Sentence Boundary Detection of Spontaneous Japanese using Statistical Language Model and Support Vector Machines

Yuya Akita[1,2] *Masahiro Saikou*[1] *Hiroaki Nanjo*[3] *Tatsuya Kawahara*[1,2]

[1] School of Informatics, Kyoto University,
[2] Academic Center for Computing and Media Studies, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

[3] Faculty of Science and Technology, Ryukoku University,
Seta, Otsu 520-2194, Japan

10.21437/Interspeech.2006-333

Abstract

This paper presents two different approaches utilizing statistical language model (SLM) and support vector machines (SVM) for sentence boundary detection of spontaneous Japanese. In the SLM-based approach, linguistic likelihoods and occurrence of pause are used to determine sentence boundaries. To suppress false alarms, heuristic patterns of end-of-sentence expressions are also incorporated. On the other hand, SVM is adopted to realize robust classification against a wide variety of expressions and speech recognition errors. Detection is performed by an SVM-based text chunker using lexical and pause information as features. We evaluated these approaches on manual and automatic transcription of spontaneous lectures and speeches, and achieved F-measures of 0.85 and 0.78, respectively.

Index Terms: sentence boundary detection, spontaneous speech, statistical language model, support vector machines.

1. Introduction

Recent advance of automatic speech recognition (ASR) technology, especially for spontaneous speech, enables various applications such as spoken document archiving and retrieval, speech summarization and speech translation. To organize a spoken document in a structured form and to give useful indices, transcriptions should be segmented into appropriate units like sentences. Moreover, these applications are usually built by combining an ASR system with natural language processing (NLP) systems such as a parser and a machine translator, which often assume that input text is a sentence. However, sentences in spontaneous speech are ill-formed, and sentence boundaries are indistinct. Output text by ASR systems is just a sequence of words and has no explicit sentence boundaries, so the further step of segmenting the ASR output is required for these applications.

Automatic boundary detection of spoken sentences has been explored mainly on broadcast news (BN) tasks[1, 2, 3] and conversational telephone speech (CTS) tasks[3, 4, 5] in English. As features for detection, pause, prosodic and linguistic information is often used. Most popular approach is a combination of prosodic and linguistic information[2, 3], which realizes high performance on BN and CTS tasks. Prosody-based approaches[1, 5] have also been investigated. Meanwhile, linguistic information is not used by itself, since most of these works were performed on English data, where cue words or expressions of sentence boundaries are not easily defined.

In Japanese, cue expressions are typically observed at the end of sentences and expected to be useful for boundary detection. However, variety of such expressions is so large in spoken Japanese, that it is hard to collect sufficient amount of data for training statistical models such as a maximum entropy (ME) model. Moreover, many of cue expressions consist of particles, which are apparently difficult to be detected in ASR. Thus, robustness for ASR errors should also be investigated.

In this paper, we address two approaches of sentence boundary detection for spontaneous Japanese. As frameworks of detection, we adopt and compare statistical language model (SLM) and support vector machines (SVM). The proposed approaches are evaluated with real lectures and speeches included in the Corpus of Spontaneous Japanese (CSJ).

2. Definition of sentence in spontaneous Japanese

Spoken Japanese is much different from written Japanese. For example, function words are sometimes inserted or omitted, and inversion of clauses within a sentence is frequently occurred. End-of-sentence expressions in spoken

Table 1: Types of clause boundaries defined in the CSJ

Type	Examples
Absolute	Predicate verbs (basic form) End-of-sentence particles: "*desu*," "*masu*" Particle "*to*"
Strong	Conjunctive particles: "*keredomo*," "*ga*" (but) "*shi*," "*(mashi/deshi) te*" (and, then)
Weak	Case particles: "*kara*," "*node*" (because) "*tara*," "*nara*," "*reba*" (if, when)

Japanese have a considerably wider variety according to speaking-styles and dialects of speakers, while those in written Japanese are typically limited to "*desu*" and "*masu*." With these phenomena, explicit definition of sentence is difficult for spoken Japanese.

In the Corpus of Spontaneous Japanese (CSJ)[6], which is a large collection of spontaneous lectures and speeches, a sentence is defined as a sequence of one or more clauses. Clause is a syntactically and semantically meaningful unit, and defined using lexical and morphological information. Then, sentence boundaries are manually selected among clause boundaries. We adopt this definition for reference in this work.

The clause boundaries can be classified into three levels as listed in Table 1. Absolute boundaries correspond to sentence boundaries in the usual meaning. For example, verbs in basic form are usually related to this boundary. Strong boundaries are the points that can be regarded as major breaks in utterances, thus proper points for segmentation. For example, clauses whose rightmost words are "*ga* (but)" or "*shi* (and)" are often related to this boundary. Weak boundaries are not regarded as proper points for segmentation because they are strongly dependent on other clauses. For example, clauses whose rightmost words are "*node* (because)" or "*tara* (if)" are often related to this boundary.

These three levels differ in their degree of completeness as a syntactic and semantic unit, and the independence from their subsequent clauses. Among these clause boundaries, absolute boundaries and strong boundaries are basically defined as sentence boundaries. However, the rule-based automatic detection system of clause boundaries used in developing of the CSJ did not work well on erroneous ASR results, and performance was significantly degraded.

Therefore, we present two approaches, which take into account the analysis of clause-boundary expressions mentioned above, but mainly rely on machine learning based on lexical information and pause information. Specifically we use statistical language model (SLM) and support vector machines (SVM). The difference of these two approaches is that the former represents linguistic likelihoods of sentence boundaries, while the latter is directly trained to classify boundaries. The former is the most standard approach to assess linguistic appropriateness of the word sequences in language modeling. As for machine learning approach like the latter, maximum entropy (ME) framework has been widely used[3, 4], because of its capability of integrating multiple information sources. In this work, however, we focus on lexical features, which have a wide variety and some of them are less frequent, together with pause duration, which ME is not good at handling. On the contrary, discriminant power of SVM is expected to offer robust classification even with sparse vectors of a large dimension. We therefore adopt SVM as a machine learning framework.

3. Sentence boundary detection based on statistical language model

In SLM-based approach, we regard sentence boundary detection as "translation" from a spoken word sequence X to a segmented word sequence Y, and apply a framework of statistical machine translation[7]. Statistical machine translation generates sentence Y of target language from sentence X of source language, which maximizes posterior probability $P(Y|X)$ based on Bayes' rule.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{1}$$

Here $P(X)$ is not actually used because it does not affect choice of Y for given X. $P(Y)$ is a linguistic likelihood provided by a word N-gram (trigram in this work) language model which contains sentence boundary. $P(X|Y)$ is usually computed by a translation model which defines relationships between X and Y.

In this approach, $P(X|Y)$ determines possible points of sentence boundaries. Candidates of sentence boundary are limited to those having typical clause-boundary expressions as shown in Table 1. Then $P(X|Y)$ is defined as 1 at these points and 0 at the other points. However, same expression is often used in the middle of sentence, for example, "... *to*" (then), "... *nai*" (not) and "*de* ..." (then). Many of them correspond to weak clause boundaries in Table 1. Therefore, we assume that a pause is observed with these expressions if they appear at the end of a sentence, and in only these cases, they can be sentence boundary candidates. On the other hand, predicate verbs in basic form, which are also typical end-of-sentence expressions and suggest absolute or strong clause boundaries in Table 1, become boundary candidates

Table 2: Features for SVM

Type	Description
Linguistic	Word surface, reading, POS tag **(basic)** Conjugation type and form **(optional)** for preceding and following 3 words Estimated clause boundary tag
Pause	Normalized duration, if exists subsequently
Dynamic	Estimated results for preceding words

regardless of pauses.

For every boundary candidate, where $P(X|Y) = 1$, detection is performed using the probability $P(Y)$ by an N-gram language model. Note that fillers are treated as words and linguistically modeled. When computing the likelihood, we incorporate a parameter of insertion penalty as in ASR:

$$\log P(Y) + \beta \times N(Y) \quad (2)$$

N is a number of words in a word sequence and β is an insertion penalty.

4. Sentence boundary detection based on support vector machines

In this approach, sentence boundary detection is regarded as a text chunking problem. We adopt IOE as a labelling scheme, where I, O and E mean inside chunk, outside chunk and end of chunk, respectively. As a text chunker, we use "YamCha"[8] which is based on SVM with polynomial kernel functions. The order of kernel functions is three. Direction of analysis is left to right.

Features given to SVM are linguistic, pause and dynamic features as shown in Table 2. For every input word, preceding and following three words and their reading, part-of-speech (POS) tags are used as linguistic features. Their conjugation types and forms are optionally added. A clause boundary tag which corresponds to the three types in Table 1 is also used as a feature. This tag is preliminarily estimated by another SVM classifier using the linguistic features. Duration of the subsequent pause to the word is an important feature. Duration of the pause is affected by the speaking rate and significantly different between speakers. In this work, therefore, duration of each pause is normalized by an average in a talk. Similar to the SLM-based approach, fillers are regarded as words and included in linguistic features. In addition to these static features, the estimated chunk tags for preceding several words, called dynamic features, are also used.

5. Experimental evaluation

5.1. Experimental setup

We evaluated the proposed SLM-based and SVM-based sentence boundary detection using lectures and speeches in the CSJ. Sentence boundaries are manually annotated for about 200 selected talks called "*core*" data, thus we used them for both training and testing. As a test-set, we used 30 talks, which are determined as a standard test-set for ASR evaluation[9]. The text size of the test-set is 71K words. Automatic transcription was prepared using the baseline speaker-independent ASR system, and average word error rate is 30.2%[9]. The rest of *core* data, namely, 168 talks were used for training of statistical language model and SVM. The text size of the training data is 424K words.

5.2. Evaluation of SLM-based approach

In the proposed SLM-based approach, occurrence of pause is required for boundary candidates for some specific linguistic expressions as described in Section 3. For comparison, we tested three cases: (1) occurrence of pause is assumed for a sentence boundary, (2) occurrence of pause is not required, (3) assumption of pause depends on linguistic expressions, i.e., proposed method.

Table 3 shows recall rates, precision rates and F-measures of respective cases. Recall rate is degraded when the pause occurrence is mandatory, since typical end-of-sentence expressions such as "*desu*" and "*masu*" not followed by pauses were not correctly detected. On the contrary, when pause information is not used, precision rate is significantly degraded. It is because some modifier words and expressions appeared in the middle of sentence, which are similar to end-of-sentence expressions, were erroneously detected. The proposed approach realized much better F-measure than these cases, thus different handling of pauses according to linguistic expressions is effective. For automatic transcription, degradation of precision rate by the proposed approach is larger than manual transcription, since ASR system often provides wrong conjugation forms, which are vital for computation of linguistic likelihood.

Degradation of F-measure for automatic transcription is 13.3% for the proposed method, which is smaller than the word error rate (30.2%). Thus, the proposed method robustly worked on automatic transcription.

5.3. Evaluation of SVM-based approach

As basic features for SVM, we used word, clause and pause information listed in Table 2. As optional features, we prepared conjugation type and form for conjugational words, and tested with and without these features.

Table 4 shows the results of experiments. Recall and

Table 3: Accuracy of SLM-based sentence boundary detection

Transcript	Pause required?	Recall	Precision	F-measure
Manual	Yes	74.9%	87.8%	0.809
	No	79.4%	80.4%	0.799
	Conditional	79.2%	84.6%	**0.818**
Automatic	Yes	66.0%	76.3%	0.708
	No	67.6%	71.9%	0.697
	Conditional	70.2%	71.6%	**0.709**

precision rates in these three cases were almost same for manual transcription, while higher accuracy was obtained by using only basic features for automatic transcription. This is because conjugation forms are often confused in ASR, and such errors cause degradation of detection performance. Degradation of F-measure in automatic transcription is 9.1% for the best case. It is significantly smaller than WER, thus SVM-based approach is also robust for ASR errors.

Comparing the results of SLM-based approach (Table 3) and SVM-based approach (Table 4), the latter realized higher performance throughout the experiment. Differences of F-measures between the two approaches are 4% and 7% for manual and automatic transcription, respectively. The results suggest that the SVM-based approach is more robust against ASR errors than the SLM-based approach. This is because the features of SVM are independent each other ("bag of words") and classification succeeds if only a key feature ("support vector") is correctly detected, while a single ASR error will severely affect linguistic likelihoods of word sequences in the N-gram model.

6. Conclusion

We have presented two approaches of sentence boundary detection for spontaneous Japanese. One uses linguistic likelihoods provided by a statistical language model together with pause information. The other is based on support vector machines (SVM) which count various linguistic and pause information. By experimental evaluation on lectures and speeches in the CSJ, F-measures of 0.82–0.85 and 0.71–0.78 were obtained on manual and automatic transcription, respectively. The robustness of proposed approaches against speech recognition errors is also confirmed. Especially, the SVM-based approach showed more robust and effective performance. These are the best figures reported in the CSJ task so far. In the future, combination of these two approaches will be investigated.

Table 4: Accuracy of SVM-based sentence boundary detection

Transcript	Features	Recall	Precision	F-measure
Manual	Word, pause, clause	83.0%	87.9%	**0.854**
	+Conjugation form	83.0%	87.9%	0.854
	+Conjugation type, form	82.9%	87.8%	0.853
Automatic	Word, pause, clause	73.9%	81.7%	**0.776**
	+Conjugation form	72.5%	81.9%	0.766
	+Conjugation type, form	72.5%	79.9%	0.760

7. References

[1] D. Wang, L. Lu, and H.-J. Zhang, "Speech Segmentation without Speech Recognition," in *Proc. ICASSP*, 2003.

[2] A. Srivastava and F. Kubala, "Sentence Boundary Detection in Arabic Speech," in *Proc. Eurospeech*, 2003.

[3] 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*, 2005.

[4] J. Huang and G. Zweig, "Maximum Entropy Model for Punctuation Annotation from Speech," in *Proc. ICSLP*, 2002.

[5] D. Wang and S. S. Narayanan, "A Multi-pass Linear Fold Algorithm for Sentence Boundary Detection using Prosodic Cues," in *Proc. ICASSP*, 2004.

[6] 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.

[7] P. Brown, S. Pietra, V. Pietra, and R. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[8] T. Kudo and Y. Matsumoto, "Chunking with Support Vector Machines," in *Proc. NAACL*, 2001.

[9] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark Test for Speech Recognition using the Corpus of Spontaneous Japanese," in *Proc. SSPR*, 2003.