INTERSPEECH 2009 BRIGHTON

Effective use of pause information in language modelling for speech recognition

Kengo Ohta[1], *Masatoshi Tsuchiya*[2], *Seiichi Nakagawa*[1]

[1]Department of Information and Computer Sciences / [2]Information and Media Center,
Toyohashi University of Technology, Japan

kohta@slp.ics.tut.ac.jp, tsuchiya@imc.tut.ac.jp, nakagawa@slp.ics.tut.ac.jp

10.21437/Interspeech.2009-126

Abstract

This paper addresses mismatch between speech processing units used by a speech recognizer and sentences of corpora. A standard speech recognizer divides an input speech into speech processing units based on its power information. On the other hand, training corpora of language models are divided into sentences based on punctuations. There is inevitable mismatch between speech processing units and sentences, and both of them are not optimal for a spontaneous speech recognition task. This paper presents two sub issues to address this problem. At first, the words of the preceding units are utilized to predict the words of the succeeding units, in order to address the mismatch between speech processing units and optimal units. Secondly, we propose a method to build a language model including short pause from a corpus with no short pause to address the mismatch between speech processing units and sentences. Their combination achieved a 4.5% relative improvement over the conventional method in the meeting speech recognition task.
Index Terms: speech recognition, language model, pause information, processing unit for speech recognition

1. Introduction

Optimal processing units of speech recognition systems must meet two conditions: the first is the number of errors in the decoding stage, and the second is its processing time. For example, a sentence is a possible optimal processing unit for the speech recognition task against read speech, because there are strong lexical and semantic cohesion between words in a sentence. In other words, it is unsuitable to use sentence fragments as the processing unit, since the language model would be unable to make full use of lexical and semantic cohesion. Likewise, it is unsuitable to use a longer processing unit consisting of several sentences because the necessary search space would become unwieldy. On the other hand, the utterance unit intended by the speaker may be the optimal processing unit for spontaneous speech recognition.

A standard speech recognition system does not use the described units, but uses a power-based unit which are extracted from the input speech based on the power information, because it is very difficult to extract the described units from the input speech [1][2][3]. However, there is dependency between power-based units. Nanjo et al.[4] investigated the relation between speech recognition accuracy and the threshold of the pause length of power-based unit boundaries. In their experiment, three cases were considered: (1) all silences were treated as a short pause, (2) silences shorter than 500msec were treated as a short pause (i.e., a silence greater than 500msec was regarded as an utterance boundary), and (3) silences shorter than 1000msec were treated as a short pause. The best result was obtained from case (3).

On the other hand, Chung et al. [5] proposed a method focusing on the *bunsetsu* unit (corresponding to "phrase" in English). In this method, *bunsetsu* boundaries are detected from word information and word N-grams are calculated separately for the two cases, namely crossing and not crossing *bunsetsu* boundaries. Hirose et al. [6] proposed a similar method focusing on the prosodic unit. In this method, prosodic boundaries are detected based on mora-F0 transition modelling.

This paper addresses mismatch between optimal processing units of spoken speech and observable units. We presents two sub issues to address this problem. As the first sub issue, we show that a decoding method which utilizes the end words of the preceding unit to predict the beginning words of the succeeding unit is effective to address the mismatch between power-based units and optimal units. As the second sub issue, this paper proposes a novel method to build a language model including short pause from a training corpus including no short pause, in order to address the mismatch between power-based units and sentence units.

2. Treatment of sentence units and utterance units

2.1. Analysis using the short pause

In standard speech recognition systems, speech processing units are automatically acquired based on the power information or zero crossing rates of the input speech. However, such power-based units often do not correspond to either sentence units or utterance units. The distributions of length of the Inter-Pausal Units in the CSJ, and the sentence units in the NDR[1] and Mainichi newspaper, are shown in Fig. 1. The Inter-Pausal Units in the CSJ are separated by silences greater than 200msec, and these almost correspond with the power-based units in standard speech recognizers. The sentence units in the NDR are separated by punctuation marks annotated according to the shorthand writers' judgement, thus ignoring the speakers intention. The sentence units in the Mainichi newspaper are separated by punctuation marks annotated by the writers. As can be seen in Fig. 1, the Inter-Pausal Units in the CSJ tend to be shorter than the sentence units in both the NDR and Mainichi newspaper. This means that power-based units are not suitable because the language model would be unable to make full use of the word cohesion. However, units based on power information can be acquired robustly, while sentence units are more difficult to detect. On the basis of this, we consider silences shorter than a certain threshold as a short pause ($< sp >$), as shown in Fig. 2. It should be noted that the word history for the language model continues across the $< sp >$. Here, $< s >$ and $< /s >$ are the symbols indicating the head and tail of the unit, respectively.

After treating all or some of the silences as short pauses in this way, we constructed and evaluated the language models.

2.2. Treatment of short pause

We conducted experiments using the CSJ, the NDR, and the Mainichi newspaper. The CSJ is separated into units based on silences, whereas the NDR and Mainichi newspaper are separated into sentences in which the words are lexically and semantically connected, and do not include any pause information.

In the experiment using the CSJ, we defined $< sp >$ based

[1]http://kokkai.ndl.go.jp/

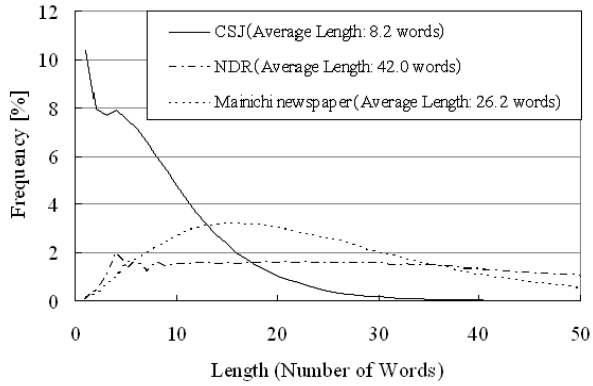

Figure 1: Comparison between Inter-Pausal Units and sentence units.

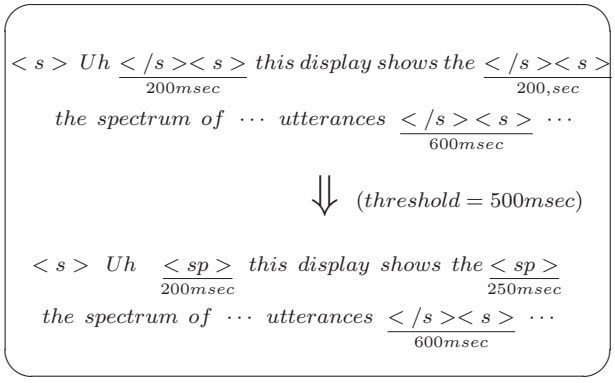

Figure 2: Example of inserting $< sp >$ into an Inter-Pausal Unit boundary (threshold: 500msec)

on the length of the silence as described above. In the experiments using the NDR and Mainichi newspaper, we treated randomly selected sentence boundaries as $< sp >$ since these corpora lack pause information.

The details of each corpus are given in Table 1. We constructed word 3-gram and 4-gram language models from each training corpus with Witten-Bell discounting[7] and a 0-0 cutoff for the CSJ, 1-1 for the NDR, and 3-3 for the Mainichi newspaper. The vocabulary size for all these language models is 20K.

We used the Short-Unit Word (SUW)[2] as the word unit in the experiments using the CSJ. On the other hand, in the experiments using the NDR and Mainichi newspaper we segmented documents into words using the morphological analyzer MeCab ver 0.96[3] (with UniDic ver1.3.8[4]) .

Three cases were compared with respect to test-set perplexity in each experiment: (1) always applying the 3-gram language model, (2) applying the 4-gram language model instead of the 3-gram when the word history includes $< sp >$, and (3) always applying the 4-gram language model.

Next we consider the calculation of perplexity. In our definition of $< sp >$, the language model probabilities of $< sp >$ become higher as the threshold of $< sp >$ increases. Because of this, the language models based on different threshold values of $< sp >$ cannot be compared fairly using the standard perplexity metric. Taking this into account, we treat $< sp >$, $< s >$, and $< /s >$ as context cues only, the occurrence probabilities of which are excluded from the calculation of the test-set perplexity. This corresponds to calculating the perplexity by regarding the test corpus as a long word sequence instead of regarding the test corpus as an information source that generates sentences. In addition, we use $P'(w|h)$ defined below as the language model probability of word w occurring in context h.

$$P'(w|h) = \begin{cases} 0 & \text{if } w \in ccs \\ \alpha(h) \cdot P(w|h) & \text{otherwise} \end{cases}, \quad (1)$$

$$\alpha(h) = \frac{1}{1 - \sum_{w \in ccs} P(w|h)}, \quad (2)$$

$$ccs = \{< s >, < /s >, < sp >\}, \quad (3)$$

where $P(w|h)$ is the conventional language model probability of word w occurring in the context h and ccs is the set of context cues.

[2]In the CSJ, the word unit is represented as a Short-Unit Word (SUW) which approximates the dictionary item from an ordinary Japanese dictionary.

[3]http://mecab.sourceforge.net/

[4]http://www.tokuteicorpus.jp/dist/

Table 1: Statistics of Experimental Data.

	CSJ		NDR		Mainichi newspaper	
	Training	Test	Training	Test	Training	Test
# of words	7M	24K	39M	196K	205M	2M
Vocab. size	66K	2K	58K	7K	298K	54K
Style	spontaneous, exact transcription		spontaneous, post-editing		written text	

2.3. Experimental results

The results of the experiments using the CSJ, the NDR, and the Mainichi newspaper are shown in Fig. 3, Fig. 4, and Fig. 5, respectively.

As shown in Fig. 3, perplexity improves as the threshold of $< sp >$ increases. This result shows a similar tendency to that of Nanjo[4]. In the best case (in which all silences are treated as $< sp >$), a relative improvement of 3.9% was achieved over the baseline model that does not contain $< sp >$. As the Inter-Pausal Unit in the CSJ does not correspond to a lexical or semantic unit, it is effective to continue the word history across the unit boundaries [5] . Examples of the contexts in which the word prediction was improved are given in Table 2. When the 4-gram model was applied, test-set perplexity deteriorated. This is probably because there is insufficient training data in the CSJ to train a 4-gram model.

At the same time, as shown in Fig. 4 and Fig. 5, the perplexity of the head words in every unit improved as the number of $< sp >$ increased in the experiments using the NDR and Mainichi newspaper, respectively. In the best case (in which all sentence boundaries are treated as $< sp >$), relative improvements of 10.0% and 11.8% were achieved over the baseline model in the NDR and Mainichi newspaper experiments, respectively. However, it should be noted that the rate of improvement over all words was much lower than in the CSJ experiment because the occurrence of sentence boundaries is very small (once every 20 to 40 words).

Table 2: Examples of contexts in which word prediction was improved (in Japanese).

History	Succeeding Word	Example
aux.v. , $< sp >$	conj.	$\cdots$ *masu* $< sp >$ *shikashi* $\cdots$
aux.v. , $< sp >$	pron.	$\cdots$ *desu* $< sp >$ *soko* $\cdots$
noun, $< sp >$	particle	$\cdots$ *kekka* $< sp >$ *wo* $\cdots$
int. , $< sp >$	noun	$\cdots$ *eh* $< sp >$ *konkai* $\cdots$
particle, $< sp >$	verb	$\cdots$ *to* $< sp >$ *iu* $\cdots$

[5]The method in which all $< sp >$ were excluded from the language model history achieved worse results than our proposed method.

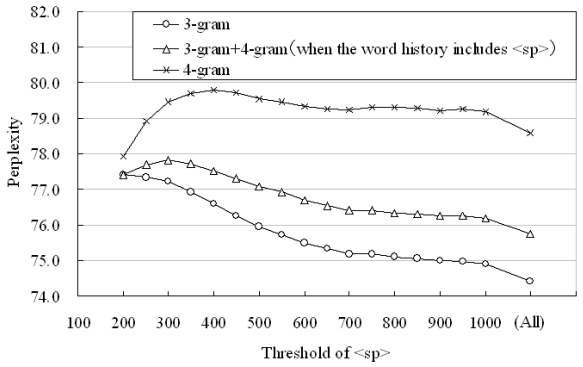

Figure 3: Relation between the length of a short pause and perplexity over all words (CSJ, $< sp >$ insertion into an Inter-Pausal Unit boundary).

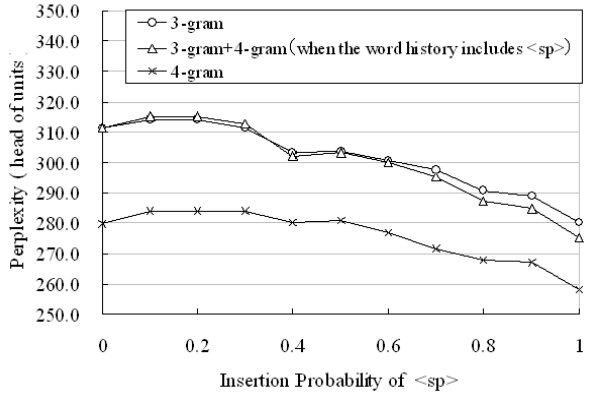

Figure 4: Relation between the insertion probability of a short pause and perplexity on the head word of a sentence (NDR, $< sp >$ insertion into a sentence boundary).

3. Construction of a language model considering short pauses

3.1. Insertion of a short pause

In general, corpora used in training a language model are separated into sentence units that correspond to lexical or semantic units, and lack information of pauses that occur in actual utterances. In a read speech recognition task, it is considered appropriate to treat pause punctuation (a comma) as a short pause. However, in a spontaneous speech recognition task, a comma in the training corpus does not necessarily correspond to a short pause in the actual utterance.

Considering this, we propose a method to insert pauses into a corpus in the same manner as our previously proposed filler prediction model [8].

3.2. Procedure to insert a short pause

Given a certain word sequence, the short pause insertion model predicts the places where short pauses would normally be inserted. We formalized this model as a sequence labelling problem as shown in Fig. 6, where BOS denotes the beginning of the sentence. The label P means that a short pause should be inserted immediately after the labelled word, whereas the label O denotes the contrary.

We use a conditional random field (CRF)[9] model for this labelling problem. A CRF is a discriminative probabilistic model that offers several advantages over hidden Markov models and is used in several statistical natural language processing

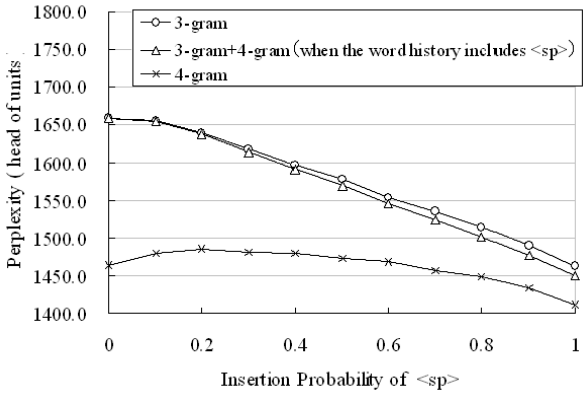

Figure 5: Relation between the insertion probability of a short pause and perplexity on the head word of a sentence (Mainichi newspaper, $< sp >$ insertion into sentence boundary).

Word sequence	Uh	this	display	shows	···
	(BOS)	int.	pron	noun	verb
Label sequence	O	P	O	O	O	···

Figure 6: Example of short pause insertion labelling.

tasks, including language modelling[10].

Given a certain word sequence X, the conditional probability of a label sequence Y is defined as follows:

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{i}^{n}\sum_{a}\lambda_a f_a(X_i, Y_i)\right), \quad (4)$$

where n is the length of X, f_a is a feature function, λ_a is the weight of the feature function, and $Z(X)$ is the normalization factor.

In this experiment, the training set of the CSJ described in Table 1 is employed as the training corpus for the short pause insertion model, and short pauses are inserted into the NDR (using the training set of the NDR described in Table 1). Here, silences shorter than 1000msec are treated as short pauses. Fillers are also inserted into the NDR in advance, using our previously proposed filler prediction model.

We also evaluated two baseline methods, where (1) all or randomly selected commas are treated as $< sp >$, and (2) $< sp >$ are randomly inserted into the sentence. In addition, the effect of corpus combination (CSJ) and treating periods as $< sp >$ were investigated. Each model was evaluated with respect to test-set perplexity (PP), word correct (Cor.), and word accuracy (Acc.).

20 minutes of test data were extracted from the test corpus of the NDR in Table 1. All fillers and other disfluencies were recovered manually by referring to the NDR video archives for creating accurate transcriptions. In all the experiments, we used the SPOJUS decoder [11] with a context-dependent syllable-based acoustic model trained from the CSJ.

3.3. Experimental results

Comparative results of the pause insertion methods are shown in Fig. 7. As can be seen in Fig. 7, the proposed method based on the short pause insertion model achieved better performance than the baseline methods.

The results of the speech recognition experiments are given in Table 3. These results show that in a spontaneous speech task, it is not effective to treat commas as short pauses, as commas in the corpus often do not correspond to actual short pauses. On the other hand, the proposed method that models a short pause in actual utterances achieves better performance than the

Table 3: Evaluation results of ASR (NDR, $< sp >$ insertion within a sentence).

Method	PP	$< sp >$ ratio (%)	Cor.(%)	Acc.(%)
$< sp >$ is not considered	60.9	0	67.4	56.6
All commas are treated as $< sp >$ (baseline)	59.6	6.5	66.7	57.9
$< sp >$ are inserted using CRF (proposed)	55.3	8.9	67.0	59.1
All commas are treated as $< sp >$ + CSJ combined	50.7	7.0	66.4	57.9
$< sp >$ are inserted using CRF + CSJ combined	49.7	9.0	66.4	58.3
All commas are treated as $< sp >$ + all periods are treated as $< sp >$	58.5	7.6	67.7	57.8
$< sp >$ are inserted using CRF + all periods are treated as $< sp >$	54.7	9.9	69.2	60.5
All commas are treated as $< sp >$ + all periods are treated as $< sp >$ + CSJ combined	50.5	8.1	68.2	58.6
$< sp >$ are inserted using CRF + all periods are treated as $< sp >$ + CSJ combined	49.4	10.1	68.2	59.3

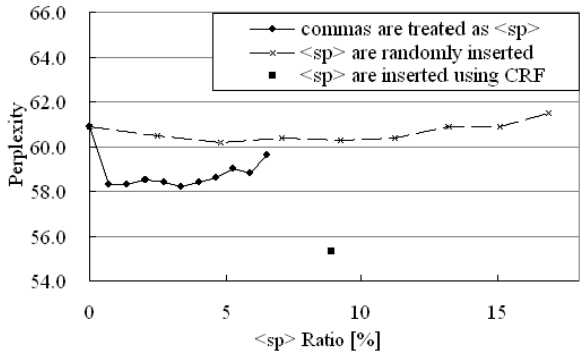

Figure 7: Comparison of insertion methods (NDR, $< sp >$ insertion within a sentence).

baseline methods. Moreover, further improvement is achieved by including periods as $< sp >$. The effect of combining the training corpus with the CSJ was also investigated. The results show improvements of 49.4 from 59.6 for the test-set perplexity, 59.3% from 57.9% for accuracy, and 68.2% from 66.7% for correct.

Based on these results we conclude that pause information which is easier to detect than punctuation marks is more effective in word prediction.

4. Conclusion

This paper addressed mismatch between optimal processing units of spoken speech and observable units.

At first, we showed that a decoding method which utilizes the end words of the preceding unit to predict the beginning words of the succeeding unit is effective to address the mismatch between power-based units and optimal units. Based on the results of experiments using the CSJ, we obtained a 3.9% relative improvement in test-set perplexity. Moreover, according to the results of experiments using the NDR and Mainichi newspaper, relative improvements in test-set perplexity (only on the head word in a unit) of 10.0% and 11.8%, respectively, were achieved.

In addition, we proposed a method to create a spoken language model that includes pauses, from a corpus without pause information, in the same manner as our previously proposed filler prediction model. According to the results of the speech recognition experiment using the NDR, combination of our proposed methods achieved a 4.5% relative improvement in word accuracy over the baseline method.

5. References

[1] Y.Liu and E.Shriberg et.al. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Trans. Audio, Speech and Language Process*, Vol. 14, No. 5, pp. 1526–1539, 2006.

[2] Y.Liu and N.V.Chawla et al. A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech and Language*, Vol. 20, pp. 468–494, 2006.

[3] Yuya Akita, Masahiro Saikou, Hiroaki Nanjo, and Tatsuya Kawahara. Sentence boundary detection of spontaneous Japanese using statistical language model and support vector machines. In *Proc. of ICSLP*, pp. 1033–1036, 2006.

[4] Hiroaki Nanjo, Kazuomi Kato, Akinobu Lee, and Tatsuya Kawahara. Lecture speech recognition using large corpus of spontaneous Japanese. *IEICE transactions on information and systems*, Vol. 86, No. 4, pp. 450–459, 2003, (in japanese).

[5] S. Chung, K. Hirose, and N. Minematsu. N-gram language modeling of Japanese using bunsetsu boundaries. In *Proc. of ICSLP*, pp. 993–996, 2004.

[6] K. Hirose, N. Minematsu, and M. Terao. Statistical language modeling with prosodic boundaries and its use for continuous speech recognition. In *Proc. of ICSLP*, pp. 937–940, 2002.

[7] I. H. Witten and T. C. Bell. The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. In *IEEE Transactions on Information Theory*, Vol. 37, pp. 1085–1094, Jul 1991.

[8] Kengo Ohta, Masatoshi Tsuchiya, and Seiichi Nakagawa. Evaluating spoken language model based on filler prediction model in speech recognition. In *Proceedings of INTERSPEECH*, pp. 1558–1561, 2008.

[9] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of ICML*, pp. 282–289, 2001.

[10] Brian Roark, Murat Saraclar, Michael Collins, and Mark Johnson. Discriminative language modeling with conditional random fields and the perceptron algorithm. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pp. 47–54, Barcelona, Spain, July 2004.

[11] J. Zhang, L. Wang, and S. Nakagawa. Lvcsr based on context dependent syllable acoustic models. In *Proc. of Asian Workshop on Speech Science and Technology*, SP2007-200, pp. 81–86, 2008.