INCORPORATING SEMANTIC INFORMATION TO SELECTION OF WEB TEXTS FOR LANGUAGE MODEL OF SPOKEN DIALOGUE SYSTEM

Koichiro Yoshino, Shinsuke Mori and Tatsuya Kawahara

School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan
E-mail: kawahara@i.kyoto-u.ac.jp

ABSTRACT

A novel text selection approach for training a language model (LM) with Web texts is proposed for automatic speech recognition (ASR) of spoken dialogue systems. Compared to the conventional approach based on perplexity criterion, the proposed approach introduces a semantic-level relevance measure with the back-end knowledge base used in the dialogue system. We focus on the predicate-argument (P-A) structure characteristic to the domain in order to filter semantically relevant sentences in the domain. Several choices of statistical models and combination methods with the perplexity measure are investigated in this paper. Experimental evaluations in two different domains demonstrate the effectiveness and generality of the proposed approach. The combination method realizes significant improvement not only in ASR accuracy but also in semantic and dialogue-level accuracy.

Index Terms— Language model, spoken dialogue system, Web, predicate-argument structure

1. INTRODUCTION

In the past decade, a number of spoken dialogue systems have been developed for real-world applications. Most recently, remarkable technical progress has been made to realize "open-domain" systems, which try to reply to any queries. However, the function of the open-domain systems is actually limited to single search or question-answering, which totally relies on the back-end system. Therefore, most of the dialogue systems that involve a sequence of interactions with users still assume a particular domain such as restaurants and sightseeing spots in a specific area. By exploiting the domain knowledge, it is possible to design a dialogue that disambiguates user queries or makes system-initiative information presentation [1]. Research on spoken dialogue systems is now focused on automating the system design including language model (LM), understanding model and dialogue model, because hand-crafting the domain knowledge is so costly. Note that even if a statistical model is adopted, it would be labor-intensive to collect a large amount of training data and annotate them.

This work is focused on effective and efficient training of an LM for ASR of a spoken dialogue system, without collecting any domain-dependent corpus. Instead, we use Web resources for collecting relevant training data. Recently, there are a number of Web sites of "wisdom of crowds", in which people can ask any questions, which are replied by other persons. Thus, we can collect a number of useful query patterns, which may be similar to those used for spoken dialogue systems.

There are a number of previous studies on selecting Web texts for training an LM for ASR systems [2, 3, 4, 5, 6]. The majority of them assume a "seed" corpus for collection and selection, and they adopt a selection criterion based on perplexity, because an LM is usually evaluated with perplexity besides ASR accuracy. In this paper, we propose to incorporate semantic information to the selection of Web texts, because the goal of spoken dialogue systems is to extract semantic meaning of user utterances. The information is expected to be effective for selecting Web texts that are semantically relevant to the task domain of the spoken dialogue system.

2. LANGUAGE RESOURCE

We assume a back-end document set that will be used for retrieval of the reply to user queries. The documents are not a relational database but natural language texts of a particular domain, for example, tourist guidebook and cooking recipe. In this work, we primarily use a set of newspaper articles of the professional baseball domain as a knowledge base, which are used for replying to user queries on baseball events. Note that the documents are not directly used for LM training, because their style is much different from user queries and the majority of content is not relevant for query.

In this work, we turn to a much larger Web resource of Yahoo! QA [1], a Web site of wisdom of crowds, in which people can ask questions according to the domain category. Note that the definition of domain categories does not match that of the spoken dialogue system, and moreover there are many irrelevant queries in the Web site.

3. SELECTION BASED ON PERPLEXITY

Previously, many studies have been conducted on selection of Web texts for LM training, but the majority of them adopt the perplexity criterion or its variants for selection. Many works assume a seed corpus to prepare a seed LM for generating a Web search query or computing perplexity [2, 3, 4, 5]. Misu et al. [6] proposed to combine the domain-dependent back-end documents with a domain-independent dialogue corpus to automatically prepare the seed corpus.

For a sentence $s = w_1, \ldots, w_l$, its perplexity by a seed LM trained with the document set D is defined by

$$H(s, D) = -\frac{1}{l} \sum_{i}^{l} \log_2 P_D(w_i). \tag{1}$$

$$PP(s, D) = 2^{H(s,D)}. \tag{2}$$

[1] http://chiebukuro.yahoo.co.jp/; The corpus is provided by Yahoo!Japan and NII, Japan.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

Here, P_D is a probability computed by the seed N-gram (3-gram in this work) model trained with D. In fact, this is equivalent to KL divergence between s and D,

$$KL(s||D) = \frac{1}{l}\sum_{i}^{l} P_s(w_i)\log_2\frac{P_s(w_i)}{P_D(w_i)}, \tag{3}$$

since we can approximate the N-gram probability trained with the sentence itself $P_s(w_i)$ as 1 when s is short and its N-gram sequences are unique.

Each sentence s is evaluated with $PP(s,D)$ or $KL(s||D)$, and selected for the use in LM training.

4. SELECTION BASED ON SEMANTIC RELEVANCE MEASURE

The perplexity criterion evaluates the word surface-level relevance of sentences. In this paper, we propose a semantic-level criterion to measure the relevance to the task domain of the spoken dialogue system.

4.1. Definition of Semantic Relevance Measure

For this purpose, we need to define a semantic relevance measure. We focus on the predicate-argument (P-A) structure. It is a classical concept for semantic analysis in natural language processing (NLP), and recently used in information extraction [7, 8].

The P-A structure consists of a predicate (w_p), which is usually defined by a verb, and an argument (w_a) and its semantic case. There are more than one P-A pairs in a sentence which share a predicate. Some of recent NLP parsers have a function to generate P-A structures, as a result of training with a large text corpus. We use the KNP parser [2] in this work. However, not every P-A pair is meaningful in spoken dialogue in a particular domain; actually only a fraction of them are useful. For example, in the baseball domain, key patterns include "[A (agent) beat B (object)]" and "[A (agent) hit B (object)]", and in the business domain, "[A (agent) sell B (object)]" and "[A (agent) acquire B (object)]". Thus, the useful semantic information is dependent on the domain. Conventionally, this kind of templates for information extraction have been hand-crafted [7], but the heuristic process is so costly that it cannot be applied to a wide variety of domains. We have proposed a method to automatically extract domain-specific templates for a flexible information navigation system [9].

In our previous work [9], it was shown that the Naive Bayes score worked better than the tf-idf score for selecting characteristic P-A patterns, thus we adopt the Naive Bayes score in this work. Here, we define a probability of word w_i being in the document set of a particular domain D by assuming the other set of documents $\bar{D}$ of different domains,

$$\begin{aligned} P(D|w_i) &= \frac{P(w_i|D) * P(D)}{P(w_i)}, & (4)\\ &\simeq \frac{C(w_i,D) + P(D)*\gamma}{C(w_i)+\gamma}, & (5)\end{aligned}$$

where $C(.)$ stands for an occurrence count and $P(D)$ is a normalization factor determined by the size of D and $\bar{D}$. γ is a smoothing factor estimated with a Dirichlet prior using the Chinese Restaurant Process (CRP).

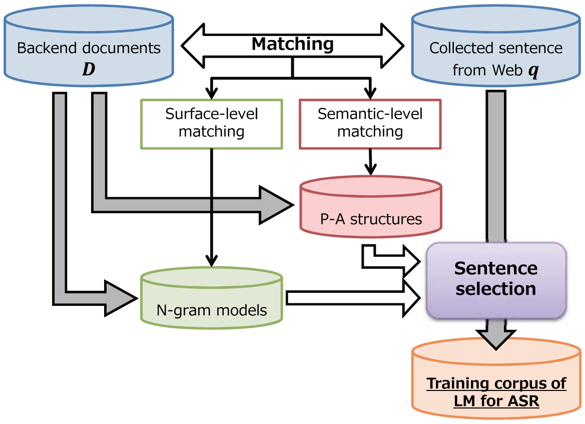

Fig. 1. Overview of the proposed approach

The above formula is a variation of unigram probability, but here we focus on P-A pairs of predicate w_p and w_a, not dealing all words uniformly. For a P-A pair PA_j consisting of w_{jp} and w_{ja}, we define $P(D|PA_j)$ as a geometric mean of $P(D|w_{jp})$ and $P(D|w_{ja})$, and use it as a semantic relevance measure in the domain defined by the document set D. For each sentence s, we compute a mean of $P(D|PA_j)$ for P-A pairs included in the sentence, defined as $P(D|s)$. According to the mean score, sentences are selected for LM training.

Alternatively, we can compute $P(D|s)$ via a discriminative model such as Logistic Regression (LR) model and Conditional Random Fields (CRF). In this case, we prepare a positive training set of P-A patterns extracted from sentences in D and a negative training set from sentences in $\bar{D}$, and train a classifier to discriminate them using the P-A features. We will compare the performance of the LR model with that of the simple Naive Bayes classifier.

4.2. Combination with Perplexity Measure

Then, we investigate combination of the proposed semantic relevance measure with the perplexity measure, since they presumably model different aspects of the relevance with the target domain. A simple combination method is to use the ranks by the two measures. We can re-order the sentences based on the sum of them.

We can also define a score-based combination. For this purpose, we convert the perplexity $PP(s,D)$ into a score dimension $[0,1]$ via a sigmoid function,

$$PP' = \frac{1}{1+\exp(-PP)} \tag{6}$$

which can be linearly-combined with the semantic relevance measure based on $P(D|s)$. In the preliminary evaluation, we did not see a significant difference between the rank-based method and the score-based method. So, we will show the results by the rank-based method in the following section.

The overall procedure is summarized in Figure 1, in which text selection is conducted based on the two relevance measures.

[2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

Table 1. Text size (number of sentences) in two domains

domain	baseball	Kyoto
back-end documents	177K	36K
Web sentences for selection	404K	680K
test utterances	2747	219

5. EXPERIMENTAL EVALUATIONS

We have evaluated the proposed approach in a speech-based navigation system in the Japanese professional baseball domain [9] and the Kyoto sightseeing domain [10]. The system can answer user's questions regarding the domain using the back-end documents of the respective domains. As the back-end document set D, we used newspaper articles (Mainichi Newspaper Corpus) tagged with the professional baseball and Wikipedia entries with a tag of Kyoto City for the respective domains.

The statistical models described in Section 3 and 4 are trained with the document set D, and used for selecting query sentences collected in the Yahoo! QA Web site. We collected sentences in the baseball domain and tourism domain, respectively. The test set of user utterances was separately collected using the dialogue system. The text size for the experimental evaluations is summarized in Table 1.

Word trigram LMs were trained with the texts selected based on the relevance measures described in Section 3 and 4. We trained a variety of LMs using the texts of different sizes relative to all available texts (3/10 through 10/10 where all texts are used) by changing the selection threshold.

5.1. Evaluation with Perplexity and ASR Accuracy

We first evaluated the test-set perplexity by each LM. Since the resultant training text set is different due to the selection process, the perplexity is adjusted by assuming the same vocabulary, which is defined by the entire training set. In the adjusted perplexity (APP), the probability of unknown words (<*UNK*>) is divided by the number of unseen lexical entries in the current training set. APP is plotted for LMs of different text sizes in Figure 2 and 3 for the baseball news domain and the Kyoto sightseeing domain, respectively. Compared with the conventional perplexity-based measure (PP), the proposed semantic relevance measure (PA) performs comparably in the baseball news domain (Figure 2) and more effectively in the Kyoto sightseeing domain (Figure 3). However, there is no significant difference between the Naive Bayes classifier (PA) and the Logistic Regression model (PA(LR)) in the two graphs, thus we adopt the simple Naive Bayes classifier.

Then, we made an evaluation in terms of ASR accuracy. Word error rate (WER) is computed for the test-set utterances. We used a speaker-independent triphone model for the acoustic model and the Julius decoder [11] [3]. WER is plotted for LMs of different text sizes in Figure 4 and 5 for the baseball news domain and the Kyoto sightseeing domain, respectively. It is shown that the text selection results in significant WER reduction. In the baseball news domain, the proposed semantic relevance measure (PA) performed significantly better than the conventional perplexity measure (PP), and the combination of the two measures (PP+PA) is not so effective. In the Kyoto sightseeing domain, however, the combination of the two measures has a synergetic effect. In both domains, the optimal point lies around 7/10.

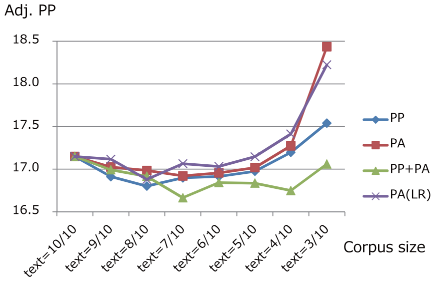

Fig. 2. APP by LMs with selected texts (baseball domain)

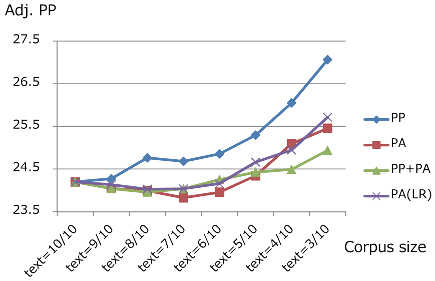

Fig. 3. APP by LMs with selected texts (Kyoto domain)

5.2. Evaluation with Semantic and Dialogue-level Accuracy

Next, we made an evaluation with the semantic and dialogue-level accuracies, which are more related with the performance of the spoken dialogue system. Semantic accuracy is measured by an error rate of P-A pairs, in which we count as correct if both the predicate and the argument are correctly extracted. The P-A error rate (PAER) is plotted for the baseball news domain in Figure 6. [4] By using the LM selected (by 7/10) by the combination method (PP+PA), the PAER is reduced to 20.4% from the baseline 21.5% without text selection.

Dialogue-level accuracy is measured by the ratio of appropriate responses against all user queries. We also observed an increase of the appropriate responses (by 0.8% absolute) as a result of the PAER improvement.

6. RELATION TO PRIOR WORK

Use of Web resources for LM training has been investigated as the Web becomes prevailing. In the early years, Zhu et al. [12] enhanced trigram statistics with frequencies of the trigram in the Web, and Nisimura et al. [13] collected texts from the Web by manually specifying keywords in the task domain.

But the most standard approach [2, 3, 4, 14, 15] is to use characteristic N-gram entries as search queries for the Web to collect relevant texts, but this requires a seed corpus to estimate a seed N-gram model. Several works used other resources for generating search

[3] http://julius.sourceforge.jp/

[4] PAER for the Kyoto sightseeing domain is not presented because the number of P-A pairs is not large enough for evaluation.

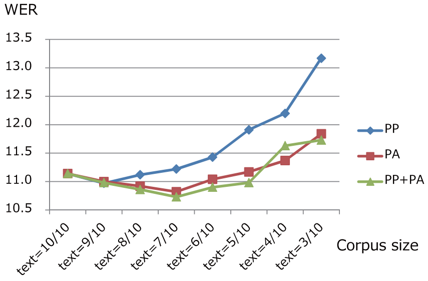

Fig. 4. WER by LMs with selected texts (baseball domain)

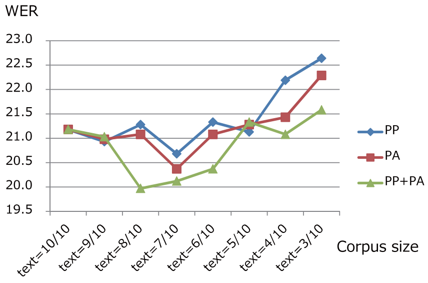

Fig. 5. WER by LMs with selected texts (Kyoto domain)

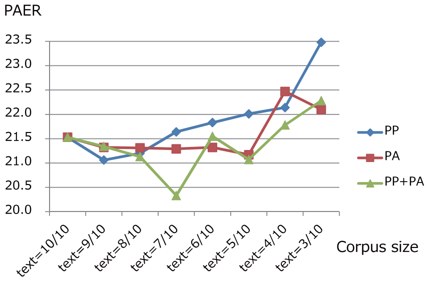

Fig. 6. PAER by LMs with selected texts (baseball domain)

queries, such as back-end documents [6], presentation slides in lectures [16, 17], or initial ASR transcripts [18].

Selection of the collected Web texts has also been investigated. The majority of the previous studies adopted the perplexity measure by the seed LM [2, 4, 6], or its variants such as BLEU score [3] and normalization by the background topic model [5] or the self model [19]. Masumura et al. [20] introduced a Naive Bayes classifier for selecting spoken-style texts. But all the previous works do not consider semantic-level information.

An exception is the work by Akbacak et al. [21], which defined characteristic noun phrases and verbs to filter Web texts. However, their method was largely heuristic and did not define a statistical semantic relevance measure. Hakkani-Tur et al. [22] and Yoshino et al. [23] introduced a P-A structure to parse sentences in the in-domain Web pages and convert them into query-style sentences. This work is different in that the P-A structure is parameterized into a semantic relevance measure, which is used for selection of large-scale Web texts.

7. CONCLUSIONS

We have presented a novel text selection approach for training LMs for spoken dialogue systems. Compared to the conventional perplexity criterion, the proposed approach introduces a semantic-level relevance measure with the back-end knowledge base used in the dialogue system. Thus, it can effectively filter semantically relevant sentences for the task domain. It can also be combined with the perplexity measure for a synergetic effect. We have made experimental evaluations in two different domains, demonstrating its effectiveness and generality. The combination method realized significant improvement not only in WER but also in semantic and dialogue-level accuracies. The proposed approach only uses the texts in the back-end system, and does not require any "seed" corpus. Therefore, it can be used for building a spoken dialogue system of a particular domain from scratch.

Acknowledgments: This work was supported by JST CREST and JSPS Grant-in-Aid for Scientific Research.

8. REFERENCES

[1] T.Kawahara. New perspectives on spoken language understanding: Does machine need to fully understand speech? In *Proc. IEEE Workshop Automatic Speech Recognition & Understanding (ASRU)*, pages 46–50 (invited paper), 2009.

[2] I.Bulyko, M.Ostendorf, and A.Stolcke. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proc. HLT*, volume 2, pages 7–9, 2003.

[3] R.Sarikaya, A.Gravano, and Y.Gao. Rapid language model development using external resources for new spoken dialog domains. In *Proc. IEEE-ICASSP*, volume 1, pages 573–576, 2005.

[4] T.Ng, M.Ostendorf, M.-Y.Hwang, M.Siu, I.Bulyko, and X.Lei. Web-data augmented language models for mandarin conversational speech recognition. In *Proc. IEEE-ICASSP*, volume 1, pages 589–592, 2005.

[5] A.Sethy, P.G.Georgiou, and S.Narayanan. Building topic specific language models from webdata using competitive models. In *Proc. INTERSPEECH*, pages 1293–1296, 2005.

[6] T.Misu and T.Kawahara. A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web texts. In *Proc. INTERSPEECH*, pages 9–12, 2006.

[7] R.Grishman. Discovery methods for information extraction. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 243–247, 2003.

[8] L.Ramshaw and R.M.Weischedel. Information extraction. In *Proc. IEEE-ICASSP*, volume 5, pages 969–972, 2005.

[9] K.Yoshino, S.Mori, and T.Kawahara. Spoken dialogue system based on information extraction using similarity of predicate argument structures. In *Proc. SIGdial Meeting Discourse & Dialogue*, pages 59–66, 2011.

[10] T.Misu and T.Kawahara. Bayes risk-based dialogue management for document retrieval system with speech interface. *Speech Communication*, 52(1):61–71, 2010.

[11] A.Lee and T.Kawahara. Recent development of open-source speech recognition engine Julius. In *Proc. APSIPA ASC*, pages 131–137, 2009.

[12] X.Zhu and R.Rosenfeld. Improving trigram language modeling with the world wide web. In *Proc. IEEE-ICASSP*, volume 1, pages 533–536, 2001.

[13] R.Nisimura, K.Komatsu, Y.Kuroda, K.Nagatomo, A.Lee, H.Saruwatari, and K.Shikano. Automatic n-gram language model creation from web resources. In *Proc. Eurospeech*, pages 5181–5184, 2001.

[14] V.Wan and T.Hain. Strategies for language model web-data collection. In *Proc. ICASSP*, volume 1, pages 1069–1072, 2006.

[15] A.Tsiartas, P.Georgiou, and S.Narayanan. Language model adaptation using www documents obtained by utterance-based queries. In *Proc. ICASSP*, pages 5406–5409, 2010.

[16] C.Munteanu, G.Penn, and R.Baecker. Web-based language modelling for automatic lecture transcription. In *Proc. Interspeech*, pages 2353–2356, 2007.

[17] T.Kawahara, Y.Nemoto, and Y.Akita. Automatic lecture transcription by exploiting presentation slide information for language model adaptation. In *Proc. IEEE-ICASSP*, pages 4929–4932, 2008.

[18] M.Suzuki, Y.Kajiura, A.Ito, and S.Makino. Unsupervised language model adaptation based on automatic text collection from WWW. In *Proc. Interspeech*, pages 2202–2205, 2006.

[19] R.C.Moore and W.Lewis. Intelligent selection of language model training data. In *Proc. ACL*, pages 220–224, 2010.

[20] R.Masumura, S.Hahm, and A.Ito. Training a language model using webdata for large vocabulary Japanese spontaneous speech recognition. In *Proc. INTERSPEECH*, pages 1465–1468, 2011.

[21] M.Akbacak, Y.Gao, L.Gu, and H.-K.J.Kuo. Rapid transition to new spoken dialogue domains: Language model training using knowledge from previous domain applications and web text resources. In *Proc. INTERSPEECH*, pages 1873–1876, 2005.

[22] D.Hakkani-Tur and M.Gilbert. Bootstrapping language models for spoken dialog systems from the world wide web. In *Proc. IEEE-ICASSP*, volume 1, pages 1065–1068, 2006.

[23] K.Yoshino, S.Mori, and T.Kawahara. Language modeling for spoken dialogue system based on sentence transformation and filtering using predicate-argument structures. In *Proc. APSIPA ASC*, 2012.