Similarity Based Language Model Construction for Voice Activated Open-Domain Question Answering

István Varga Kiyonori Ohtake Kentaro Torisawa Stijn De Saeger
Teruhisa Misu Shigeki Matsuda Jun'ichi Kazama
Institute of Information and Communications Technology (NICT)
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan
{istvan, kiyonori.ohtake, torisawa, stijn, teruhisa.misu,
shigeki.matsuda, kazama}@nict.go.jp

Abstract

This paper describes a novel method of constructing a language model for speech recognition of inputs with a particular style, using a large-scale Web archive. Our target is an open domain voice-activated QA system and our speech recognition module must recognize relatively short, domain independent questions. The central issue is how to prepare a large scale training corpus with low cost, and we tackled this problem by combining an existing domain adaptation method and distributional word similarity. From 500 seed sentences and 600 million Web pages we constructed a language model covering 413,000 words. We achieved an average improvement of 3.25 points in word error rate over a baseline model constructed from randomly sampled Web sentences.

1 Introduction

This paper presents a novel language model construction method for speech recognition, which is to be utilized by *Ikkyu*, an open-domain voice-activated Japanese QA system. Ikkyu takes relatively short spoken questions concerning a broad range of topics as input through a smartphone and provides the answers retrieved from a large scale Web archive. The challenge we tackle here is to provide a language model for a large-vocabulary speech recognition module to recognize such questions.

The widespread practice in language model construction is to train the model with a corpus that matches the domain and style of the target application. "Domain" usually refers to a set of utterances that are strongly related to a particular topic (e.g., travel). Ikkyu does not have any domain in this sense. The final QA system is expected to answer all the questions concerning every possible topic as far as our QA module can find the answers from the Web.

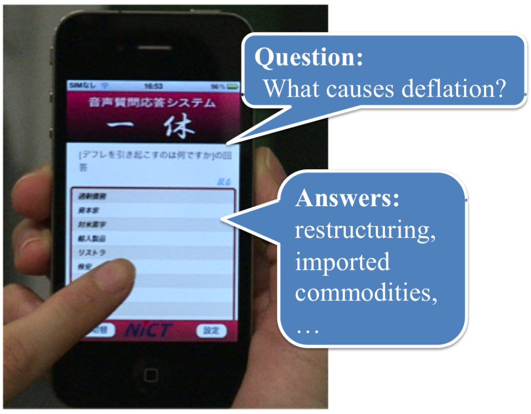

Figure 1: Screenshot of our QA system (*Ikkyu*) with the answers for *"What causes deflation?"*.

Since current state-of-the-art speech recognition systems cannot recognize long sentences through a smartphone with a high sentence accuracy, we decided to focus on relatively short questions, roughly consisting of a noun, an interrogative pronoun, and a predicate. Also, our current QA module can deal with only relatively short questions and this restricts the answerable questions. We call the restrictions on possible questions due to these factors *style* hereafter. (See Section 2 for a detailed description of style.) Our challenge is to prepare a large number of questions over various topics that match this style. Manual construction of such a corpus is impossible considering the necessary vocabulary coverage, thus an automatic method for collecting questions is needed.

Our method starts from an extremely small seed

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 536–544,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

(1)	Hayabusa ha nannen buri ni chikyuu ni kikan shita? *After how many years did Hayabusa (Japanese space probe) return to the Earth?*
(2)	Saikin hatsubai sareta Sony no gakushuu rimokon no kataban ha? *What's the model ID of Sony's recent universal remote control device?*
(3)	Itazuke iseki ha doko ni arimasu ka? *Where are the Itazuke ruins?*
(4)	Minamoto Yoritomo no otouto no namae ha nani desu ka? *What is the name of the brother of Yoritomo Minamoto (Japanese feudal lord)?*
(5)	Tokyo Disneyland no moyori no eki ha doko desu ka? *Which station is closest to Tokyo Disneyland?*
(6)	Gogatsu no tanjouseki wo oshiete kudasai. *Tell me the birthstones of May.*
(7)	Netchuushou no shoki shoujou ha? *(What is) the first symptom of hyperthermia?*
(8)	Kokusei chousa ha nannen oki ni jisshi sareru? *How long is the interval (in year) between each national census?*
(9)	Suteroido no fukusaiyou ni ha donna mono ga arimasu ka? *What are the side effects of steroids?*
(10)	Kaiketsu Zorori no sakusha ha dare? *Who is the author of Kaiketsu Zorori (cartoon)?*
(11)	Wimbledon de yuushou ssita hito ha dare? *Who is the champion at Wimbledon?*
(12)	Rui 14 sei no gyouseki ha nan desu ka? *What are the achievements of Louis XIV?*
(13)	Nihon de iPhone ha dore kurai urete imasu ka? *How many iPhones have been sold in Japan?*
(14)	Posutomodan to ha nani desu ka? *What is postmodern?*
(15)	Java no saishin baajon ha? *What is the latest version of Java?*

Table 1: Correctly recognized question examples from the test data. (May contain questions that can not be answered by our QA module.)

(1)	Defure wo hikiokosu no ha nani desu ka. *What causes deflation?*
(2)	Yanaacheku ga sakkyoku shita no ha nan desu ka? *What [musical pieces] are composed by Janáček?*
(3)	Heisokusei doumyaku koukashou wo fusegu no ha nan desu ka? *With what can peripheral artery disease be prevented?*
(4)	Kawazugawa de nani ga tsuremasu ka? *What [kind of fish] can you catch in the Kawazu river?*

Table 2: Answerable questions of the QA module.

corpus consisting of hundreds of sentences that are manually tailored so that they match the style while also covering a wide range of topics. Next it selects sentences similar to ones in the seed corpus from a large Web archive (Shinzato et al., 2008). The selection is done by a combination of distributional similarity for nouns (Kazama et al., 2010) and an existing automatic *domain* adaptation method, intended to construct a relatively large *in-domain* training corpus (Misu and Kawahara, 2006). We show that this adaptation technique can be useful even in constructing an open-domain but style restricted corpus.

The major problem we tackle is the following: since constructing a large seed corpus is not affordable, we need to deal with the inevitable sparsity of a smaller seed corpus. Since this seed corpus needs to cover as many topics as possible, the number of questions for each topic is too scarce. Although the domain adaptation method is designed to deal with a similar problem, i.e., data sparseness caused by small seed corpora, it is questionable whether it would be equally efficient in domain independent, style restricted settings with a seed corpus of the same size. Our basic idea to overcome this difficulty is to expand the seed corpus by replacing nouns with distributionally similar nouns in the seed corpus, adding the resulting new sentences to the seed corpus. As a result, the domain adaptation method can pick up sentences referring to a wider range of topics from the Web more efficiently.

In the experiments using an existing speech recognition engine, ATRASR (Matsuda et al., 2006), our proposal's best model covering 413,000 words achieved an average word error rate of 15.49% and an average sentence error rate of 54.73%. The obtained improvement by our method over a baseline language model constructed from randomly sampled Web sentences was 3.25 points in word error rate and 4.28 points in sentence error rate.

Table 1 shows some questions correctly recognized by our best model. These suggest that our speech recognition module can actually recognize questions concerning a wide range of topics.

2 Our Open-Domain QA System *(Ikkyu)*

Before presenting the proposed method, we explain about Ikkyu, our open domain QA system, to clarify the motivation behind the task settings and the requirements for our speech recognition module. Figure 1 is a screenshot of Ikkyu, answering the question *"What causes deflation?"*. The answers are extracted from a large Web archive (6.0×10^8 Web pages) (Shinzato et al., 2008) in just a few seconds and each answer is linked to the original Web page from which it is extracted. Users can check more information regarding the answers just by following these links. Table 2 lists some examples of the questions the current prototype can actually answer.

For the example question of Figure 1 (*"What causes deflation?"*), the system unexpectedly presented a major Japanese automobile manufacturer as an answer. The blog entry, from which the system extracted the answer, claimed that the company kept its huge profit (more than tens of billion US dollars) and this shrinked public demand and worsened the deflation. The same story was reported in an authoritative economic magazine after we found this answer.

The ultimate objective in this speech driven QA project is to offer a platform with which users can easily broaden their viewpoint by discovering valuable information including unexpected ones, like the example above, at any time, any place, which may result in more efficient decision making in all circumstances. In achieving this goal, we believe flexible speech recognition can be a great help since it can allow users to ask anytime, anywhere, any questions that come to mind.

The QA module is an extension of a pattern-based relation extraction method (De Saeger et al., 2009). It converts the input question, such as *"What causes deflation?"*, into the lexico-syntactic binary pattern "X causes Y" and automatically computes its paraphrases as well, such as "X triggers Y" and "Y is a cause of X". X and Y are variables, corresponding to the topic and interrogative pronoun of the question. These patterns are then matched against the Web archive after one of the variables is filled with the corresponding noun in the original question (Y = "deflation" in the above example). The nouns matching the unfilled variable (X) are provided as answers.

An important point here is that the form of the answerable questions are restricted by the patterns utilized by the QA module. These were automatically extracted beforehand from frequently observed dependency paths of a Web archive, amounting to 70 million in number, thus covering virtually all topics found in the archive. Because of this pattern extraction scheme, most patterns consist of a predicate, two variables (to be filled with nouns) and some postpositions connecting the predicate and the variables. Due to this tendency in the patterns, most answerable questions consist of a predicate, a noun, an interrogative pronoun and some additional function words. This is the *style* we assumed for our QA system.

Further elaboration of this notion of *style* may lead to the idea that the language model may be derived by converting all possible patterns into questions. We attempted this approach, but found that it was extremely difficult, as will be discussed in Section 5.3.

Instead, we start from a small corpus prepared in a relatively independent way from the architecture of our QA module. A positive side-effect of this approach is that it can be applied to speech recognition for other QA systems as far as the style of the questions match that of our QA module.

3 Background

This section describes the statistical domain adaptation method and the distributional word similarity to be used in the proposed method.

3.1 Statistical adaptation method

We combine our similarity based expansion method with an existing statistical adaptation method proposed by Misu and Kawahara (2006). In their method a seed corpus S is used to generate search queries for retrieving similar, relevant text from the Web. The search queries are generated by automatically extracting keywords with large TF-IDF values. Next, for each sentence from the retrieved text a $score$ is calculated, i.e. the word perplexity relative to the seed language model.

$$score = 2^{-\frac{1}{n}\sum_{i=1}^{n}\log_2 p(w_i|w_{i-1},w_{i-2})}$$

Here w_i represents the i-th word in a sentence with n words. For trigrams that contain unknown words, the minimum trigram probability of the seed language model is assigned as a penalty. The sentences whose $score$ is smaller than a threshold θ are selected to form the training corpus T, together with the seed corpus.

Note that this method selects sentences from Web search results using the keyword queries determined according to TF-IDF. However, our main goal is to efficiently identify questions covering a wide range of topics while matching a certain style, often represented by colloquial textual fragments and therefore consisting of frequent words. We judged that Web search using TF-IDF score is harmful because it is likely to filter out such frequent words, therefore we skipped it. Instead, $score$ was computed for all the sentences of the Web archive and sentences were selected solely based on the $score$ values. We fixed the unknown trigram probability at 10^{-10}. Compound nouns (or noun sequences) are treated as single noun.

We call our implementation of this statistical adaptation method "Misu's method" hereafter.

3.2 Distributional word similarity

Distributional word similarity measures the semantic similarity between words based on the distributional hypothesis (Harris, 1954), which states that words that occur in the same contexts tend to have similar meanings. Based on this hypothesis, many similarity measures have been proposed. We adopt a recent method of Kazama et al. (2010).

Kazama et al. (2010) applied the Bayesian approach for the similarity calculation to alleviate the problem of data sparseness. The method starts from the base similarity measure, the Bhattacharyya coefficient, which is defined over probability distributions p_1 and p_2 as follows:

$$BC(p_1, p_2) = \sum_{k=1}^{K} \sqrt{p_{1k} \times p_{2k}}$$

p_1 and p_2 are the conditional context distributions for two given words, $p(f_k|w_1)$ and $p(f_k|w_2)$. The contexts, f_k, used in (Kazama et al., 2010) are dependency relations such as "subj-of-swim" for the word "tuna". Instead of using $p(f_k|w_1)$ directly, their method estimates the distribution of $p(f_k|w_1)$ itself using the Bayesian method to capture the unreliability of data and calculates the expectation of the above base similarity under those distributions. They showed that this method outperforms many existing similarity measures through the experiments using large-scale Japanese Web data.

4 Proposed Method

The proposed method can be described as follows. The inputs are the seed corpus S and a Web archive W.

Step 1 For each sentence s in S, pick up every noun w that is not in a stop-word list L and replace w in s with the most similar k words according to Kazama's distributional similarity (Kazama et al., 2010). The resulting sentences are added to S.

Step 2 Apply Misu's method (Misu and Kawahara, 2006) to S and W and construct a training corpus T.

Step 3 Construct a language model from T using existing tools for speech recognition.

Assume the question *"What are the symptoms of gout?"* is in S, "gout" and "symptom" are not in L, "osteoporosis" and "cause" are among the k most similar words to "gout" and "symptom" respectively. Then, in Step 1 the new sentences *"What are the symptoms of osteoporosis?"* and *"What are the causes of gout?"* are added to the seed corpus S. In Step 2, we can expect that Misu's method picks up sentences such as *"What are the treatments of osteoporosis?"* in the Web archive and adds it to the training corpus T. This is because the new question shares two trigrams ("What are the" and "of osteoporosis ?") with the newly added seed sentence *"What are the causes of osteoporosis?"*, and is likely to have a relatively low (thus better) $score$ value. Note that the question is less likely to be added to the training corpus if the noun replacement of "gout" with "osteoporosis" was not conducted since it has less common trigrams with the original seed question. This is a benefit obtained by our method.

In our experiments, we used approximately 500 questions as seed corpus S. These were manually crafted according to the instructions described in Section 5.1.2. The Web archive W consists of 6.0×10^8 Japanese Web pages (Shinzato et al., 2008). The distributional similarity between nouns was computed from another Web archive consisting of 1.0×10^8 Japanese Web pages. The stop-word list L contains about 2,000 nouns whose frequency exceed 10^7 in our Web archive. This list was devised because highly frequent nouns often behave as function words, and replacing such nouns often yield ill-formed sentences.

While the style of the initial seed corpus necessarily reflects the underlying task (i.e. recognizing question sentences), the various steps of our proposed method do not rely on any explicit or implicit stylistic elements of the input seed corpus, so we consider our method to be task-independent.

5 Evaluation

5.1 Evaluation settings

5.1.1 Corpora

We prepared two Web archives as the starting point of language modeling: the first (`www`) is unfiltered in regards of style, while the second (`wwwq`) is a subset of the first, attempting to follow the style requirements of the QA system.

- **www** The first archive consists of 6.0×10^8 Web pages (Shinzato et al., 2008). We use this Web archive as training data for language modeling. After sanity check, we retained a corpus of 1.79×10^{10} words in 1.35×10^9 sentences. We call this corpus `www`.

- **wwwq** While the `www` corpus preserves the open-domain characteristics of the Web archive, it completely ignores the requirements of Ikkyu regarding style, containing various sentence types besides questions. Using simple heuristics, we selected questions from www. Since Japanese doesn't necessarily use question marks with questions, we identified questions as sentences that end with question marks or question markers (ka, kai, kashira, kana). We also extracted requests, which end with kudasai (≈*please*) or continuative verb + gerund te (≈Japanese colloquial request). This filtering retained a question corpus of 1.23×10^9 words in 1.01×10^8 sentences. We call this corpus wwwq.

Note that these regular expressions allow many sentences that don't match the style of Ikkyu, such as *why*, *how* or polar questions. Removing such problematic sentences from the corpus is not a trivial task in Japanese. For instance, interrogative pronouns can be omitted in Japanese.

The benefit of our proposed method is that we can rely on statistics and a small seed corpus without dealing with such numerous minor problems.

5.1.2 Evaluation sets

We recorded the questions uttered by 50 people (25 female and 25 male). Each subject was presented with 100 questions that Ikkyu can accept as input, after which each subject uttered approximately 50 of her/his own spontaneous questions which were recorded using smartphones. They were instructed to formulate simple questions, consisting of a noun, a predicate and an interrogative pronoun ("what", "who", "where" or "when") for a wide range of topics as far as possible, with the interrogative pronoun possibly being replaced by an expression to formulate a request, or being omitted altogether. Note that in spite of these instructions the recorded data contains some questions which do not conform.

The group was randomly split into three, with g0 containing 10 speakers, g1 and g2 containing 20 speakers each. Table 3 shows their details. The corpus of g0 was used as seed corpus, whereas g1 and g2 were used for two-fold cross validation.

Group	**g0** (S)	**g1**	**g2**
# of sentences	498	1000	999
# of words	4043	7671	8322
average sentence length	8.118	7.671	8.330

Table 3: Seed corpus and evaluation data.

5.2 Results

5.2.1 Best distributional similarity rank (k)

Firstly, we tried to find the best similarity rank k and training data size combination in terms of word-error-rate (WER), using the following values: k=1, 2, 3, 4, 5, 10, 15, 20, 100. After the seed corpus expansion, we incrementally increased the size of the training data, starting from a 10 million word corpus, gradually changing the threshold θ on Misu's score. We determined the best setting by two-fold cross validation, that is, we used the g1 set as development set with g2 as evaluation set, and vice versa. Figure 2 presents the WER curves for each parameter k, trained on the www corpus. Training on the www corpus provided with the best absolute performances, with $k = 10$ being the best setting in terms of WER consistently in both evaluation data, when the corpus size was approximately 160 million words (Figure 2). The vocabulary of this model has 413,000 entries. Note that $k = 10$ achieved best or 2nd best WER at most data points. This consistency suggests the effectiveness of our noun replacement. An interesting point is that the performance is saturated when k is relatively small, suggesting that the noun replacement is a relatively sensitive operation.

Note that when $k = 0$, the method is equivalent to Misu's method. The difference in WER between the best performance of Misu's method and our method when $k = 10$ is 0.19 points in g1 and 0.72 points in g2. Using McNemar's test (McNemar, 1947), we found that these differences are statistically significant ($p < 0.05$). Although the difference in g1 is quite small, (1) the best performance of our method is achieved using a smaller training corpus (50% of Misu's method); (2) if we use the same corpus size achieving the best performance in all the settings, i.e., 160 million words, the difference between Misu's method and $k = 10$ is even larger: 0.66 points in g1 and 0.95 points in g2; (3) there is also a large difference between the peak performances regarding sentence error rate (SER) in both of g1 and g2 (0.90 and 1.90 points). These observations suggest that our method is more effective than Misu's original method.

When we used the question archive wwwq instead of www, the same tendency was observed. The best performance was obtained when $k = 10$ for both g1 and g2. The peak performance was obtained when the corpus size had 80 million words consistently in both test data. However, the best

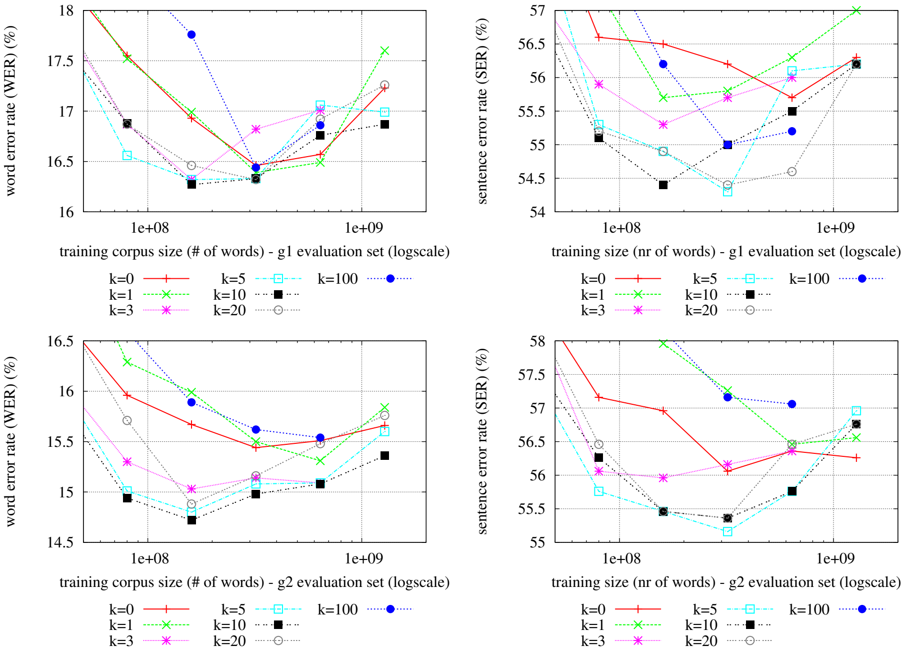

Figure 2: WER curves with various distributional similarity rank k, trained on the www corpus.

performance was worse than when we trained on www (16.35% in g1 and 15.28% in g2).

5.2.2 Comparison with the baselines

Next, we confirm that our method with the best setting ($k = 10$, 160 million words, trained on www) outperforms other baseline methods by comparing the following methods:

- **www.X** Our method applied to www.
- **wwwq.X** Our method applied to wwwq.
- **www.R** Random sampling from www.
- **wwwq.R** Random sampling from wwwq.

The results are presented in Figure 4.

Here we must mention that due to memory constraints, our language model training module (ATRASR) could not handle the entire www corpus. (We used machines with 72 GB memory.) The largest training data we could experiment on was 3.10×10^9 words. In case of the wwwq corpus such limitations don't apply, since 1.23×10^9 words can be handled by the module.

The peak performance of the baselines were achieved when the training data was largest, with wwwq showing lower WERs. Our method outperformed both baselines: regarding WER, we achieved 16.27% and 14.72% on g1 and g2, with an average improvement of 3.25 points, over the best baseline (www). These differences are also statistically significant ($p < 0.01$). Regarding SER, we achieved 54.30% and 55.16% on g1 and g2, respectively, with an average improvement of 4.28 points over the best baseline (wwwq), with the difference being statistically significant ($p < 0.01$). The performances of the baseline methods are not saturated and it may show much lower WER and SER if we can avoid the memory limitation problem somehow. However, the corpus size becomes so large (about 8 times the size of our best method), that would cause a significant slowdown of speech recognition. Figure 3 shows the real time factor[1] measured in our experiments. The best baseline method (trained on www) is already 2.8 times slower than the proposed method with the best setting. These observations suggest that our method is most suitable for our purpose.

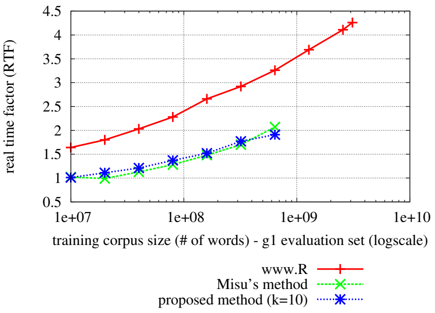

Figure 3: RTF in function of training corpus size.

[1] RTF ("real time factor"): defined as the processing time of the input divided by the duration of the input.

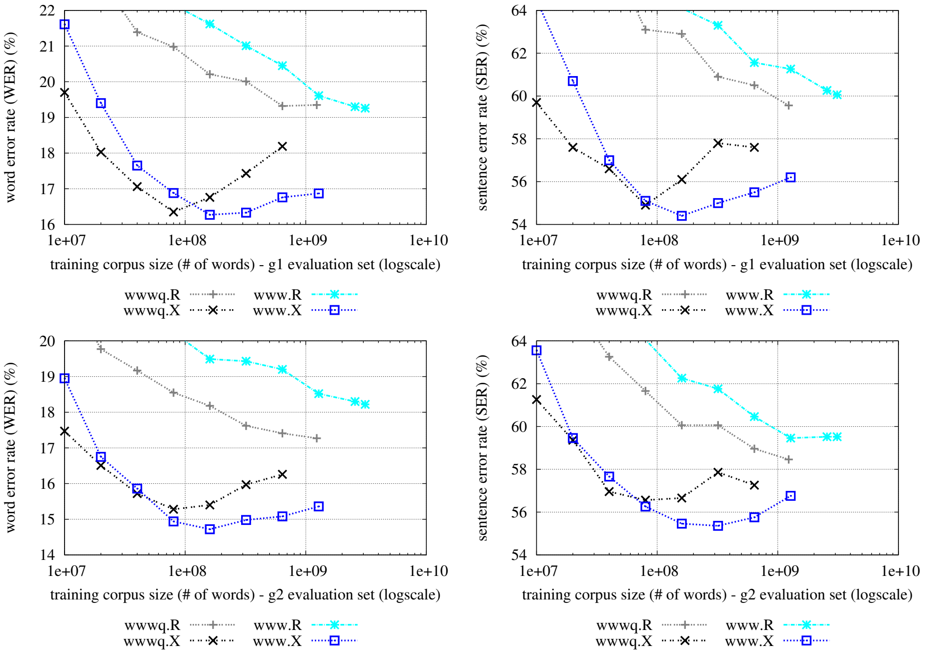

Figure 4: Our proposal versus the baselines.

5.2.3 N-best evaluation

The user interface of our QA system on smartphones has a user-friendly editing environment for word lattices provided by the speech recognition module. The user can correct the recognition results just by selecting some alternative words if the correct ones are included in the best N recognition results. The chance of a correct recognition is estimated as the probability of the correct question being in the N-best recognition results. We calculated SER over the top 100 recognition results. We found that between 59% and 62% (on g2 and g1, respectively) of speech inputs can be easily retrieved either by the top speech recognition result, or utilizing the error-recovery interface. This is important for the QA module in order to properly interpret and answer the questions.

5.3 Discussions

A concern is whether noun replacement really works as we intended. The improvement may have been achieved only for the words other than nouns, particularly the ones frequently observed in questions, such as verb suffixes indicating interrogative mood. This would not lead to topic expansion, as we intended. We conducted another series of experiments to investigate this supposition. Figure 5 shows the WER and SER of only nouns in the test data. Regarding noun-SER, a sentence is correct if all its nouns are correctly recognized. Interrogative pronouns were not considered. Using two-fold cross validation, on both g1 and g2 our method consistently achieved a lower noun-WER and noun-SER than Misu's method (without noun replacement), resulting in more accurate noun recognition. The difference between the best performances was 0.70 and 0.56 points in terms of noun-WER (on g1 and g2, respectively); 1.20 and 1.80 points in terms of noun-SER. These differences are statistically significant ($p < 0.05$). This implies that our concern was unfounded. Another observation is that the performance of the noun recognition does not saturate as the corpus size grows as far as we have tested. Analyzing this phenomenon more deeply and further improving the performance is our future work.

Another point is whether the noun replacement must be done before applying Misu's method. We may be able to achieve the same effect by performing noun replacement on the corpus obtained by Misu's method only using the original seed corpus. We experimentally confirmed this is not the case, although we don't present the experimental results for the sake of space. A possible explanation is that noun replacement generates many semantically ill-formed questions. If these are used as training corpus, they may be harmful. However, as seeds they only derive trigram probabilities for sentence selection, the selected sentences being natural, real life sentences from the Web.

We also attempted to build a language model by converting the patterns used by the QA module

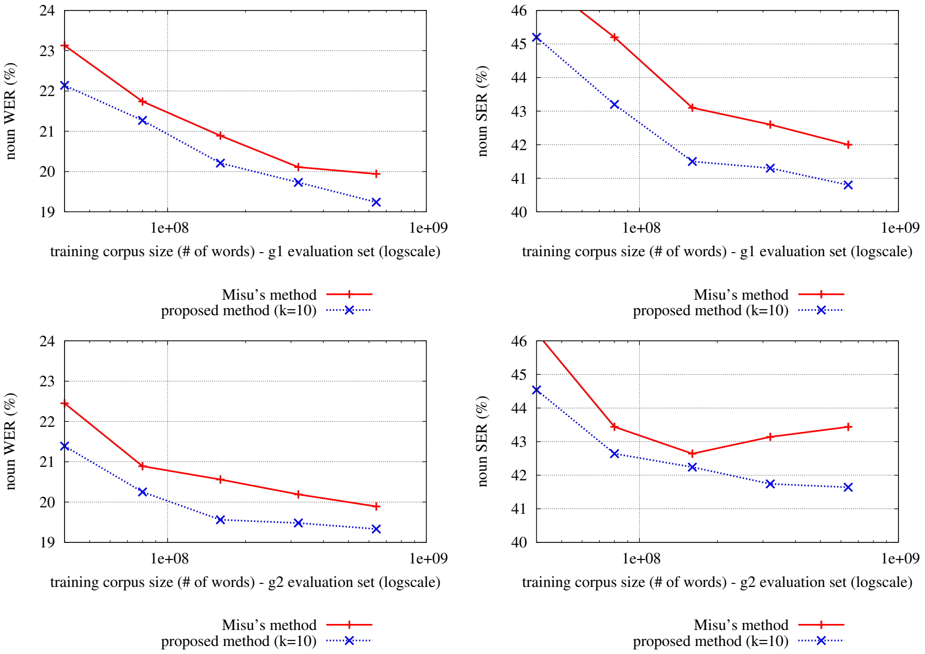

Figure 5: Noun evaluation on `www`.

to questions by replacing one of its variables with interrogative pronouns and attach some words indicating the interrogative mood. For instance, we could generate *"What causes deflation?"* from the pattern "X causes Y", but also less natural questions, such as *"What causes destruction?"*, which is extremely vague and is unlikely to be asked without a specific context. The conversion procedures generated a large number of such unnatural questions, and the resulting speech recognition performance was worse than that obtained by a corpus randomly sampled from our Web archive. To achieve the performance compatible to our method, thorough research must be conducted on generating only natural questions.

6 Related Work

The Web has been used as a relatively inexpensive source of large-scale data. "Just-in-time" language modeling (Berger et al., 1998) submits content words from previous user sentences as queries to a web search engine, Zhu et al. (2001) use a search engine to update the probabilities of already existing n-grams. More recently Bulyko et al. (2003) use frequent n-grams of the seed corpus as queries to retrieve similar text from the Web. Sarikaya et al. (2005) retrieves relevant text based on the BLEU score. Word perplexity is another frequently used similarity measure (Misu and Kawahara, 2006; Creutz et al., 2009). Some of these frameworks can be substituted for Misu's method in our framework, such replacement being a primary candidate of our future work.

Another category of related work relevant to our method is class-based language modeling (Brown et al., 1992). Many attempts refine this framework (Yamamoto et al., 2001; Chen and Chu, 2010; Emami et al., 2010). By replacing word trigrams in Misu's method by class trigrams, we may be able to achieve an effect similar to that obtained by our method without conducting noun replacement as additional procedure. However, the granularity of word classes may be a problem, considering that, in our framework, the optimal number of similar nouns replacing a noun in the seed corpus is relatively small (just 10). The comparison of class-based language models and our framework would be an interesting future work.

7 Conclusions

We have proposed a similarity based language model construction method for Ikkyu, a voice driven open-domain QA system. We used the combination of a distributional similarity based noun replacement method and a statistical domain adaptation method. Our best language model outperformed the baseline model constructed from random sampling of a Web archive by 3.25 points in word error rate and 4.28 points in sentence error rate, while using 8 times less training data.

References

Adam Berger, Robert Miller. 1998. Just-in-time language modeling. In *Proceedings of ICASSP-98*, pages 705–708.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, Robert L. Mercer. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics* 18(4), pages 467–479.

I. Bulyko, M. Ostendorf, A. Stolcke. 2003. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proceedings of Human Language Technology 2003 (HLT2003)*, pages 7–9.

Stanley F. Chen, Stephen M. Chu. 2010. Enhanced Word Classing for Model M. In *Proceedings of Interspeech 2010*, pages 1037–1040.

Mathias Creutz, Sami Virpioja, Anna Kovaleva. 2009. Web augmentation of language models for continuous speech recognition of SMS text messages. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 157–165.

Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata. 2009. Large Scale Relation Acquisition using Class Dependent Patterns. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'09)*, pages 764–769.

Ahmad Emami, Stanley F. Chen, Abraham Ittycheriah, Hagen Soltau, Bing Zhao. 2010. Decoding with shrinkage-based language models. In *Proceedings of Interspeech 2010*. pages 1033–1036.

Zellig Harris. 1954. Distributional Structure. In *Word* 10(23), pages 142–146.

Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, Kentaro Torisawa. 2010. A Bayesian Method for Robust Estimation of Distributional Similarities. In *Proceedings of ACL 2010*, pages 247–256.

S. Matsuda, T. Jitsuhiro, K. Markov, S. Nakamura. 2006. ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles *IEEE Transactions on Information and Systems* vol. E89-D(3), pages 989–997.

I. McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. In *Psychometrika* 12, pages 153–157.

Teruhisa Misu and Tatsuya Kawahara. 2006. A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts. In *Proceedings of Interspeech 2006*. pages 9–13.

R. Sarikaya, A. Gravano, Y. Gao. 2005. Rapid Language Model Development Using External Resources for New Spoken Dialog Domains. In *Proceedings of ICASSP 2005*, vol I, pages 573–576.

Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, Sadao Kurohashi. 2008. TSUBAKI: An open search engine infrastructure for developing new information access. In *Proceedings of IJCNLP*, pages 189–196.

Xiaojin Zhu, R. Rosenfeld. 2001. Improving trigram language modeling with the world wide web. In *Proceedings of ICASSP*, pages 533–536.

Hirofumi Yamamoto, Shuntaro Isogai, Yoshinori Sagisaka. 2001. Multi-Class Composite N-gram Language Model for Spoken Language Processing Using Multiple Word Clusters. In *Proceedings of ACL-2001*, pages 6–11.