Recovering Capitalization and Punctuation Marks for Automatic Speech Recognition: Case Study for Portuguese Broadcast News

F. Batista [a,b] D. Caseiro [a,c] N. Mamede [a,c] I. Trancoso [a,c]

[a] *L^2F – Spoken Language Systems Laboratory - INESC ID Lisboa*
R. Alves Redol, 9, 1000-029 Lisboa, Portugal

[b] *ISCTE – Instituto de Ciências do Trabalho e da Empresa, Portugal*

[c] *IST – Instituto Superior Técnico - Technical University of Lisbon, Portugal*

Abstract

The following material presents a study about recovering punctuation marks, and capitalization information from European Portuguese broadcast news speech transcriptions. Different approaches were tested for capitalization, both generative and discriminative, using: finite state transducers automatically built from language models; and maximum entropy models. Several resources were used, including lexica, written newspaper corpora and speech transcriptions. Finite state transducers produced the best results for written newspaper corpora, but the maximum entropy approach also proved to be a good choice, suitable for the capitalization of speech transcriptions, and allowing straightforward on-the-fly capitalization. Evaluation results are presented both for written newspaper corpora and for broadcast news speech transcriptions. The frequency of each punctuation mark in BN speech transcriptions was analyzed for three different languages: English, Spanish and Portuguese. The punctuation task was performed using a maximum entropy modeling approach, which combines different types of information both lexical and acoustic. The contribution of each feature was analyzed individually and separated results for each focus condition are given, making it possible to analyze the performance differences between planned and spontaneous speech. All results were evaluated on speech transcriptions of a Portuguese broadcast news corpus. The benefits of enriching speech recognition with punctuation and capitalization are shown in an example, illustrating the effects of described experiments into spoken texts.

Key words: Rich transcription, punctuation recovery, sentence boundary detection, capitalization, truecasing, maximum entropy, language modeling, weighted finite state transducers.

Email addresses: `Fernando.Batista@inesc-id.pt` (F. Batista),

1 Introduction

Enormous quantities of digital audio and video data are daily produced by TV stations, radio, and other media organizations. Automatic Speech Recognition (ASR) systems can now be applied to such sources of information in order to enrich them with additional information for applications, such as: indexing, cataloging, subtitling, translation and multimedia content production. Automatic Speech Recognition output consists of raw text, often in lower-case format and without any punctuation information. Even if useful for many applications, such as indexing and cataloging, for other tasks, such as subtitling and multimedia content production, the ASR output benefits from the correct punctuation and capitalization. In general, enriching the speech output aims to improve legibility, enhancing information for future human and machine processing. Apart from the insertion of punctuation marks and capitalization, enriching speech recognition covers other activities, such as detection and filtering of disfluencies, not addressed in this paper.

Depending on the application, punctuation and capitalization tasks may be required to work online. For example, on-the-fly subtitling for oral presentations or TV shows demands a very small delay between the speech production and the corresponding transcription. In these systems, both the computational delay and the number of words to the right of the current word that are required to make a decision, are important aspects to be taken into consideration. One of the goals behind this work consists of building a module for integration on an on-the-fly subtitling system, and a number of choices were taken with this purpose, for example, all subsequent experiments avoid a right context longer than two words for making a decision.

This paper describes a set of experiments concerning punctuation and capitalization recovery for spoken texts, providing the first joint evaluation results of these two tasks on Portuguese broadcast news. The remaining of this section describes related work, both on capitalization and punctuation. Section 2 describes the performance measures used for evaluation. Section 3 describes the main corpus and other resources. Section 4 is centered on the capitalization task, presenting the multiple employed methodologies and results achieved. Section 5 focus on the punctuation task, describing how the corpus was processed, the feature set used by the maximum entropy approach, and results concerning punctuation insertion. Section 6 presents a concrete example, showing the benefits of punctuation and capitalization over spoken texts. Sections 7 and 8 present some final comments and address the future work.

Diamantino.Caseiro@inesc-id.pt (D. Caseiro), Nuno.Mamede@inesc-id.pt (N. Mamede), Isabel.Trancoso@inesc-id.pt (I. Trancoso).

The capitalization task, also known as truecasing (Lita et al., 2003), consists of rewriting each word of an input text with its proper case information given its context. Different practical applications benefit from automatic capitalization as a preprocessing step: many computer applications, such as word processing and e-mail clients, perform automatic capitalization along with spell corrections and grammar check; and while dealing with speech recognition output, automatic capitalization provides relevant information for automatic content extraction, named entity recognition, and machine translation.

Capitalization can be viewed as a lexical ambiguity resolution problem, where each word has different graphical forms. Yarowsky (1994) presents a statistic procedure for lexical ambiguity resolution, based on decision lists, that achieved good results when applied to accent restoration in Spanish and French. The capitalization and accent restoration problems can be treated using the same methods, given that a different accentuation can be regarded as a different word form. Mikheev (1999, 2002) also presents an approach to the disambiguation of capitalized common words, but only where capitalization is expected, such as the first word of the sentence or after a period.

The capitalization problem may also be seen as a sequence tagging problem (Chelba and Acero, 2004; Lita et al., 2003; Kim and Woodland, 2004), where each lower-case word is associated to a tag that describes its capitalization form. Chelba and Acero (2004) study the impact of using increasing amounts of training data as well as a small amount of adaptation. This work uses a Maximum Entropy Markov Model (MEMM) based approach, which allows to combine different features. A large written newspaper corpora is used for training and the test data consists of Broadcast News data. Lita et al. (2003) builds a trigram language model (LM) with pairs (word, tag), estimated from a corpus with case information, and then uses dynamic programming to disambiguate over all possible tag assignments on a sentence. A preparatory study on the capitalization of Portuguese broadcast news has been performed by Batista et al. (2007b).

Other related work includes a bilingual capitalization model for capitalizing machine translation (MT) outputs using conditional random fields (CRFs) reported by (Wang et al., 2006). This work exploits case information both from source and target sentences of the MT system, producing better performance than a baseline capitalizer using a trigram language model.

Spoken language is similar to written text in many aspects, but is different in many others, mostly due to the way these communication methods are produced. Current ASR systems focus on minimizing the WER (word error rate), making no attempts to detect structural information which is available in written texts. Spoken language is also typically less organized than textual material, making it a challenge to bridge the gap between spoken and written material. The insertion of punctuation marks into spoken texts is a way of approximating such texts, even if a given punctuation mark may assume a slightly different behavior in speech. For example, a sentence in spontaneous speech does not always correspond to a sentence in written text.

A large number of punctuation marks can be considered for spoken texts, including: *comma*; *period* or *full stop*; *exclamation mark*; *question mark*; *colon*; *semicolon;* and *quotation marks*. However, most of these marks rarely occur and are quite difficult to insert or evaluate. Hence, most of the available studies focus either on *full stop* or in *comma*, which have higher corpus frequencies. Previous work on other punctuation marks, such as *question mark* and *exclamation mark*, have not shown promising results (Christensen et al., 2001).

Comma is the most frequent punctuation mark, but it is also the most problematic because it serves many different purposes. It can be used to: introduce a word, phrase or construction; separate long independent constructions; separate words within a sentence; separate elements in a series; separate thousands, millions, etc. in a number; and also prevent misreading. Beeferman et al. (1998) describe a lightweight method for automatically inserting intra-sentence punctuation marks into text. This method relies on a trigram LM built solely using lexical information, and uses the Viterbi algorithm for classification. The paper focus the *comma* punctuation mark and presents a qualitative evaluation based on user satisfaction, concluding that the system performance is qualitatively higher than sentence accuracy rate would indicate.

When dealing with conversational speech the notion of utterance (Jurafsky and Martin, 2000) or sentence-like unit (SU) is often used (Strassel, 2004) instead of "sentence". A SU may correspond to a grammatical sentence, or can be semantically complete but smaller than a sentence. Detecting a SU consists of finding the limits of that SU, which roughly corresponds to the task of detecting the *period* or *full stop* in conversational speech. SU boundary detection has gained increasing attention during recent years, and it has been part of the NIST rich transcription evaluations. It provides a basis for further natural language processing, and its impact on subsequent tasks has been recently analyzed in many speech processing studies (Harper et al., 2005; Mrozinsk et al., 2006).

The work conducted by Kim and Woodland (2001) and Christensen et al. (2001) uses a general HMM framework that allows the combination of lexical and prosodic cues for recovering punctuation marks. A similar approach was also used by Gotoh and Renals (2000) and Shriberg et al. (2000) for detecting sentence boundaries. Another approach, based on a maximum entropy model, was developed by Huang and Zweig (2002) to recover punctuation in the Switchboard corpus, using textual cues. Different modeling approaches, combining different prosodic and textual features have also been recently investigated by other authors, such as Liu et al. (2006) for sentence boundary detection, and Batista et al. (2007a) for punctuation recovery on Portuguese broadcast news.

2 Performance measures

The following well-known performance measures are used in punctuation and capitalization tasks: Precision, Recall, and Slot Error Rate (SER) (Makhoul et al., 1999), defined in equations (1) to (3). For the punctuation task, a slot corresponds to the occurrence of a punctuation mark in the corpus. For the capitalization task, a slot corresponds to all words not written as a lower-case form.

$$Precision = \frac{C}{H} = \frac{C}{C+S+I} \tag{1}$$

$$Recall = \frac{C}{R} = \frac{C}{C+S+D} \tag{2}$$

$$SER = \frac{total\ slot\ errors}{R} = \frac{I+D+S}{C+D+S} \tag{3}$$

In the equations, C is the number of correct slots; I is the number of insertions (spurious slots / false acceptances); D is the number of deletions (missing slots / false rejections); S is the number of substitutions (incorrect slots); R is the number of slots in reference; and H is the number of slots in hypothesis. Precision and Recall are often combined in a single value (F-Measure).

Applying the performance measures to both examples of figure 1, a 50% Precision and Recall are achieved. While the F-Measure is also 50%, the SER is 100%, which may be a more meaningful measure, given that the number of slot errors in the example is greater than the number of correct ones. The work of Makhoul et al. (1999) shows that “this measure implicitly discount

```
Ref: w1 w2    w3 w4 . w5 w6 . w7
Hyp: w1 w2 . w3 w4    w5 w6 . w7
          ins     del     cor

Ref: here is an Example of a big SER
Hyp: here Is an example of a big SER
          ins   del              cor
```

Figure 1. Example of correct and incorrect slots.

the overall error rate, making the systems look like they are much better than they really are". Hence, this work will not include F-measure values.

The previously defined SER for punctuation corresponds to the NIST error rate for sentence boundary detection, which is defined as the sum of the insertion and deletion errors per number of reference sentence boundaries.

Despite the performance metrics here presented being widely used by the scientific community, other performance metrics could be exploited for an improved analysis. A recent study conducted by Liu and Shriberg (2007) shows the advantages of curves over a single metric for sentence boundary detection.

3 Information sources

Both capitalization and punctuation tasks described here share the same spoken corpus, however for the capitalization task other information sources were used, including a written newspaper corpus and two small lexica containing case information. The following subsections provide more details about each one of the data sources.

3.1 "Speech Recognition" Corpus

The Speech Recognition corpus (SR) is an European Portuguese broadcast news corpus, collected in the scope of the ALERT European project (Meinedo et al., 2003). Table 1 presents details for each part of the corpus.

The manual orthographic transcription of this corpus constitutes the reference corpus, and includes information such as punctuation marks, capital letters and special marks for proper nouns, acronyms and abbreviations. Each file in the corpus is divided into segments, with information about their start and end locations in the signal file, speaker id, speaker gender, and focus conditions. The orthographic transcription process follows the LDC Hub4 (Broad-

Table 1
Different parts of the Speech Recognition (SR) corpus

Sub-corpus	**Recording period**	**Duration**	**Tokens**	
train	2000 - Oct. and Nov.	61h	467k	81%
development	2000 - December	8h	64k	11%
test	2001 - January	6h	46k	8%

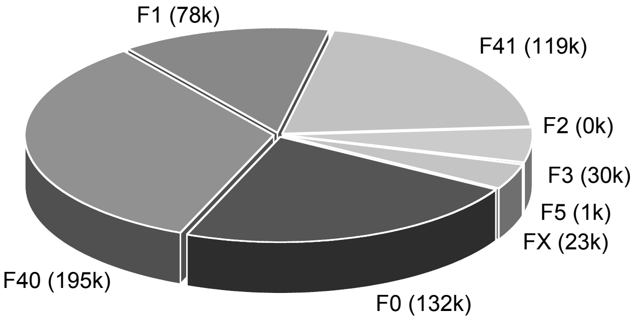


Figure 2. Distribution of words in the SR corpus by focus condition. The number of words is shown next to the label.

cast Speech) transcription conventions [1]. Each segment in the corpus is marked as: planned speech with or without noise (F40/F0); spontaneous speech with or without noise (F41/F1); telephone speech (F2); speech mixed with music (F3); non-native speaker (F5); all other speech (FX). As shown in Figure 2, most of the corpus consists of planned speech (F0+F40). Nevertheless, 34% is still a large percentage of spontaneous speech (F1+F41).

Besides the manual orthographic transcription, we also have available the automatic transcription produced by the ASR module, and other information automatically produced by the Audio Preprocessor (APP) module namely, the speaker id, gender and background speech conditions (Noise/Clean). Each word has a reference for its location in the audio signal, and includes a confidence score given by the ASR module.

3.2 "Recolha do Público" corpus

RecPUB is a written corpus, created from the Portuguese "Público" newspaper. It contains about 130 million words, and can be used to provide information about the capitalization of words. Table 2 provides details on each part of the corpus.

[1] http://www.ldc.upenn.edu/ Projects/ Corpus_Cookbook/ transcription/ broadcast speech/ english/ conventions.html

Table 2
Different parts of the RecPUB corpus

Corpus	**Period**	**Words**	
train	1995 to 2000	97.9 M	76%
development	1st sem. 2001	15.7 M	12%
test	2nd sem. 2001	16.4 M	12%

Table 3
The different sources of information used for building LEX

List	**Words**
Acronyms and abbreviations	72
Anthroponyms	466
Names of countries and cities	357
Nouns and abbreviations (POS selection)	652
Acronyms (POS selection)	14

The properties of a written newspaper corpus are quite different from what can be found in speech transcriptions. For example, a speech transcription, specially the spontaneous part, contains phenomena, such as filled pauses and disfluencies, not found in a written corpus. However, word co-occurrence in written corpora may be a valuable resource for the capitalization task.

3.3 Lexica

Capitalization experiments here described use a limited vocabulary of 57k words, which is also the vocabulary used by the ASR module. The SR corpus information is clearly insufficient to provide enough training material for all words in the vocabulary. In order to mitigate the small training data size, two lexica were built, from predefined gazetteers:

LEX – gathers information coming from existent lists of words, and unambiguous proper nouns, abbreviations and acronyms identified within the vocabulary using a part-of-speech (POS) tagger. Table 3 shows the different information sources that were used for building the lexicon. After merging all the separated components, a lexicon of about 1500 unique entries is achieved.

PubLEX – built from information coming from the RecPUB training data, covers all words in the vocabulary. The lexicon information consists of the most frequent written form of each word found on the RecPUB training data.

4 Capitalization task

The present study explores three ways of writing a word: lower-case, all-upper, and first-capitalized, not covering mixed-case words such as “McDonald’s” and “SuSE”.

The experiments were conducted both on written newspaper corpora and on spoken transcriptions, making it possible to analyze the impact of the different methodologies over these two different data. Written newspaper corpus, lexica and spoken transcriptions were combined in order to provide richer training sets and reduce the problem of having small quantities of spoken data for training. The evaluation on spoken data is performed over the SR manual transcriptions, because the current automatic speech transcription does not include case information. The following subsections describe the different methods employed and achieved results.

4.1 Methods

Different approaches were exploited for the capitalization task, including: (1) an HMM-based tagger, as implemented by the `disambig` tool from the SRILM toolkit (Stolcke, 2002); (2) a transducer, built from a previously created language model (LM); and (3) maximum entropy models. The first two modeling approaches are generative (joint), while the last one is discriminative (conditional). The following subsections provide details on each of the methods.

4.1.1 HMM-based tagger

Both generative approaches depend on n-gram language models, therefore the initial step of these approaches consists of creating n-gram LMs from the training corpus. The trigram language models were created using backoff estimates, as implemented by the `ngram-count` tool of the SRILM toolkit, without n-gram discounts.

The HMM-based tagger, implemented by the `disambig` tool, uses a hidden-event n-gram LM (Stolcke and Shriberg, 1996), and can be used to perform capitalization directly from the LM. Figure 3 illustrates the process, where each cloud represents a process and ellipses represents data. *Map* represents a file that contains all possible graphical forms of words in the vocabulary. The idea consists of translating a stream of tokens from a vocabulary L (lower-case words) to a corresponding stream of tokens from a vocabulary C (capitalized words), according to a 1-to-many mapping. Ambiguities in the mapping are resolved by finding the C sequence with the highest probability given the L

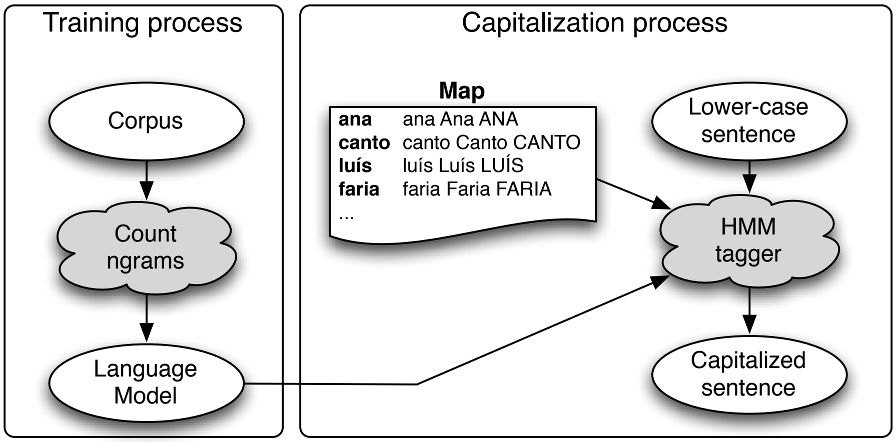


Figure 3. Using the HMM-based tagger.

sequence. This probability is computed from the LM[2].

This implementation of the HMM-based tagger can use different algorithms for decoding. However, results in this paper are achieved using 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. For example, it was part of the baseline suggested in the IWSLT2006 workshop competition[3].

4.1.2 Transducers

The capitalization based on Weighted Finite State Transducers (WFST) is illustrated in figure 4. This approach makes use of the LM previously built for the HMM-based tagger, which is converted into an automaton (FSA), corresponding to a WFST having the input equal to the output. The capitalization transducer T is created from this last WFST by converting every word in the input to its lower-case representation. Notice that the input of the transducer T uses a lower-case vocabulary while the output includes all graphical forms. In order to capitalize a given input sentence, it must be firstly converted into an FSA (S) and then composed with the transducer T. The resultant transducer contains all possible sequences of capitalized words, given the input lower-case sequence. The *bestpath*() operation over this composition returns the most probable sequence of capitalized words.

In a more theoretical point of view, the capitalization process consists of cal-

[2] see `disambig` manual for more information.
[3] http://www.slt.atr.jp/ IWSLT2006/downloads/case+punc_tool_using_SRILM-.instructions.txt

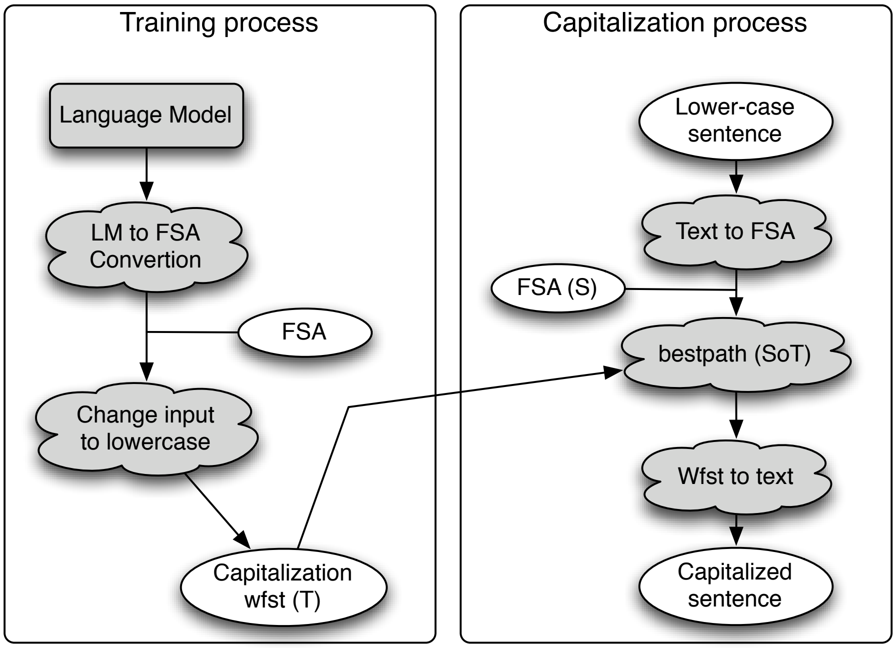


Figure 4. Using a WFST to perform capitalization.

culating the best sequence of capitalized tokens $c \in C^*$ for the lower-case sequence $l \in L^*$, as expressed in equation 4.

$$\hat{c} = \underset{c \in C^*}{argmax}\, P\left(c|l\right) \tag{4}$$

using Bayes' rule:

$$P(c|l) = \frac{P(l|c)\,P(c)}{P(l)} = \frac{P(l,c)}{P(l)} \tag{5}$$

assuming that $P(l)$ is a constant, the capitalization process consists of maximizing the result of $P(l|c) * P(c)$ or $P(l,c)$ as expressed by equation 6.

$$\hat{c} = \underset{c \in C^*}{argmax} P\left(l,c\right) \tag{6}$$

In terms of transducers, the prior $P(c)$ can be computed from the FSA built from the LM, and $P(l|c)$ is computed from the FSA built from the sentence. The composition SoT contains all possible capitalization sequences c for the input sequence l, and the $P(l,c)$ can be computed from all paths associated with sequence c. The Viterbi approximation is used, therefore *bestpath*() operation over the composition returns the c sequence that maximizes the $P(l,c)$ probability.

4.1.3 Maximum entropy

The discriminative modeling approach here described is based on maximum entropy (ME) models, firstly applied to natural language problems in (Berger et al., 1996). A ME model estimates the conditional probability of the events given the corresponding features. Considering a sequence of events E and features F, the ME model takes the form:

$$P(E_i|F) = \frac{1}{Z_\lambda(F)} exp\left(\sum_k \lambda_k f_k(E_i, F)\right) \tag{7}$$

where $Z_\lambda(F)$ is a normalizing term determined by the requirement that

$$\sum_{E_i} P(E_i|F) = 1$$

for all E_i:

$$Z_\lambda(F) = \sum_{E_i} exp\left(\sum_k \lambda_k f_k(E_i, F)\right) \tag{8}$$

$f_k(E_i, F)$ are feature functions corresponding to features defined over events. The index k indicates different features, each of which has an associated weight λ_k. The ME model is estimated by finding the parameters $\lambda_k$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 and also maximize the conditional likelihood $\prod_i P(E_i|F)$ over the training data. In the ME model, decoding is conducted for each sample individually and the correct graphical form of a given word is calculated by means of a weighted sum of values of its corresponding features.

Figure 5 illustrates the ME approach for the capitalization task, where the left side of the picture represents the training process using a set of predefined features, and the right side corresponds to predicting results using previously trained models. 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. This constitutes a training problem, making it difficult to train with large corpora, such as RecPUB corpus. However, classification is straightforward, making it interesting for on-the-fly usage. This framework provides a very clean way of expressing and combining several sources and different aspects of the information, such as word identification and POS tagging information.

The experiments described in this paper use the `MegaM` tool (Daumé III, 2004), which uses conjugate gradient and a limited memory optimization of logistic regression. The `MegaM` tool includes an option for predicting results

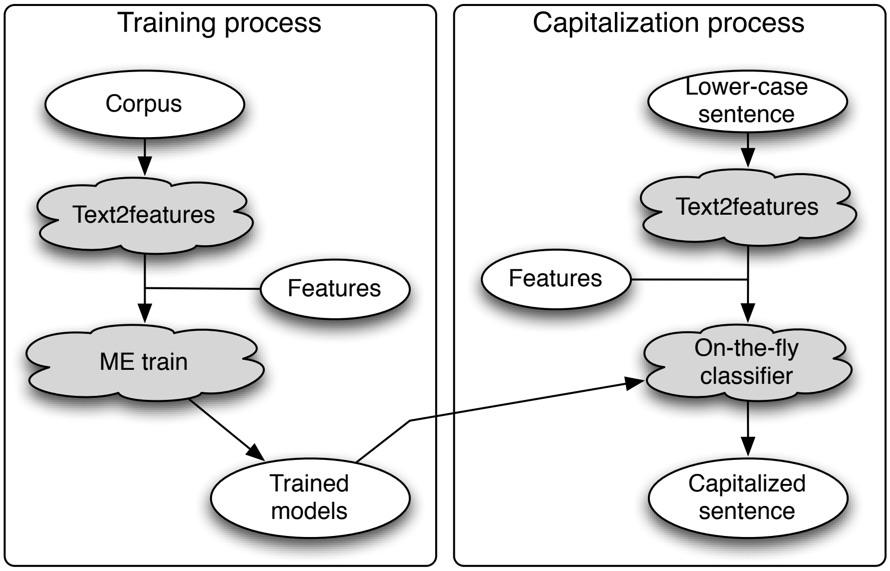


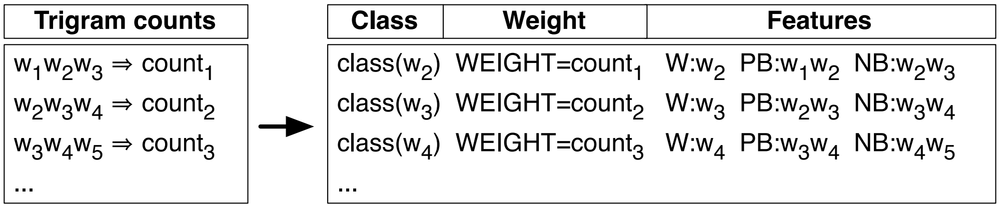
Figure 5. The maximum entropy approach.

from previously trained models. Unfortunately, by the time these experiments started, it was not prepared to deal with a stream of data, producing results only after completely reading the input. An on-the-fly predicting tool was created, that uses the models in the original format and overcomes this problem.

The current implementation of `MegaM` tool also has limitations concerning the size of the corpus (number of observations), so the corpus dimension also constitutes a problem for using ME. This problem occur in the capitalization task and is minimized using a modified training strategy, based on the fact that scaling the event by the number of occurrences is equivalent to multiple occurrences of that event. This strategy consists of counting all n-gram occurrences in the training data and then using such counts for producing the input features. Figure 6 illustrates this process considering trigram counts. The class of each word corresponds to the type of capitalization observed for that word. Each trigram provides feature information for its middle word, namely: W (current word), PB (previous bigram) and NB (next bigram). This strategy maps all the occurrences of a given event into a single input line, allowing to remove less frequent n-grams if desired. It and can be used with higher order n-grams, nevertheless, it is not possible to produce all the desirable representation from n-gram counts, for example, sentences containing less than n words are discarded in n-gram counts, which may conduct to defective results.

4.2 Results

The following experiments assume that the capitalization of the first word of each sentence is performed in a separated processing stage (after punctuation



Trigram counts
$w_1w_2w_3 \Rightarrow count_1$
$w_2w_3w_4 \Rightarrow count_2$
$w_3w_4w_5 \Rightarrow count_3$
...

Class	Weight	Features
class(w_2)	WEIGHT=$count_1$	W:w_2 PB:w_1w_2 NB:w_2w_3
class(w_3)	WEIGHT=$count_2$	W:w_3 PB:w_2w_3 NB:w_3w_4
class(w_4)	WEIGHT=$count_3$	W:w_4 PB:w_3w_4 NB:w_4w_5
...		

Figure 6. Conversion of trigram counts into features.

Table 4
Different LM sizes when dealing with RecPUB corpus

LM options	**unigrams**	**bigrams**	**trigrams**
LM size	3.2MB	31MB	92MB

for instance), since its correct graphical form depends on its position in the sentence. Evaluation results may be influenced when taking such words into account (Kim and Woodland, 2004). As a closed vocabulary is used, all words outside the vocabulary were marked “unknown” and punctuation marks were also removed from corpus. This brings the written newspaper corpus closer to a speech transcription, without recognition errors or disfluencies. The out-of-vocabulary (OOV) words include proper nouns and domain-specific words, but their capitalized form is usually fixed. Hence, they can be handled with domain-specific and periodically updated lexica. The information used in these experiments comprises only the word identification, sometimes combined as bigrams and trigrams.

The next subsections show results achieved with both generative and discriminative approaches. The two approaches are applied to both written newspaper corpora and speech transcriptions. However, the ME approach memory requirements impose limitations in the amount of training data. The first set of experiments are performed on written newspaper corpora, using the RecPUB corpus both for training and evaluation, allowing to establish an upper-bound for capitalization. Results achieved using only the most common graphical form are included in all experiments, which is a popular baseline for similar work (Lita et al., 2003; Chelba and Acero, 2004; Kim and Woodland, 2004).

4.2.1 The generative approaches

A LM created from a big written newspaper corpus may include spelling errors and rare words, which combined as bigrams and trigrams increases the size of the LM without much gain. Thus, all bigrams and trigrams occurring less than 5 times were removed from LMs built from the RecPUB training data. Doing so, a significant reduction in the LM size is achieved without much impact in the results. For example, the size of a bigram LM decreased from 149MB to 31MB (about 15%). Table 4 shows the size of each LM, after this restriction,

Table 5
Results over RecPUB corpus. The left side of the table shows results achieved by using the HMM-based tagger, and the right side shows equivalent results using transducers depending on the building options.

LM options	**HMM-based tagger**			**WFST**		
	Prec	**Recall**	**SER**	**Prec**	**Recall**	**SER**
unigrams	88%	75%	0.345	88%	76%	0.344
bigrams	91%	85%	0.224	91%	86%	0.220
trigrams	92%	87%	0.205	93%	88%	0.189

Table 5 shows results achieved by training and testing on written newspaper corpus. The left side of the table shows results produced by the HMM-based tagger, while the right side shows results produced using the WFST approach, for the same training and testing data. Similar results were expected from both methods, since the transducers were built from exactly the same LM, nevertheless the WFST method achieves a slightly better performance in all experiments. As expected, results improve as the LM order increases: the best results were achieved using trigram models, however the largest difference occurs when moving from unigrams to bigrams. While the ASR output does not contain spelling errors, recognition errors and disfluencies are quite frequent, specially in spontaneous speech. For this reason, results on a written newspaper corpus should be taken as an upper-bound for the capitalization over spoken text.

The remaining experiments concern capitalization of speech transcriptions. The spoken training data is insufficient for training, so both RecPUB and SR training data were combined in order to provide a richer LM. The final LM is a linear interpolation between: LM1 - built from RecPUB training data; and LM2 - built from the SR training data, where the interpolation parameter lambda was 0.759379 for trigrams ($perplexity = 169.2$) and 0.730531 for bigrams ($perplexity = 234.7$). Previous lambda values, calculated using the compute-best-mix tool (included in the SRILM toolkit), minimize the perplexity of the interpolated model, considering the development SR corpus subset (not previously used for training).

Table 6 shows results for capitalization of speech transcriptions. These results reveal the expected decrease of performance when moving from written newspaper corpora to speech transcriptions, specially in terms of precision. The best results are produced with bigrams instead of trigrams, given the weaker linguistic structure of spoken texts, in opposition to written corpora. Since the written newspaper corpora has properties different from speech transcriptions, the availability of more spoken training data would certainly improve these

Table 6
Results over the SR corpus. The left side of the table shows results achieved by the HMM-based tagger and the right side shows equivalent results achieved using transducers

LM options	**HMM-based tagger**			**WFST**		
	Prec	**Recall**	**SER**	**Prec**	**Recall**	**SER**
unigrams	84%	74%	0.401	84%	74%	0.397
bigrams	80%	84%	0.369	80%	85%	0.364
trigrams	78%	85%	0.385	79%	86%	0.369

Table 7
Results of using ME models to capitalize the written newspaper corpus

Training data	**Features**	**Prec**	**Recall**	**SER**
last three months of RecPUB	w_i $2w_{i-1}$ $2w_i$	**93%**	**83%**	**0.229**
all RecPUB corpus, Freq≥5	w_i $2w_{i-1}$ $2w_i$	93%	68%	0.369

results.

Previous results have shown that the WFST method consistently produces better results than using the `disambig` tool. Nevertheless, the current implementation of the WFST method implies loading, composing and searching a large non-deterministic transducer, thus being the most computationally expensive method here proposed.

4.2.2 The discriminative approach

The ME-based approach requires all the information to be expressed in terms of features. The following features are used for a given word w in the position i of the corpus: w_i, w_{i+1}, $2w_{i-1}$, $2w_i$, $3w_{i-2}$, $3w_{i-1}$, $3w_i$, 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. For example: the trigram (w_{i-1}, w_i, w_{i+1}) corresponds to $3w_{i-1}$.

The memory limitations mentioned in subsection 4.1.3 make it difficult to use all written newspaper corpus for training. Therefore, the following experiments use two different strategies: (1) use only the last three months of data for training (about 6 million words); (2) use all training data, by extracting n-gram counts and then producing features for each corresponding n-gram (see Section 4.1.3). Table 7 shows the corresponding results for written newspaper corpora. The first row corresponds to using the first strategy and reveal the best performance in terms of SER. Even if only a small corpus subset is used, results are almost as good as results achieved with generative approaches and

Table 8
Results of using ME models to capitalize the BN transcriptions

Training data	**Features**	**Prec**	**Recall**	**SER**
last three months of RecPUB + SR	$w_i\ 2w_{i-1}\ 2w_i$	81%	83%	0.365
all RecPUB corpus, Freq≥5 + SR	$w_i\ 2w_{i-1}\ 2w_i$	**82%**	**82%**	**0.352**

Table 9
Results of the maximum entropy approach for the SR corpus

Exp	**Features**	**Prec**	**Rec**	**SER**
1	w_i	80%	78%	0.414
2	$w_i\ 2w_{i-1}\ 2w_i$	82%	74%	0.418
3	$w_i\ 2w_{i-1}\ 2w_i\ 3w_{i-2}\ 3w_{i-1}\ 3w_i$	83%	74%	0.413
4	PubLEX	80%	78%	0.414
5	$w_i\ 2w_{i-1}\ 2w_i$ + LEX	82%	76%	0.402
6	$w_i\ 2w_{i-1}\ 2w_i$ + PubLEX	83%	82%	**0.350**
7	$w_i\ 2w_{i-1}\ 2w_i$ + LEX+PubLEX	83%	82%	**0.348**

bigrams. The second strategy uses all corpus by means of trigram counts, but a significant reduction in the recall shows that some phenomena contained in the original text were not correctly captured.

The evaluation of the capitalization task over BN transcriptions also follows the two previously described strategies. In this case, however, the SR training data was used together with the RecPUB training data in order to create the ME models. Table 8 shows the corresponding results. While the first strategy was more adequate for capitalizing written newspaper corpora, the second produces better results for the BN transcriptions, corresponding to the best results seen so far. The second strategy learns the most common capitalization combinations appearing in the corpus, being suitable for the less syntactic restrictions found in the speech transcriptions.

The final experiments uses lexica instead of written corpus in order to minimize the problem of small training datasets. Promising results are expected, while using smaller linguistic resources. Table 9 shows results of several experiments combining different feature sets and lexicon information. Experiment 1 establishes a baseline for what can be achieved using only unigrams and the SR corpus, assuming that if no training material is available for a given word it will be kept lower-case (otherwise a poor 80% SER could be achieved). Experiments 2 and 3 show that using bigrams or trigrams does not improve the SER if the corpus is the only resource used. Experiment 4 shows that using only the most common way of writing a word works better that using the SR

Table 10
Frequency of each punctuation mark in written newspaper corpora. Wall Street Journal (WSJ) results extracted from (Beeferman et al., 1998)

Corpus	**tokens**	"."	","	"?"	"!"
WSJ (English)	42M	4.17%	4.66%	0.04%	0.01%
RecPUB (Portuguese)	130M	3.22%	6.36%	0.10%	0.02%

training corpus, which again indicates that the SR corpus is far from sufficient. Experiment 5 shows that a small lexicon of known words (LEX) contributes to the SER enhancement. The best results were achieved with experiments 6 and 7, combining bigram information from the training corpus with PubLEX.

The best results are achieved by combining the maximum entropy with the PubLEX lexicon. These results are about 1.5% better than best results achieved using the generative approaches, in terms of SER, while much less training data is used. The classification method also provides a fast way of performing capitalization directly from an input stream.

5 Punctuation task

In order to better understand the usage of each punctuation mark, their occurrence was counted in written newspaper corpora, using RecPUB and published statistics from WSJ. Results are shown on table 10, revealing that *comma* is the most common punctuation mark for Portuguese written corpora. The *full-stop* frequency is lower for Portuguese, revealing that the Portuguese written language contains longer sentences when compared to English.

An equivalent study was also performed in Europarl (Koehn, 2005), a multilingual parallel corpus covering 11 languages and extracted from the proceedings of the European Parliament. On this corpus, *comma* is the most frequent punctuation mark in all languages, achieving one of the highest frequency scores for Portuguese (6.75%). Results also confirm that, from all languages, the Portuguese language contains the lowest percentage of *full stops* (3.30% vs. 3.56% for English). All other punctuation marks have shown lower and similar frequencies for all languages.

The previous study was also extended to BN transcriptions. Table 11 shows the corresponding results, performed for Portuguese, English and Spanish languages. The most frequent punctuation mark for Portuguese and Spanish languages is also *comma*, however, this is not the case for English where the *full-stop* punctuation mark is now the most frequent. The Portuguese BN transcriptions present the highest frequency of *comma*. The *full-stop* frequency is

Table 11
Frequency of each punctuation mark in broadcast news speech transcriptions

Broadcast News Transcript	**tokens**	"."	","	"?"
LDC98T28 (Hub4 English)	854k	5.08%	3.52%	0.29%
LDC98T29 (Hub4 Spanish)	350k	4.03%	5.07%	0.14%
SR (Portuguese)	682k	5.02%	8.07%	0.23%

equivalent for English and Portuguese BN transcriptions, and about 1% lower for the Spanish language. The frequency of other punctuation marks on BN corpora is very low.

Previous analysis confirm that spoken text sentences, corresponding to utterances or SUs, are much smaller than written text sentences, specially for the Portuguese language. Intra-sentence punctuation marks also occur more often in spoken texts, concerning the Portuguese language.

The punctuation task benefits from lexical and acoustic information, found in speech transcriptions but unavailable in written corpora. Features, such as pause duration and pitch contour, may be used together with word identification in order to provide clues for punctuation insertion. Thus, spoken data will be the only source of information for the punctuation task. The following subsection will present the steps taken to produce the data, suitable for the training, developing and testing.

5.1 Corpus preparation

The spoken corpus, described in subsection 3.1, provides manual and automatic transcriptions, each one of them in a different format and containing complementary information. For that reason, two different data sources were created, using the same XML format, suitable for experiments both on manual or automatic transcriptions:

MAN – built from manual transcriptions, where part-of-speech data is added to each word.
AUT – built from both manually annotated and automatic transcriptions.

The resultant files of both data sources include information of the APP/ASR output: time intervals to be ignored in scoring, focus conditions, speaker information, punctuation marks, part-of-speech of each word and the word confidence score. These two data sources have exactly the same type of information, allowing the application of the same procedures and tools. The diagram of figure 7 illustrates the creation process of the MAN and AUT data sources. The

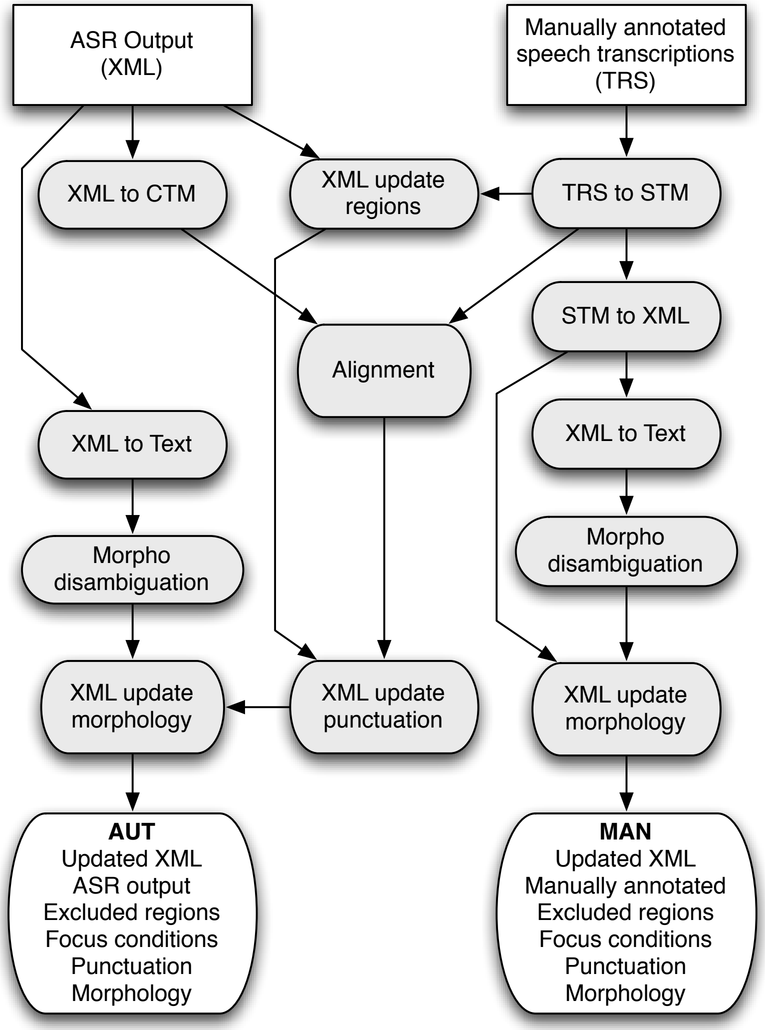


Figure 7. Creation of the MAN and AUT data sources. The following file formats are used: CTM (time marked conversation scoring), STM (segment time mark), TRS (XML-based standard Transcriber), and XML (Extensible Markup Language).

punctuation information was included in the AUT data source by means of a previous automatic alignment between the manual and automatic transcriptions, performed using the NIST SCLite[4] tool. The morphological information was added using the morphological analyzer Palavroso (Medeiros, 1995), followed by the ambiguity resolver MARv (Ribeiro et al., 2004).

Figure 8 shows a transcription segment, extracted from a AUT file, where the focus condition, punctuation and excluded regions information is updated with information coming from the manual transcriptions.

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

```
<TranscriptSegment>
  <TranscriptGUID>2</TranscriptGUID>
  <AudioType start="970" end="1472">Clean</AudioType>
  <Time start="970" end="1472" reasons=""/>
  <Speaker id="1000" name="Homem" gender="M" known="F"/>
  <SpeakerLanguage native="T">PT</SpeakerLanguage>
  <TranscriptWList>
    <W start="970" end="981" conf="0.765016" focus="F0" pos="S.">em</W>
    <W start="982" end="997" conf="0.525857" focus="F0" pos="Nc">boa</W>
    <W start="998" ... conf="0.982816" focus="F0" punct="." pos="Nc">noite</W>
    <W start="1050" end="1064" conf="0.904695" focus="F0" pos="Td">os</W>
    <W start="1065" end="1113" conf="0.974994" focus="F0" pos="Nc">centros</W>
    <W start="1114" end="1121" conf="0.938673" focus="F0" pos="S.">de</W>
    <W start="1122" end="1173" conf="0.993847" focus="F0" pos="Nc">emprego</W>
    <W start="1174" end="1182" conf="0.951339" focus="F0" pos="S.">em</W>
    <W start="1183" end="1229" conf="0.999291" focus="F0" pos="Np">portugal</W>
    <W start="1230" end="1283" conf="0.979457" focus="F0" pos="V.">continuou</W>
    <W start="1284" end="1285" conf="0.967095" focus="F0" pos="Td">a</W>
    <W start="1286" end="1345" conf="0.996321" focus="F0" pos="V.">registar</W>
    <W start="1346" end="1399" conf="0.946317" focus="F0" pos="R.">menos</W>
    <W start="1400" ... "0.851160" focus="F0" punct="." pos="V.">inscritos</W>
  </TranscriptWList>
</TranscriptSegment>
```

Figure 8. Example of a transcript segment extracted from AUT data source.

5.2 Maximum entropy and the feature set

These experiments use real valued features for expressing information, such as word identification, morphological class, pauses, speaker gender and speaker id, sometimes combined as bigrams or trigrams. The following features are used for a given word w in position i of the corpus:

Lexical features:

Word: Captures word identification.
Used features: w_i, w_{i+1}, $2w_{i-2}$, $2w_{i-1}$, $2w_i$, $2w_{i+1}$, where w_i is the current word, w_{i+1} is the word that follows and $2w_{i+x}$ is the word bigram that starts x positions after i.

POS tag: Captures part-of-speech information.
Used features: p_i, p_{i+1}, $2p_{i-2}$, $2p_{i-1}$, $2p_i$, $2p_{i+1}$, where p_i is the part of speech of the word at position i, and $2p_i$ is the POS bigram that starts at position i of the corpus.

Acoustic features:

Speaker changes: Captures speaker id changes.
Used feature: $SpeakerChgs_{i+1}$, true if the speaker id changes before w_{i+1}.

Gender changes: Captures speaker gender changes.
Used feature: $GenderChgs_{i+1}$, true if speaker gender changes before w_{i+1}.

Table 12
Recovering the *full stop* over the ASR output, using only the $Segmchg_{i+1}$ feature. The SER is shown as an absolute value

Focus		**Ref. Slots**	**Prec**	**Rec**	**SER**
All		2470	45%	79%	1.161
F0	planned, clean	391	56%	83%	0.810
F1	spontaneous, clean	111	29%	64%	1.918
F40	planned, noise	930	56%	83%	0.825
F41	spontaneous, noise	791	33%	74%	1.738
F0+F40	all planned	1321	56%	83%	0.821
F1+F41	all spontaneous	902	33%	72%	1.760

Acoustic segments: Captures acoustic segment changes.
Used feature: $SegmChgs_{i+1}$, true if the word w_{i+1} starts a new segment, as previously defined by the APP (Audio Preprocessor) module.
Time: Captures time difference between words.
Used feature: $TimeGap_{i+1}$, the amount of time from the end of word w_i to the start of w_{i+1}.

The score of each word, given by the ASR module, is used for both Word and POS features. For all other features a score of 1.0 is used.

5.3 Results

Several punctuation marks could be considered for this task but, according to the arguments mentioned in subsection 1.2, the following experiments put their focus only on *full stop* and *comma*.

5.3.1 Recovering the full stop *(".")*

This work is now being used in a System for Selective Dissemination of Multimedia Information – SSNT (Amaral et al., 2007), which has been deployed since 2003, and has important requirements concerning the legibility of the results. The system previously used the APP segmentation as the only clue for detecting the sentence boundary, i.e., inserting the *full stop* mark. The performance of the previous system is shown in table 12, where only the APP segmentation is used. These results succeed in terms of recall, but the low precision achieved is translated into an overall SER above 100%. Results for planned speech are better than for spontaneous speech, but no significant difference occurs from noisy to clean speech.

Table 13
Recovering the *full stop* in the MAN data source. The left side of the table shows results, using all the MAN training data, while the right side shows results of training only with the planned speech portion of the MAN training data

Train:	**MAN Training Data**			**Planned Speech only**		
Focus	**Prec**	**Rec**	**SER**	**Prec**	**Rec**	**SER**
All	75%	70%	0.532	73%	72%	0.544
F0	85%	74%	0.383	84%	78%	0.361
F1	65%	60%	0.719	58%	55%	0.842
F40	80%	71%	0.463	79%	75%	0.445
F41	67%	63%	0.679	63%	64%	0.730
F0+F40	82%	72%	0.439	81%	76%	0.420
F1+F41	67%	63%	0.684	63%	63%	0.743

The upper-bound estimate for the methods is achieved with MAN data source, since a manual transcription does not contain ASR errors. Table 13 shows the corresponding results, either using all features and the whole MAN training data or just the planned speech subset. Table 13 shows that an SER of about 53% can be achieved, and that using all training data leads to better results. Better results were expected by reducing the frequency of phenomena, such as disfluencies, from the training data, but results from the right side of the table do not support this assumption and the overall performance decreased about 1.2%. These worse results are related with the reduction of the training material to about 56% of all available training material and also because, by removing the spontaneous part of the training corpus, some spontaneous speech phenomena are not captured. As expected, the performance is much better for planned speech, but no significant differences exist between the clean speech and the noisy one.

The final experiments are evaluated on the AUT data source, which means that they are performed directly on the ASR output. Table 14 shows the corresponding results, either training with MAN or with AUT data source. The best SER performance is achieved when the train is performed with AUT data source, revealing that models trained directly with the ASR output become more suitable for ASR output, although ASR data includes recognition errors, since training and testing data share the same conditions.

The upper-bound SER calculated with the MAN evaluation data is about 19% better than these last results computed over the ASR output, reflecting the performance of the ASR system. The version of the ASR module used has a WER of about 21.5% (15% for planned speech and 30% for sponta-

Table 14
Recovering the *full stop* over real ASR

Train:	**Training using MAN**			**Training with AUT**		
Focus	**Prec**	**Rec**	**SER**	**Prec**	**Rec**	**SER**
All	66%	56%	0.723	70%	53%	0.696
F0	77%	66%	0.534	83%	61%	0.508
F1	62%	41%	0.846	55%	34%	0.936
F40	72%	57%	0.648	77%	55%	0.613
F41	53%	47%	0.943	56%	45%	0.901
F0+F40	74%	60%	0.614	79%	57%	0.582
F1+F41	54%	46%	0.931	56%	43%	0.905

neous speech) (Neto et al., 2008). The performance difference is higher for spontaneous speech, where the ASR WER is also higher.

Results cannot be directly compared with other related work, mainly because data sets are different. Even so, the recent work of Liu et al. (2006) presents several approaches for SU boundary detection on BN RT-04 Eval test data, and some results are quite similar. The paper reports an SER of 47% for broadcast news manual transcriptions, using a maximum entropy approach in combination with a HMM approach, while the maximum entropy approach alone yields an SER of 50%. Using similar training and testing conditions, our results are about 3.2% worse (53.2%). For automatic transcriptions, the same authors report a minimum SER of 57% using HMM and maximum entropy in combination, and 59% for the ME approach alone. For this similar task our experiments show a SER of 69.6% (about 10% higher) in the overall corpus. Liu's work achieves a difference of about 9% between manual and automatic transcriptions. Our experiments reveal a difference of about 19%, mainly because of the large percentage of spontaneous speech in the corpus (34%) and the higher WER of the ASR system, specially for spontaneous speech.

5.3.2 Feature contribution for full stop *results*

Previous results were produced with the combination of all available features, both lexical and acoustic. To assess the contribution of each type of feature, previously described in section 5.2, some experiments were performed and results are illustrated on figure 9. The graph shows that lexical features have less impact than acoustic features on the final performance, however the combination of all features consistently produces the best results.

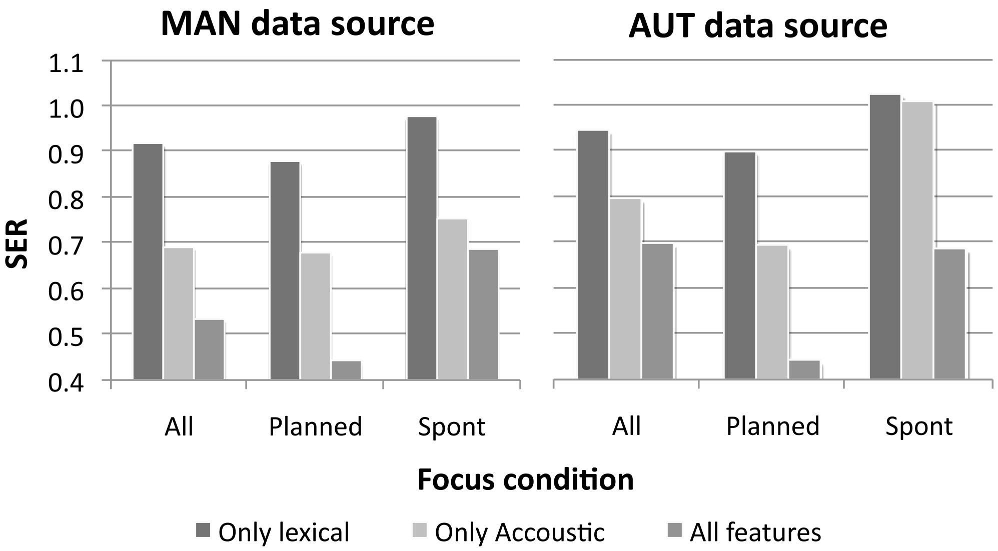


Figure 9. Influence of each feature type in the reduction of the SER by focus condition, for MAN and AUT data sources.

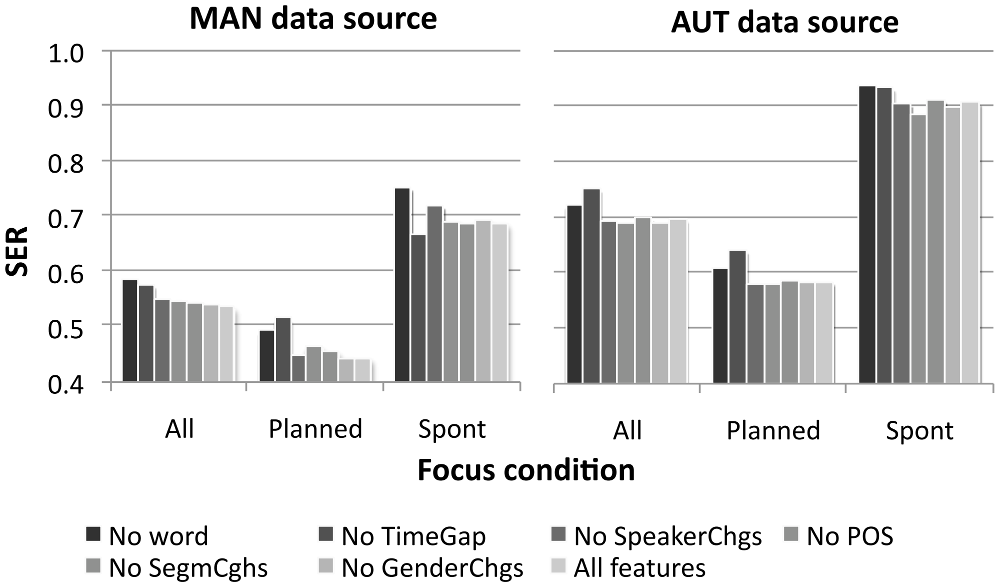


Figure 10. Influence of each feature in reducing the SER by focus condition.

The contribution of each one of the six features, introduced in section 5.2, is also illustrated in figure 10. The figure shows results when using all but a given feature, where: *No Word* means results achieved without word related features; *No TimeGap* means results excluding time intervals between words; *No SpeakerChgs* means results excluding speaker change information; *No POS* means results achieved without part-of-speech related features; *No SegmChgs* means results achieved without segmentation information coming from the ASR; and *No GenderChgs* means results excluding speaker gender change information. The combination of all features produces the best results for the

Table 15
Recovering *comma* in the AUT data source

Focus	Slots	Prec	Rec	SER
All	3727	45%	16%	1.035
F0+F40	1382	41%	17%	1.071
F1+F41	2054	48%	16%	1.010

Table 16
Recovering *full stop* and *comma* in the AUT data source

Focus	Cor	Ins	Del	Sub	Prec	Rec	SER
All	2455	1089	2837	905	55%	40%	0.779
F0+F40	1241	418	1062	400	60%	46%	0.695
F1+F41	958	597	1564	434	48%	32%	0.877

manual transcriptions, but for the automatic transcriptions only the Word and TimeGap information has a relevant influence in the results. Concerning the manual transcriptions, all the features have shown to improve results for all focus conditions. With respect to the automatic transcriptions, results are even improved by removing part-of-speech information, mainly because the POS tagger was not specially adapted for spoken data. The biggest contribution for the automatic transcription results comes from the Word and TimeGap information. However, notice that gender and speaker, as well as time and segmentation information are related, and for this reason their removal has a small impact in the results.

5.3.3 Recovering the comma *(“,”)*

Comma is one of the most frequent and unpredictable punctuation mark appearing in the corpus, its use is highly dependent of the corpus and most of the times there is weak human agreement on a given annotation. For this task, the same approach previously used for recovering the *full stop* is followed, using the same feature set, and results are shown in table 15 for the AUT data. An SER of about 100% is achieved, characterized by a very low recall. Results are consistent with the work reported in (Christensen et al., 2001) for recovering the *comma* on Hub-4 broadcast news corpora, which shows an SER above 80% and for some cases around 100%. This evaluation, however, may not reflect the real achievements of this work, and would benefit from a human evaluation (Beeferman et al., 1998).

In order to better understand the relation between the two punctuation marks, some experiments have also been conducted for recovering *full stop* and *comma* simultaneously. Table 16 shows the achieved results, revealing a significant

a vodafone que controla a telecel em portugal
vá pagar cerca de duzentos milhões de contos por cerca de um terço
do segundo maior operador móvel do méxico
as acções da vodafone registaram hoje forte queda com esta notícia
na bolsa de lisboa e porto o dia foi negativo
jorge pereira tem as notas do diário
início de semana marcado por um fraco volume de negócios pouco mais de dezanove milhões de contos
e pela queda dos títulos da nova economia duas excepções a portugal telecom o título mais negociado
conseguiu inverter a queda nos últimos minutos da sessão
fechou com um ganho muito ligeiro nos dez euros e um cêntimos
a estreia do novo canal da sic permitiu a impresa de pinto balsemão suster
as fortes quedas dos últimos dias
a impresa subiu zero vírgula sete por cento para os seis euros e dez cêntimos

A Vodafone que controla a Telecel em Portugal vá pagar cerca de duzentos milhões de contos por cerca de um terço do segundo maior operador móvel do México.
As acções da Vodafone registaram hoje forte queda com esta notícia.
Na bolsa de Lisboa e Porto o dia foi negativo.
Jorge Pereira tem as notas do diário.
Início de semana marcado por um fraco volume de negócios pouco mais de dezanove milhões de contos.
E pela queda dos títulos da nova economia duas excepções a Portugal Telecom o título mais negociado conseguiu inverter a queda nos últimos minutos da sessão fechou com um ganho muito ligeiro nos dez euros e um cêntimos.
A estreia do novo canal da SIC permitiu a impresa de Pinto Balsemão suster.
As fortes quedas dos últimos dias.
A impresa subiu zero vírgula sete por cento para os seis euros e dez cêntimos.

A Vodafone, que controla a Telecel em Portugal, vai pagar cerca de duzentos milhões de contos por cerca de um terço, do segundo maior operador móvel do México.
As acções da Vodafone registaram hoje forte queda com esta notícia.
Na Bolsa de Lisboa e Porto dia foi negativo.
Jorge Pereira tem as Notas do Diário.
Início de semana marcado por um fraco volume de negócios, pouco mais de dezanove milhões de contos, e pela queda dos títulos da Nova Economia.
Duas excepções, a Portugal Telecom, o título mais negociado, conseguiu inverter a queda nos últimos minutos da sessão, fechou com um ganho muito ligeiro nos dez Euros e onze cêntimos.
A estreia do novo canal da SIC permitiu à empresa de Pinto Balsemão suster, as fortes quedas dos últimos dias.
A empresa subiu zero vírgula sete por cento para os seis Euros e dez cêntimos.

Figure 11. Excerpt of transcribed text (top), automatically enriched (middle), and manually annotated (bottom).

number of substitutions, thus indicating that these two punctuation marks are very close to each other.

6 Concrete example

Figure 11 shows an example of text extracted from an automatic speech transcription, where the first word of each sentence is marked on bold for better identification of the beginning of each sentence. The text at top is splitted into sentences, according to the segmentation proposed by the APP/ASR module. The text at middle was automatically enriched with *full stops* and capitalization information, and the segmentation was performed accordingly to the *full stop* prediction. The first word of each sentence was capitalized in a post-processing stage, as a consequence of the punctuation results. The text at bottom shows the corresponding manual transcription, revealing the recognition, segmentation and capitalization problems.

The text from the example, automatically capitalized and segmented accordingly to the *full stop* prediction, is closer to the manual transcription, and offers a more comfortable reading when compared with the original transcription. The example illustrates one of the most common punctuation problems, resulting from the confusion between the *full stop* and *comma*. In terms of capitalization, the example also illustrates one interesting problem related with the different training and testing periods. All the training data was collected until the end of 2000, but the evaluation data was collected after this period, when Portugal was preparing for the "Euro" currency. Thus, the two occurrences of this word in the text were badly capitalized. Concerning this subject, Mota (2008) has shown that as the time gap between training and test data is increased, the performance of a named tagger based on co-training (Collins and Singer, 1999) decreases.

7 Concluding remarks

This paper addresses two tasks that contribute to the enrichment of the output of an ASR system.

Concerning the capitalization task, three different methods were described and results were presented, both for manual transcriptions of speech and written newspaper corpora. The experiments show that the used speech recognition corpus is too small to cover much of the vocabulary. Another conclusion is that manually built lexica can contribute to enhance the results when the training dataset is reduced, and that, in these conditions, using trigrams does not significantly improve the performance. Finite state transducers produced the best results for written newspaper corpora, but the maximum entropy approach also proved to be a good choice, suitable for the capitalization of speech transcriptions, and allowing straightforward on-the-fly capitalization.

Concerning the punctuation task, a set of statistics on the frequency of each punctuation mark in corpora have been computed. Results show that Portuguese broadcast news transcriptions 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, comparing to newspaper written corpora, revealing shorter sentences. Only the most common punctuation marks were considered for the experiments: *full stop* and *comma*, however results shows that *comma* is quite difficult to predict. Separate results for both spontaneous and planned speech were shown, and the influence of each type of feature in the final result was also analyzed. Achieved results for the MAN data source are similar to other reported work for English broadcast news corpora, however the performance is considerably lower when dealing with the real ASR output, mainly due to possible alignment problems and the higher WER of our ASR module.

An integrated on-the-fly module for punctuation and capitalization recovery has been developed, following the discriminative approach. This module is an important asset in an automatic subtitling system, and has been included in the fully automatic subtitling module for broadcast news, deployed at the national television broadcaster since early March, in the scope of the national TECNOVOZ[5] project.

8 Future work

For the capitalization task, only three ways of writing a word were explored: lower-case, all-upper, first-capitalized, not covering mixed-case words such as McDonald's and SuSE. These words are now being addressed by a small lexicon, but no evaluation was performed so far in order to assess the performance improvement.

The train and test corpora used in these experiments consisted of manually corrected and annotated speech transcriptions. A strategy must be defined in order to perform the evaluation directly on the automatic ASR output, permitting to produce comparative results between manual and automatic transcriptions, and to study the impact of ASR errors in capitalization. Other features, such as word prefix and suffix, number of vowels and consonants shall also be explored. The introduction of information coming from a part-of-speech tagger in the ME models, which has already showed to improve results in (Mikheev, 2002), is also planned.

The broadcast news subtitling module currently uses a baseline vocabulary of 100K words, combined with a daily modification of the vocabulary (Martins et al., 2007) and re-estimation of the language model. This dynamic vocabulary provides an interesting scenario for the capitalization task and is now being addressed.

Concerning punctuation recovery, this study covers the two most common punctuation marks: *full stop*, equivalent to detecting sentence boundaries; and *comma*. Different lexical and acoustic features were combined, but the introduction of other prosodic features is also planned, such as pitch con-

[5] http://www.tecnovoz.com.pt/

tour, already proven to enhance results for detecting sentence boundaries (Liu et al., 2006). The near future plans include a qualitative evaluation based on user satisfaction, specially for intra-sentence punctuation marks, given that a quantitative evaluation may not reflect the real achievements of an automatic system. A study of other marks is also planned, specially the *question mark*, which is also related to one of the SU subtypes in the Metadata Extraction task of the NIST RT-04F evaluation, concerning SU detection.

This work will also be extended to the recognition of classroom lectures, in the scope of LECTRA project (Trancoso et al., 2006), where the use of spontaneous speech in a technical domain poses very interesting problems.

Acknowledgements

This paper has been partially funded by the FCT projects LECTRA (POSC/-PLP/58697/2004) and DIGA (POSI/PLP/41319/2001), and by the PRIME National Project TECNOVOZ number 03/165. INESC-ID Lisboa had support from the POSI program of the "Quadro Comunitário de Apoio III".

References

Amaral, R., Meinedo, H., Caseiro, D., Trancoso, I., Neto, J. P., May 2007. A prototype system for selective dissemination of broadcast news in european portuguese. EURASIP Journal on Advances in Signal Processing 2007 (37507).

Batista, F., Caseiro, D., Mamede, N. J., Trancoso, I., August 2007a. Recovering punctuation marks for automatic speech recognition. In: Interspeech 2007. Antwerp, Belgium, pp. 2153 – 2156.

Batista, F., Mamede, N. J., Caseiro, D., Trancoso, I., September 2007b. A lightweight on-the-fly capitalization system for automatic speech recognition. In: Proceedings of the RANLP 2007. Borovets, Bulgaria.

Beeferman, D., Berger, A., Lafferty, J., 1998. Cyberpunc: a lightweight punctuation annotation system for speech. Proceedings of the IEEE ICASSP, 689–692.

Berger, A. L., Pietra, S. A. D., Pietra, V. J. D., 1996. A maximum entropy approach to natural language processing. Computational Linguistics 22 (1), 39–71.

Chelba, C., Acero, A., 2004. Adaptation of maximum entropy capitalizer: Little data can help a lot. EMNLP '04.

Christensen, H., Gotoh, Y., Renals, S., 2001. Punctuation annotation using statistical prosody models. In: Proceedings of the ISCA Workshop on Prosody in Speech Recognition and Understanding. pp. 35–40.

Collins, M., Singer, Y., 1999. Unsupervised models for named entity classification. In: Proceedings of the Joint SIGDAT Conference on EMNLP.

Daumé III, H., August 2004. Notes on CG and LM-BFGS optimization of logistic regression, `http://hal3.name/megam/`.

Gotoh, Y., Renals, S., 2000. Sentence boundary detection in broadcast speech transcripts. In: Proceedings of the ISCA Workshop: Automatic Speech Recognition: Challenges for the new Millennium ASR-2000. pp. 228–235.

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

Huang, J., Zweig, G., 2002. Maximum entropy model for punctuation annotation from speech. In: Proceedings of the ICSLP. pp. 917 – 920.

Jurafsky, D., Martin, J. H., 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR.

Kim, J., Woodland, P. C., 2001. The use of prosody in a combined system for punctuation generation and speech recognition. In: Proceedings of Eurospeech. pp. 2757–2760.

Kim, J.-H., Woodland, P. C., 2004. Automatic capitalisation generation for speech input. Computer Speech & Language 18 (1), 67–90.

Koehn, P., 2005. Europarl: A parallel corpus for statistical machine translation. MT Summit 2005.

Lita, L. V., Ittycheriah, A., Roukos, S., Kambhatla, N., 2003. tRuEcasIng. In: Proceedings of the 41^{st} annual meeting on ACL. ACL, Morristown, NJ, USA, pp. 152–159.

Liu, Y., Shriberg, E., 2007. Comparing evaluation metrics for sentence boundary detection. In: Proceedings of the IEEE ICASSP. Honolulu, Hawaii.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper, M., 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. IEEE Transaction on Audio, Speech and Language Processing 14 (5), 9.

Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R., Feb. 1999. Performance measures for information extraction. In: Proceedings of the DARPA Broadcast News Workshop. Herndon, VA.

Martins, C., Teixeira, A., Neto, J. P., December 2007. Dynamic language modeling for a daily broadcast news transcription system. In: ASRU 2007.

Medeiros, J. C., 1995. Processamento morfológico e correcção ortográfica do português. Master's thesis, IST/ UTL, Portugal.

Meinedo, H., Caseiro, D., Neto, J. P., Trancoso, I., 2003. Audimus.media: A broadcast news speech recognition system for the european portuguese language. In: PROPOR'2003 - 6th International Workshop on Computational Processing of the Portuguese Language. Vol. 2721 of Lecture Notes in Computer Science. Springer, pp. 9–17.

Mikheev, A., 1999. A knowledge-free method for capitalized word disambiguation. In: Proceedings of the 37^{th} annual meeting of the ACL. ACL, Morristown, NJ, USA, pp. 159–166.

Mikheev, A., 2002. Periods, capitalized words, etc. Computational Linguistics 28 (3), 289–318.

Mota, C., 2008. How to keep up with language dynamics? A case study on Named Entity Recognition. Ph.D. thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

Mrozinsk, J., Whittaker, E. W., Chatain, P., Furui, S., 2006. Automatic sentence segmentation of speech for automatic summarization. In: Proceedings of the ICASSP.

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D., 2008. Broadcast news subtitling system in portuguese. Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, 1561–1564.

Ribeiro, R., Mamede, N. J., Trancoso, I., 2004. Language Technology for Portuguese: shallow processing tools and resources. Edições Colibri, Lisbon, Ch. Morpho-syntactic Tagging: a Case Study of Linguistic Resources Reuse, pp. 31–32.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., Tür, G., 2000. Prosody-based automatic segmentation of speech into sentences and topics. Speech Communications 32 (1-2), 127–154.

Stolcke, A., 2002. SRILM - An extensible language modeling toolkit. In: Proceedings of the ICSLP. Vol. 2. Denver, CO, pp. 901–904.

Stolcke, A., Shriberg, E., 1996. Automatic linguistic segmentation of conversational speech. In: Proceedings of ICSLP '96. Vol. 2. Philadelphia, PA, pp. 1005–1008.

Strassel, S., 2004. Simple Metadata Annotation Specification V6.2. Linguistic Data Consortium.

Trancoso, I., Nunes, R. J. F., Neves, L. M. L., do Céu Guerreiro Viana Ribeiro, M., Moniz, H. G. S., Caseiro, D. A., da Silva, A. I. M., 2006. Recognition of classroom lectures in european portuguese. In: Proceedings of the ISCA conf. Interspeech 2006.

Wang, W., Knight, K., Marcu, D., 2006. Capitalizing machine translation. In: HLT-NAACL. ACL, Morristown, NJ, USA, pp. 1–8.

Yarowsky, D., 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In: Proceedings of ACL '94. pp. 88–95.