Development and Evaluation of Automatic Punctuation for French and English Speech-to-Text

Jáchym Kolář, Lori Lamel

Spoken Language Processing Group, LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
{jachym,lamel}@limsi.fr

Abstract

Automatic punctuation of speech is important to make speech-to-text output more readable and to facilitate downstream language processing. This paper describes the development of an automatic punctuation system for French and English. The punctuation model uses both textual information and acoustic (prosodic) information and is based on adaptive boosting. The system is evaluated on a challenging speech corpus under real-application conditions using output from a state-of-the-art speech-to-text system and automatic audio segmentation and speaker diarization. Unlike previous work, automatic punctuation is scored on two independent manual references. Comparisons are made for the two languages and the performance of the automatic system is compared with inter-annotator agreement.

Index Terms: automatic punctuation, rich transcription, prosody

1. Introduction

Automatic punctuation of speech is important to make speech-to-text (STT) output more readable for humans and more accessible for downstream language processing modules. This paper describes recent efforts towards improving automatic punctuation of speech transcripts under the Quaero project. The Quaero program focuses on the development of multimedia and multilingual indexing and management tools for professional and general public applications. Although Quaero deals with a number of European languages, the current effort on automatic punctuation focuses on two of them – English and French. Other languages are planned to be addressed later. Our initial steps described in [1] proposed guidelines for punctuating speech transcripts, created speech resources with careful punctuation annotation, and analyzed inter-labeler agreement on this annotation. This paper deals with the development and evaluation of an automatic system for punctuation of the STT output.

Methods for automatic punctuation of speech have been investigated in various studies – for example [2, 3, 4, 5, 6, 7]. Although these studies give interesting insight into the problem, most of them did not have the same goal as this paper – to use automatic punctuation within a state-of-the-art STT system working in a realistic setup. Some of the past punctuation studies focused only on comma prediction under the condition that true sentence boundaries are given, or only analyzed sentence segmentation without generation of commas. Some notable exceptions are papers by Christensen et al. [8] and Kim and Woodland [9] investigating automatic punctuation generation for English STT systems, and the more recent work of Batista et el. [10] on punctuating automatic closed-captions for Portuguese TV broadcasts.

The goal of this work is to develop and evaluate an automatic punctuation system working under real-world conditions. The task is to determine the location of punctuation marks for a given word sequence using both textual information (recognized words) and acoustic information (prosody). The punctuation marks are limited to a period, a comma and a question mark. Working with challenging corpora containing large portions of interactive conversational speech, the features are extracted fully automatically using information output by a state-of-the-art STT system (automatic speech/non-speech segmentation, speaker diarization and transcription). The STT word error rate on the data used is relatively high (17–19%), but still usable for some applications in which automatic punctuation can help.

An additional contribution of this work is that the French system presented herein is, to our best knowledge, the first end-to-end punctuation system for spoken French. The only published study about automatic punctuation in French [11] employs lexical and syntactic features in the CRF framework to recover commas in French newspaper text.

Two fundamental information sources – prosodic and textual (recognized words) are used in this work. The next subsection briefly describes the prosodic features, while the representation of the textual information is discussed in Section 2.2 along with a description of the statistical model. A lot of emphasis is also put on proper evaluation of automatically generated punctuation. The system is tested on carefully punctuated transcripts using special speech punctuation guidelines. Two independent punctuation references are used for evaluation. Section 3 discusses choices for the scoring method. System performance is also compared with human inter-annotator agreement.

2. Features and Models

2.1. Acoustic–Prosodic Features

The prosodic features are associated with interword boundaries and capture pause, duration, pitch and energy information. The features are extracted directly from the speech signal using word-level and phone-level time alignment information determined by the speech recognizer. Since the recognizer does not output phone-level time alignment, this information is generated from forced alignment of the 1-best hypothesis.

The pause information is simply the raw pause duration at the word boundary. The duration features include the duration of vowels and final rhymes, aiming mainly to capture the phenomenon of preboundary lengthening. These are normalized using phoneme duration statistics from the whole training set. The pitch features included minimal, maximal and mean f_0 values, f_0 slopes, and differences and ratios of values across word boundaries. These features are extracted from the f_0 contour stylized by a piece-wise linear function. To cut down speaker-dependency, values are normalized based on automatic speaker

IDs. The energy patterns are captured in terms of features representing per-channel normalized frame-level RMS statistics. In order to capture local prosodic dynamics, we also use features associated with the previous and the following word boundaries for some of the feature types. In addition, binary features indicating speaker change at the word boundary according to the automatic speaker diarization are also included in the model. In total there are 30 feaures in the acoustic-prosodic feature set.

2.2. Models and Textual Features

The statistical punctuation model is based on adaptive boosting. The principle of boosting is to combine many weak learning algorithms to produce an accurate classifier. Each weak classifier is built based on the outputs of previous classifiers, focusing on the samples that were formerly classified incorrectly. The general boosting method can be combined with any classifier. In this work, the algorithm called *AdaBoost.MH* [12] as implemented in the Icsiboost toolkit [13] was employed. It combines weak classifiers having a basic form of one-level decision trees (stumps) using confidence-rated predictions. The test at the root of each tree can check for the presence or absence of an N-gram, or for a value of a continuous feature. Hence, the approach allows a straightforward combination of lexical and prosodic features in a single statistical model. An additional advantage of this model is that it is able to handle features that may take undefined values. This is very useful for the punctuation task where the problem of undefined feature values arises on the edges of speech segments or in regions where the forced phone-level alignment failed and the prosodic features could not be extracted.

Textual information is captured by word N-grams. That means, all N-grams up to 4-grams containing the word before the boundary of interest, plus the unigram right after the boundary are extracted at each interword boundary. To capture word repetitions, a binary flag indicating whether the two words across the boundary are identical or not is also included in the feature vector.

The boosting-based approach does not allow the system to directly benefit from additional large text corpora since there are no prosodic features associated with words and the learning algorithm assumes that all features are available during training. To overcome this problem, we trained an additional large LM in an HMM framework (hidden-event language model [14]) and the posteriors from this model were subsequently used as extra features during training and testing of the overall model. Thus, the boosting training only iterates over the data for which the acoustic information is available. The auxiliary LM was trained on standard text corpora distributed by LDC (e.g., Gigaword) containing 3.3G running words for English and 0.7G for French.

When using the classifier, we do not perform the 4-way classification directly but first join end-of-sentence marks (i.e., periods and question marks) into a single class. Then, in a second pass, a separate question/statement classifier is used to distinguish periods and question marks. The second pass classifier uses the first pass prosodic features at the boundary plus N-gram features from the whole sentence. The reason for this setup is that some cue N-grams characterizing questions typically appear at the beginning of the sentence (for example, wh-pronouns or the bigram "do you"). This information has been shown to be very important for question detection [15]. However, since the question classifier used relies on inaccurate automatic sentence boundaries, its performance cannot be as good as when using manual sentence segmentation.

3. Evaluation Method

The choice of the evaluation metric for the automatic punctuation task is not straightforward. There is no commonly agreed-upon metric. In the following paragraphs, we review some metrics proposed to evaluate the quality of automatic punctuation. We use the following notation: C – number of correctly assigned punctuation marks, D – number of deletion errors, I – number of insertion errors, S – number of substitution errors, N – number of words in the test set.

Classification Error Rate views the punctuation task as a classification at each word boundary.

$$CER = \frac{D + I + S}{N} \times 100\%$$

This view corresponds well to the way an automatic system works, but the measure also has drawbacks. The main problem is that it does not reflect the target class imbalance inherent to the punctuation task. As most words (roughly 75–90% depending on the genre and language) are not followed by any punctuation mark, using this metric results in seemingly modest error rates even when the classifier always outputs the majority class, i.e. no punctuation. The metric is also not convenient for system comparisons across domains or languages.

Precision, Recall & F-measure are metrics well known from information retrieval. The punctuation task is viewed more like an event detection problem – correctly classified "no-punctuation" interword boundaries do not increase the figure of merit.

$$P = \frac{C}{C + I + S}, R = \frac{C}{C + D + S}, F = \frac{2PR}{P + R}$$

While the P and R pair gives informative insight about the system performance, the use of F-measure as a single metric in a multiclass classification task is rather problematic. As argued by Makhoul et al. [16], the deletion and insertion errors are deweighted by a factor of two in comparison with the substitution errors.

Slot Error Rate was proposed in the above cited paper to be used in place of the F-measure.

$$SER = \frac{D + I + S}{C + D + S} \times 100\%$$

Similar to F, it gives no credit for correctly classified no-punctuation boundaries, but it weights all error types equally. An unnatural aspect of this metric is that it may be greater than 100% – if the number of insertion errors was higher than the number of correctly detected marks. The chance performance (i.e. classifying all test samples as no punctuation) corresponds to $SER = 100\%$.

Cohen's Kappa is a measure of inter-annotator agreement defined as

$$K = \frac{A_o - A_e}{1 - A_e}$$

where A_o denotes the observed agreement and A_e stands for the expected agreement (i.e., based on label priors). If the annotators are in complete agreement, then $K = 1$, while agreement expected by chance corresponds to $K = 0$. Although the metric was proposed to measure agreement between two humans, it can also be used to measure agreement between a human and an automatic system.

Among the above stated options, the SER was chosen as the most convenient metric for the automatic punctuation task. In

Table 1: Data set sizes (#words) and STT word error rate

	French	English
#Train	431.1K	417.4K
#Development	49.2K	42.5K
#Test	34.3K	43.8K
WER	19.0%	17.3%

Table 2: Class distribution [%] in dev (D) and test (T) data

	FR-D	FR-T	EN-D	EN-T
No punctuation	81.4	79.3	86.4	86.1
Comma	11.2	13.7	7.7	8.0
Period	6.7	6.2	5.2	5.3
Question mark	0.6	0.8	0.6	0.6

this paper, we also employ the K metric to compare the quality of the system with human inter-annotator agreement.

In contrast to previous work, two independent manual punctuation annotations were available for all test sets [1]. There are two general ways how to use the dual references to get more robust error estimates. The first one is to calculate SER for each of the two references separately and then compute their mean. We call this approach *average* (SER_AVG). It decreases the error rate variance but seems to be too strict for the automatic system since at word boundaries where the two references differ, the system is always wrong with respect to at least one of them. In such cases, the system should not be penalized for using the other possible solution. To moderate the number of unjust score penalizations, another scoring method is proposed. In the *lenient* (SER_LEN) approach, a classification is considered as incorrect only if it does not match either of the two references.

4. Experimental Conditions

This work uses about 50 hours of speech in French and English selected and transcribed for the Quaero project. The data for each language are split between Broadcast News (BN) and more varied data including talk shows, debates and web podcasts collectively called Broadcast Conversation (BC). The ratio between BN and BC is approximately 30% to 70%. Several speakers appear in each show and the BC data are usually very interactive, containing lots of spontaneous and overlapping speech. All the original transcripts were manually repunctuated based on special guidelines as described in [1].

The data split between training, development and test sets is shown in Table 1. The development data were used to tune the prosodic features and the auxiliary language model, and to optimize the number of iterations in boosting. The experiments reported below were then performed on the test data. The test sets were used in the Quaero P3 (2010) speech recognition evaluation. Table 1 also shows the word error rates of the LIMSI-Vocapia STT system [17] used in this work. To investigate the impact of word recognition errors, all experiments were evaluated using both forced alignment of human-generated reference transcripts (REF) and automatic (STT) transcripts. To generate the "reference" punctuation for the STT words, the reference setup was aligned to the recognition output in terms of minimum-cost edit distance taking into account word similarities and time spans. Target class priors in the development and test sets are shown in Table 2.

5. Results and Analysis

To give more insight about the importance of the types of information, the system is tested with three different feature sets. In the first (Text), the classifier only employs textual features, in the second (Prosody) only prosodic features, and in the third both. The overall punctuation error rate is given along with the partial results for the individual punctuation marks. The results are presented in terms of SER_AVG and SER_LEN defined in Section 3. The latter is taken as the main error metric.

Tables 3 and 4 present results for French and English, respectively. Of the two individual information sources, the models only using textual features work much better than the prosody-only model, however, the Text+Prosody systems always perform significantly better than the partial systems. The relative SER_LEN reduction by adding prosodic features ranges from 11.0 to 19.6% depending on the test condition and language. The error reduction is higher for the REF condition.

The comparison of the two metrics, SER_AVG and SER_LEN, shows that the latter is on average lower by 11% relative for all the overall values, and by 15% relative for the Text+Prosody systems. In general, the relative difference between the two numbers is higher for systems that perform at lower error rates. The comparison of the overall SER between the two languages shows that while for the REF conditions, the French system works significantly better, for the STT conditions, the error rates are almost equal. The results for the individual marks indicate that the better performance of the French model in REF is from a major part given by the good predictivity of commas by the textual model. French commas seem to be easier to detect based on words, but this advantage is diminished in the presence of transcription errors. The period detection performs about the same for both languages with the REF, but for STT the English model is again more robust to the recognition errors. The detection of question marks works better for English in both conditions. Despite the two-pass approach, SER for this type of punctuation stays rather high. Since question marks are sparse in the data (their proportion is only around 0.7%), it is more difficult to train a robust model for their detection.

We also compare the performance of the automatic systems with the inter-annotator agreement in terms of the K statistic. This comparison gives nice insight since the inter-annotator agreement corresponds to an upper bound of the possible system performance. K for the human-computer agreement is computed as the average of K between the system output and each of the two manual references. The values of K are displayed in Table 5. The direct comparison can only be made on REF because we do not have manual annotations based on STT words, however, K is also reported on STT for illustration. For REF, the K for agreement between the automatic system and a human is lower than between two humans, by 23.4% relative for French and 26.6% relative for English.

6. Summary and Conclusion

We explored the task of automatic punctuation from speech in French and English under real-world conditions. A boosting-based punctuation model relying on both textual and prosodic information was developed. The method was evaluated on difficult speech data, still challenging for the current STT technology. Unlike previous work, the evaluation was performed using two independent reference punctuation annotations. To investigate the impact of word recognition errors, all experiments were performed using both forced alignments of human-generated

Table 3: Automatic punctuation Slot Error Rates [%] on the French test set

	Overall		Period		Comma		Quest. mark	
FR-REF	**SER_AVG**	**SER_LEN**	**SER_AVG**	**SER_LEN**	**SER_AVG**	**SER_LEN**	**SER_AVG**	**SER_LEN**
Text	67.0	56.6	84.5	80.6	67.3	55.7	67.0	56.6
Prosody	86.3	80.9	71.5	67.6	94.6	89.2	100.0	100.0
Text+Prosody	56.6	**45.2**	59.6	52.3	63.1	50.0	83.4	81.7
FR-STT								
Text	84.0	75.5	92.2	89.1	83.5	73.8	99.3	99.1
Prosody	94.1	89.4	88.2	84.3	97.6	93.1	100.0	100.0
Text+Prosody	76.5	**67.2**	77.1	71.4	79.9	69.3	95.1	94.6

Table 4: Automatic punctuation Slot Error Rates [%] on the English test set

	Overall		Period		Comma		Quest. mark	
EN-REF	**SER_AVG**	**SER_LEN**	**SER_AVG**	**SER_LEN**	**SER_AVG**	**SER_LEN**	**SER_AVG**	**SER_LEN**
Text	74.0	63.4	73.7	68.3	80.7	68.0	87.5	86.4
Prosody	89.7	86.0	78.3	75.3	98.9	96.9	99.6	99.6
Text+Prosody	65.3	**53.5**	58.7	52.7	78.3	64.0	71.9	68.9
EN-STT								
Text	85.5	78.1	83.8	80.2	89.7	80.7	92.9	92.2
Prosody	91.0	88.1	79.9	77.2	100.3	99.4	100.4	100.4
Text+Prosody	77.2	**68.0**	69.8	65.1	88.2	76.4	78.5	76.4

Table 5: Comparison of Inter-Human and Human-Automatic system agreement on punctuation annotation [K]

	Inter-Human	Auto-REF	Auto-STT
French	0.819	0.627	0.492
English	0.800	0.587	0.503

transcripts and STT transcripts.

With manual transcriptions, the French system performed better than the English, but under the STT conditions, the SERs were very similar for both languages – 67.2% for French and 68.0% for English. The relative increase in SER due to speech recognition errors was 25.8% for English and 47.7% for French. For the reference conditions, the automatic system performance could also be compared with the inter-annotator agreement. The K statistic for the agreement between a human and the automatic system was lower than the agreement between two humans by 23.4% relative for French and by 26.6% relative for English. As these numbers indicate, automatic punctuation in the real-application conditions is a very difficult task.

In the future work, we plan to explore different modeling techniques. Moreover, we want to investigate the impact of automatic punctuation on transcript readability by performing perceptual tests.

7. Acknowledgments

This work was achieved as part of the Quaero Program funded by OSEO, French State Agency for Innovation.

8. References

[1] J. Kolář and L. Lamel, "On development of consistently punctuated speech corpora," in *Proc. INTERSPEECH*, Florence, 2011.

[2] D. Beeferman, A. Berger, and J. Lafferty, "Cyberpunc: A lightweight punctuation annotation system for speech," in *Proc. ICASSP*, Seattle, WA, USA, 1998.

[3] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. ICSLP*, Denver, CO, USA, 2002.

[4] Y. Liu et al., "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[5] B. Favre, D. Hakkani-Tür, and E. Shriberg, "Syntactically-informed models for comma prediction," in *Proc. ICASSP*, Taipei, Taiwan, 2009.

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

[7] J. Kolář, Y. Liu, and E. Shriberg, "Speaker adaptation of language and prosodic models for automatic dialog act segmentation of speech," *Speech Communication*, vol. 52, no. 3, pp. 236–245, 2010.

[8] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," in *Proc. ITRW Prosody in Speech Recognition and Understanding*, 2001.

[9] J. H. Kim and P. Woodland, "A combined punctuation generation and speech recognition system and its performance enhancement using prosody," *Speech Communication*, vol. 41, no. 4, pp. 563–577, 2003.

[10] F. Batista et al., "Extending the punctuation module for European Portuguese," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010.

[11] C. Cerisara, P. Král, and C. Gardent, "Commas recovery with syntactic features in French and in Czech," in *Proc. INTERSPEECH*, Florence, Italy, 2011.

[12] R. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2–3, pp. 135–168, 2000.

[13] B. Favre, D. Hakkani-Tur, and S. Cuendet, "Icsiboost," http://code.google.come/p/icsiboost, 2007.

[14] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *Proc. ICSLP*, Philadelphia, USA, 1996.

[15] K. Boakye, B. Favre, and D. Hakkani-Tur, "Any questions? Automatic question detection in meetings," in *Proc. ASRU*, Merano, Italy, 2009.

[16] J. Makhoul et al., "Performance measures for information extraction," in *Proc. DARPA Broadcast News Workshop*, Herndon, 1999.

[17] L. Lamel et al., "Speech recognition for machine translation in Quaero," in *Proc. IWSLT*, San Francisco, CA, USA, 2011.