. *Cor*, *Ins* and *Del* represent the number of correct, inserted and deleted sentences, respectively. As expected, question marks are easier to detect for written English, since this language has more lexical cues, mainly quite frequent n-grams related with "auxiliary verb plus subject inversion" (e.g., *do you?*, *can you?*, *have you?*). The difference of about 24% SER is mostly related with the high number of deletions (non-identified sentences) for Portuguese. This is due to the fact that yes/no questions, corresponding to almost 50% of all the questions in the corpus, are mainly disambiguated from a

declarative sentence using prosodic information. Concerning the force aligned transcripts, results are again better for English. The difference between force aligned and automatic transcripts is bigger in English (21.4%) than in Portuguese (16.6%), reflecting the impact of the recognition errors in this task. Although n-grams related with "auxiliary verb plus subject inversion" are relevant features for correctly identifying question marks in English, the auxiliary verbs (e.g., *do*, *can*, *have*) are often misrecognized, particularly in spontaneous speech, causing that bigger impact.

When using only this limited set of features, the recall percentages are correlated with specific types of questions, namely, wh-questions for both languages; and yes/no questions almost exclusively for English. Due to language specific properties, namely, "auxiliary verb plus subject inversion", the recall percentages for English are always higher than for Portuguese. Not surprisingly then, the bigram *"do you"*, for instance, is fairly well associated with a yes/no question. For Portuguese, the recall percentage of the aligned data is comparable to the one of wh-questions for BN and newspapers. However there is still a small percentage of this type of interrogative not accounted for, mainly due to very complex structures hard to disambiguate automatically. *Tag* and *alternative* questions in either language are not easily identified with lexical features only.

2) Experiments with all available features: This section adds different kinds of acoustic information, available in speech, to the previous feature set based on lexical features. The following extended features are analogous to the features used in Section VI-A2, except that for *question marks* they are always extracted at the sentence level: $GenderChgs$ and $SpeakerChgs$ correspond to changes in speaker gender, and speaker clusters from the current to the next sentence; $TimeGap$ corresponds to the time period between the current and the following sentence. The remaining features were calculated for each *sentence transition*, with or without a pause, using the same analysis scope as [54] (last word, last stressed syllable and last voiced phone from the *current sentence*, and the first word, and first voiced phone from the *following sentence*). The following set of features was also used: f_0 and energy slopes in the words before and after a silent pause, f_0 and energy differences between these units, and duration of the last syllable and the last phone. With this set of features, we aim at modeling nuclear and boundary tones, amplitude, pitch reset, and final lengthening.

Table IX shows performance results, considering two groups of features: *word-based* (WB), and *syllable-based* (SL) features. There is an effective gain for the recognized Portuguese and for the aligned English data, but results are not very significant, due to the relatively small number of *question marks* found in the corpora. Results partially agree with the ones reported in [54], regarding the contribution of each prosodic parameter, and also the set of discriminative features used, where the most influential feature turned out to be f_0 slope in the words and between word transitions for Portuguese. As stated by [48], these features are language independent. Language specific properties in our data are related to the fact that *word-based* features are more useful for the Portuguese corpus, while *syllable-based* ones give the best results for the English data. This result may be interpretable by language specific syllabic properties, i.e., English allows for more segmental material in the syllabic skeleton. Thus, for Portuguese, the *word-based* features give us more context. Moreover, we may find different durational patterns at the end of an intonational unit (e.g., in European Portuguese post-tonic syllables are quite often truncated). Also different pitch slopes may be associated with discourse functions beyond sentence-form types.

Table IX
RECOVERING THE QUESTION MARK, ADDING PROSODY.

Corpus		Feat	Cor	Ins	Del	Prec	Rec	F	SER
PT	Align	WB	147	28	268	84.0	35.4	49.8	71.3
		SL	148	29	267	83.6	35.7	50.0	71.3
		All	146	31	269	82.5	35.2	49.3	72.3
	Recog	WB	76	22	313	77.6	19.5	31.2	86.1
		SL	71	26	318	73.2	18.3	29.2	88.4
		All	75	23	314	76.5	19.3	30.8	86.6
EN	Align	WB	152	21	112	87.9	57.6	69.6	50.4
		SL	151	19	113	88.8	57.2	69.6	50.0
		All	149	19	115	88.7	56.4	69.0	50.8
	Recog	WB	100	31	152	76.3	39.7	52.2	72.6
		SL	100	27	152	78.7	39.7	52.8	71.0
		All	102	33	150	75.6	40.5	52.7	72.6

Summing up, when training only with lexical features, *wh-questions* are effectively identified in both languages, and *yes/no* questions in the English data. When training with all the features, *yes/no*, *tag* and *alternative* questions are then identified for Portuguese (the English data had no *tag* questions). We have also verified that prosodic features increase the identification of interrogatives in Portuguese BN spontaneous speech, e.g., yes/no question with a request to complete a sentence (e.g., *recta das?/lines of?*), tag questions (such as *não é?/isn't it?*), and alternative questions as well (*contava com esta decisão ou não?/were you expecting this decision or not?*). Even when all the information is combined, we still have questions for both languages that are not well identified, due to the following aspects: i) many questions occur in the transition between newsreader and reporter, with noisy background (such as war scenarios); ii) frequent elliptic expressions, e.g., *Eu?*/"me?"; iii) sequences with disfluencies, e.g., *<é é é> como é que se consegue?*, contrasted with a similar question without disfluencies that was classified: *Como é que conseguem isso?/How do you manage that*?; iv) sequences starting with the copulative conjunction *e/and* or the adversative conjunction *mas/but*, which usually do not occur at the start of sentences; v) false insertions of question marks in sequences with subordinated questions, which are not marked with a question mark; vi) sequences with more than one consecutive question, e.g., *(...) nascem duas perguntas: quem? e porquê?/two questions arise: who? and why?*; and vii) sequences integrating parenthetical comments or vocatives, e.g., *Foi acidente mesmo ou atentado, Noé?/Was it an accident or an attack, Noé?*.

VII. CONCLUSIONS

This paper addresses the tasks of recovering capitalization and punctuation marks from spoken transcripts. These two

practical RT tasks were performed using the same discriminative approach, based on maximum entropy, adequate for on-the-fly integration, and of great importance for tasks such as online subtitling, which may have strict latency requirements. Reported experiments were conducted both over Portuguese and English BN data. Both force aligned and automatic transcripts were used in the experiments, allowing the measurement of 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. 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 typical approach, that can be easily applied to this problem. While the HMM-based approach captured the structure of written corpora better, the ME-based approach proved to be suitable for dealing with speech transcripts, and also more robust to ASR errors.

Regarding 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 first, 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 focus on the usage of additional information sources and diverse linguistic structures that can be found in the speech data. Two different methods were explored for improving the baseline results for *full stop* and *comma*. The first makes use of the punctuation information that can be found in large written corpora. The second consists of introducing prosodic features, besides the initial lexical, time-based and speaker-based features. We have observed that the linguistic structure in both languages is captured in different ways for distinct punctuation marks: *commas* are mostly identified by lexical features, while *full stops* are mostly depending on prosodic ones. The most significant gains come from combining all the available features. Although the relative small number of *question marks* does not allow us to observe significant differences, there is a small gain in combining all features both for recognized Portuguese and for English aligned data.

In terms of capitalization, an interesting future direction would be the fusion of the generative and the discriminative approaches, since they perform better for written corpora and speech transcripts, respectively. In terms of punctuation, there are many interesting research directions, particularly in what concerns prosodic features (for instance, by using pseudo-syllable information directly derived from the audio data). Extending this study on interrogatives to other domains, besides BN, will allow better modeling of different types of interrogatives not well represented in this corpus. Further experiments must be performed in order to assess to what extent our prosodic features are language-based or language-independent features.

ACKNOWLEDGMENT

The authors would like to thank the three anonymous reviewers for their very helpful comments. The authors would also like to thank Thomas Pellegrini for his support with the English BN corpora, Vera Cabarrão for her revision of the Portuguese BN transcripts, and Hugo Meinedo for his support with the speech recognition system. 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. Helena Moniz is supported by the FCT grant SFRH/BD/44671/2008.

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*, 2003, pp. 1585–1588.

[2] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed. Prentice Hall PTR, 2009.

[3] S. Strassel, *Simple Metadata Annotation Specification V6.2*, online, Linguistic Data Consortium, 2004.

[4] 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.

[5] M. Harper, B. Dorr, J. Hale, B. Roark, I. Shafran, M. Lease, Y. Liu, M. Snover, L. Yung, A. Krasnyanskaya, and R. Stewart, "Parsing and spoken structural event detection," in *2005 Johns Hopkins Summer Workshop Final Report*, 2005.

[6] J. Mrozinsk, E. Whittaker, P. Chatain, and S. Furui, "Automatic sentence segmentation of speech for automatic summarization," in *ICASSP*, 2006.

[7] 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, and C. Wooters, "Speech segmentation and spoken document processing," *IEEE Signal Processing Magazine*, vol. 25(3), pp. 59–69, 2008.

[8] L. V. Lita, A. Ittycheriah, S. Roukos, and N. Kambhatla, "tRuEcasIng," in *41^{st} annual meeting on ACL*. ACL, 2003, pp. 152–159.

[9] E. Blaauw, *On the Perceptual Classification of Spontaneous and Read Speech*. Research Institute for Language and Speech, 1995.

[10] R. Amaral, H. Meinedo, D. Caseiro, I. Trancoso, and J. P. Neto, "A prototype system for selective dissemination of broadcast news in European Portuguese," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 37507, 2007.

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

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

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

[14] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *ICASSP 2009*, 2009.

[15] E. Brill, "Some advances in transformation-based part of speech tagging," in *Proc. of AAAI '94*, vol. 1, 1994, pp. 722–727.

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

[17] 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.

[18] W. Wang, K. Knight, and D. Marcu, "Capitalizing machine translation," in *HLT-NAACL*. ACL, 2006, pp. 1–8.

[19] C. Martins, A. Teixeira, and J. P. Neto, "Dynamic language modeling for a daily broadcast news transcription system," in *ASRU 2007*, 2007.

[20] C. Mota and R. Grishman, "Is this NE tagger getting old?" in *Proc. of the LREC'08*, ELRA, Ed., 2008.

[21] C. Mota and R. Grishman, "Updating a Name Tagger using contemporary unlabeled data," in *ACL-IJCNLP (Short Papers)*, 2009, pp. 353–356.

[22] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *11^{th} annual conference on Computational Learning Theory*. ACM, 1998, pp. 92–100.

[23] M. Collins and Y. Singer, "Unsupervised models for Named Entity classification," in *Proc. Joint SIGDAT Conference on EMNLP*, 1999.

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

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

[26] I. Duarte, *Língua Portuguesa, Instrumentos de Análise*. Universidade Aberta, 2000.

[27] M. H. Mateus et al., *Gramática da Língua Portuguesa*. Caminho, 2003.

[28] H. Moniz, F. Batista, I. Trancoso, and A. I. Mata, *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, 1st ed., ser. LNCS. Springer, 2011, vol. 6456, ch. Analysis of interrogatives in different domains, pp. 136–148.

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

[30] 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*, 2001, pp. 35–40.

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

[32] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," in *ASR-2000: Challenges for the new Millennium*, 2000, pp. 228–235.

[33] 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.

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

[35] D. Wang and S. Narayanan, "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues," in *ICASSP*, vol. 1, 2004, pp. 525–528.

[36] B. Favre, D. Hakkani-Tur, and E. Shriberg, "Syntactically-informed Models for Comma Prediction," in *Proc. ICASSP 2009*, 2009.

[37] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra, "A Maximum Entropy approach to Natural Language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

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

[39] J. P. Neto, H. Meinedo, R. Amaral, and I. Trancoso, "A system for selective dissemination of multimedia information," in *MSDR*, 2003.

[40] H. Moniz, "Contributo para a caracterização dos mecanismos de (dis)fluência no Português Europeu," Master's thesis, University of Lisbon, 2006.

[41] J. Carletta, "Assessing agreement on classification tasks: The kappa statistic," *Computational Linguistics*, vol. 22, pp. 249–254, 1996.

[42] R. Ribeiro, L. Oliveira, and I. Trancoso, "Using morphosyntactic information in TTS systems: comparing strategies for European Portuguese," in *PROPOR 2003*. Springer, 2003, pp. 26–27.

[43] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *International Conf. on New Methods in Language Processing*, 1994.

[44] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Broadcast News Worksh.*, 1999.

[45] A. Stolcke, "SRILM - An extensible language modeling toolkit," in *Proc. of the ICSLP*, vol. 2, 2002, pp. 901–904.

[46] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *ICSLP '96*, vol. 2, 1996, pp. 1005–1008.

[47] M. C. Viana, L. C. Oliveira, and A. I. Mata, "Prosodic phrasing: Machine and human evaluation," *International Journal of Speech Technology*, vol. 6, no. 1, pp. 83–94, 2003.

[48] J. Vassière, "Language-independent prosodic features," in *Prosody: modules and measurements*. Springer, 1983, pp. 55–66.

[49] M. C. Viana, "Para a síntese da entoação do Português," Ph.D. dissertation, University of Lisbon, 1987.

[50] A. I. Mata, "Para o estudo da entoação em fala espontânea e preparada no Português Europeu," Ph.D. dissertation, University of Lisbon, 1999.

[51] S. Frota, *Prosody and Focus in European Portuguese. Phonological Phrasing and Intonation.* Garland Publishing, 2000.

[52] M. Vigário, *The prosodic word in European Portuguese.* Mouton de Gruyter, 2003.

[53] I. Falé, "Percepção e reconhecimento da informação entoacional em Português Europeu," Ph.D. dissertation, University of Lisbon, 2005.

[54] E. Shriberg, B. Favre, J. Fung, D. Hakkani-Tur, and S. Cuendet, "Prosodic similarities of dialog act boundaries across speaking styles," *Linguistic Patterns in Spontaneous Speech*, no. A25, pp. 213–239, 2009.

Fernando Batista is graduated in Mathematics and Computer Sciences, at Universidade da Beira Interior, in 1997. He received a Masters degree in Electrical and Computer Engineering from Instituto Superior Técnico (IST), in 2003, and finished his PhD in Computer Science and Engineering also at Instituto Superior Técnico, in 2011. Since 2000, he is a lecturer at the Lisbon University Institute (ISCTE-IUL), and since 2001, he is also a researcher at the L^2F - INESC-ID. He was a researcher at INESC's Natural Language Group from 1996 to 2000, participating in several European and National projects, such as LE-PAROLE, LE-SIMPLE, and Hypermuseum. His current research interests include Rich Transcription, working to generate readable transcriptions of speech, by integrating knowledge-based information with data-driven methods.

Helena Moniz is graduated in Modern Languages and Literature – Portuguese Studies, at Faculty of Letters, University of Lisbon (FLUL), in 1998. She took a Teacher Training graduation course in 2000, also at FLUL. She was a high school teacher from 2000 to 2006. She received a Masters degree in Linguistics at FLUL, in 2007, and she is currently taking her PhD in Linguistics at FLUL in cooperation with the Technical University of Lisbon (IST). She has been working at INESC-ID since 2000, in several national and international projects (DIXI+, Tecnovoz, LECTRA, PoSTPort, COST-2102, PT-STAR), involving multidisciplinary teams of linguists and speech processing engineers. She is a member of SProSIG and SIG-IL. Her current research interests include the analysis of disfluencies, working on the integration of prosodic information on the output of the speech recognition.

Isabel Trancoso received the Licenciado, Mestre, Doutor and Agregado degrees in Electrical and Computer Engineering from Instituto Superior Técnico, Lisbon, Portugal, in 1979, 1984, 1987 and 2002, respectively. She has been a lecturer at this University since graduation, being currently a Full Professor. She is also a senior researcher at INESC ID Lisbon, having launched the speech processing group, now restructured as L^2F. Her first research topic was medium-to-low bit rate speech coding, a topic where she worked at AT&T Bell Laboratories, Murray Hill, New Jersey. Her current scope is much broader, encompassing many areas of spoken language processing. She was Editor in Chief of the IEEE Transactions on Speech and Audio Processing, and Member-at-Large of the IEEE Signal Processing Society Board of Governors. She is currently the President of ISCA. She chaired the Organizing Committee of the INTERSPEECH'2005 Conference. She has received the 2010 IEEE Signal Processing Society Meritorious Service Award, and was elevated to IEEE fellow in 2011.

Nuno Mamede received is graduation, MSc and PhD degrees in Electrical and Computer Engineering by the Instituto Superior Técnico, Lisbon, in 1981, 1985 and 1992, respectively. In 1982 he started as lecturer and since 2006 he holds a position of Associate Professor in Instituto Superior Técnico, where he has taught Digital Systems, Object Oriented Programming, Programming Languages, Knowledge Representation and Natural Language Processing. He has been a researcher at INESC-ID Lisboa, since its creation in 1980, and participated in the foundation of L^2F. His activities have been in the areas of Written Natural Language Processing, namely on Syntactic Processing, Named Entity Recognition, and Natural Language Interfaces to Data Bases. He has authored a significant number of scientific papers. Member of AAAI, ACM and ACL.