Automatic Comma Insertion of Lecture Transcripts Based on Multiple Annotations

Yuya Akita, Tatsuya Kawahara

Academic Center for Computing and Media Studies, Kyoto University,
Kyoto 606-8501, Japan

Abstract

To enhance readability and usability of speech recognition results, automatic punctuation is an essential process. In this paper, we address automatic comma prediction based on conditional random fields (CRF) using lexical, syntactic and pause information. Since there is large disagreement in comma insertion between humans, we model individual tendencies of punctuation using annotations given by multiple annotators, and combine these models by voting and interpolation frameworks. Experimental evaluations on real lecture speech demonstrated that the combination of individual punctuation models achieves higher prediction accuracy for commas agreed by all annotators and those given by individual annotators.

Index Terms: Automatic punctuation, lecture speech, conditional random fields, multiple annotations

1. Introduction

Automatic speech recognition (ASR) research has recently been focused on various spontaneous speech such as public speeches [1], classroom lectures [2] and congressional meetings [3]. ASR of these kinds of spontaneous speech is useful for speech translation, captioning and documentation. To ensure readability of resulting captions and documents, ASR output should be punctuated into proper units. Moreover, these units are essential for succeeding language processing such as machine translation, which assumes punctuated texts as inputs. However, an ASR system usually produces a sequence of words without any punctuation. For human reading and automated systems, automatic punctuation is an important issue.

Many studies on automatic punctuation of speech transcripts have mainly focused on periods, i.e., sentence boundary detection, and it has been explored on broadcast news and conversation tasks. Popular approaches to automatic punctuation adopt machine learning frameworks such as maximum entropy, support vector machines (SVM) and conditional random fields (CRF), with prosodic, pause and linguistic information [4]. We also proposed sentence boundary detection of Japanese lectures using SVM with linguistic and pause information [5]. On the other hand, previous work on prediction of commas has been limited [6, 7], and its accuracy is much lower than that for periods. Compared to periods, there is much disagreement in commas between humans, since insertion of commas is more frequent and subjective. Commas given by a single annotator, which were often used as references in previous studies, are not always reliable, therefore we use different punctuation annotations given by multiple annotators.

In this paper, we address automatic punctuation of Japanese lecture speech using multiple punctuation annotations. First, we analyze manual punctuation of lecture speech in terms of variations between annotators. Then, an automatic punctuator is designed based on the CRF framework. Specifically, we train CRF which independently model the tendency of punctuation of each human annotator, then combine these models to make more reliable prediction. We train general and personalized punctuators, and evaluate the performance over real lecture speech.

2. Corpus and annotations

For analysis of punctuation, we used lecture speech in "the Corpus of Spontaneous Japanese" (CSJ) [8], which was a collection of speech and transcripts of academic presentations and extemporaneous public speeches. We chose 70 presentations and 107 speeches for our analysis. The CSJ includes audio data, transcripts and annotations such as pauses and disfluencies. Since no punctuation marks were given to transcripts in the CSJ, we conducted manual annotation of periods and commas by three professional stenographers independently. As a result, three punctuated transcripts were obtained for each of 177 lectures. Note that simple edits, such as removal of fillers, substitution of colloquial expressions and modification of end-of-sentence expressions, were performed on the faithful transcripts in the CSJ before punctuation, by other annotators. The total size of 177 lectures after the editing was 365,305 words. The three annotators did not listen to the corresponding speech, i.e., the annotation was made by referring to the transcripts only.

The transcripts are automatically split into lexical units by a parser, as there is no word boundaries in Japanese texts. In Japanese, a sentence can be broken into several syntactic units called *bunsetsu*, each of which is comprised of one or a few words, and used as a basic unit of dependency analysis and parsing. Basically, commas are put at *bunsetsu* boundaries, however, not all *bunsetsu* units can have commas. Typical purposes of commas in Japanese texts are (1) sign of an end of a phrase, (2) listing up several elements, like "A, B, C," (3) clarification of syntactic dependency on lexical elements (i.e., which *bunsetsu* modifies what), and (4) segmentation of a word sequence to make it easy to read, as Japanese texts do not have any spaces between words. Purposes (1), (2) and (3) are similar to those in other languages such as English, while (4) is peculiar to Japanese. Commas for (3) and (4) are subjective and these can be inserted in various ways, however, too many commas are not preferred. Consequently, *bunsetsu* units which modify the next unit tend to have no commas.

3. Analysis of punctuation marks

In this section, we investigate differences of punctuations given by multiple annotators. We also investigate how linguistic and pause information is associated with commas.

Table 1: Numbers of punctuation marks by annotators

Annotator	Commas	Periods
A	29,393	16,958
B	23,371	16,972
C	19,854	16,969

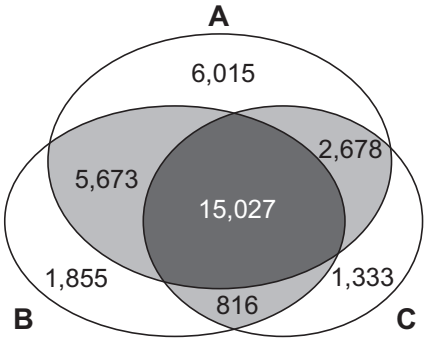

Figure 1: Overlaps of commas given by three annotators

3.1. Comparison of punctuation marks between annotators

First, we compared the numbers of commas and periods in the punctuated transcripts and overlaps between the annotators. Table 1 lists the numbers of commas and periods given by annotators A, B and C. The numbers of periods are almost same among three annotators. Actually, 97% of periods are common to all annotators. In contrast, the numbers of commas are significantly different between the annotators. Annotator C gave only two-thirds of commas given by annotator A. Figure 1 shows overlaps of commas by the three annotators. The number of commas jointly given by all annotators is 15,027, which is 51%, 64% and 76% of commas given by A, B and C, respectively. On the other hand, 20% of A's (6,015), 8% of B's (1,855) and 7% of C's (1,333) commas are inserted only by a single annotator. The statistics suggest that the number and position of commas are affected by human subjects, even if they are professional stenographers.

3.2. Typical linguistic expressions around commas

It is hypothesized that people have their own punctuation points, especially for commas, when writing sentences. To verify individual tendencies of comma insertion, we investigate linguistic expressions which appear with commas. By counting words followed by/following commas inserted only by a single annotator (i.e., white areas in Figure 1), we found typical expressions to each annotator. For example, annotator A often inserted commas after postpositional particles, such as *wa* and *ga*, which indicate grammatical cases. In contrast, annotators B and C inserted many commas after conjunctive words, however, specific words were different; annotator B gave commas after *soshite* (then), *tsumari* (that is) and *sunawachi* (that is), while annotator C gave after *aruiwa* (or/possibly) and *sorekara* (then).

3.3. Correlation with pauses

Pauses are often used as key features for automatic punctuation of speech recognition results. Although annotators did not listen to the lecture speech, i.e., they did not refer to pause information for punctuation, we extract pause information from the CSJ and investigate co-occurrence with punctuation marks. Figure 2 shows a histogram of duration of pauses detected in the lecture speech, together with counts of periods and commas given at the pauses. In the CSJ, pauses were measured by hand, but pauses shorter than 0.2 seconds were not annotated, thus not included in Figure 2. Here we used transcripts punctuated by annotator A to calculate statistics in Figure 2. The distribution of pauses with punctuation marks is almost common among the three annotators. Pauses longer than 1.0 second are likely to be followed by periods or commas, but they do not account for majority of punctuations. On the other hand, the majority associated with commas are shorter than 0.5 seconds, but 47%–55% of them are not associated with any punctuation marks. The ratio is almost constant regardless of the pause duration, however, it is apparently higher than that in the case of no pauses. Hence, the occurrence of pauses helps comma prediction, while the duration information of pauses is not expected to do so.

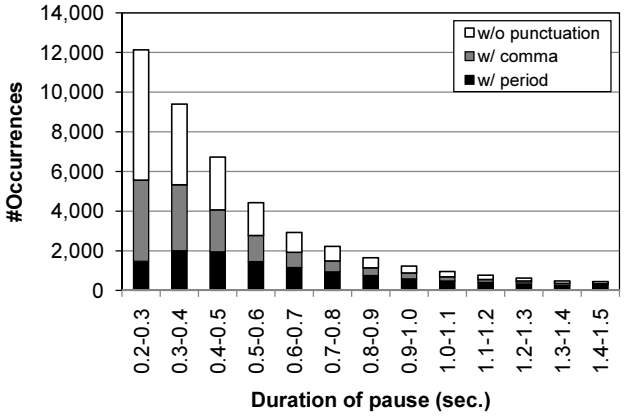

Figure 2: Correlation of pauses and punctuation marks

4. Automatic punctuation method

4.1. CRF-based modeling

Based on the analysis described above, we design an automatic punctuation method. As a modeling framework, conditional random fields (CRF) are adopted. We used CRF++[1] to train and test models for punctuation. Features of CRF were word surface form, part-of-speech (POS) tag and boundaries of *bunsetsu*. We also used local syntactic dependency which was defined only at adjacent pairs of *bunsetsu* units, because a *bunsetsu* unit which has such dependency is strongly associated with the next, and thus no commas are usually put there. Estimation of local dependency is expected to be robust, while the long dependency structure is often hard to identify. All of these lexical and syntactic features were automatically extracted by a Japanese morphological analyzer and a parser. As for pause features, pauses longer than 0.2 seconds were extracted and used. We did not use the duration information of pauses, because the correlation of the pause duration and commas is not strong as shown in Figure 2. CRF used these features of previous and following three words, as well as those of the current input word. In the

[1] http://crfpp.sourceforge.net

Table 2: Comparison of various combinations of features for automatic punctuation

Features used	Periods			Commas		
	Recall	Precision	F-measure	Recall	Precision	F-measure
Word	0.972	0.969	0.971	0.611	0.729	0.665
Word+bunsetsu boundary	0.975	0.974	0.975	0.647	0.764	0.700
Word+bunsetsu+dependency	0.978	0.983	0.981	0.698	0.768	0.731
Word+POS	0.974	0.973	0.974	0.624	0.764	0.687
Word+POS+bunsetsu	0.976	0.973	0.975	0.679	0.768	0.721
Word+POS+bunsetsu+dependency	0.979	0.983	0.981	0.713	0.774	0.742
Word+POS+bunsetsu+dependency+pause	0.975	0.984	0.980	0.734	0.784	0.758

experiments described below, every evaluation metric was averaged over results of 10-fold cross validation on 177 lectures mentioned in Section 2, unless otherwise indicated.

4.2. Effects of various features

First, we evaluated the effect of each feature, by changing combination of features. In this experiment, we conducted training and evaluation using periods and commas given jointly by at least two annotators. Table 2 shows recall rate, precision rate and F-measure of punctuation by various CRF models trained with different sets of features. For periods, high performance was achieved even with the single word feature. This is because the evaluation was done over manually edited transcripts, where typical end-of-sentence expressions can be easily detected. On the other hand, there were no dominant features for comma prediction. All features synergistically improved performance of comma prediction, as each feature represented different aspects of comma insertion.

4.3. Prediction of general commas

Next, we evaluated the performance of the CRF-based comma prediction over several punctuation labels. As features for experiments hereafter, all features used in the previous experiment were adopted. We prepared labels of six types. As general punctuation labels, "3," "2+" and "1+" were defined based on the commonness of punctuation marks among the three annotators. The label "3" was made from punctuation marks given by all of the three annotators, "2+" given by at least two annotators, and "1+" given by at least one annotator. These labels are chosen based on multiple human subjects and hence considered to be general. In contrast, punctuation labels given by each annotator were used as personalized labels "A," "B" and "C."

For prediction of general commas (i.e., general labels), we directly trained CRF models using the respective general labels. Furthermore, we trained personalized models using "A," "B" and "C" labels, then conducted voting using the results of these models. Here, we performed three types of voting; "Any" (punctuation adopted if at least one model votes), "Majority" (at least two models vote) and "Consensus" (all three models vote). These should be compared with the direct modeling of "1+," "2+" and "3," respectively. In other words, voting was done for training labels in the direct modeling, while voting by personalized models were done at the time of prediction.

Table 3 shows the results of comma prediction by the direct modeling and voting by personalized models. In the case of "3" test labels, where commas were common to all annotators, F-measure was 0.620 by the "3" model. On the other hand, in case of "1+" test labels where every possible point of commas should be predicted, F-measure was 0.822 by the "1+" model. These results suggest that prediction of possible points is relatively easier than common (i.e., essential) commas. As for voting results, "Consensus" voting achieved higher F-measure than that of the "3" model. "Majority" voting was almost comparable to the "2+" model, and "Any" voting did not improve the performance. This result suggests that combination of multiple models trained independently by different labels is effective for prediction of commas based on a certain criterion, while the direct modeling better models arbitrary commas.

Table 3: Results of comma insertion for general labels

Direct modeling

Test label	1+	2+	3
Training label	1+	2+	3
Recall	0.814	0.734	0.559
Precision	0.830	0.784	0.695
F-measure	0.822	0.758	0.620

Voting by A,B and C models

Test label	1+	2+	3
Training label Voting type	A,B,C Any	A,B,C Majority	A,B,C Consensus
Recall	0.774	0.729	0.633
Precision	0.849	0.786	0.652
F-measure	0.810	0.756	0.642

1+: Labels given by at least one annotator,
2+: Labels given by at least two annotators,
3: Labels given by all annotators,
A/B/C: Labels given by each annotator

4.4. Prediction of personalized commas

Next, we investigated personalized modeling of commas. For each of personalized labels "A," "B" and "C," personalized models are tested. We also tested the "1+" model, which realizes high recall and precision rates for possible commas, as shown in Table 3. Moreover, we introduced interpolation of the personalized and the "1+" models. In the CRF framework, a probability is calculated for every prediction result, and classification is performed based on the probabilities, i.e., the result which has the largest probability is selected as an output. Here, the personalized and the "1+" models give probabilities $P_{\mathrm{personal}}(C|X)$ and $P_{1+}(C|X)$, respectively, to a classification output C for an input feature vector X. Then, we interpo-

Table 4: Results of comma insertion for personalized labels

Test label		A	B	C
Personalized (A/B/C) model only	Recall	0.772	0.712	0.617
	Precision	0.799	0.776	0.711
	F-measure	0.785	0.743	0.661
1+ model only	Recall	0.832	0.877	0.859
	Precision	0.758	0.635	0.529
	F-measure	0.793	0.737	0.655
Weighted interpolation (A/B/C & 1+)	Recall	0.803	0.793	0.741
	Precision	0.786	0.725	0.644
	F-measure	0.795	0.758	0.689

For A, B, C and 1+, refer to Table 3.

Table 5: Results of comma insertion on automatic transcripts

Test label		1+	2+	3
Manual transcripts	Recall	0.821	0.735	0.525
	Precision	0.827	0.775	0.715
	F-measure	0.824	0.754	0.605
Automatic transcripts	Recall	0.601	0.493	0.315
	Precision	0.494	0.435	0.354
	F-measure	0.542	0.462	0.334

For 1+, 2+ and 3, refer to Table 3.

late these two probabilities to make a final decision:

$$P(C|X) = \lambda P_{\text{personal}}(C|X) + (1-\lambda)P_{1+}(C|X). \quad (1)$$

The interpolation weight λ was set as 0.6 in this experiment. Here, the best value was chosen a posteriori.

Table 4 shows the results of comma prediction for annotator A's, B's and C's labels by the corresponding personalized models, the general "1+" model and the weighted interpolation of these models. The interpolated model achieved the highest performance among the three types of models. The combination with other annotator's information is useful for enhancing the personalized model.

4.5. Evaluation on automatic transcripts

Finally, we tested the prediction model on ASR results of lectures. In this experiment, we used eight lectures as a test set. The number of words in this test set is 17,925, and the word error rate is 17.1%.

Table 5 lists the results for automatic transcripts together with corresponding manual transcripts, in case of "1+," "2+" and "3" labels used for both training and testing. Compared to the results over manual transcripts, the performance was significantly lower for automatic transcripts. One reason for degradation is ASR errors, but the other major reason is the editing process conducted on the transcripts, as described in Section 2. Here we applied simple rule-based transformation of end-of-utterance expressions to the ASR results, which was not sufficient for automatic punctuation. We need to improve this transformation to realize higher punctuation performance.

5. Conclusions

We have addressed automatic punctuation of lecture speech using CRF with lexical, pause and syntactic information. We first confirmed different tendencies in comma insertion between professional annotators. Therefore, we adopted an approach to make personalized models and combine them. Using different punctuation labels given by multiple annotators, punctuation models dedicated to respective annotators are trained. By combining these personalized models, the performance of comma prediction was improved for both general and personalized criteria.

Acknowledgment: This work was supported by JST CREST and JSPS Grant-in-Aid for Scientific Research.

6. References

[1] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan, "An Audio Indexing System for Election Video Material," in *Proc. ICASSP*, 2009, pp. 4873–4876.

[2] J. Glass, T. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent Progress in the MIT Spoken Lecture Processing Project," in *Proc. Interspeech*, 2007, pp. 2553–2556.

[3] Y. Akita, M. Mimura, G. Neubig, and T.Kawahara, "Semi-automated Update of Automatic Transcription System for the Japanese National Congress," in *Proc. Interspeech*, 2010, pp. 338–341.

[4] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper, "Structural Metadata Research in the EARS Program," in *Proc. ICASSP*, vol. 5, 2005, pp. 957–960.

[5] Y. Akita, M. Saikou, H. Nanjo, and T. Kawahara, "Sentence Boundary Detection of Spontaneous Japanese Using Statistical Language Model and Support Vector Machines," in *Proc. Interspeech*, 2006, pp. 1033–1036.

[6] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, "Recovering Punctuation Marks for Automatic Speech Recognition," in *Proc. Interspeech*, 2007, pp. 2153–2156.

[7] B. Favre, D. Hakkani-Tur, and E. Shriberg, "Syntactically-informed Models for Comma Prediction," in *Proc. ICASSP*, 2009, pp. 4697–4700.

[8] S. Furui, K. Maekawa, and H. Isahara, "Toward the Realization of Spontaneous Speech Recognition —Introduction of a Japanese Priority Program and Preliminary Results—," in *Proc. ICSLP*, 2000, pp. 518–521.