TRIGGER-BASED LANGUAGE MODEL CONSTRUCTION BY COMBINING DIFFERENT CORPORA

Carlos TRONCOSO† Tatsuya KAWAHARA† Hirofumi YAMAMOTO‡ and Genichiro KIKUI‡
†School of Informatics, Kyoto University
Yoshida Nihonmatsu-cho, Sakyo-ku, Kyoto, 606-8501 Japan
‡Spoken Language Translation Research Laboratories, ATR
2-2-2 Hikaridai, Seika-cho, Kyoto, 619-0288 Japan
†{carlos, kawahara}@ar.media.kyoto-u.ac.jp, ‡{hirofumi.yamamoto, genichiro.kikui}@atr.jp

ABSTRACT

In this paper we study the trigger-based language model, which can model dependencies between words longer than those modeled by the n-gram language model. Generally in language modeling, when the training corpus matches the target task, its size is typically small, and therefore insufficient for providing reliable probability estimates. On the other hand, large corpora are often too general to capture task dependency. The proposed approach tries to overcome this generality-sparseness trade-off problem by constructing a trigger-based language model in which task-dependent trigger pairs are first extracted from the corpus that matches the task, and then the occurrence probabilities of the pairs are estimated from both the task corpus and a large text corpus to avoid the data sparseness problem. We report evaluation results in the Corpus of Spontaneous Japanese (CSJ).

1. INTRODUCTION

Statistical language models try to capture regularities of natural language to improve the performance of many different natural language applications such as automatic speech recognition, machine translation, document classification, information retrieval, handwriting recognition, spelling correction, etc. Statistical language models estimate the probability distribution of linguistic units from large amounts of text data.

The most popular and widely used model is the n-gram language model, where *n* typically ranges from 2 (bigram) to 4 (4-gram). The n-gram language model estimates the occurrence probability of *n* consecutive words in the text, and its parameters are usually estimated from a large text corpus. This model is known to be effective, but it is apparently limited in scope, because it is unable to model dependencies longer than *n*.

Some works in the literature, such as the trigger-based language model [1][2] and the cache-based language model [3], tried to broaden the scope of the n-gram model by modeling long-distance dependencies between words. The trigger-based language model uses a set of correlated word pairs, known as trigger pairs, to raise the probability of the words "triggered" by others in the word history. When training this model, however, we usually find a fundamental problem, depending on the nature of the training data. When the trigger pairs are trained from a large corpus, many of the pairs are not task-dependent, because the corpus is usually too general. Therefore, the effectiveness of the trigger-based language model is undermined by the specificity of the target task. On the other hand, when the training data set is from the same domain as the target task, its size is usually insufficient and the probability estimates are unreliable.

To overcome this trade-off between generality and sparseness, we propose an approach that takes advantage of two different corpora to create a trigger-based language model so that the trigger pairs are dependent on the target task and have reliable estimates.

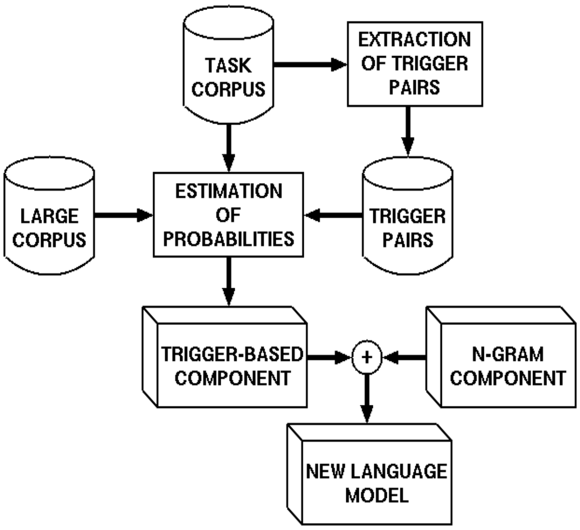

Figure 1: *Outline of the proposed approach*

The rest of this paper is organized as follows. Section 2 introduces the proposed trigger-based language model in detail. Its experimental evaluation with the Corpus of Spontaneous Japanese is reported in section 3. Finally, the conclusions and some discussion are given in section 4.

2. PROPOSED APPROACH

Figure 1 illustrates the outline of the proposed approach. First, the trigger pairs are extracted from a text corpus that matches the target task (task corpus). Then the probabilities of the pairs are estimated, based on their co-occurrence frequency within a text window, from two different corpora: the mentioned task corpus and a large text corpus, providing us with two different sets of trigger pairs with their corresponding probabilities. Finally, the resulting trigger-based component is combined with the n-gram component to produce a new language model.

The proposed model uses a combination of the probabilities from the two trigger pair sets when the trigger pairs can be found in the set trained from the task corpus. Otherwise, the probabilities from the set trained from the large corpus are used.

By extracting the trigger pairs from the target domain, we solve the generality problem, while we avoid the data sparseness problem by using the set of trigger pairs whose probabilities are estimated from the large text corpus.

2.1. Extraction of trigger pairs from task corpus

A trigger pair is a pair of content words that are semantically related to each other. Trigger pairs can be represented as $A \rightarrow B$, which means that the occurrence of A "triggers" the appearance of B, that is, if A appears in a document, it is likely that B will come up afterwards.

The trigger pairs are first extracted from a text corpus that matches the target domain. In this way, we can obtain task-dependent trigger pairs. For the selection of pairs, instead of the average mutual information used in [1], we use the term frequency/inverse document frequency (TF/IDF) measure [4]. We employ this measure because it is document-based, that is, it lets us extract the trigger pairs from a whole document, rather than from a text window of the target corpus. In this way, we can capture global, topic constraints from each document. This measure is also chosen because of its simplicity.

The TF/IDF value of a term t_k in a document D_i is computed as follows:

$$v_{ik} = \frac{tf_{ik} \log(N / df_k)}{\sqrt{\sum_{j=1}^{T} (tf_{ij})^2 [\log(N / df_j)]^2}}, \qquad (1)$$

where tf_{ik} is the frequency of occurrence of t_k in D_i, N is the total number of documents, df_k is the number of documents that contain t_k, and T is the number of terms in D_i.

For each document, we create all possible word pairs, including pairs of the same words (self-triggers), with the base forms and parts of speech (POS) of all the words with a TF/IDF value above a threshold. POS-based filtering is introduced to discard function words, as well as a word stop list to ignore words of very frequent appearance. By using base forms we avoid same-root triggers, and we can apply the trigger pair when a word is presented with any inflection, while by using the POS information

Table 1: Example of trigger pairs extracted from the CSJ

Triggering word	Triggered word
machi (town)	*sumu* (to live)
oya (parent)	*kodomo* (child)
mujintou (desert island)	*shima* (island)
hontouni (really)	*sugoi* (amazing)
haha (mom)	*chichi* (dad)
nihon (Japan)	*amerika* (America)
taberu (to eat)	*oishii* (delicious)
shigoto (job)	*kaisha* (company)
nihonjin (Japanese)	*nihon* (Japan)
ryokou (travel)	*kaigai* (abroad)
sensei (teacher)	*gakkou* (school)
byouin (hospital)	*nyuuin* (hospitalization)
daigaku (university)	*koukou* (high school)

we distinguish between homonyms with different POS when applying the trigger pairs.

Table 1 shows some examples of trigger pairs extracted from the target task.

2.2. Probability estimation from two corpora

The probabilities of the trigger pairs are then estimated from two different corpora by using a text window to calculate the co-occurrence frequency of the pairs inside it. This text window consists of the 20 words previous to the one being processed.

The two distinct corpora used are the text corpus that matches the target task and a large text corpus. The probability estimation stage results in two different sets of trigger pairs: the trigger pairs with the probabilities estimated from the task corpus (hereafter trigger set TC), and the trigger pairs whose probabilities are estimated from the large corpus (hereafter trigger set LC). The trigger set TC provides a probability distribution more faithful to the target domain, whereas the trigger set LC offers a more reliable distribution that can cope with the problem of data sparseness that we discussed above.

The probability of each trigger pair $w_1 \rightarrow w_2$ is computed as follows:

$$P_{TP}(w_2 \mid w_1) = \frac{N(w_1, w_2)}{\sum_j N(w_1, w_j)}, \tag{2}$$

where $N(w_1, w_2)$ denotes the number of times the words w_1 and w_2 co-occur within the text window, and j runs throughout all words triggered by w_1.

2.3. Proposed trigger-based language model

The proposed trigger-based language model is then constructed by linearly interpolating the probabilities of the trigger pairs with those of the baseline trigram (3-gram) model, so that both long and short-distance dependencies can be captured at the same time.

The probability of the new language model for a word w_i given the word history H is computed in the following way:

$$P_{LM}(w_i \mid H) = \frac{1}{|H|} \sum_{h \in H} P_{LM}(w_i \mid h), \tag{3}$$

where $|H|$ means the number of words the history comprises, and $P_{LM}(w_i \mid h)$ is calculated as follows:

$$\begin{cases} P_{NG}(w_i \mid H), \quad \text{if } P_{TP}^{TC}(w_j \mid h) = 0, P_{TP}^{LC}(w_j \mid h) = 0, \forall j \\ \lambda P_{NG}(w_i \mid H) + (1-\lambda) P_{TP}^{LC}(w_i \mid h), \quad \text{if } P_{TP}^{TC}(w_j \mid h) = 0, \forall j \\ \lambda P_{NG}(w_i \mid H) + (1-\lambda)\left(\delta P_{TP}^{LC}(w_i \mid h) + (1-\delta) P_{TP}^{TC}(w_i \mid h)\right), \\ \hfill \text{otherwise} \end{cases} \tag{4}$$

Here P_{NG} is the probability of the n-gram component; P_{TP}^{TC} is the probability of the trigger set TC; P_{TP}^{LC} is the probability of the trigger set LC; λ is the language model interpolation weight; and δ is the trigger set interpolation weight.

When there are no words triggered by h in either of the two sets, the trigram model alone is used. When there are no trigger pairs for h in the trigger set TC, the trigram probabilities and the probabilities from the trigger set LC are linearly interpolated. Otherwise, the probabilities of the trigram are linearly interpolated with a linear interpolation between the probabilities from both trigger sets.

Table 2: Specification of used corpora

Corpus name	Contents	Type of language	Size
Corpus of Spontaneous Japanese	Extemporaneous speeches	Spoken language	3.5M words
Mainichi Shimbun	Newspaper articles	Written language	289M words
Web corpus	Chat logs	Spoken language	270M words

3. EXPERIMENTAL EVALUATION

3.1. Corpora and procedure

The Corpus of Spontaneous Japanese (CSJ) [5] is a conversational corpus consisting of lectures on various academic topics and extemporaneous speeches about different matters such as hobby and travel. We used the extemporaneous speeches, which are divided into 1705 speeches of training data, comprising 3.5 million words, and 10 speeches of evaluation data, containing 18 thousand words.

The trigger pairs were extracted from the CSJ training data. We used the lecture as the document unit. The threshold for the TF/IDF value was initially chosen to be 0.015 based on a subjective judgment of the goodness of the pairs from a sample taken at random, and it was later tuned empirically, producing the value 0.031.

For estimating the probabilities, we used two different corpora: the Mainichi Shimbun newspaper corpus and a conversational text corpus extracted from the World Wide Web (WWW) [6] (hereafter web corpus). We used 11 years (1991-2001) of articles from the Mainichi Shimbun corpus, consisting of 289 million words. The web corpus consists of conversational texts that can be found on the WWW, such as chat logs, and comprises 270 million words. Being conversational, the web corpus is closer in style to the CSJ than the Mainichi Shimbun newspaper corpus, so we expected to get better experimental results with the former. Table 2 summarizes the corpora used in this work.

The language model interpolation weight and the trigger set interpolation weight were empirically tuned to produce the values 0.7 and 0.76, respectively.

The baseline language model was a back-off trigram model trained from the CSJ training set. The size of the vocabulary was 30K words.

3.2. Perplexity evaluation

We evaluated the test-set perplexity by the proposed language model for different values of the coverage of the trigger pairs in the evaluation data, determined by the threshold for the frequency of the words in the stop list. We compared four different models: the model that was constructed by using the CSJ and the web corpus (CSJ + Web), the model constructed with the CSJ and the Mainichi Shimbun corpus (CSJ + Mainichi), a model that used only the CSJ corpus (CSJ), both to extract the trigger pairs and to calculate their probabilities, and a model that used only the Mainichi Shimbun corpus (Mainichi), extracting the trigger pairs from the portion corresponding to year 2001 and estimating their probabilities from the whole corpus. We did not create a model only from the web corpus because it is not divided into documents, so it is not suitable for the TF/IDF computation.

The values of the threshold for the stop list were 500, 1000, 2000, 3000, 5000, and no stop list, for the first three mentioned models, and 100000, 200000, 400000 and no stop list, for the last one.

The number of extracted trigger pairs varied from 11,483,557 to 12,048,275 for the CSJ + Web model, from 11,109,675 to 11,804,186 for the CSJ + Mainichi model, from 3,838,096 to 3,907,486 for the CSJ model, and from 22,774,387 to 23,810,712 for the Mainichi model.

The results are illustrated in Figure 2. The highest perplexity reduction was 12.8%. We can

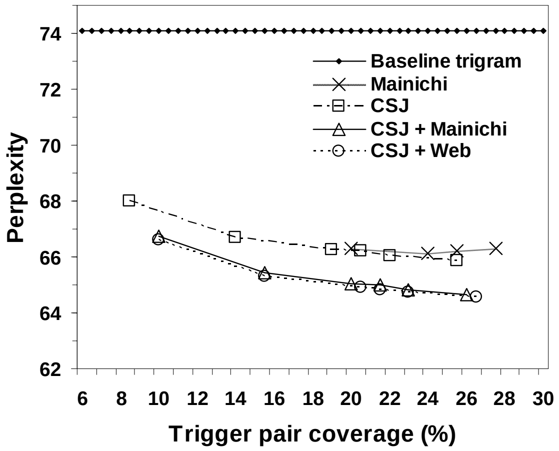

Figure 2: *Perplexity against coverage of trigger pairs for different sets of trigger pairs*

see that the CSJ + Web model and the CSJ + Mainichi model resulted in very similar perplexity results. Furthermore, the perplexity of the models that used two corpora was always lower than that of the models that used only one corpus.

4. CONCLUSION AND DISCUSSION

We presented a novel approach to the trigger-based language model based on two different corpora. We take advantage of the task corpus in order to obtain task-dependent trigger pairs, while we use large corpora to cope with the data sparseness problem. A significant improvement in perplexity was achieved when using the two corpora for constructing the model, as compared with the baseline trigram and the models that use only one corpus.

We found out that, contrary to our expectations, the performance of the web corpus was almost identical to that of the Mainichi Shimbun. The corpus size seems to supersede the differences in style.

The proposed approach is particularly useful in tasks where large amounts of training data are not readily available, since we have observed that, with the proposed method, general corpora such as the Mainichi Shimbun can be used to complement small task corpora. This is specifically true for spoken language tasks, where corpora are usually small.

ACKNOWLEDGEMENTS

We would like to thank Mr. Nobuhiro Kaji and Professor Sadao Kurohashi of University of Tokyo for providing us with the conversational web corpus, and also Dr. Shinsuke Mori of IBM Japan, who gave us very useful comments.

REFERENCES

[1] R. Rosenfeld, “A Maximum Entropy Approach to Adaptive Statistical Language Modeling,” Computer, Speech and Language, vol. 10, pp. 187–228, 1996.

[2] C. Tillmann, H. Ney, “Selection Criteria for Word Trigger Pairs in Language Modeling,” International Colloquium on Grammatical Inference, pp. 95-106, 1996.

[3] R. Khun, R. De Mori, “A Cache-Based Natural Language Model for Speech Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 6, pp. 570–583, 1990.

[4] G. Salton, “Developments in Automatic Text Retrieval,” Science, vol. 253, pp. 974–980, 1991.

[5] K. Maekawa, H. Koiso, S. Furui, H. Isahara, “Spontaneous speech corpus of Japanese,” Proceedings LREC, vol. 2, pp. 947–952, 2000.

[6] N. Kaji, M. Okamoto, S. Kurohashi, “Paraphrasing Predicates from Written Language to Spoken Language Using the Web,” Proceedings HLT-NAACL, pp. 241–248, 2004.