Improved Chinese Broadcast News Transcription by Language Modeling with Temporally Consistent Training Corpora and Iterative Phrase Extraction

Pi-Chuan Chang, Shuo-Peng Liao, Lin-Shan Lee

Dept. of Computer Science and Information Engineering, National Taiwan University
Taipei, Taiwan, Republic of China
{pcchang,deanl}@speech.ee.ntu.edu.tw
lslee@gate.sinica.edu.tw

Abstract

In this paper an iterative Chinese new phrase extraction method based on the intra-phrase association and context variation statistics is proposed. A Chinese language model enhancement framework including lexicon expansion is then developed. Extensive experiments for Chinese broadcast news transcription were then performed to explore the achievable improvements with respect to the degree of temporal consistency for the adaptation corpora. Very encouraging results were obtained and detailed analysis discussed.

1. Introduction

In broadcast news transcription, a well estimated language model (LM) is a key element to obtain satisfactory recognition results. Although the ultimate goal of LM estimation may be to obtain a general LM that can model the behavior of the language, in spite of the subject domains of the contents, such a goal is actually impractical for broadcast news transcription, because the news always includes new subjects with new words very often serving as the keywords of the contents. Therefore the linguistic behavior and statistical characteristics for news articles could be very diverse and difficult to predict in advance.

When dealing with Chinese language, the problem becomes more serious. Different from many western languages, Chinese is not alphabetic. Each of the thousands of different Chinese characters is a morpheme with its own meaning, and therefore plays very independent linguistic roles in sentences. A Chinese word is composed of one to several characters, but with very flexible wording structure because each component character has its own meaning. Furthermore , there are no "blanks" in written Chinese sentences between two words serving as word boundaries as in western languages. As a result, the "word" in Chinese is very often not well defined, there doesn't exist a commonly accepted lexicon, and the segmentation of a sentence into words is usually not unique. In particular, a proper noun (personal name, organization name, event name, etc.) can be easily constructed by combining several arbitrary characters. Such new words are usually the keywords in new articles, specially new stories. All these make the out-of-vocabulary (OOV) problem very serious for Chinese broadcast news transcription, which in turn, together with many other issues, make the accurate estimation of a LM for broadcast news transcription very difficult.

In this paper we propose an iterative approach to extract Chinese new words or phrases from temporally consistent text corpora, considering both the association between words and the context variation statistics for segments of words, to extract contemporary new words and estimate contemporary LMs. Further analysis for the performance of this approach with respect to the temporal consistency of the training corpora is also presented. The overall framework for this work is illustrated in Fig. 1. The temporally consistent sub-corpus is first selected from a large corpus. The iterative new phrase extraction algorithm is then applied and the contemporary LM estimated based on the temporally consistent sub-corpus. This contemporary (foreground) LM is then used to adapt a background LM. Although the discussions here are concentrated on broadcast news transcription, the concept of temporal consistency may be directly extended to more general area of corpus homogeneity, such as subject domain homogeneity and task scenario homogeneity. For those cases the application of the approaches proposed here are definitely not limited to broadcast news transcription. The rest of the paper is organized as follows. The iterative new phrase extraction algorithm is described in Section 2, and the considerations for selection of homogeneous sub-corpus in Section 3. In Section 4, we present experimental results including analysis with respect to the temporal consistency of the training corpus. The conclusion is finally given in Section 5.
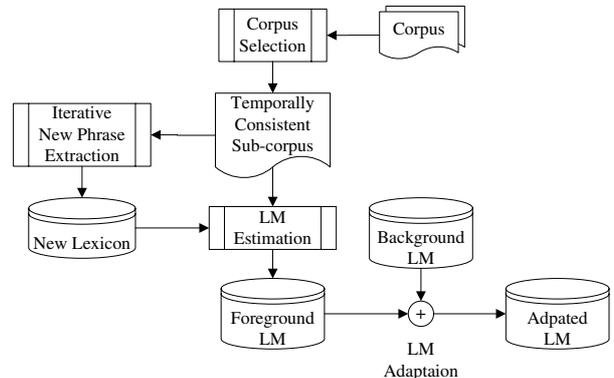

Figure 1: *The overall framework of LM adaptation and iterative new phrase extraction*

2. Iterative Chinese New Phrase Extraction

New word or phrase extraction has been an important problem for Chinese language processing, and several useful approaches have been proposed. These approaches may be classified as character-based (i.e., processed character by character) [1, 2] and word-based (i.e., processed word by word) [3, 4] approaches. The basic idea for word-based approaches is very

similar to the compound-word extraction approaches which were also discussed for western languages [5, 6]. The iterative approach proposed in this paper is primarily word-based, considering both the intra-phrase association as well as the context variation statistics.

The intra-phrase association is a metric that describes the "association" (or the "stickiness") between the component words (or segments of words) within a phrase. The intra-phrase association used in this paper, $A(w_i, w_j)$, for two component words (or segments of words) w_i and w_j appearing adjacent to each other in the corpus is defined as

$$A(w_i, w_j) \equiv \sqrt{P_f(w_j|w_i)P_r(w_i|w_j)}$$
$$= \frac{N(\langle w_i, w_j \rangle)}{\sqrt{N(w_i)N(w_j)}}, \tag{1}$$

where $P_f(w_j|w_i) \equiv N(\langle w_i, w_j \rangle)/N(w_i)$ is the forward bigram and $P_r(w_i|w_j) \equiv N(\langle w_i, w_j \rangle)/N(w_j)$ is the reverse bigram, $N(\cdot)$ represents the frequency counts in the corpus, and $\langle w_i, w_j \rangle$ is the concatenation of the two consecutive words (or segment of words) w_i and w_j.

This intra-phrase association $A(w_i, w_j)$ can in fact be related to the mutual information $MI(w_i, w_j)$ between the two component words w_i and w_j [1, 2, 3, 4, 5]. So the concatenation of two words $\langle w_i, w_j \rangle$ (or segments of words) can be a candidate new phrase if the intra-phrase association $A(w_i, w_j)$ is higher than a threshold, and the frequency counts of $\langle w_i, w_j \rangle$ is high enough.

Context variation statistics, on the other hand, indicates whether the concatenated new phrase has good lexical boundaries. Given a candidate phrase X (X can be the concatenation $\langle w_i, w_j \rangle$ mentioned above), the right context R_X of X is defined as

$$R_X = \{\beta | \langle X, \beta \rangle \in \text{corpus}\}, \tag{2}$$

where β is a word (or segment of words) appearing to the right of X in the corpus. Similarly, the left context L_X of X can be defined as

$$L_X = \{\alpha | \langle \alpha, X \rangle \in \text{corpus}\}. \tag{3}$$

Let $|R_X|$ and $|L_X|$ denote the number of distinct words (or segment of words) in R_X and L_X. The context variation statistics criterion proposed here is then simply considering X as a new phrase if

$$|R_X| > a \text{ or } |R_X| = 0, \tag{4}$$

and

$$|L_X| > b \text{ or } |L_X| = 0, \tag{5}$$

where a, b are two thresholds. The basic idea is that if X is a new phrase, it should behave quite independently and therefore have large enough number of distinct right/left context units. On the other hand, $|R_X| = 0$ or $|L_X| = 0$ occur if X always appears at the sentence end or the sentence beginning. This means the right or left side of X is a trivial phrase boundary. Similar concept of context variation statistics for phrase boundary was previously proposed and found useful in Chinese keyword extraction and information retrieval [1], but some modifications were made here. Preliminary experiments showed that such a modification offered better results.

So the iterative new phrase extraction process is shown in Fig. 2. A corpus $\mathcal{C}$, and a seed lexicon V_0 are given first. In the i-th iteration, the corpus $\mathcal{C}$ is word segmented using a maximum matching segmentation algorithm [7] based on the given lexicon V_i. The intra-phrase association $A(w_i, w_j)$ is then evaluated and the frequency counted for each concatenation $\langle w_i, w_j \rangle$

which appears in the corpus, where w_i, w_j are the words or segments of words obtained in the i-th iteration. Those concatenation $\langle w_i, w_j \rangle$ with intra-phrase association and occurrence frequency exceeding the thresholds are put in a candidate priority queue Q sorted by $A(w_i, w_j)$. The top-n concatenations are then selected from Q based on the context variation statistics for them. Candidates satisfying the conditions are extracted as new phrases and put in the new phrase set T_i. So the new lexicon for the next iteration is $V_{i+1} \leftarrow V_i \cup T_i$. After the terminal condition is met, we have all new phrases extracted and included in the final lexicon.

Although this approach is primarily word-based, all characters can be taken as mono-character words and included in the initial seed lexicon V_0. In this way, a new word (such as a personal name) may first be segmented into a series of isolated mono-character words, then concatenated into the new word and added to the lexicon. This is why the proposed approach is able to extract all new words or phrases from proper nouns, special terminologies for special domains to compound words or phrases.
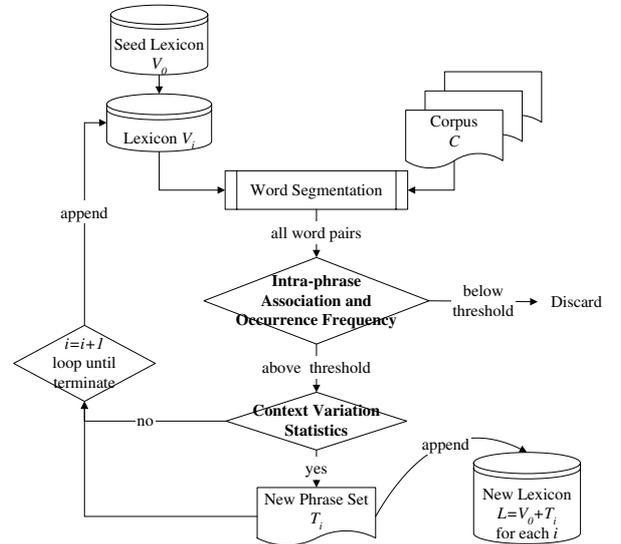

Figure 2: *Iterative Chinese new phrase extraction*

3. Improved Homogeneity between Language Models and Speech Data

In Chinese news the new terms, new words and new phrases are generated every day. With the iterative new phrase extraction procedure presented above, a natural approach is to use temporally consistent corpus, or contemporary text news to update the lexicon and the LM [8]. In other words, in the news broadcast news task, it's very likely that the text news and broadcast news are discussing similar events with similar sentence patterns if they appear near in time. Since we can obtain news texts labeled by dates from the web very easily, the LMs can be adapted to fit the current broadcast news contents better. The achievable improvements with such temporally consistent adaptation data will be analyzed in detail below. In fact, new contents appear on the web everyday, therefore the application of this approach of LM adaptation with temporally consistent corpus may be directly extended to areas other than broadcast news.

Another natural corpus homogeneity comes from the sub-

ject domains. In the news domain, the text news on web are usually categorized into subject domains already. These corpora classified by subject domains can be used as training set for extracting domain-specific key phrases and training domain-specific LMs using exactly the same framework as in Fig. 1. This will be a natural extension of the approach discussed above.

4. Experimental Results

A series of experiments was performed to verify the feasibility of the approaches presented here. Since the goal is to improve the ASR performance, the performance metric used here is the ASR character accuracy.

For all the ASR experiments, the recognizer developed at National Taiwan University [9] was used. The acoustic models consist of 151 initial-final sub-syllabic units. They are 112 Initials, 38 Finals, and a silence. Here Initial is the initial consonant of a syllable, and Final is everything in the syllable following the Initial. Each Initial has three states. Each Final has four states. In each state, 16 Gaussian-mixtures are used. The baseline background 60K-word trigram LM (**BSL**) is estimated on a 40M-character corpus from the Central News Agency (CNA) at Taipei for 1997-1999 selected news, smoothed with Good-Turing discounting by SRI Language Modeling Toolkit (SRILM) [10]. The test set is the Chinese broadcast news (CBN) collected from News98 radio station at Taipei [11] in September, 2002. It includes news of 18 days with total length of 3.7 hours. The adaptation corpus is collected from Yahoo! News [12], which collects major news sources of Taiwan. We use news of one month (from the mid-August 2002 to the mid-September). It is for new phrase extraction and LM enhancement. The time period of adaptation text corpus and testing speech data is shown in Fig. 3. As can be seen in Fig. 3, about half of the testing data are temporally consistent with about half of the adaptation data (labeled as "overlapped"), and the other half is "non-overlapped".

4.1. Preliminary Experiment on New Word Extraction

In the first experiment, the results with the iterative new phrase extraction proposed here are compared with the PATtree method previously proposed [1] which considers the context variation statistics only and is not iterative. The results are listed in Table 1. The first column is for the baseline LM (**BSL**), while the next two columns are for the foreground LM (not yet adapted with the background LM) trained with the adaptation corpus only, one with the lexicon extracted using the previously proposed PATtree method [1] (**PAT**) and the other with the lexicon extracted using the iterative method proposed here (**ITR**). Because the latter is consistently better, the last column is the results for the iterative foreground LM (**ITR**) adapted with the baseline LM (**BSL**). For each case the results for overlapped and non-overlapped test data are listed separately. The results include character accuracy and its improvements with respect to the baseline (**BSL**)

From Table 1, we can see that the baseline LM (**BSL**) gives consistent performance for both overlapped and non-overlapped test data, while the PATtree (**PAT**) and iterative (**ITR**) approaches perform better in the overlapped part, but the performance drop significantly in the non-overlapped part. This is because the LMs for **PAT** and **ITR** approaches are estimated with only one month of news text, thus are overfitted and weak to unseen events. The baseline LM (**BSL**) was estimated with
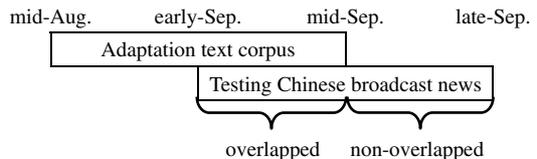

Figure 3: *Period of adaptation text corpus and testing speech data*

character accuracy	**BSL**	**PAT**	**ITR**	**ITR+BSL**
overlapped	76.50	76.92 (0.55)	78.54 (2.67)	82.40 (7.74)
non-overlapped	77.06	72.77 (-5.57)	73.77 (-4.27)	79.78 (3.53)

Table 1: *Character accuracy (percentage improvements) for the baseline LM (**BSL**), PATtree method (**PAT**), iterative approach (**ITR**) and iterative approach adapted with the baseline LM (**ITR+BSL**)*
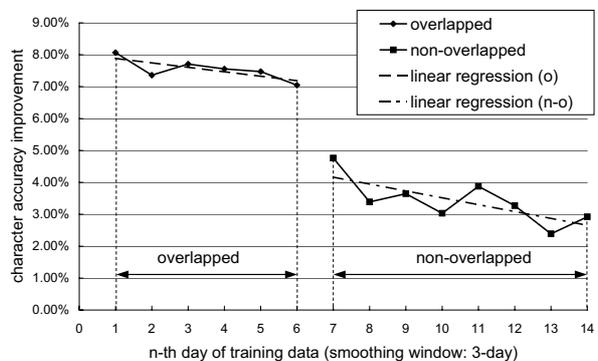

Figure 4: *Character accuracy improvements for 3-day windows with respect to the degree of temporal consistency for adaptation corpus*

much wider range of corpus, so it is much more robust to unseen event. Moreover, **ITR** approach gives apparently better performance than **PAT** approach in both overlapped and non-overlapped cases. This indicates that the new phrase extraction method described in Section 2 is actually useful. In the last column of Table 1, when **ITR** is adapted with **BSL** (**ITR+BSL**), we can see significant improvements achieved with the proposed approach for both overlapped or non-overlapped cases. It should be pointed out that the events, and therefore key phrases and sentence patterns in the news may propagate for some time. Therefore even for the non-overlapped case, significant improvements were still obtained.

4.2. Analysis of LM Enhancement with respect to the Degree of Temporal Consistency for Adaptation Corpus

Two sets of experiments were performed to analyze the LM enhancement with respect to the degree of temporal consistency in the adaptation corpus. The results are presented below. The distribution of the improvements of 7.74% (for overlapped case) and 3.53% (for non-overlapped case) in the last column of Table 1 over time is shown in Fig. 4. In this figure, every point is the average percentage improvement for the test data within a 3-
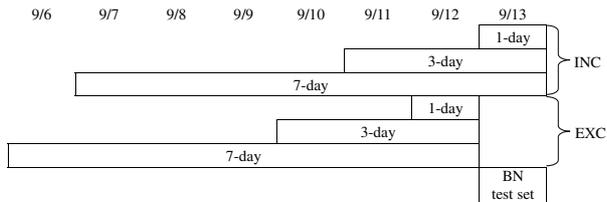
Figure 5: *Time period for the six sets of adaptation corpora (Sep. 13th is the date for test set as an example)*

char. accuracy (improvement)	1-day	3-day	7-day
INC	81.77(6.21)	81.89(6.36)	81.65(6.05)
EXC	78.11(1.45)	78.49(1.95)	78.84(2.41)

Table 2: *Single-day broadcast news transcription accuracy (relative improvements compared to* **BSL** *of 76.99%) for 6 different sets of adaptation corpora, averaged for 6 single-day testing corpora*

day window. The left half of 6 points are those in the overlapped case whose average is 7.74% and the right half of 8 points are those in the non-overlapped case whose average is 3.53%. Considering the time periods for the training and testing corpora as shown in Fig. 3, we can see the clear trend in Fig. 4 that the performance improvements fade out as the temporal distance between the training and testing corpora increases, specially from the regression lines. Also, there is a clear gap or discontinuity between the regression lines for overlapped (left half) and non-overlapped (right half) parts. This time fading phenomenon and discontinuity indicates the role of temporal consistency in LM enhancement, and how the temporal distance between the training and testing corpora should be taken care of for broadcast news transcription.

The next experiment considers the possible approach to achieve high accuracy with LM enhancement by very limited adaptation corpora. Only the broadcast news for a single day was taken as the test set, while 6 different sets of adaptation corpora were considered. 3 sets of the adaptation corpora include the news corpora for exactly the same date as the testing set (**INC**), while the other 3 excludes the news corpora for the same day (**EXC**). In both cases (**INC** and **EXC**) the 3 sets of adaptation corpora used include those for 1-day, 3-day and 7-day, counted backward from the date of the testing set, as shown more clearly in Fig. 5. The experiment was performed on six sets of single-day broadcast news on Sep. 13, 17, 20, 23, 26 and 30 respectively, and the character accuracy (and relative improvements as compared to **BSL** of 76.99%) listed in Table 2 for the 6 sets of adaptation corpora are those averaged over the 6 days of testing single-day broadcast news.

From the first row of Table 2 **INC**, we can see if the adaptation data includes the text news of the same day as the testing set of broadcast news (**INC**), significant improvements can be immediately achieved even using only 1-day news as the adaptation data, and the improvements almost saturated with just 1-day adaptation data for the **INC** case. In other words, the data for the several previous days didn't help for the **INC** case. In the second row of Table 2 **EXC**, on the other hand, we observe that the performance improvements gradually increases as the size of the adaptation corpora increases, but always lower than the first row **INC**. The performance improvements behave quite consistently as the average results for all the six days as listed in Table 2. So the data in Table 2 are in fact representative.

We may conclude here that the adaptation corpus for the same day as the testing data (i.e., **INC** with 1-day) contains highly homogeneous linguistic information (similar words, phrases, sentence patterns, etc.) though they are from different sources. Of course adaptation with such highly homogeneous corpora may lead to overfitted LM, but the overfitted LM is used to recognize the highly homogeneous test data only. This can be very helpful in broadcast news archiving and retrieval which requires more precise transcriptions.

5. Conclusion

In this paper, we propose an iterative Chinese new phrase extraction method, which selects the new phrases based on two criteria: intra-phrase association and context variation statistics. With a LM enhancement framework including the expanded lexicon and LM adaptation, detailed analysis for the achievable transcription accuracy for broadcast news with respect to the degree of temporal consistency from adaptation corpora is also presented.

6. References

[1] L.-F. Chien, "PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval," in *Information Processing and Management*, vol. 35, no. 4, 1999, pp. 501–521.

[2] J. Gao, J. Goodman, M. Li, and K.-F. Lee, "Toward a unified approach to statistical language modeling for Chinese," in *ACM Transactions on Asian Language Information Processing*, vol. 1, no. 1, 2002, pp. 3–33.

[3] J. Zhang, J. Gao, and M. Zhou, "Extraction of Chinese compound words - an experimental study on a very large corpus," in *The Second Chinese Language Processing Workshop attached to ACL2000*, 2000.

[4] C.-J. Wang, "Chinese speech information retrieval – data-driven / predefined indexing features, different retrieval models and improved approaches," Master's thesis, National Taiwan University, 1999.

[5] G. Saon and M. Padmanabhan, "Data-driven approach to designing compound words for continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, May 2001.

[6] C. Beaujard and M. Jardino, "Language modeling based on automatic word concatenations," *Proc. EUROSPEECH*, 1999.

[7] P.-K. Wong and C. Chan, "Chinese word segmentation based on maximum matching and word binding force," in *Proc. of Computational Linguistics*, 1996, pp. 200–203.

[8] M. Federico and N. Bertoldi, "Broadcast news LM adaptation using comtemporary texts," in *Proc. EUROSPEECH*, 2001.

[9] Y.-C. Pan, "One-pass and word-graph-based search algorithms for large vocabulary continuous mandarin speech recognition," Master's thesis, National Taiwan University, 2001.

[10] A. Stolcke, "SRI language modeling toolkit," version 1.3.3, http://www.speech.sri.com/projects/srilm/.

[11] "News 98 FM-98.1," http://www.news98.com.tw/.

[12] "Yahoo! Kimo News portal," http://tw.news.yahoo.com/.