Sentence Detection Using Multiple Annotations

Ann Lee, James Glass

MIT Computer Science and Artificial Intelligence Laboratory,
Cambridge, Massachusetts 02139, USA
{annlee, glass}@mit.edu

Abstract

In this paper, we develop a sentence boundary detection system which incorporates a prosodic model, word and preterminal-level language models, and a global sentence-length model. An important aspect of this research was the investigation of crowdsourced punctuation annotations as a source of multiple references for evaluation purposes. In order to evaluate the system we propose a BLEU-like metric which compares a hypothesis to multiple references. Experiments on both transcription and ASR output show that the global sentence length model can improve the performance by 7.2% on reference transcripts and 3.8% on ASR output.

Index Terms: sentence boundary detection, prosody, finite-state transducer, amazon mechanical turk

1. Introduction

The output of an automatic speech recognizer (ASR) typically does not contain punctuation. However, many natural language processing (NLP) applications, such as in machine translation, parsing, and information retrieval, assume that text has been partitioned into sentence-like units (SUs). For these reasons, automatic sentence boundary detection is often used to insert breaks into ASR-generated text, which not only improves the readability of the output, but also bridges the gap between ASR and subsequent NLP applications [1, 2].

One challenging aspect for evaluation of sentence boundary detection systems is to determine appropriate reference annotations. In many cases for spontaneous speech, punctuation decisions can be somewhat arbitrary, so it may be reasonable to collect multiple alternatives rather than having a single gold standard. Such an approach has been employed for machine translation [3]. A related aspect to data annotation is the evaluation metric. F-score and Slot Error Rate (SER) are the two most commonly used evaluation metrics. F-score is the harmonic mean of precision and recall, where precision is the fraction of retrieved documents (sentence breaks in our case) that are relevant to the search, and recall is the fraction of the retrieved documents that are relevant to the query. SER is the ratio between the number of punctuation generation errors and the number of punctuation marks in the reference. Both of these metrics fail to capture the variable nature of annotations because they compare the generated result to only one reference.

In this paper, we describe our efforts to detect sentence boundaries in a corpus of spoken restaurant reviews. We have explored the use of multiple sources of punctuation that are obtained via crowdsourcing on Amazon Mechanical Turk. We also introduce a BLEU-like scoring metric that enables a system output to be compared to multiple annotations. After describing our data collection methods and evaluation metric in the next section, we describe the sentence boundary detection system, which incorporates both global constraints of sentence length, and local constraints containing prosodic and language model information based on a finite-state transducer (FST) implementation. After presenting evaluation results of the system using the new metric, we summarize and suggest directions for future research.

2. Annotating and Evaluating Punctuation

2.1. Restaurant Review Corpus

The experiments performed in our research are based on a previously recorded and transcribed corpus of 512 spoken restaurant reviews of up to one minute in length from 135 subjects [1]. The content includes specific comments on food, service and atmosphere, as well as overall reviews. The data were recorded using cell-phones, and are spontaneous in nature.

2.2. Crowdsourcing Punctuation Annotation

We published tasks on Amazon Mechanical Turk (AMT) for collecting punctuation annotations. Each turker (the person performing the task) was asked to listen to the audio and then insert commas or periods at appropriate places in the transcripts. Each review was annotated by ten different turkers. The ten annotations for each review were assessed in a jack-knifing fashion for quality check. Each individual annotation obtained a score which was the average of the F-score obtained by comparing it to each of the nine others. When the average score fell below a threshold, it was manually checked to ascertain if it was reasonable or not.

Ultimately we collected 3,713 annotations from 177 turkers for the 375 reviews that were longer than ten words. Table 1 shows some statistics of the annotated re-

Table 1: *Statistics from Punctuation Annotation*

type	# reviews	Avg. # words	commas	periods
Ambience	69	56.1	5.8%	4.9%
Cuisine type	38	16.6	10.4%	7.7%
Food quality	69	44.8	5.4%	5.9%
Service	70	47.3	5.9%	5.2%
General	129	86.2	5.3%	5.2%

sults. We can see that longer reviews (all except cuisine type) have a similar percentage of commas and periods, which implies that subjects tend to pause after saying a certain number of words (i.e., to take a breath or think of what to say next). The "cuisine type" review was the only outlier in this regard, and we believe this is because of the short nature of the responses (e.g., "Indian food" or "American traditional breakfast").

To examine the variation among annotations, we randomly chose one as the reference and another one as the hypothesis from the multiple annotations, repeated this procedure 100 times and computed the average F-score. When we treat commas and periods differently, the average F-score is 0.48, while it is 0.64 if we treat them the same. These results indicate that the variation between different annotating styles is not something we can completely ignore. Table 2 shows an example of two different annotations on the same review.

2.3. Evaluation Metric

Given the availability of multiple reference annotations and inspired by the fact that the BLEU score can achieve high correlation with human evaluation when judging a machine translation system [3], we present a modified version for evaluating a sentence boundary detection system. The BLEU score is defined as:

$$BLEU = BP \times \exp(\sum_{n=1}^{N} w_n \log p_n), \qquad (1)$$

where p_n is the modified n-gram precision, BP the sentence brevity penalty, N the maximum n-gram length that is considered, and w_n the weighting factor, which is usually set to $1/N$. The modified n-gram precision can be computed as:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}, \qquad (2)$$

where $Count(n\text{-}gram)$ is the total number of appearance of an n-gram pattern in the hypothesis texts, and $Count_{clip}(n\text{-}gram)$ is the smaller one between $Count(n\text{-}gram)$ and the maximum number of times such n-gram pattern occurs in any single reference texts. The sentence brevity penalty can be computed as:

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c \leq r \end{cases}, \qquad (3)$$

where c is the total length of all the hypotheses, and r is the sum of the length of references whose length best match that of the corresponding hypothesis.

Table 2: *Different Annotations on the Same Review*

(a) the food quality was great. we started with a salad that was really good. for the meal we had a philly sandwich. the mushrooms were cooked perfectly, and the fries were great as well.
(b) the food quality was great. we started with a salad. that was really good for the meal. we had a philly sandwich, the mushrooms were cooked perfectly and the fries were great as well.

For sentence boundary detection evaluation, we only consider n-gram patterns that reflect the places of SU breaks. We keep the computation of p_n, since it works well in capturing opinions from different references and punishing the results that over-generate breaks. Now the denominator of p_n counts all n-gram sequences consisting of locations of the hypothesized SU breaks, and the numerator counts all hypothesized n-gram sequences containing locations of SU breaks that also occurred in at least one of the references. However, instead of looking for a reference annotation that is the closest to the hypothesis in length when computing r in BP, we choose the reference that best matches the hypothesis in terms of F-score. In other words, we should consider the reference with the annotation style that is the closest to that of the hypothesis.

3. Sentence Boundary Detection System

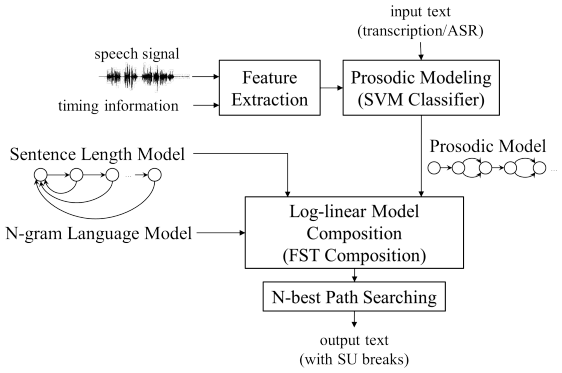


Figure 1: *System diagram*

3.1. Related Work

There has been considerable prior research in sentence boundary detection. The problem of sentence segmentation can be modeled as a tagging problem, and a variety of machine learning techniques, such as a hidden Markov model (HMM), maximum entropy (Maxent) and conditional random field (CRF) models, have been explored [4, 5, 6]. Many studies have shown that combining both textual and prosodic features produces better results than utilizing each of them in isolation [4, 5, 7].

One problem with the tagging approach is that the prosodic features and the language models only consider local information. To solve this problem, Matusov et. al [2] processed the texts from left to right, and con-

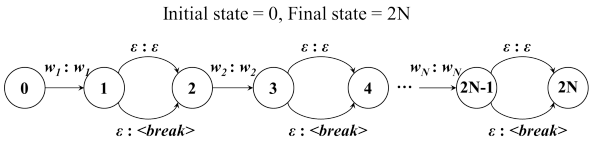


Figure 2: *Prosodic model of an input string of length N*

strained the search space between the pre-defined minimum and maximum sentence length. Within this search space, a log-linear combination of language model, prosodic model and a log-normal prior on sentence length, was adopted to compute the score of each inter-word boundary.

Our system, as shown in Figure 1, is most similar to that in [2]. The difference is that we formulate our system by using Finite-State Transducers (FSTs), and thus we can convert the problem to searching for a path that leads to a globally optimal result from the linear combination of different models.

3.2. Prosodic Model

Table 3 shows the prosodic features we extracted, following those that have been explored previously [4]. Timing information can be obtained through forced alignment on transcript or directly from a recognizer's output along with ASR results. The pause duration value can be represented as raw time, or normalized by the maximum pause length in the same utterance. Phone duration is normalized by subtracting the mean and dividing by the standard deviation from the training data. For fundamental frequency, we used an autocorrelation method to compute the F0 contour, and smooth the result using a five point median filter.

Table 3: *Prosodic Features*

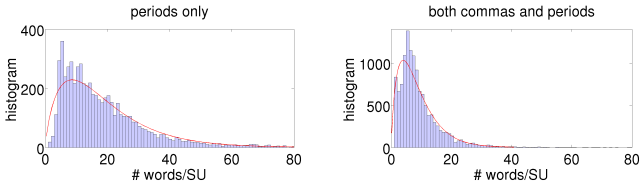
category	features
pause duration	pause length between the current word and the next word, the duration of any pause that precedes the current word
phone duration	the duration of the last phone in the current word, the durations of the vowels in the current word, and the maximum normalized phone duration
F0	difference across an inter-word boundary, difference between the end/the mean of the current word or the beginning/the mean of the next word and the minimum F0 value in the utterance, maximum/minimum F0 value within the current/next word
energy	features are extracted using the same fashion as for F0

Given the prosodic features, we trained an SVM classifier and used its posterior probability output to build the prosodic model. The FST implementation can be done by inserting two arcs between each neighboring words (Figure 2). One arc outputs an empty string, while the other outputs a sentence break ($\langle break \rangle$). The weights on these arcs are the negative log probability output from the SVM classifier.

3.3. Language Model

We collected text-based restaurant reviews from the web to train both a tri-gram and a four-gram LM, with "$\langle break \rangle$" as a word. The FST implementation of a LM takes any word in the vocabulary as input, and will output the same word itself.

In addition to a word-level language model, we also take advantage of a Context Free Grammar (CFG) parser, in which we have a set of self-defined grammars and 391 preterminals, including "$\langle break \rangle$". From the parsable sentences in the training data, we can train up a preterminal-level LM.



(a) sentence-level SU (b) clausal-level SU

Figure 3: *Histograms of the number of words per SU and the fitted gamma distributions*

3.4. Sentence Length Model

Figure 3 contains two histograms of the number of words per SU for the restaurant review corpus, including all types except the general reviews, which will be our test data. We can fit the histogram by a gamma distribution, which is often used in NLP for modeling word length. On the basis of the fitted gamma distribution, a sentence length model can be built in the format shown in Figure 4.

In this FST, the penalty on the arcs from *state n* to *state 0* is the negative log probability from the fitted gamma distribution, since it represents a SU with length n, while the weight of the arc from *state k* to *state k+1* is $-\log[1 - \sum_{i=1}^{k} p(l = i)]$. There is a pre-defined maximum length, L, which is restricted to be 80 in our case.

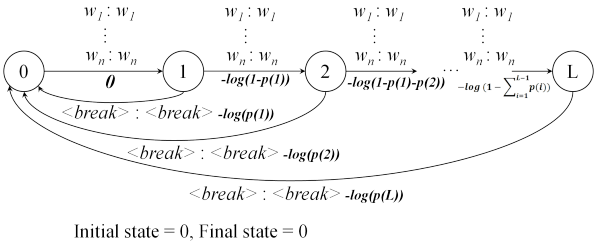


Figure 4: *Sentence length model*

3.5. Model Combination

We compose our models in the order of prosodic model, language model and sentence length model. Since the weights in our FST models are all in a negative log format, the entire composition process is a log-linear combination of the three models.

4. Experimentation

4.1. Experimental Settings

In our experiments we have focused on detecting clausal SU breaks, i.e. breaks that are indicated by commas or periods. The data for training the prosodic model includes all types of spoken reviews except the general ones, resulting in 383 spoken reviews containing 13.2K words. The extra text reviews from the web for training LMs contain around 12.3M words. We chose the 129 general reviews as test data, which contains 4.8K words. The experiments were done on both reference transcripts and ASR output. We try different combinations of the models we have, and for each scenario, the experiment was carried out 100 iterations. During each iteration, we randomly chose five annotations from each review as the reference and one as the hypothesis for human labeling. An RBF kernel was used for the SVM classifier, and the parameters were tuned for each scenario. The modified BLEU metric was used for evaluation. We computed n-gram precision up to tri-gram, and set all w_n's to $\frac{1}{3}$. Different weightings on the models were also explored.

4.2. Performance on Transcription

Table 4 shows the experimental results on transcripts. The choice of LM is the one that works the best under each scenario. In this case, using both word-level and preterminal-level LMs can achieve better performance, since the latter incorporates some sentence-level syntactic information. Utilizing a prosodic model achieves a relatively low performance because it over-generates sentence breaks. Combining the prosodic model with LMs can filter out inappropriate breaks and greatly improves the performance. Introducing a global sentence length model further improves performance by 7.2%. From the results we can see that p_n of prosodic+LM is similar to that of prosodic+LM+sent-len. The reason why the former achieves a lower BLEU-like score is that it generates too few breaks, and thus is penalized by the exponential brevity penalty.

Table 4: *Performance on Transcription*

	p_1	p_2	p_3	BLEU-like
Human	0.91	0.67	0.47	0.61
Prosodic only	0.36	0.13	0.05	0.13
LM only	0.87	0.57	0.33	0.50
Prosodic*0.6 + LM	0.88	0.59	0.35	0.53
Prosodic*2.5 + LM + sent-len	0.84	0.58	0.37	0.56

4.3. Performance on ASR outputs

The automatic speech recognizer that was available to process the spoken reviews obtained a 22.0% WER. We considered the one-best ASR outputs only, and the reference annotations were generated by aligning them with a reference transcription in time, and then inserting breaks at appropriate locations. The results of evaluating sentence boundary detection on ASR outputs is shown in Table 5. The performance of the prosodic model is similar to that for the reference transcriptions, since the timing information from the recognizer is similar to the forced alignment results. However, unlike the case for reference transcripts, a word-level trigram LM works the best, since a four-gram model considers the relationship between words in a larger range, in which there are more errors. Nevertheless, the global sentence length model still improves performance by 3.8% because it can filter out short SUs.

Table 5: *Performance on ASR outputs*

	p_1	p_2	p_3	BLEU-like
Human	0.91	0.68	0.48	0.62
Prosodic only	0.32	0.10	0.04	0.11
LM only	0.66	0.31	0.14	0.30
Prosodic*1 + LM	0.64	0.33	0.17	0.33
Prosodic*1 + LM + sent-len	0.65	0.35	0.18	0.34

5. Summary and Future Work

In this paper, we first described our efforts to collect multiple punctuation annotations via Amazon Mechanical Turk, and suggested the use of a modified BLEU-metric for assessing sentence boundary detection. An FST-based implementation of a sentence detection system that incorporates prosodic and language information, as well as a global sentence length model to find a globally optimal solution was also introduced. We believe one advantage of adopting an FST framework is its flexibility and compatibility to other systems, such as a word graph output from a recognizer.

In future work, we would like to explore to what extent our model can help downstream NLP applications. We would also like to measure the correlation between the modified BLEU-like score and human judgements, and also consider different weighting schemes that were taken from the original BLEU metric.

6. References

[1] Polifroni, J. et. al, "Good grief, i can speak it! preliminary experiments in audio restaurant reviews", Proc. SLT, 2010, pp. 177-186

[2] Matusov, E., Mauser, A. and Ney, H., "Automatic sentence segmentation and punctuation prediction for spoken language translation", Proc. IWSLT, 2006, pp. 158-165

[3] Papineni, K. et. al, "Bleu: a method for automatic evaluation of machine translation", Proc. ACL, 2002, pp. 311-318

[4] Huang, J. and Zweig, G., "Maximum entropy model for punctuation annotation from speech", Proc. ICSLP, 2002, pp. 917-920

[5] Shriberg, E. et. al, "Prosody-based automatic segmentation of speech into sentences and topics", Speech Communication, 2000, vol. 32, pp. 127-154

[6] Liu, Y. et. al, "Using conditional random fields for sentence boundary detection in speech", Proc. ACL, 2005, pp. 451-458

[7] Kim, J.-H. and Woodland, P. C., "A combined punctuation generation and speech recognition system and its performance enhancement using prosody", Speech Communication, 2003, vol. 41, pp. 563-577