



Recovering Punctuation Marks for Automatic Speech Recognition

Fernando Batista^{1,2}, Diamantino Caseiro^{1,3}, Nuno Mamede^{1,3}, Isabel Trancoso^{1,3}

¹L2F – Laboratório de Sistemas de Língua Falada - INESC ID Lisboa
R. Alves Redol, 9, 1000-029 Lisboa, Portugal

²ISCTE – Instituto de Ciências do Trabalho e da Empresa, Portugal

³IST – Instituto Superior Técnico - Universidade Técnica de Lisboa, Portugal
{fmmb, dcaseiro, njm, imt}@l2f.inesc-id.pt

Abstract

This paper shows results of recovering punctuation over speech transcriptions for a Portuguese broadcast news corpus. The approach is based on maximum entropy models and uses word, part-of-speech, time and speaker information. The contribution of each type of feature is analyzed individually. Separate results for each focus condition are given, making it possible to analyze the differences of performance between planned and spontaneous speech.

Index Terms: rich transcription, punctuation recovery, sentence boundary detection, maximum entropy.

1. Introduction

Large quantities of digital and video data are daily produced by media organizations, such as radio and TV stations. Automatic speech recognition systems (ASR) can now be applied to such sources of information in order to enrich it with additional information for applications, such as: retrieval, indexing, captioning, translation, and multimedia content production. The ASR output consists of raw text, often in lower-case format. Even if useful for many applications, such as retrieval and classification, the ASR output benefits from other information, such as punctuation and correct capitalization, for other tasks, such as captioning and multimedia content production. In general, enriching the speech output aims at improving legibility, enhancing information for future human and machine processing. Besides the insertion of punctuation marks, enriching speech recognition covers other activities, such as capitalization and the detection and filtering of disfluencies, not addressed in this paper.

The paper starts by describing our corpus and how we processed it. Section 3 defines the performance measures used for evaluation. Section 4 describes the feature set used by the maximum entropy approach. Section 5 presents results and comments concerning punctuation insertion. The paper ends with some final comments and remarks concerning future work.

2. Corpus description and preparation

The data used for our experiments is a subset of the broadcast news corpus in European Portuguese collected in the scope of the ALERT international project¹. The training data was recorded during October and November 2000, the development data was recorded during December, and the evaluation data was recorded during January 2001². Table 1 presents details for

Corpus	Duration	Tokens	
train	61h	467k	81%
development	8h	64k	11%
test	6h	46k	8%

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

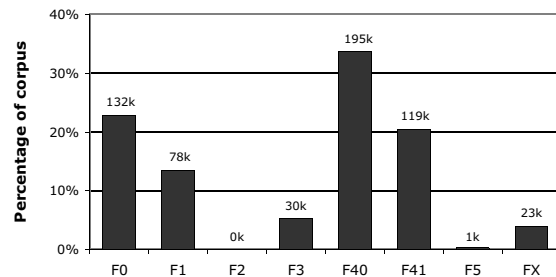


Figure 1: Distribution of words in the SR corpus by focus condition. The number of words is shown on top of each bar.

each part of this corpus, henceforth abbreviated as SR corpus (Speech Recognition).

The manual orthographic transcription of this corpus constitutes our 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 the start and end locations in the signal file, speaker id, speaker gender and focus condition. The orthographic transcription process follows the LDC Hub4 transcription conventions. 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 1, most of the corpus consists of planned speech (F0+F40), nevertheless, 34% is still a large percentage of spontaneous speech (F1+F41).

The corpus was also automatically processed by two modules: the Audio Preprocessor (APP) module and the ASR module. Each segment of the automatically generated transcription has information concerning the background condition (Noise/Clean), speaker index and gender. Each word has a reference for its location in the audio signal, and includes a confidence score given by the ASR module.

¹<https://www.l2f.inesc-id.pt/wiki/index.php/ALERT>

²https://www.l2f.inesc-id.pt/wiki/index.php/ALERT_Corpus

```

<TranscriptSegment>
<TranscriptGUID>2</TranscriptGUID>
<AudioType start="970" end="1472">Clean</AudioType>
<Time start="970" eend="1472" reasons=""/>
<Speaker id="1000" name="Homem" geer="M" known="F"/>
<SpeakerLanguage native="T">PT</SpeakerLanguage>
<TranscriptWList>
<W s="970" e="981" conf="0.765016" focus="F0" pos="S.">em</W>
<W s="982" e="997" conf="0.525857" focus="F0" pos="Nc">boa</W>
<W s="998" e="1049" conf="0.982816" focus="F0" punct="."
pos="Nc">noite</W>
<W s="1050" e="1064" conf="0.904695" focus="F0" pos="Td">os</W>
<W s="1065" e="1113" conf="0.974994" focus="F0" pos="Nc">centros</W>
<W s="1114" e="1121" conf="0.938673" focus="F0" pos="S.">de</W>
<W s="1122" e="1173" conf="0.993847" focus="F0" pos="Nc">emprego</W>
<W s="1174" e="1182" conf="0.951339" focus="F0" pos="S.">em</W>
<W s="1183" e="1229" conf="0.999291" focus="F0" pos="Np">portugal</W>
<W s="1230" e="1283" conf="0.979457" focus="F0" pos="V.">continou</W>
<W s="1284" e="1285" conf="0.967095" focus="F0" pos="Td">a</W>
<W s="1286" e="1345" conf="0.996321" focus="F0" pos="V.">registar</W>
<W s="1346" e="1399" conf="0.946317" focus="F0" pos="R.">menos</W>
<W s="1400" e="1503" conf="0.851160" focus="F0" punct="."
pos="V.">inscritos</W>
</TranscriptWList>
</TranscriptSegment>

```

Figure 2: Example of a transcript segment from HYP data source.

2.1. Corpus preparation

The data source used for training, developing and testing consists of XML files, which gather information from the APP/ASR output and from the manually annotated transcriptions. Two different data sources are created: reference data source (REF), built from the manually annotated transcriptions, where part-of-speech data was added to each word; hypothesis data source (HYP), built from both manually and automatic transcriptions. Each data source comprehends 3 files corresponding to the train, development and test parts of the corpus. The resulting files contain information about regions to be ignored in scoring, focus and speaker information, punctuation marks, and the part-of-speech (POS) of each word together with its confidence score. These two data sources have exactly the same type of information, allowing the application of the same procedures and tools.

The sequence of words of the ASR output is different from the one found in manual transcriptions, since the word error rate (WER) of our ASR system is approximately 12% for planned speech and 24% for spontaneous speech [1]. The NIST SCLite software was used to perform the alignment and for adding the punctuation information to the ASR output. Morphological information was added both for HYP and REF words, using the sequence Palavroso-MARv, where Palavroso [2] is the morphological analyzer and MARv [3] is the ambiguity resolver. No special retrain or adaptation was made in POS tagging procedure for processing the spontaneous speech transcripts.

Figure 2 shows an example of a transcription segment of a HYP file where the focus condition, punctuation and part-of-speech information was updated with information coming from the manual transcriptions.

3. Performance measures

Besides the well-known performance measures: precision, recall, and F-measure, all results will be presented in terms of Slot Error Rate (SER) [4]. For the punctuation task, a slot corresponds to the occurrence of a punctuation mark in the corpus.

$$SER = \frac{\text{total slot errors}}{\text{ref}} = \frac{I + D + S}{C + D + S} \quad (1)$$

Where: C = number of correct slots; I = number of insertions (spurious slots / false acceptances); D = number of deletions (missing slots / false rejections); S = number of substitutions (incorrect slots); ref = number of slots in reference.

Reference: w1 w2 w3 w4 . w5 w6 . w7
Hypothesis: w1 w2 . w3 w4 w5 w6 . w7
ins del cor

Figure 3: Example of slot errors.

Applying the performance measures to the example of Figure 3, a 50% Precision, Recall and F-Measure is achieved, but the SER is 100%, which may be a more meaningful measure.

4. Maximum entropy and the feature set

A maximum entropy approach is followed for combining several sources of information, such as word identification, morphological class, pauses and speaker id. We used the MegaM tool - Maximum Entropy (GA) Model Optimization Package [5] for producing the results. Several different types of features were combined and real valued features were used, some of them combined in bigrams. The following are the features for a given word w in the position i of the corpus:

- Word: Captures word information: $w_i, w_{i+1}, 2w_{i-2}, 2w_{i-1}, 2w_i, 2w_{i+1}$, 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: $p_i, p_{i+1}, 2p_{i-2}, 2p_{i-1}, 2p_i, 2p_{i+1}$, 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.
- Speaker: Captures speaker changes: $Spkrchg_{i+1}$ (true if the speaker id changes before w_{i+1}).
- Acoustic segments: Captures acoustic segment changes: $Segmchg_{i+1}$ (true if the word w_{i+1} starts a new segment, previously defined by the APP module).
- Time: Time difference between words: $TimeGap_{i+1}$ (time interval from word i to word $i+1$). For each x starting at 250ms and double until x overlaps the difference of time between the two words, a feature $TimeGap_{i+1}(x)$ is used.

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

5. Punctuation results

Even if several punctuation marks could be considered for this task, only results about the “full stop” and “comma” will be presented. Other punctuation marks, such as “?” and “!”, are rarely found on broadcast news corpora and previous work on the area has not yet shown promising results.

5.1. Recovering the full stop (“.”)

This work is currently being applied to a System for Selective Dissemination of Multimedia Information (SSNT) [6], which has been operating since 2003. The previous version of this system used the APP segmentation information as the only clue for detecting sentence boundaries, i.e., inserting the full stop mark. Table 2 shows the results achieved when using only the APP segmentation information. The results shows a decrease of performance when dealing with spontaneous speech. However no significant difference occurs from noisy to clean speech.

Focus	Ref. Slots	Prec	Rec	F-measure	SER
All	2526	43%	70%	53%	1.23
F0	397	55%	77%	64%	0.87
F1	111	25%	54%	34%	2.05
F40	944	53%	72%	61%	0.92
F41	817	31%	64%	42%	1.77
F0+F40	1341	54%	74%	62%	0.90
F1+F41	928	30%	63%	41%	1.80

Table 2: Recovering the *full stop*, using only the $Segmchg_{i+1}$ feature. The SER is shown as an absolute value.

Our first tests using additional features were done on the REF data, thus providing the upper-bound limit, since no ASR errors are present in the manual transcription. Table 3 shows that an overall SER of 52% can be achieved, with better precision than recall. As expected, the performance is better for planned speech. Nevertheless, the clean planned speech achieves less 8% SER than the noisy one.

Focus	Prec	Rec	F-measure	SER
All	76%	69%	73%	0.52
F0	85%	74%	79%	0.39
F1	67%	60%	63%	0.69
F40	80%	70%	75%	0.47
F41	69%	64%	67%	0.65
F0+F40	82%	71%	76%	0.45
F1+F41	69%	64%	66%	0.65

Table 3: Recovering the *full stop* in the REF data source.

For the second experiment, only planned speech was used for training the models. We expected to achieve better results by removing phenomena such as disfluencies from the training data. Results are shown on Table 4, revealing an increase of performance when evaluating over planned speech, yet the overall performance decreased about 1%, caused by the reduced training data (56% of all the training material) and because some phenomena, mainly found in spontaneous speech, were not included in the training and thus not captured.

Focus	Prec	Rec	F-measure	SER
All	73%	71%	72%	0.55
F0	84%	77%	80%	0.38
F1	57%	49%	53%	0.89
F40	78%	75%	76%	0.46
F41	64%	63%	64%	0.72
F0+F40	80%	75%	77%	0.44
F1+F41	63%	62%	63%	0.74

Table 4: Recovering the *full stop* in the REF data source, training over planned speech.

The next experiment consisted of applying previously trained models directly to the HYP data (ASR output). Table 5 shows the corresponding results. Taking into account results from table 3, the expected increase of SER can be observed for all focus conditions, mainly caused by a significant decrease of the recall measure. These worse results reflect the impact of moving from manually annotated data to automatic data. Despite that, precision increases for planned speech.

Focus	Prec	Rec	F-measure	SER
All	75%	38%	50%	0.75
F0	91%	39%	54%	0.65
F1	60%	25%	36%	0.92
F40	86%	35%	50%	0.70
F41	58%	37%	45%	0.90
F0+F40	88%	36%	51%	0.69
F1+F41	58%	35%	44%	0.90

Table 5: Recovering the *full stop* over real ASR, training performed with all the training REF data source.

Focus	Prec	Rec	F-measure	SER
All	69%	48%	56%	0.74
F0	81%	58%	68%	0.55
F1	52%	31%	39%	0.98
F40	78%	47%	59%	0.66
F41	53%	42%	47%	0.95
F0+F40	79%	50%	62%	0.63
F1+F41	53%	40%	46%	0.95

Table 6: Recovering the *full stop* over real ASR, training performed with the ASR data source.

In our final experiment, the models were trained and tested using the HYP data source. Results from this experiment are shown on Table 6 where the best performance was achieved so far in what concerns SER. Even though the HYP data includes recognition errors, training with the HYP data source causes the training and testing conditions to be the same, thus making the models more suitable for such data.

Results achieved cannot be directly compared with other related work, mainly because language and data sets are different. Nevertheless, several approaches for SU (Sentence-like Units) Boundary detection on BN RTF-04 Eval test data were found in [7]. Such task is similar to inserting the *full stop* punctuation mark and some similarities can be observed. For the broadcast news REF data, the paper reports an SER of 47% using a maximum entropy approach in combination with an HMM approach, while the maximum entropy approach alone yields a 50% SER. For the Broadcast News ASR output, the same paper reports 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 an SER of 74% in the overall corpus. Nevertheless, for the planned speech 63% were achieved. Yang Liu’s work achieves a difference of about 9% between REF and ASR data. Our experiments reveal a difference of about 22%, mainly because of the proportion of spontaneous speech in our corpus (34%) and the transcription errors made by the ASR module.

5.2. Individual feature contribution for *full stop* results

Previous results were produced by combining all the available features. To assess the contribution of each individual feature some experiments were performed, considering the five types of features previously introduced in section 4. Figure 4 illustrates results achieved when using all but a given type of features, where: All=combining all the features; No-POS=excluding the POS related features; NoWord=excluding the Word related features; NoSpkr=excluding speaker change information; NoTime=excluding time intervals between words;

NoSegm=excluding segmentation information coming from the ASR. Usually, the combination of all features produces the best results, however spontaneous speech contains some exceptions: removing POS information (NoPOS) improves all spontaneous speech results, mainly because the part-of-speech tagger was not trained for such information; also for spontaneous speech, the use of segmentation information produced by the APP system has not shown any improvement in results.

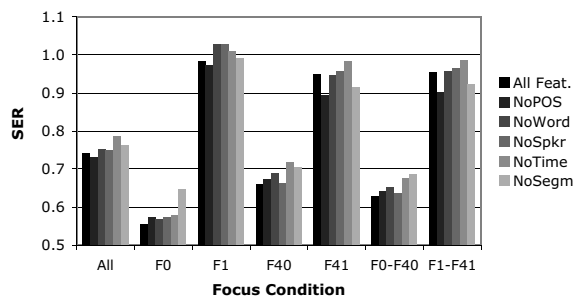


Figure 4: Influence of each feature type, by focus condition.

The features that mostly contribute to our *full stop* results over the ASR output are: the time interval between words, and the segmentation provided by the APP module. The smallest contributions comes from the POS and speaker information.

This study was also performed for the REF data, and results are quite different: word information and speaker information are the most contributing features, followed by the time information. Moreover, all the features have shown to improve results for all focus conditions.

5.3. Recovering the comma (“,”)

Comma is one of the most frequent and unpredictable punctuation marks 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, we follow the same approach previously used for recovering the *full stop*, using the same feature set. Results are shown on table 7 for the REF data. An SER of approximately 100% is achieved with a very low recall. Results are consistent with the work reported in [8] for recovering the *comma* over Hub-4 broadcast news corpora, which shows an SER above 80% and for some cases around 100%.

Focus	Slots	Prec	Rec	F-measure	SER
All	3841	50%	27%	35%	1.00
F0+F40	1382	44%	25%	32%	1.07
F1+F41	2126	52%	29%	37%	0.97

Table 7: Recovering *comma* in the REF data source.

6. Concluding remarks and Future Work

Results for recovering punctuation marks over the ASR output for a Portuguese Broadcast News corpora were presented. Only the most common punctuation marks, *full stop* and *comma*, were considered. Separate results both for spontaneous and planned speech are shown and the influence of each type of feature in the final result is also analyzed. Achieved results for the

REF 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 high proportion of spontaneous speech in our corpus, with a higher word error rate.

The punctuation that humans put on spontaneous speech seems highly subjective specially for the *comma*. This may explain the high SER observed for this punctuation mark. We plan to further research this issue by performing a human evaluation of the acceptability of automatic and human *comma* placements.

In the scope of the national TECNOVOZ³ project, large amounts of broadcast news annotated data are now being daily produced. Soon more training material will be available, which will hopefully provide more accurate results. In the near future, we also plan to introduce other prosodic features, such as the pitch contour and energy, which have already proved to enhance results [7].

We also plan to create an on-the-fly module for punctuation recovery, using the maximum entropy approach. Such a module will be integrated on an automatic subtitling system, currently deployed in the Portuguese national broadcaster RTP.

7. Acknowledgments

This work was partially funded by the FCT project POSC/PLP/58697/2004.

8. References

- [1] R. Amaral, H. Meinedo, D. Caseiro, I. Trancoso, and J. Neto, “Automatic vs. manual topic segmentation and indexation in broadcast news,” pp. 123–128, November 2006.
- [2] J. C. Medeiros, “Processamento morfológico e correção ortográfica do português,” Master’s thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Portugal, 1995.
- [3] R. Ribeiro, N. J. Mamede, and I. Trancoso, *Language Technology for Portuguese: Shallow Processing Tools and Resources*, ch. Morphosyntactic Tagging as a Case Study of Linguistic Resources Reuse. Edições Colibri, October 2004.
- [4] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, “Performance measures for information extraction,” 1999.
- [5] H. Daumé III, “Notes on CG and LM-BFGS optimization of logistic regression.” Implementation available at <http://hal3.name/megam/>, August 2004.
- [6] J. P. Neto, H. Meinedo, R. Amaral, and I. Trancoso, “A system for selective dissemination of multimedia information,” April 2003.
- [7] 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, p. 9, 2006.
- [8] H. Christensen, Y. Gotoh, and S. Renals, “Punctuation annotation using statistical prosody models,” *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, pp. 35–40, 2001.

³<http://www.tecnovoz.com.pt/>