Detecting missing sentence boundaries in learner English

Ryo Nagata [1], Junichi Kakegawa [2], Takafumi Kutsuwa [1]

[1] Konan University – 8-9-1 Okamoto, Higashinada, Kobe – Japan

[2] Tokyo University of Science, Yamaguchi – 1-1-1 Daigakudori, Sanyo-Onoda – Japan

Abstract

Learners of English mistakenly omit sentence-final punctuation marks (called missing sentence boundaries in this paper). Detection of missing sentence boundaries is useful in automated essay scoring and learner corpus creation. This paper proposes a novel method for detecting sentences that contain missing sentence boundaries. It first automatically generates training data. Then, from the training data, it creates a model for detecting sentences that contain missing sentence boundaries. It further uses some options such as sampling of training data to reduce bias and noise that inevitably exist in the training data. Experiments show that it achieves an F-measure of 0.716 and outperforms the method used for comparison. In addition, it has following two advantages: (i) it does not require manually annotated training data and (ii) the performance is adjustable.

Keywords: sentence boundary, missing sentence boundary, punctuation, learners of English, English writing

1. Introduction

Surprisingly, learners of English mistakenly omit sentence-final punctuation marks (., *?* and *!*), which will be referred to as missing sentence boundaries in this paper. According to our investigation, about 1.9% of punctuation marks are missing in essays written by Japanese junior high school students. An example is:

I'm praktis foot boll. I like foot ball very much. I went to kobe to eat sweet samar vacation homewook is very hard. Yesterday I study math

where a missing sentence boundary exists between the word *sweet* and the word *samar* which is correctly *Summer*. Note that learner English contains a wide variety of errors in capitalization, spelling, and grammar as well as missing sentence boundaries, which indicates that learner English is noisy.

Detection of missing sentence boundaries is useful in at least two natural language processing applications. One is automated essay scoring such as e-rater (Burstein et al., 1998). In automated essay scoring, it is necessary to detect missing sentence boundaries in order to evaluate learners' essays in terms of mechanics that involves the use of punctuation. In other words, automated essay scoring systems have to evaluate whether or not punctuation is correctly used in learner English. The other is learner corpus creation. Normally, sentence boundaries are annotated in corpus data such as in the British National Corpus (Burnard, 1995). This means that corpus annotators have to identify sentences with or without sentence-final punctuation marks in learner corpus creation. It takes time to identify missing sentence boundaries as we can see in the example above. Besides, the fact that learner English is noisy makes the task harder. For example, a sentence may begin with a lowercase letter or a middle of a sentence may begin with an upper case letter,

which hinders the reader from identifying sentences. Automated missing sentence boundary detection should greatly reduce the effort taken to annotate sentence boundaries.

At first sight, it might seem that the problem is trivial since there has been a great deal of work on sentence boundary identification techniques (Mikheev, 1998; Reynar and Ratnaprkhi, 1997) for written text. However, they are not capable of detecting missing sentence boundaries because they identify sentence boundaries by disambiguating punctuation marks. Namely, they require sentence boundaries to have punctuation marks.

More related to missing sentence boundary detection is sentence boundary identification in speech (Liu et al., 2004). Unlike written text, there is no explicit information available on punctuation and capitalization in speech. This makes the task similar to missing sentence boundary detection. The differences between the two are: (i) acoustic information is available in speech and (ii) learner English is noisy. The first difference implies that conventional sentence boundary identification methods for speech data are not directly applicable to our task. For example, the method (Liu et al., 2004) identifies sentence boundaries based on acoustic information including duration, pitch and energy patterns as well as textual information. The second difference means that learner English contains a wide variety of errors that are unlikely to appear in speech data.

In view of this background, this paper proposes a method for detecting missing sentence boundaries in learner English. The proposed method solely relies on textual information unlike sentence boundary identification in speech. It first automatically generates training data. In other words, it does not require manually annotating missing sentence boundaries whereas the sentence boundary identification methods for speech data. It then creates a classifier from the training data that detects sentences that contain missing sentence boundaries. Although it does not specify exactly where missing sentence boundaries exist, it should be still useful for automated essay scoring and learner corpus creation. In automated essay scoring, counting missing sentence boundaries is enough to evaluate how accurately punctuation is used. In corpus creation, it is not so hard for a corpus annotator to identify missing sentence boundaries, given sentences that contain them.

2. Basic idea

The proposed method assumes that the target texts consist of a number of sentences. This is easily satisfied in both corpus creation and automated essay scoring. A corpus normally consists of a number of sentences. In automated essay scoring, each essay may contain few sentences. However, considering that automated essay scoring is used in a situation where there are a number of users (often examinees), the whole essays are likely to consist of a number of sentences. Conversely, if the target sentences are few, there is no need for automated essay scoring in the first place.

The proposed method uses the target texts themselves as a source of training data. An existing sentence splitter such as OAK system is likely to identify most of the sentence boundaries in the target texts considering the reported performances of the conventional methods. For example, an existing sentence splitter would divide the example essay in Section 1 into the sentences:

I'm praktis foot boll
I like foot ball very much
I went to kobe to eat sweet samar vacation homewook is very hard
Yesterday I study math

These split sentences are negative instances for the training data. Here, *negative instance* refers to sentences that do not contain missing sentence boundaries. To be precise, they sometimes contain missing sentence boundaries as in the third sentence above. However, the split sentences are treated as negative instances in the proposed method. Note that our task is to find sentences containing missing sentence boundaries from the negative instances. Positive instances can easily be generated from the negative instances as follows (positive instance refers to sentences that contain missing sentence boundaries). First, punctuation marks at the end of each divided sentence are stripped off. Then, adjacent sentences are put together into one. These are positive instances. For example, the negative instances above would give the positive instances:

I'm praktis foot boll I like foot ball very much
I like foot ball very much I went to kobe to eat sweet samar vacation homewook is very hard
I went to kobe to eat sweet samar vacation homewook is very hard Yesterday I study math

With the training data, missing sentence boundary detection can be solved as a classification problem using a machine learning algorithm.

However, this way of obtaining training data has a drawback. The training data are noisy and biased. They are noisy in that negative instances erroneously include missing sentence boundaries that are originally in the target texts as exemplified in the third negative instance above. Because of the noise, classifiers trained on the data may not perform well. This is especially true when the classifier has a low generalization ability. For instance, 1-nearest neighbor classifier based on the bag of words model always fails to detect missing sentence boundaries. All sentences containing missing sentence boundaries are erroneously included in the negative instances in the training data. Because of the noise, the 1-nearest neighbor classifier always classifies them as NOT-missing sentence boundary. This implies that generalization ability is a key to success. The training data are biased because every positive instance contains the exact same words in the same order as the corresponding negative instances. We can easily observe it in the example of positive instances above. The bias may affect the performance of the classifier.

3. Proposed method

It is crucial to carefully select features used for the classifier since the training data are noisy and biased as discussed in Section 2. We have already seen that the bag of words model is not suitable for missing sentence detection because of the noise.

One of the promising features is sentence length. Sentences that contain missing sentence boundaries tend to be longer as exemplified in Fig. 1; the third sentence that contains a missing sentence boundary is much longer than the others. To be precise, their lengths are expected to be 2μ where μ denotes mean length of sentences written by learners of English.

To use sentence length as a feature, it is assumed that sentence length has a Gaussian distribution. Under this assumption, the probability of a sentence whose length is equal to or shorter than l is given by

$$p(l) = \int_{-\infty}^{l} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}dx \qquad (1)$$

where μ and σ^2 are the mean and variance of sentence length in learner English. Fortunately, both μ and σ^2 can be estimated from the negative instances in the training data. The probability is one of the features used in the proposed method. Note that the longer the sentence in question is, the higher the probability is. Thus, the proposed method considers a sentence with a high

probability of Equation (1) to likely have a missing sentence boundary. Also it is expected that the noise in the training data scarcely affect the probability estimate (i.e., estimate of μ and σ^2) since the ratio of sentences that contain missing sentence boundaries to the others is very low as experiments will show.

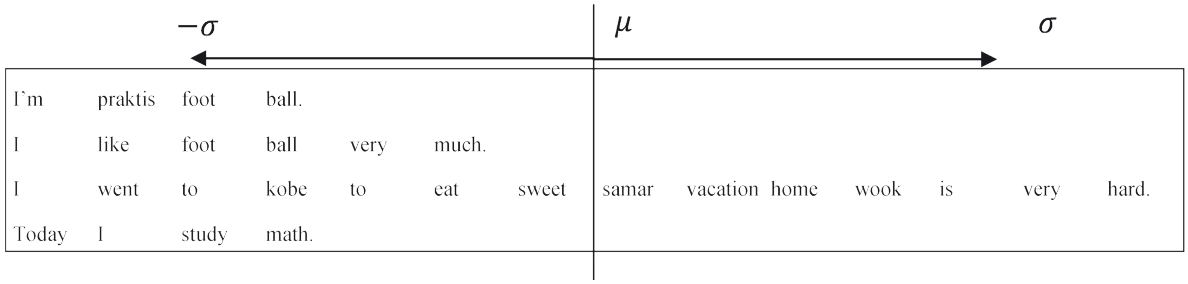

Figure 1: Length of sentences containing a missing sentence boundary

The number of words beginning with a capital letter is another source of evidence. Normally, a sentence begins with a capital letter. Therefore, sentences that contain several words beginning with a capital letter might have missing sentence boundaries. Exceptions are proper nouns and the word *I*. Considering these, the number of words beginning with a capital letter in the sentence in question is used as a feature, excluding proper nouns and the word *I*. A Part-of-Speech (POS) tagger is used to determine whether a given word is a proper noun or not (it is also used to obtain other POSs in feature extraction described below). It should be noted that in learner English a sentence may begin with a lowercase letter or a middle of a sentence may begin with a capital letter since learners make errors in capitalization. This means that the information on capitalization is not sufficient by itself.

Verbs are also informative for missing sentence boundaries. A sentence is assumed to consist of a single verb unless it contains conjunctions and/or clauses. In other words, a sentence containing a missing sentence boundary tends to have more than one verb. This is especially true in the writing of beginning learners such as junior high school students who have not acquired complicated sentence structures and who are most likely to make missing sentence boundary errors. Considering this, the total number of verbs is used as a feature; here, present and past participles are excluded from verbs. To handle conjunctions and clauses, their frequencies are included in the features. Frequencies of each conjunction, such as *and* and *or*, are included in the features as well as the total number of conjunctions. Likewise, frequencies of wh-pronouns (*e.g.*, *which*) and wh-adverbs (*e.g.*, *when*) that indicate a clause are included in the features as well as the total number. Also, frequencies of certain verbs [1] that can take a clause are included in the features.

Other function words are also used as features. Like the verb feature, a sentence is assumed to consist of a subjective personal pronoun, if any, unless it contains conjunctions and/or clauses. In other words, a sentence that has several subjective personal pronouns may contain a missing boundary. So, frequencies of personal pronouns are used as features as well as the total number.

[1] The following verbs that seem to be used frequently by learners of English according to grammar books (Hirota, 1992; Huddleston and Pullum 2006; Iizuka and Hagino, 1997) are selected in this paper: *admit*, *ask*, *believe*, *claim*, *expect*, *hear*, *hope*, *know*, *notice*, *say*, *suppose*, *tell*, *think*.

Prepositions can make a sentence longer without a clause or a conjunction implying that a sentence with prepositions might not contain a missing sentence boundary even if it is a long one. Like other features, frequencies of prepositions are used as features as well as the total number. For the same reason, frequencies of the punctuation marks *:*, *;* and *,* are used as features as well as the total number.

With these features, the detection of missing sentence boundaries is done as follows. First, training data are automatically generated from the obtained sentences as described in Section 2. Then, a classifier is trained on the training data. Support Vector Machines (SVMs) are used as a classifier which have been shown to be effective in a wide variety of natural language processing applications [2]. SVMs are suitable for our task because it is a binary classification problem and the features are real-valued. Finally, each negative instance in the training data is examined whether or not it contains a missing sentence boundary. If it does, it is outputted.

There are two options in the proposed method to reduce the noise and the bias. One is post-processing based on a rule that prevents false positives that are due to errors made by the sentence splitter or the writer (learners). The rule blocks the detection if the target sentence in question contains a period mark (excluding the one at the end of the sentence), a comma, or the word *Because* in the middle of the target sentence. A pre-experiment with development data revealed that one of the major causes of false positives was period marks that were erroneously identified as a not-sentence boundary by the sentence splitter. The proposed method tended to recognize such sentences to contain a missing sentence boundary since the information on the sentence boundary was not available to the proposed method because of the error. Similarly, commas and the word *Because* were often erroneously used and were misleading. The above rule is used to prevent these false positives.

The other option is sampling of positive instances. As already mentioned, positive instances are biased in that every positive instance contains the exact same words in the same order as the corresponding negative instances. Also, the ratio of positive instances to negative ones in the training data is much higher than the true ratio. Because of the bias, the proposed method excessively detects missing sentence boundaries, which results in undesirable false positives. To reduce the bias, positive instances are randomly chosen and included in the training data. Doing so, some positive instances are included in the training data and others are discarded, which is expected to reduce the bias. Here, the problem is how many positive instances should be included in the training data. If some manually-annotated training data are available, they can be used as development data to determine the sampling rate. Otherwise, all negative instances are included in the training data.

4. Experiments

4.1. Experimental conditions

For evaluation, 4533 essays written by Japanese second and third year junior high school students were collected and transcribed by professional transcribers. The topics were either *winter holiday* or *school trip*. Missing sentence boundaries were manually identified. Missing sentence boundaries at the end of the essays were excluded from the evaluation because it is trivial to detect them; they can be detected by checking if there is a punctuation mark at the

[2] The second order polynomial kernel was used in the experiments.

end or not. The target essays were divided into development and test sets. The development set was used to determine the features. It was also used to determine the sampling rate of positive instances. The proposed method was tested on the development set with sampling rate ranging 1% to 100% (increasing 1% at a time). The best performing sampling rate (71%) was used in the evaluation. The test set was used to evaluate the performance of the proposed method. Tab. 1 shows statistics on the two sets.

Set	*# of sentences*	*# of words*	*μ(σ)*	*# of MSBs*
Dev.	1595	9303	5.8 (5.3)	26
Test	11321	65757	5.8 (4.8)	214
Total	12916	75060	5.8 (4.9)	240

Table 1: Statistics on target essays

Liu et al. (2004)'s method, which was based on the maximum entropy framework, for detecting sentence boundaries in speech was implemented for comparison. The features concerning the acoustic information were excluded in the method because it was not available in our task where the target data were written essays. Also, information on chunk boundaries were excluded because they report that it is better not to use the information if the chunking performance is not good; the existing chunkers or parsers are not designed for learner English but native-speaker English. The number of iterations in learning was determined by using the development set. The performance was evaluated by recall, precision, and *F*-measure.

4.2. Experimental results

Tab. 2 shows the experimental results. The results reveal that the proposed methods outperform Liu et al. (2004)'s method in terms of *F*-measure. Note that both the proposed method with sampling and Liu et al. (2004)'s method use the development set to optimize their parameters. By contrast, the proposed method without the sampling does not depend on parameter tuning. Nevertheless, it performs as well as Liu et al. (2004)'s method does. Its performance further improves to an *F*-measure of 0.716, when the sampling of positive instances is used. In addition to its performance, the proposed methods has an advantage over Liu et al. (2004)'s method. The original Liu et al. (2004)'s method requires manually annotated-training instances (i.e., supervised) whereas the proposed methods automatically generate training instances (i.e., semi-supervised) [3].

Method	*R*	*P*	*F*
Proposed (Sampling)	0.808	0.643	0.716
Proposed	0.803	0.538	0.644
Liu's	0.868	0.503	0.637

Table 2: Experimental results

[3] Although the original Liu et al. (2004)'s method required manually annotated-training data, it was revised to be semi-supervised in the experiments. In other words, it was trained on the same (automatically generated) training instances that were used in the proposed methods.

5. Discussion

5.1. Effect of sampling

To investigate the effect of the sampling, the proposed method was further tested on the test set with different sampling rate settings (ranging 1% to 100%, increasing 1% at a time). Fig. 2 shows the relation between the sampling rate and the performance.

Fig. 2 clearly shows that the overall performance (*F*-measure) is always better when the sampling is used unless the sampling rate is too low. The estimated optimal sampling rate (71%) achieves an *F*-measure of 0.716 which is near to the best *F*-measure of 0.718. These results imply that the sampling reduces the bias as we expected. Thus, it is better to use sampling especially when a data set is available to estimate the optimal sampling rate. Even if not, it is still better to set the sampling rate to somewhere that is not too small. Fig. 2 also shows that the performance of the proposed method is adjustable. Namely, if the sampling rate is high, the proposed method is recall-oriented. If low, it is precision-oriented. In practice, this feature of the proposed method is useful. In learner corpus creation, it is preferable to detect as many missing sentence boundaries as possible (i.e., recall-oriented) to annotate all sentence boundaries. Then, a human corpus annotator can relatively easily discard false positives. In automated essay scoring, false positives should be avoided (i.e., precision-oriented); rating an essay as poorer than it is actually is worse than the opposite.

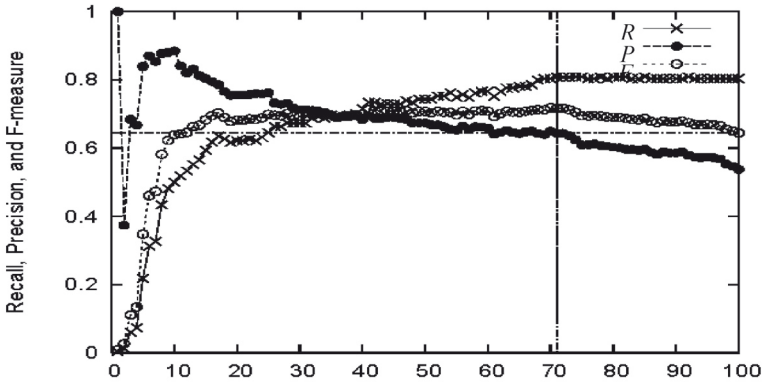

Figure 2: Relation between sampling rate and performance

5.2. Analyzing false positives and false negatives

One of the major causes of false positives is related to conjunctions such as *and* and *or* (36.5% of false positives; *e.g.*, *I like rice* ***and*** *I eat it every day*). Their frequencies were included in the feature vectors to handle the case that a sentence has more than one verb. However, they did not work well. A possible reason is that conjunctions connecting sub-sentences were not distinguished from those connecting phrases (*e.g.*, *bread and butter*) when their frequencies were counted. Another major cause of false positives is spelling and grammatical errors (31.3% of all). If informative words or phrases for not-missing sentence boundaries contain spelling or grammatical errors, they are not effective in prediction. For instance, *The dog is I like the best* (correctly, *The dog is what I like the best*?) was mistakenly detected as missing sentence boundary because it contains two verbs *is* and *like* without the word *what* (or some other words). It seems to be difficult even for human readers to identify whether such cases are missing sentence boundaries or not. Errors in capitalization come next (18.8% of all). A middle of a

sentence sometimes begins with an upper case letter in learner English. The proposed method sometimes mistakenly detected such cases as missing sentence boundaries (*e.g.*, *Thanks to our teachers **We** can practice football*).

False negatives are mostly due to the rule that prevents false positives discussed in Section 3 (63.4% of all false negatives). This kind of false negative is inevitable as long as the proposed method uses the rule. However, the proposed method performs better with the rule; it achieved an *F*-measure of 0.661 without the rule (estimated optimal sampling rate: 71%). Spelling and grammatical errors come next (22.0% of all). As in false negatives, if informative words or phrases for missing sentence boundaries contain spelling or grammatical errors, they are not effective in prediction.

6. Conclusions

This paper proposed a method for detecting sentences containing missing sentence boundaries in learner English. The experiments showed that it achieved an *F*-measure of 0.716 and outperformed the method used for comparison. In addition to its performance, it has following two advantages: (i) it does not require manually annotated training data and (ii) the performance is adjustable.

In future work, we will explore other features to improve the proposed method. We will also investigate how much improvement in sentence boundary identification can be achieved by using the proposed method.

References

Burnard L. (1995). *Users Reference Guide for the British National Corpus. version 1.0*. Oxford: Oxford University Computing Services.

Burstein J., Kukich K., Wolff S., Lu C., Chodorow M., Braden-Harder L. and Harris M.D. (1998). Automated scoring using a hybrid feature identification technique. In *Proceedings of 36th Annual Meeting of ACL*, pp. 206-210.

Hirota S. (1992). *Mastery (in Japanese)*. Kirihara Shoten.

Huddleston R. and Pullum G.K. (2006). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Iizuka S. and Hagino S. (1997). *Prestige (in Japanese)*. Buneido.

Liu Y., Stolcke A., Shriberg E. and Harper M. (2004). Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech. In *Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 64-71.

Mikheev A. (1998). Feature lattices for maximum entropy modeling. In *Proceedings of 17th International Conference on Computational Linguistics*, pp. 848-854.

Reynar J.C. and Ratnaprkhi A. (1997). A maximum entropy approach to identifying sentence boundaries. In *Proceedings of 5th Conference on Applied Natural Language Processing*, pp. 16-19.