TIME DEPENDENT LANGUAGE MODEL FOR BROADCAST NEWS TRANSCRIPTION AND ITS POST-CORRECTION

Akio Kobayashi, Kazuo Onoe, Toru Imai, and Akio Ando

NHK (Japan Broadcasting Corp.) Sci. & Tech. Res. Labs.
1-10-11 Kinuta Setagaya, Tokyo 157-8510, Japan

{akio, onoe, imai, ando}@strl.nhk.or.jp

ABSTRACT

This paper presents two linguistic techniques to improve broadcast news transcription. The first one is an adaptation of a language model which reflects current news content. It is based on a weighted mixture of long-term news scripts and latest scripts as training data. The mixture weights are given by the EM algorithm for linear interpolation and then normalized by their text sizes. Not only n-grams but also the vocabulary are updated by the latest news. We call it the Time Dependent Language Model (TDLM). It achieved a 4.4% reduction in perplexity and 0.7% improvement in word accuracy over the baseline language model. The second technique is correction of the decoded transcriptions by their corresponding electronic draft scripts. The corresponding drafts are found by using a sentence similarity measure between them. Parts to be considered as recognition errors are replaced with the original drafts. This post-correction led to a 6.7% improvement in word accuracy.

1. INTRODUCTION

In news shows, coverage of the same topics extends over several days. It can be considered that new words and word-pairs in the latest news are likely to be used frequently in upcoming news also. On the other hand, words and word-pairs which appeared frequently in old news may not appear in the latest news any more. There are high correlations between the latest news and upcoming news to be decoded. In this paper, we show how latest news is relevant to upcoming news and discuss the adaptation of a language model (LM) for broadcast news transcription.

When an LM from large texts of long-term news is adapted by small texts of short-term latest news, MAP estimation[1] can be applied. The small texts are added to the large texts with an appropriate weight to construct an adapted LM. If the adapted LM is expressed as a mixture of two LMs made from each data, the mixture weights can be determined by the EM algorithm[1]. The vocabulary is determined by the two different texts in advance without regard to the mixture weights. Our proposed method basically follows this mixture model, but its vocabulary is set by the weighted word frequencies. The adapted model constructed by this method is called a time dependent language model (TDLM) since it reflects current news. We examine the efficiency of the TDLM with various lengths of a training period on the latest news and show experimental results on perplexity and word accuracy.

For broadcast news transcription, original electronic scripts written by reporters could be used to correct the recognition results, though they are not exactly same as those eventually read by announcers. We propose a method to find the closest original sentence to a decoded sentence if it exists, and correct it automatically by replacing words with original ones.

2. BROADCAST NEWS SCRIPTS

For transcribing broadcast news, the decoder should take advantage of the characteristics of news scripts. In this section, we show how 'long-term' news scripts and 'short-term' (latest) scripts are important to train the LM.

2.1. Long-Term News

A statistical language model needs a large training data for accurate estimation. Figure 1 shows perplexity results for some LMs of different lengths of training term. They represent periods of one year through five years (Apr. 1, 1991 through Jun. 3, 1996, 780k sentences or 38M words) of news scripts backwards from the evaluation date (Jun. 4, 1996) of 83 news sentences (2,974 words). The LMs are bigrams with a vocabulary size of 20k and are estimated by the CMU-Cambridge SLM Toolkit[2] with Good-Turing back-off smoothing. As expected, the longest five-year-period LM gave the smallest perplexity of all LMs. In the following experiments, we call it the baseline language model.

2.2. Short-Term Latest News

It is one of the characteristics of news that many new words appear in the latest news. In other words, the number of different words also increases when the training period becomes long as shown in Figure 1. There is an approximately linear correlation between them. The results show occurrences of new words day by day.

Then we examined how new words and word-pairs affect upcoming news. Perplexity tests were performed using two different LMs. One was constructed from the latest one-year-period (June, 1995 through May, 1996) training data including relatively recent words. The other constructed from the oldest one-year-period (June, 1991 through May, 1992) training data. As shown in Table 1, the latest-one-year LM gave less perplexity than the oldest-one-year LM and it was closer to the value of the baseline model. It is considered that up-to-date words and word-pairs led to a reduction of perplexity. The out

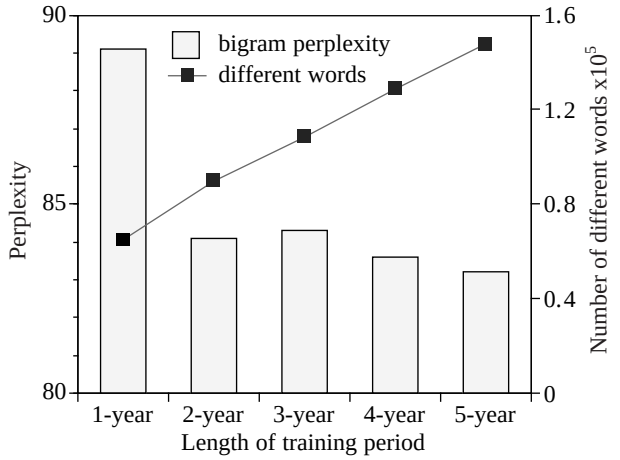

Figure 1: Perplexity results and the number of different words with various lengths of training period

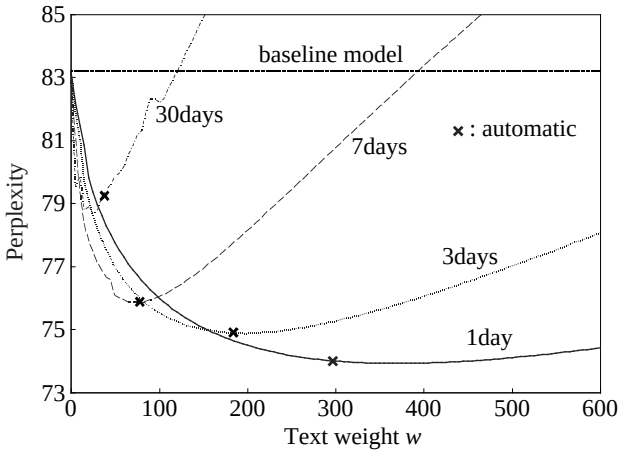

Figure 2: Perplexity results for TDLMs with various adaptation data and text weight *w*

	perplexity	OOV rate (%)
baseline (5 years)	83.2	3.4
LM (latest 1 year)	89.1	3.2
LM (oldest 1 year)	121.4	4.0

Table 1: Perplexity results and out of vocabulary (OOV) rates for LMs.

of vocabulary (OOV) rate for the latest-one-year LM was 0.8% lower than for the oldest-one-year LM. It demonstrated that some words in training data were frequently used at a certain period but not used often at other periods. These results make clear that we should use the latest training data rather than old ones.

3. TIME DEPENDENT LANGUAGE MODEL (TDLM)

In the previous section, we showed both long-term news scripts and latest short-term news scripts are important for LM training. We propose an LM adaptation method where long-term news scripts and latest short-term scripts are mixed with appropriate weights. A heuristic method and and an automatic method are proposed in this section.

3.1. Heuristically Adapted LM

We mixed large (five years) news texts and latest small (some days) news texts with a heuristically found text weight, *w*. Various lengths of the small texts were selected: 1 day, 3 days, 7 days and 30 days. Perplexity for those adapted LMs was examined as shown in Figure 2. When 1-day-long training data were added about 350 times to the large data, the adapted LM gave an 11% perplexity reduction over the baseline model. The perplexity exceeded that of the baseline model as the text weight *w* became larger.

3.2. Automatically Adapted LM

The text weight *w* should be determined automatically for practical applications. We propose an automatic method using linear interpolation with the EM algorithm and normalization by text sizes to get the text weight *w*. Unlike a previous adaptation method[1,3,4], the proposed method makes it possible to update not only n-grams but also its vocabulary by using latest news. The procedure is as follows.

Let P_0 denote an LM trained from long-term large texts and P_1 denote an LM trained from the latest small texts. We define the adapted LM as a linear interpolation of the two LMs:

$$P(y|z) = \lambda P_0(y|z) + (1 - \lambda)P_1(y|z)$$

where *y*, *z* are elements (words) in the union of each vocabulary for P_0 and P_1. The model mixture weight λ is estimated by the EM algorithm.
This procedure accepts the pre-defined vocabulary without regard to the model mixture weights. Our propose is to reflect weighted word frequencies in the vocabulary. Then we convert the model mixture weight λ into the text weight *w* in proportion to their text sizes:

$$w = \frac{(1 - \lambda)m_0}{\lambda m_1}$$

where m_0 and m_1 are sizes of the training data. New vocabulary is defined by the mixed texts according to orders of word frequencies. The above procedure is repeated until the text weight *w* converges. The final LM from the mixed texts with the updated vocabulary is the TDLM.

A cache-based model[5] is proposed to reflect decoded transcriptions. However, it just enhances decoded words with an assumption that they are correct and appear again. The TDLM is trained with an appropriate weight to correct latest news (not to the doubtful decoded texts), and its vocabulary also reflects up-to-date words.

We examined perplexity with the TDLM and obtained results as

shown in Figure 2. The automatically derived text weight *w* was close to the heuristically derived value. It gave the same perplexity reduction rate of 11% over the baseline model when the TDLM was adapted with 1-day-long latest news texts.

3.3. Pseudo-Weight for Open Test

The TDLM requires three kinds of data for training: the long-term training texts (*base**), the latest adaptation texts (*adpt**) and the evaluation texts (*news**) as illustrated in Figure 3. In practical applications, however, *news** is always unknown and the mixture weights cannot be determined. Instead of evaluating *news**, the texts of the previous day (*news*) is evaluated to get the model weight by previous day's TDLM constructed by *base* and *adpt*. The model weight is converted to the text weight (pseudo-weight) in proportion to data sizes of *base** and *adpt**. Then the final TDLM is constructed from *base** and *adpt** with the pseudo-weight.

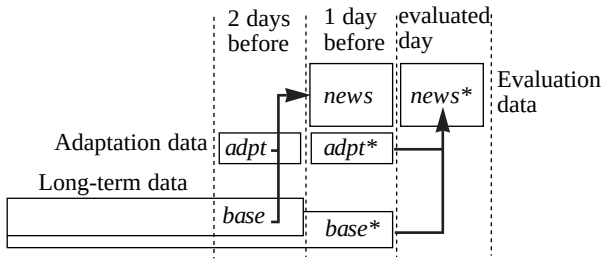

Figure 3: The way of evaluating open data *news** with a pseudo-weight derived from *base* and *adpt*

3.4. Experiment

The automatically adapted TDLM was examined to confirm its effectiveness on perplexity reduction over a series of test sets. The test sets were picked from news scripts by each five days during June, 1996. We selected one-day-long texts as the latest adaptation data. The result is shown in Figure 4. The perplexity value was reduced by the TDLM for every test set. For example, on test set from June 5, the TDLM gave the largest (11.6%) perplexity reduction over the baseline model in the series. In both the baseline model and the TDLM, perplexity values vary with each test set. One reason for this is that some topics in the test sets (e.g. traffic jams due to heavy rain) didn't appear in the training data. It is also considered that unexpected topics (e.g. aircraft accidents) in the test sets obstructed the reduction of perplexity.

The TDLM was applied to broadcast news transcription by a decoder developed at NHK[6]. The test set was NHK's Japanese news broadcast on July 11, 1996. It consists of 100 sentences (50 sentences each for male and female) of 4,231 words that were uttered by three anchors for each gender. The TDLM was constructed from one-day-long adaptation data. It showed a 4.4% perplexity reduction and 80.0% word accuracy compared to 79.3% for the baseline model.

The result was examined from a viewpoint of topic relations. There were 14 sentences in the test set with topics represented in the adaptation data and which gave a reduction in perplexity:

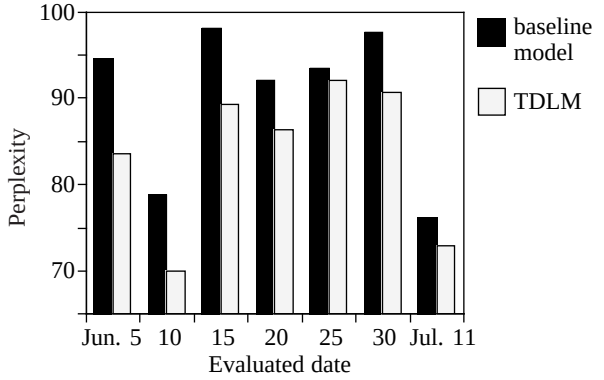

Figure 4: Perplexity results for the baseline model and the TDLM over a series of test sets

they gave a 19.2% perplexity reduction and 3.0% improvement in word accuracy. There were 7 sentences in the test set that were not related to any topics in the adaptation data, but they gave good results (5.7% in perplexity reduction and 3.4% improvement in word accuracy). It is considered that common words and word-pairs in both test set and the adaptation data caused perplexity reduction. Though there were 19 sentences related to the topics in the adaptation data with 9.5% perplexity reduction, the word accuracy got worse by 1.0%. It is considered that the news story represented in the test set evolved even if the topic remained the same, and new words or word-pairs led to worse word accuracy.

4. POST-CORRECTION OF DECODED TRANSCRIPTIONS

4.1. Manuscript in Broadcasting

In Japanese news shows, two types of manuscripts exist. One is an original electronic draft manuscript by news reporters. The manuscript is printed out and modified just before a news show by directors and announcers. The modified manuscript is read by an announcer. Sometimes two or more draft manuscripts are mixed and arranged into one news item. The reason why the draft manuscript cannot be directly used for automatic captioning is that there is no time to correct the draft manuscript electronically with word processors, and it is not possible to handle ad-libs that is not included in the draft manuscript. The types of relationship between the manuscripts actually read by announcers and the draft manuscripts written by reporters are classified as follows.

1. One sentence of a draft manuscript corresponds to one utterance of an announcer.

2. Two or more sentences of a draft manuscript correspond to one utterance of an announcer.

3. There is no draft manuscript which corresponds to an utterance.

The ratios of occurrence of these correspondences are about 7:1:2. This suggests that 80% of decoded sentences have original corresponding manuscripts which can be used for post-

correction.

4.2. Correction Method

We define a sentence similarity between a decoded sentence and an original draft sentence to find the best match. The similarity is calculated by word-based dynamic programming. Unmatched words in the decoded sentence are replaced with the original words if they are regarded as decoding errors.

4.2.1. Detection of an Original Sentence

All the original draft manuscripts from the same day are morphologically analyzed and divided into words. To detect an original sentence corresponding to a decoded sentence, word-based dynamic programming (DP) is performed. Its score is defined as the sentence similarity, and the largest value gives the corresponding sentence. The similarity is the sum of the word similarities and the transition score in the DP. The transition score T from word W_i to W_j in a sentence $W_1W_2W_3...W_n$ is modeled as follows:

$$T(W_i,W_j) = \begin{cases} S - \alpha \bullet (|\, j - i - 1\,|)\,; & \\ \qquad S - \alpha \bullet (|\, j - i - 1\,|) \geq \alpha, & \\ \alpha\,; \quad S - \alpha \bullet (|\, j - i - 1\,|) < \alpha\,. & \end{cases} \quad (i \leq j)$$

Constant S is the maximum value of the transition score and α is a value which decreases with the distance. The word similarity $W(A,B)$ of word A and B is estimated as $W(A,B) = mH$, where m is the number of matched characters between word A and B, and H is a constant value. Whether or not sentences correspond is judged by a threshold. If the number of words of a decoded sentence is N, the threshold θ is modeled by $\theta = NH$.

4.2.2. Correction of Words

The non-matched words between matched words in a decoded sentence are replaced with the corresponding part of the original one. To prevent a wrong correction for unread words, the replacement is performed only if $|\, C_1 - C_2\,| < L$, where C_1 and C_2 are the numbers of words in an original sentence and a decoded sentence, respectively, and constant L is the threshold.

4.3. Experiment

The test data were 100 sentences (3,987 of total words) from NHK news broadcasts between June 11 and June 14, 1996. The decoded sentences showed 83.5% word accuracy before the post-correction. We set the matching score H to equal to the transition score S, and the constant L to 5.

Original sentences corresponding to the decoded sentences were correctly detected from drafts manuscripts with 93.0% accuracy including the case of no corresponding original sentence. It was not possible to find the corresponding original sentence for decoded sentences with low word accuracy. It was shown that the shorter the decoded sentence, the poorer was the detection rate of corresponding sentence. In the case that two or more original sentences were combined into a read sentence, no correspondent sentence was correctly detected.

Decoded sentences were improved from 83.5% to 90.2% (+6.7%) word accuracy by the proposed technique as shown in Table 2. Unknown words for the decoder were correctly added by the post-correction. Many wrong words in the decoded sentences could be replaced with the correct words in the automatically detected original sentences. However, there are some problems in the post-correction. Words with spelling errors and some errors in the original electronic manuscripts are inserted in the final sentence. It may be pointed out that our correction method does not use an acoustic score. The acoustic score does not help the word correction quality because the decoder has already used the acoustic score and finally produced the sentence. In addition to the correction of decoded words, the proposed sentence similarity can also be used as a measure of the confidence of the decoded words.

	word accuracy (%)
Before the correction	83.5
After the correction	90.2

Table 2 : Results of post-correction

5. CONCLUSION

An adaptation method was proposed to reflect the content of the most recent news broadcast. The time dependent language model (TDLM) showed perplexity reduction and word accuracy improvement. Continuation of topics in the LM adaptation is under consideration. We also proposed a correction method for the decoded sentence by using the original draft sentence, and showed the improvement in the word accuracy. The case where plural sentences are combined into one read sentence will be reported in the future.

REFERENCES

1. M. Federico, "Baysian Estimation Methods for N-gram Language Model Adaptation," Proc. ICSLP-96, pp.240-243, 1996

2. P. Clarkson, R. Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit," Proc. EUROSPEECH-97, 1997

3. H. Masataki, Y. Sagisaka, K. Hisaki, T. Kawahara,"Task Adaptation Using MAP Estimation in N-gram Language Modeling," Proc. ICASSP-97, pp.783-786, 1997

4. R. Iyer, M. Ostendorf, J. Rohlick, "Language Modeling with Sentence-Level Mixtures," In Proceedings of the ARPA Workshop on Human Language Technology, pp.82-87, 1994

5. R. Kuhn, R. de Mori, "A Cache-Based Language Model for Speech Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.12, No.6, pp.570-583, 1990

6. T. Imai, K. Onoe, A. Kobayashi, A. Ando, "Decoder for News Transcription, "Autumn Meeting of Acoustic Society of Japan, 1998 (in Japanese)