Domain-Independent Spoken Dialogue Platform using Key-Phrase Spotting based on Combined Language Model

Kazunori Komatani, Katsuaki Tanaka, Hiroaki Kashima and Tatsuya Kawahara

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
komatani@kuis.kyoto-u.ac.jp

Abstract

We present a portable platform for spoken dialogue systems and its experimental evaluation. Conventional development of speech interfaces involves much labor cost in either describing a task grammar or collecting a task corpus. Our platform automatically generates a lexicon and a language model of key-phrases based on task description and structure of the domain database. By spotting key-phrases using both the generated grammar and word 2-gram model trained with dialogue corpora of similar domains, we realize flexible speech understanding on a variety of utterances. Furthermore, adopting a GUI that explicitly displays acceptable utterance patterns is effective in guiding user utterances within the system's capability. We evaluate the generated spoken dialogue system using 24 novice users. The number of unacceptable utterances are significantly reduced with the simple phrase grammar and GUI. And the phrase spotter using the combined language model improves the semantic accuracy by 15.5% compared with the conventional method decoding the whole sentence with a fixed grammar.

1. Introduction

With improvement of speech recognition technology, many kinds of spoken dialogue systems have been developed. Information query is regarded as one of the most promising tasks, because the majority of operations in this task are to select from huge entries, in which speech interface is advantageous. But spoken dialogue systems are not ubiquitous yet. One of the cause is lack of portability [1][2] . It is necessary to set up appropriate linguistic constraints and semantic interpretation rules, but it requires a great deal of labor with expertise. When we use elaborate corpus-based statistic models, it is necessary to collect large amount of a corpus that matches the task and domain. Thus, rapid prototyping technique is important in data collection as well as system development.

Therefore, we develop a domain-independent platform for information query tasks. Domain-dependent information is extracted from the domain database. General rules for information query are retained as domain-independent information. By limiting the task to typical information query, a lexicon and grammar rules for speech recognition are extracted from the domain database based on a simple task description.

The other serious problem in spoken dialogue system is recognition errors caused by out-of-vocabulary and out-of-grammar utterances. A novice user often makes utterances beyond system's capacity, since they do not know the acceptable vocabulary and grammar. In order to accomplish the system's task successfully, users should be guided into acceptable expressions. We adopt a GUI that shows patterns of acceptable utterances and current query status explicitly, so that user utterance can be guided into the system's capacity. Moreover, our key-phrase spotter that utilizes the generated grammar and 2-gram model trained with similar domain corpora can cope with various utterances and extract key-phrases flexibly.

2. Approach to Domain-Independency

A typical spoken dialogue system consists of speech recognition, semantic interpretation and dialogue management modules. In order to realize full domain-independency, these three steps must be domain-independent. But unlimited-vocabulary spontaneous speech recognition and universal semantic interpretation for any domains are very difficult problems[3].

In our platform, we assume that the system is to perform information query using multi-modal interfaces. Information query can involve a lot of domains, but provides constraint to a semantic analyzer so that key-phrase-based understanding is feasible. Use of multi-modal interfaces with a display eases the problems of speech recognition and dialogue management.

2.1. Model of Information Query

For generality and simplicity, we regard information query as filling a query form that consists of a set of search keys. Thus, user utterances are modeled as setting and retracting search keys. Domain of the query is not limited, but includes trains, flights and hotels.

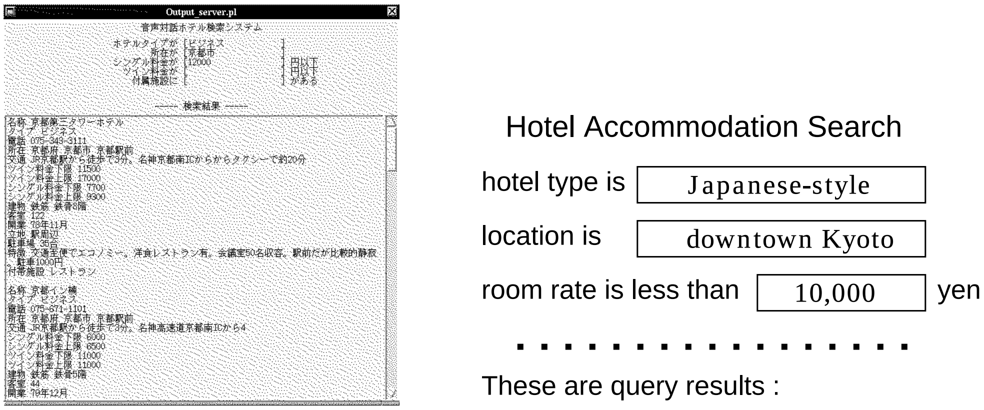
The platform semi-automatically generates a lexicon and grammar rules that cover possible expressions of search keys. They are derived from the domain database, i.e. database of trains or hotels. Search keys are generally made of a set of search items and their values, for example, "location is Tokyo", which usually correspond to database fields and entries, respectively. Thus, a lexicon is automatically derived from the database fields and entries. A baseline grammar within key-phrase is also set up to accept typical expressions used in the query.

In this framework, the semantic analysis is achieved as filling query slots with keywords and translating key-phrases into them. It is independent of domains. We also prepare several universal patterns for expressions to retract or clear search keys.

2.2. Use of GUI

Use of a GUI (Graphical User Interface) constrains user utterances to be recognized and complements a dialogue manager.

Among the major problems in speech understanding is out-of-vocabulary or out-of-grammar expressions. On the other hand, we have observed that users often hesitate to speak to



```
Output_server.pl
                    音声対話ホテル検索システム
        ホテルタイプが [ビジネス              ]
                所在が [京都市                ]
      シングル料金が [12000                 ] 円以下
          ツイン料金が [                      ] 円以下
          付属施設に [                      ] がある

                      ----- 検索結果 -----
名称 京都第三タワーホテル
タイプ ビジネス
電話 075-343-3111
所在 京都府 京都市 京都駅前
交通 JR京都駅から徒歩で3分。名神京都南ICからからタクシーで約20分
ツイン料金下限 11500
ツイン料金上限 17000
シングル料金下限 7700
シングル料金上限 9300
建物 鉄筋 鉄骨8階
客室 122
開業 78年11月
立地 駅周辺
駐車場 35台
特徴 交通至便でエコノミー。洋食レストラン有。会議室50名収容。駅前だが比較的静寂
。駐車1000円
付帯施設 レストラン

名称 京都イン楠
タイプ ビジネス
電話 075-671-1101
所在 京都府 京都市 京都駅前
交通 JR京都駅から徒歩で3分。名神高速道京都南ICから4
シングル料金下限 6000
シングル料金上限 6500
ツイン料金下限 11000
ツイン料金上限 11000
建物 鉄筋 鉄骨5階
客室 44
開業 79年12月
```

(a) A real system in Japanese

Hotel Accommodation Search

hotel type is [Japanese-style]

location is [downtown Kyoto]

room rate is less than [10,000] yen

. . . .

These are query results :

(b) Upper portion translated in English

Figure 1: An outlook of GUI (Graphical User Interface)

machines simply because they do not know which forms of expressions are acceptable. Our platform displays key-phrase patterns as a visual form of the query, which guides users how to speak and reduces the variation of input utterances.

Moreover, the system promptly displays recognition results in the slots of the visual query form as well as the query results. The feature lets the users know recognition errors and eliminates the necessity of confirmation through spoken dialogue. Instead, the users simply make "undo" commands in case of errors. This feature will avoid possible crashes in dialogue. It also enables users specify preferences incrementally by reviewing the current (number of) matched entries. It is useful for those who do not have in mind a definite preference beforehand.

An outlook of the GUI is shown in Figure 1.

2.3. Portability of Language Model for Speech Recognizer

N-gram model is a powerful language model of speech recognizer if sufficient training corpus of particular domain is available, but it is difficult to collect sufficient amount of corpus for every specific domain. On the other hand, a manual grammar is often used as language model of spoken dialogue systems. It does not need training corpus and can introduce domain-specific knowledge easily. But it is nearly impossible to describe all expression patterns with grammar rules because user utterances have enormous variations especially in filler portions[1]. Speech recognition using a described grammar only is often too rigid.

We have proposed a method based on key-phrase spotting to recognize spontaneous speech flexibly[4]. The grammar rules solely for the key-phrase portions are definite and simple, so can be written with less labor accordingly. It realizes robustness against ill-formed utterances. However, the constraint of key-phrase grammars without statistics is so loose that false alarms consisting of short words appear frequently.

Considering such issues, we present a novel phrase-spotting method based on combined language model, which consists of both grammar rules for domain-dependent key-phrases and 2-gram constraint derived from similar domain corpora for domain-independent fillers. The model puts proper constraint over the whole sentence by applying N-gram to filler portions, and consequently can improve the recognition accuracy. As word N-gram model is trained with similar domain corpora that do not necessarily match the query domain, our method realizes portability for various domains.

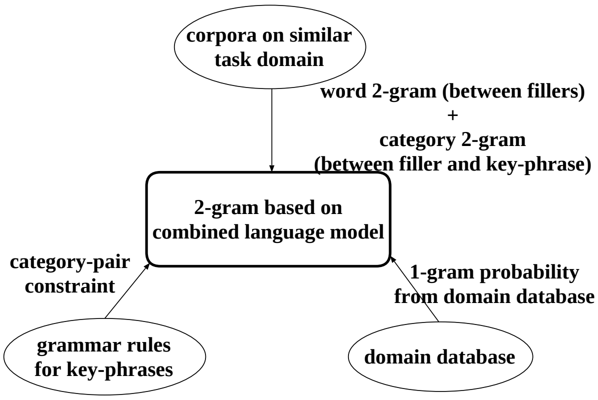

corpora on similar task domain
word 2-gram (between fillers)
+
category 2-gram
(between filler and key-phrase)
2-gram based on combined language model
category-pair constraint
1-gram probability from domain database
grammar rules for key-phrases
domain database

Figure 2: Concept of combined language model

3. Key-Phrase Spotting based on Combined Language Model

3.1. Combination of Grammar Rules and Statistical Model

We assume that the corpus is not perfectly matched to the domain but similar to the system's task. Based on this assumption, we construct the linguistic constraint from three information sources (Figure 2).

As a constraint between fillers which does not affect key-phrase portions directly, we apply 2-gram probabilities estimated with similar domain corpora, since filler portions are regarded as domain-independent. In key-phrases, word transitions are defined based on the category-pair constraint which is automatically derived from the generated phrase grammar.

As 2-gram model between a filler and a key-phrase, we introduce a class 2-gram that consists of nouns, which make up key-phrases. For words in key-phrases that have relatively domain-independent concepts (price, date, name of place), a specific class is prepared. Moreover, when this class 2-gram probabilities is transformed into word 2-gram probability, 1-

[1] We regard the portions other than key-phrases as filler, such as "I would like".

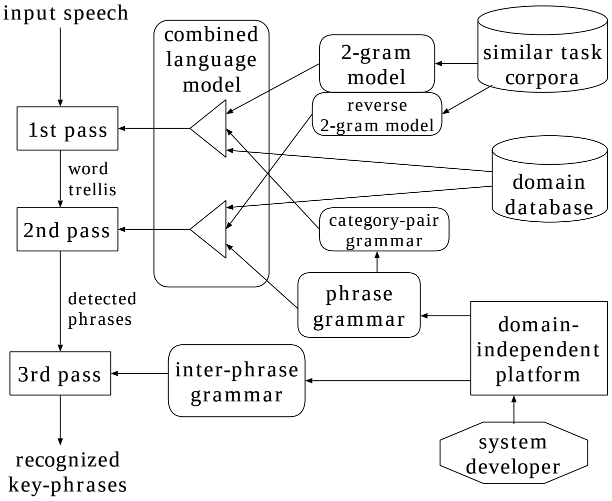


Figure 3: Overview of key-phrase spotting method

gram probability is provided based on the distribution of domain database entries.

Specifically, the constructed language model is formulated as below. Here, $p_{co}()$, $p_{db}()$ and $p_{gr}()$ denote the probability derived by similar domain corpora, a domain database and a phrase grammar, respectively. The probability assigned to the phrase grammar ($p_{gr}(c_2|c_1)$) is 1 if the concatenation c_1c_2 is defined in grammar rules, otherwise 0.

- for filler portions
  $p(w_2|w_1) \quad = \quad p_{co}(w_2|w_1)$
- between key-phrase and filler
  $p(w_2|w_1) \quad = \quad p_{db}(w_2|c_2) \cdot p_{co}(c_2|w_1)$
- inside key-phrase

$$\begin{aligned} p(w_2|w_1) &= p_{db}(w_2|c_2) \cdot p_{gr}(c_2|c_1) \\ &= \begin{cases} p_{db}(w_2|c_2) & \text{(if } c_1c_2 \text{ is defined)} \\ 0 & \text{(otherwise)} \end{cases} \end{aligned}$$

For example, a probability between a key-phrase “Kyoto” and a filler “in” $p(\text{Kyoto} \mid \text{in})$ is calculated by multiplying $p_{co}(\text{PLACE} \mid \text{in})$ derived from similar corpora by $p_{db}(\text{Kyoto} \mid \text{PLACE})$ derived from the domain database.

3.2. Key-Phrase Spotting based on Combined Model

We adopt a progressive search strategy focused on key-phrases portions. It applies more strict linguistic constraint incrementally on key-phrases in the three steps (Figure 3).

1st pass word 2-gram model and category-pair grammar

2nd pass phrase grammar (inside key-phrase) and word 2-gram model

3rd pass inter-phrase grammar (when connecting key-phrases)

In the 1st and 2nd passes, key-phrases are spotted based on the combined language model derived from word 2-gram and grammar rules. In the 3rd pass, we connect spotted phrases and recognize as a sentence. In this step, phrase candidates are connected according to their scores and semantic constraint, which is defined as inter-phrase grammar rules. Because spotting methods do not assume parsing the whole sentence, we put a constant penalty value when there is a skipped portion between key-phrases.

Table 1: Comparison with handcrafted grammar

	generated grammar	handcrafted grammar
user utterances within system capacity	93%	55%
recognition accuracy	69%	31%

4. Experimental Evaluation

We have applied the platform to hotel database and literature database, and constructed a hotel search system and a literature search system, respectively. The hotel database contains 2040 entries and has seven items such as location, upper limit of room rate, facilities and so on. Users can specify and retract slot values corresponding to the items. It is possible to specify and retract multiple items in a single utterance. The literature database contains 1913 entries and has five items users can specify, such as title, author(s), journal name and published year. Experimental evaluation is made on the hotel search system.

4.1. Initial Performance of Generated System for Hotel Domain

Performance of portable platforms can be measured by the coverage of the system over user utterances. We evaluate robustness of our platform by counting how user utterances are accepted by the prototype system. We make two experiments using the same speech recognizer Julian[5]. The vocabulary of the prototype system contains 982 words, which have been extracted from the hotel database automatically The grammar used in this experiment is a finite state automaton that is repetition of key-phrases.

In the first experiment, we use two hotel search systems: one is the prototype system generated by our platform and the other uses a carefully handcrafted grammar for speech recognizer. Subjects are students in our department. In the prototype system, acceptable patterns of utterances are displayed explicitly. The ratio of acceptable utterances is shown in Table 1. Better performance is achieved by the generated grammar. This result demonstrates that both robustness and portability can be achieved by using a simple generated grammar of key-phrases and guiding user utterances rather than writing a complex sentence-level grammar manually.

In the second experiment, we evaluate the effect of guidance by GUI for novice users using the generated grammar. Subjects are 28 novices users (19 males and 9 females) who have not used a spoken dialogue system before. We set up dialogue condition #1 and #2 described below. For about five minutes, each user tries the system without given scenario.

condition #1 users are given a manual showing search keys (location, hotel type, room rate, facilities and so on) beforehand.

condition #2 acceptable utterance patterns and the recognition results are displayed to the user through the GUI (Figure 1).

There are 518 and 429 utterances in condition #1 and #2, respectively. Table 2 compares the condition #1 and #2. The ratio of user utterances within the system’s capacity is significantly larger in condition #2. This result confirms that the proposed GUI works very effectively as a guidance of user utterances.

Table 3: Comparison with two conventional methods

	vocabulary size	in-grammar	nearly-in-grammar	out-of-grammar	total
# of correct keywords		561	116	120	797
grammar rules for the whole sentence	942	14.8%	51.4%	175.2%	45.8%
phrase spotting (without 2-gram)	1290	16.8%	34.2%	154.2%	37.9%
phrase spotting (with 2-gram)	6124	**10.8%**	**24.7%**	**140.9%**	**30.3%**

FA: ratio of incorrectly recognized keywords
SErr: ratio of keywords that are not recognized

Table 2: Classification of user utterances

	condition #1	condition #2
utterances within system's capacity	44.6%	76.4%
# of out-of-vocabulary	21.8%	11.4%
# of out-of-grammar	2.1%	1.2%
# of out-of-task	31.5%	11.0%

In summary, the portability is maintained by generating a simple phrase grammar without spoiling the robustness, which is enhanced by the use of GUI. The system is proved to suffice the prototype for data collection.

4.2. Improvement by Combined Language Model

Next, we implement the combined language model and compare it with the conventional methods. We construct the combined language model using 2-gram model trained with the ATR and RWC corpora. The corpora consist of dialogues at travel agents, hotel receptions and car dealers, and the tasks are not identical to the system's one. The text size is about 208 thousands and the vocabulary size is 5432 in total.

Subjects are 24 novice users (19 males and 5 females). As a test set, we use 665 utterances collected using the prototype hotel search system. Gender-dependent triphone model is used as the acoustic model. As an evaluation measure, we use the sum of the ratio of incorrectly recognizing keywords (False Acceptance; FA) and the ratio of slots that are not filled with correct values (Slot Error; SErr). Namely, FA and SErr are defined as the complements of the precision rate and the recall rate, respectively.

$$FA = \frac{\text{\# of incorrectly recognized keywords}}{\text{\# of recognized keywords}}$$

$$SErr = \frac{\text{\# of incorrectly recognized keywords}}{\text{\# of all correct keywords}}$$

The test set samples are classified into three types: in-grammar, nearly-in-grammar and out-of-grammar. Table 3 lists the "FA+SErr[2]" of our proposed method and two conventional methods. One uses grammar rules for the whole sentence and the other adopts key-phrase spotting without 2-gram model.

For in-grammar samples, the proposed method using the combined language model gets the best performance. Use of 2-gram model suppresses the false alarms. Since this 2-gram model is trained with similar domain corpora, the phrase spotting even outperforms the full sentence grammar while maintaining the portability. For both nearly-in-grammar and out-of-grammar samples, the best performance was also achieved by the combined language model. Even for ill-formed utterances, the proposed method realizes robust understanding. In total, the semantic accuracy is improved by 15.5%.

The key-phrase spotting approach is superior to grammar-based approach, especially for ill-formed utterances. And the superiority is enhanced by introducing 2-gram model that does not necessarily match the query domain. Thus, this improvement does not spoil the system's portability.

5. Conclusions

We have presented a portable spoken dialogue platform for information query and its experimental evaluation. The platform can be applied to various domains because it generates domain-dependent lexicon and grammar rules extracted from the domain database automatically. This portability of language model is realized by adopting simple key-phrase spotting strategy. It is also enhanced by incorporating statistics derived from similar domain corpora and the domain database. Moreover, we make use of GUI that displays typical acceptable patterns, which guides users within the system's lexicon and grammar. Overall strategy is demonstrated to work as a reasonable prototype system and realize even robust understanding on ill-formed utterances. The proposed framework does not need collecting domain-specific corpus nor writing grammar rules. Thus, it is a domain-independent platform.

6. References

[1] Stephen Sutton, David G. Novick, Ronald Cole, Pieter Vermeulen, Jacques de Villiers, Johan Schalkwyk, and Mark Fanty, "Building 10,000 spoken dialogue systems," in *Proc. Int'l Conf. on Spoken Language Processing*, 1996.

[2] Stefan Kaspar and Achim Hoffmann, "Semi-automated incremental prototyping of spoken dialog systems," in *Proc. Int'l Conf. on Spoken Language Processing*, 1998.

[3] Rovert C. Moore, "The challenge of domain-independent speech understanding," in *Proc. of ICASSP*, 1998.

[4] T. Kawahara, C.-H. Lee, and B.-H. Juang, "Flexible speech understanding based on combined key-phrase detection and verification," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 6, pp. 558–568, 1998.

[5] A. Lee, T. Kawahara, and S. Doshita, "Large vocabulary continuous speech recognition parser based on A* search using grammar category-pair constraint (*in Japanese*)," *Trans. Information Processing Society of Japan*, vol. 40, no. 4, pp. 1374–1382, 1999.

[2] Substitution errors are counted twice as FA and SErr.