Extending the punctuation module for European Portuguese

Fernando Batista[1,2], *Helena Moniz*[1,3], *Isabel Trancoso*[1,4],
Hugo Meinedo[1], *Ana Isabel Mata*[3], *Nuno Mamede*[1,4]

[1]INESC-ID, Lisbon, Portugal
[2]ISCTE-IUL - Lisbon University Institute, Lisbon, Portugal
[3]FLUL/CLUL, University of Lisbon, Portugal
[4]IST, Lisbon, Portugal

{fmmb;helenam;isabel.trancoso;meinedo;njm}@l2f.inesc-id.pt and aim@fl.ul.pt

Abstract

This paper describes our recent work on extending the punctuation module of automatic subtitles for Portuguese Broadcast News. The main improvement was achieved by the use of prosodic information. This enabled the extension of the previous module which covered only *full stops* and *commas*, to cover *question marks* as well. The approach uses lexical, acoustic and prosodic information. Our results show that the latter is relevant for all types of punctuation. An analysis of the results also shows what type of interrogative is better dealt with by our method, taking into account the specificities of Portuguese. This may lead to different results for different types of corpora, depending on the types of interrogatives that are more frequent.

Index Terms: automatic punctuation, sentence boundary detection, rich transcription, broadcast news subtitling.

1. Introduction

The main motivation of this work is the improvement of the punctuation module of our automatic broadcast news captioning system. Like the speech recognition module and the modules that precede it, the punctuation and capitalization modules share low latency requirements, given their on-the-fly usage.

Although the use of prosodic features in automatic punctuation methods is well studied for some languages, the first implemented version for European Portuguese (henceforth EP) deals only with *full stop* and *comma* recovery, and explores a limited set of features, mostly lexical and acoustic, simultaneously targeting at low latency, and language independence [1].

The aim of this work is to improve the punctuation module, first by exploring additional features, namely prosodic, and later by weighting the lexical and prosodic features impact on the baseline system when encompassing interrogatives. To the best of our knowledge, this is the first study to quantify the distinct interrogative types and also to discuss the weight of the lexical and prosodic properties of these structures, based on planned and spontaneous speech data for EP.

The next section reviews related work. The baseline module and its improvement are described in Sections 3 and 4, respectively. Section 5 deals with the implementation of prosodic features to detect *question marks*. Conclusions and future work are presented in Section 6.

2. Related work

Recent studies (e.g., [2, 3, 4, 5, 6, 7]) show that the analysis of prosodic features is crucial to improve sentence boundary detection systems. Prosodic features, such as pause, final lengthening, pitch reset, and energy, are among the most salient cues in these studies. This is supported on evidences that such cues are language-independent [8] and that languages have prosodic strategies to delimit sentence-like units (SU).

Recent studies have also pointed out that there are prosodic similarities across speaking styles [3]. The authors extracted prosodic features and ranked them accordingly to their impact on the identification of dialog act boundaries. The scaled features are (in order): pause, pitch (log ratio of the pitch at the end of the last word and at the beginning of the first word after a boundary), energy (similar to the analysis window for pitch) and duration.

Detecting positions where a punctuation mark is missing, roughly corresponds to the task of detecting a SU, or finding SU boundaries. SU boundary detection has gained increasing attention during recent years, and it has been part of the NIST rich transcription evaluations. A general HMM (Hidden Markov Model) framework that allows the combination of lexical and prosodic clues for recovering *full stop*, *comma* and *question marks* is used by [9] and [10]. A similar approach was also used for detecting sentence boundaries by [11, 2, 12]. [10] also combines 4-gram language models with a CART (Classification and Regression Tree) and concludes that prosodic information highly improve the results. [13] describes a maximum entropy (ME) based method for inserting punctuation marks into spontaneous conversational speech, where the punctuation task is considered as a tagging task and words are tagged with the appropriate punctuation. It covers three punctuation marks: *comma*, *full stop*, and *question mark*; and the best results on the ASR output are achieved using bigram-based features and combining lexical and prosodic features. [7] proposes a multi-pass linear fold algorithm for sentence boundary detection in spontaneous speech, which uses prosodic features, focusing on the relation between sentence boundaries, break indices and duration, covering their local and global structural properties. Other recent studies have shown that the best performance for the punctuation task is achieved when prosodic, morphologic and syntactic information are combined [6, 12, 14].

Table 1: Portuguese BN corpus properties.

	#Words	Dur. (h)	Planned	Spont.	WER
Train	477k	46	55%	32%	14%
Devel	66k	6	51%	38%	19%
Test	135k	18	56%	36%	19%

3. Baseline punctuation module

Our on-line broadcast news processing system consists of a pipeline of modules that starts with jingle detection, audio diarization, and automatic speech recognition (ASR) [15]. Our ASR system follows the connectionist paradigm. The hybrid models have been recently improved by the inclusion of multiple-state phone units, and a fixed set of phone transition units aimed at specifically modeling the most frequent intra-word phone transitions [16]. Although the system can be improved by using dynamic vocabulary and language models updated daily, the experiments reported in this paper used a fixed vocabulary of 100k words.

Next in our pipeline, come the punctuation and capitalization modules [1, 17]. Both use a discriminative approach, based on maximum entropy (ME) models, which provides a clean way of expressing and combining different properties of the information. This is specially useful for the punctuation task, given the broad set of available lexical, acoustic and prosodic features. This approach requires all information to be expressed in terms of features, causing the resultant data file to become several times larger than the original one. The classification is straightforward, making it interesting for on-the-fly usage.

The experiments described in this paper used the `MegaM` tool [18] for training the maximum entropy models, using conjugate gradient and logistic regression. Our baseline experiments targeted only the two most frequent punctuation marks: *full stop* and *comma*. The following features were used for a given word w in the position i of the corpus: w_i, w_{i+1}, $2w_{i-2}$, $2w_{i-1}$, $2w_i$, $2w_{i+1}$, $3w_{i-2}$, $3w_{i-1}$, p_i, p_{i+1}, $2p_{i-2}$, $2p_{i-1}$, $2p_i$, $2p_{i+1}$, $3p_{i-2}$, $3p_{i-1}$, $GenderChgs_1$, $SpeakerChgs_1$, and $TimeGap_1$, where: w_i is the current word, w_{i+1} is the word that follows and $nw_{i\pm x}$ is the n-gram of words that starts x positions after or before the position i; p_i is part-of-speech of the current word, and $np_{i\pm x}$ is the n-gram of part-of-speech of words that starts x positions after or before the position i. $GenderChgs_1$, and $SpeakerChgs_1$ correspond to changes in speaker gender, and speaker clusters; $TimeGap_1$ corresponds to the time period between the current and following word.

The corpus used in these experiments is the speech recognition subset of the BN (Broadcast News) European Portuguese Corpus, collected during 2000 and 2001 [19]. The manual orthographic transcriptions of this corpus were recently revised by an expert linguist, thereby removing many inconsistencies in terms of punctuation marks that affected our previous results. Table 1 shows some properties of this corpus, where *Dur.* values represent the duration of all speech sequences (silences not included). Although most of the corpus corresponds to planned speech, spontaneous speech is still a significant part. Given the restriction to the two most frequent punctuation marks, all the other marks were converted into one of these two, according to the following rules: “.”: “;”, “!”, “?”, “...” => *full stop*; “,”, “-” => *comma*. The training and evaluation experiments, described here, use the automatic transcripts, produced by our recognition system. The reference punctuation was provided by the corresponding manual transcriptions of the corpus, by means of an alignment between the manual and the automatic transcriptions. This is a non-trivial task mainly because of the recognition errors. The NIST SCLite tool [1] was used for this task, followed by a post-processing step, either by aligning words which can be written differently or by correcting some SCLite basic errors. The data was automatically annotated with part-of-speech information, using MARv [20].

The results achieved with this baseline version are shown in the first line of Table 2. The evaluation used the performance metrics Precision, Recall and SER (Slot Error Rate) [21]. Only punctuation marks are considered as slots and used by these metrics. Hence, the SER is computed by dividing the number of punctuation errors by the number of punctuation marks in the reference data.

4. Improved model for *full stop* and *comma*

In order to introduce prosodic features for detecting SUs, we have performed a number of additional steps. The first step consisted of extracting the pitch and the energy from the speech signal, which was achieved using the Snack Sound Toolkit[2]. Durations of phones, words, and interword-pauses are extracted from the recognizer output. By combining the pitch values with the phone boundaries, we have removed micro-intonation and octave jump effects from the pitch track. Another important step consisted of marking the syllable boundaries as well as the syllable stress. A set of syllabification rules was designed and applied to the lexicon. The rules account fairly well for native words, but need improvements for words of foreign origin. Finally, we have calculated the maximum, minimum, median and slope values for pitch and energy in each word, syllable, and phone. Duration was also calculated for each one of the previous units.

As previously mentioned, our experiments aim at analyzing the weight and contribution of each prosodic feature *per se* and the impact of the combination of prosodic features. Underlying the feature extraction process are linguistic evidences that pitch contour, boundary tones, energy slopes, and pauses are crucial to delimit sentence-like units across languages. First, we have tested if the features would perform better on different units of analysis: phones, syllables and/or words. Supported on linguistic findings for EP [22, 23], we hypothesized that the stressed and post-stressed syllables would be relevant units of analysis to automatically identify punctuation marks. When considering the word as a window of analysis, we are also accounting for the information in the pre-stressed syllables as well.

Features are calculated for each word transition, with or without a pause, using the same analysis window as [3]. The following are the features used: F0 and energy slopes in the words before and after a silent pause, F0 and energy differences between these units and also duration of the last syllable and the last phone. With this set of features we aim at capturing nuclear and boundary tones, amplitude, pitch reset, and final lengthening.

Table 2 shows the results for the *full stop* and *comma* recovery, where each prosodic parameter was analyzed separately. We have found that the feature that most contributes to this task is pitch values for the word (see W_P), that was further improved by adding the energy values also for the word (W_PE). The syllables and phone base features did not show a substantial improvement for this specific task. Moreover, combining words

[1] http://www.itl.nist.gov

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

Table 2: Punctuation performance by combining different prosodic features.

Type of Info	Added features	Exp. ID	Full-stop				Comma				All			
			Prec.	Rec.	F	SER	Prec.	Rec.	F	SER	Prec.	Rec.	F	SER
Baseline		Baseline	68.6%	64.9%	66.7%	64.8%	62.1%	41.4%	49.7%	83.9%	65.2%	50.7%	57.1%	64.6%
Word based	Pitch	W_P	73.1%	66.4%	69.6%	58.1%	63.5%	41.4%	50.1%	82.4%	68.1%	51.3%	58.5%	62.7%
	Energy	W_E	67.2%	67.5%	67.3%	65.5%	63.9%	39.4%	48.7%	82.9%	65.6%	50.5%	57.1%	64.3%
	Pitch & Energy	W_PE	74.0%	65.9%	69.7%	57.2%	63.3%	42.5%	50.9%	82.1%	68.3%	51.8%	58.9%	62.2%
Syllables & phones	Pitch	SP_P	69.2%	67.7%	68.4%	62.5%	62.9%	41.4%	49.9%	83.0%	66.0%	51.9%	58.1%	63.6%
	Energy	SP_E	69.5%	64.3%	66.8%	63.9%	62.5%	40.2%	49.0%	83.9%	65.9%	49.8%	56.7%	64.5%
	Duration	SP_D	69.4%	64.6%	66.9%	63.9%	62.4%	41.3%	49.7%	83.6%	65.8%	50.5%	57.2%	64.2%
	Pitch, Energy, Dur.	SP_PED	70.4%	66.0%	68.2%	61.7%	62.6%	41.3%	49.8%	83.4%	66.4%	51.1%	57.8%	63.8%
All Combined		ALL	72.9%	67.6%	70.1%	57.6%	62.9%	42.7%	50.9%	82.5%	67.6%	52.6%	59.1%	62.3%

Table 3: Performance results recovering the question mark.

	Precision	Recall	F-measure	SER
Baseline	76.5%	16.0%	26.4%	88.9%
+ prosody	79.0%	16.5%	27.3%	87.9%

plus syllables (ALL) achieved results similar to using only word based features (W_PE). The duration parameter is of interest in EP, since three particular strategies are used at the end of an intonational phrase: epenthetic vowel, elongated segmental material or elision of post-stressed segmental material. To the best of our knowledge, no quantifiable measures were reported for our language and little has been said about these strategies so far. Then, not surprisingly, the durational parameter did not add a substantial improvement to our model, although it did contribute to a slightly better result in the spontaneous speech data. In this specific set of data, there is a tendency to elongate the last phone or the last syllable of the word in a potential location for a punctuation mark, making duration an informative cue for this specific context.

Our results partially agree with the ones reported in [3], regarding the contribution of each prosodic parameter and also the set of discriminative features used, where the most expressive feature turned out to be F0 slope in the words and between word transitions. As stated by [8], these features are language independent. Language specific properties in our data are related with different durational patterns at the end of an intonational unit and also with different pitch slopes that may be associated with discourse functionalities beyond sentence-form types.

5. Extension to *question marks*

Detecting *full-stops* and *commas*, depends mostly on a local context, usually two or three words, and corresponds to detecting sentence boundaries. In the other hand, most *interrogatives,* specially the Wh-questions, depend on the words that are used at the begin and at the end of the sentence, which means that sentence boundaries must be previously known. Experiments here reported use the manual sentence segmentation. The same acoustic and prosodic features used for *full stop* and *comma* were also applied to *question mark*. However, lexical features are extracted from the whole sentence, and each event corresponds to a sentence instead of a word.

Given that the BN training corpus contains a small amount of sentences when comparing with newspaper corpora, we have combined two models: the first model created from written corpora (about 150M words) and the other built directly from the speech transcriptions. The achieved results are reported in Table 3, where the first row was achieved using only lexical and acoustic features, and the second one corresponds to combining lexical, acoustic and prosodic features. The results reveal an absolute increase of 1% in F-measure and SER. These results are encouraging, but still far from the ones obtained for *full stop* and *comma*. Nevertheless, other related work also show a lower performance in the detection of *question marks*. For example, [24] reports about 47% precision and 24% recall for English BN, using a very large written corpus for training, but only lexical features. This suggests that a larger corpus would contribute further improvements.

These results motivated an extended analysis of the type of interrogative that is present in our BN corpus. EP has different interrogatives: yes/no questions, alternative questions, wh- and tag questions. A yes/no question requests a yes or no answer, and in EP they generally presents the same syntactic order as a statement (contrarily to English that may encode the y/n interrogative with auxiliary verb and subject inversion). An alternative question presents two or more hypothesis expressed by the disjunctive conjunction “ou”/or. A wh-question have a wh interrogative particle, such as “what”, “who”, “when”, “where”, etc., corresponding to what is being asked about. In a tag question an interrogative clause (most frequently, “não é?” / isn’t it?) is added to the end of a statement.

The overall frequency of interrogatives in our BN corpus is 2.1% . Yes/no questions account for 47.0% of all interrogatives, wh-questions for 40.4%, while tags and alternative questions for 10.0% and 2.6%, respectively. These percentages compare well with the ones for newspapers, but not with the ones found in other corpora analyzed.

Based on language dependency effects (fewer lexical cues in EP than in other languages, such as English) and also on the statistics presented, one can say that ideally around 40.0% of all interrogatives in broadcast news would be mainly identified by lexical cues – corresponding to wh-questions – while the remaining ones would imply the use of prosodic features to be correctly identified. Results pointed out in this direction. A recent study focusing on the detection of *question marks* in meeting transcriptions [25] analyses the relevance of various features in this task, showing that the lexico-syntactic features are the most useful. When training only with lexical features, wh-questions are expressively identified, whereas y/n questions are quite residual, exception made for the bigram “acha que” (do you think). They are still wh-questions not accounted for, mainly due to very complex structures hard to disambiguate automatically. When training with all the features, y/n and

tag questions are better identified. We also have verified that, prosodic features increase the identification of interrogatives in BN spontaneous speech.

6. Conclusions and future work

This paper describes our efforts in extending our punctuation module to a better detection of the basic punctuation marks, *full stop* and *comma*, and to deal with the *question mark*. Reported experiments were performed directly over the automatic speech recognition output, using lexical, acoustic and prosodic features. Results pointed out that combining all the previous features lead to the best performance for the punctuation task. We were also able to discriminate the most relevant prosodic features for this task, being the pitch related ones the most significant *per se*. Though, the best results were obtained when combining pitch and energy. The *full stop* detection consistently achieves the best performance, followed by the *comma*, and finally by the *question mark*. The latter, however, its still in an early stage and can be further improved either by using larger training data and by extending the analysis of pitch slopes with discourse functionalities beyond sentence-form types.

We also aim at extending our work to other corpora in order to see if the weight of the features is dependent on the nature of the corpus and on the most characteristic types of interrogatives in each. In fact, the percentage of interrogatives is highly dependent on the nature of the corpus. For our map-task corpus, for instance, interrogatives represent 22.0% of all the punctuation marks and similar values (20.4%) are found in a university lectures corpus – in both corpora the proportion is ten times more than in broadcast news. This difference is not related only with the percentage of interrogatives across different corpora, but also with their subtypes. The most notorious differences concern i) the highest percentage of tag questions in the university lecture corpus (40.4%), interpretable by the teacher's need to confirm if the students are understanding what has been said; ii) the highest percentage of y/n questions in the map-task corpus (73.6%), related mostly with the description of a map made by a giver and the need to ask if the follower is understanding an instruction.

7. Acknowledgments

The order of the first two authors was chosen randomly. This work was funded by FCT projects PTDC/PLP/72404/2006 and CMU-PT/HuMach/0039/2008. The PhD thesis of Helena Moniz is supported by FCT grant SFRH/BD/44671/2008.This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and by DCTI - ISCTE-IUL – Lisbon University Institute.

8. References

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

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

[3] E. Shriberg, B. Favre, F. J., D. Hakkani-Tur, and S. Cuendet, "Prosodic similarities of dialog act boundaries across speaking styles," in *Linguistic Patterns in Spontaneous Speech*, S.-C. Tseng, Ed. Language and Linguistics Monograph Series A25. Taipei: Institute of Linguistics, Academia Sinica, 2009, pp. 213–239.

[4] R. Carlson, B. Granström, M. Heldner, D. House, B. Megyesi, E. Strangert, and M. Swerts, "Boundaries and groupings - the structuring of speech in different communicative situations: a description of the grog project," in *Fonetik 2002*, 2002, pp. 65–68.

[5] E. Campione and J. Véronis, "A large-scale multilingual study of silent pause duration," in *Speech prosody*, 2002.

[6] M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tur, M. Harper, J. Hillard, J. Hirschber, J. Heng, J. G. Kahn, Y. Liu, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wang, and C. Wooters, "Speech segmentation and spoken document processing," *IEEE Signal Processing Magazine*, no. 25, pp. 59–69, 2008.

[7] D. Wang and S. Narayanan, "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues," in *ICASSP*, 2004.

[8] J. Vassière, "Language-independent prosodic features," in *Prosody: models and measurements*, A. Cutler and R. Ladd, Eds. Berlin: Springer, 1983, pp. 55–66.

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

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

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

[12] 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*, no. 14, pp. 1526–1540, 2006.

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

[14] B. Favre, D. Hakkani-Tür, and E. Shriberg, "Syntactically-informed modules for comma prediction," in *ICASSP*, 2009.

[15] H. Meinedo, M. Viveiros, and J. Neto, "Evaluation of a live broadcast news subtitling system for portuguese," in *Interspeech*, 2008.

[16] A. Abad and J. Neto, "Incorporating acoustical modelling of phone transitions in an hybrid ann/hmm speech recognizer," in *Interspeech*, 2008.

[17] F. Batista, I. Trancoso, and N. J. Mamede, "Comparing automatic rich transcription for portuguese, spanish and english broadcast news," December 2009.

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

[19] J. Neto, H. Meinedo, R. Amaral, and I. Trancoso, "The development of an automatic system for selective dissemination of multimedia information," in *International Workshop on Content-Based Multimedia Indexing*, 2003.

[20] R. Ribeiro, L. Oliveira, and I. Trancoso, "Using morphossyntactic information in TTS systems: comparing strategies for european portuguese," in *Proc. of PROPOR 2003*, 2003, pp. 26–27.

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

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

[23] S. Frota, *Prosody and Focus in European Portuguese. Phonological Phrasing and Intonation.* New York: Garland Publishing, 2000.

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

[25] K. Boakye, B. Favre, and D. Hakkani-Tür, "Any Questions? Automatic Question Detection in Meetings," in *ASRU, Merano (Italy)*, 2009.