Spoken Dialogue System based on Information Extraction using Similarity of Predicate Argument Structures

Koichiro Yoshino, Shinsuke Mori and Tatsuya Kawahara
School of Informatics, Kyoto University
Sakyo-ku, Kyoto, 606-8501, Japan

Abstract

We present a novel scheme of spoken dialogue systems which uses the up-to-date information on the web. The scheme is based on information extraction which is defined by the predicate-argument (P-A) structure and realized by semantic parsing. Based on the information structure, the dialogue system can perform question answering and also proactive information presentation. Feasibility of this scheme is demonstrated with experiments using a domain of baseball news. In order to automatically select useful domain-dependent P-A templates, statistical measures are introduced, resulting to a completely unsupervised learning of the information structure given a corpus. Similarity measures of P-A structures are also introduced to select relevant information. An experimental evaluation shows that the proposed system can make more relevant responses compared with the conventional "bag-of-words" scheme.

1 Introduction

Recently, a huge amount of information is accumulated and distributed on the web day by day. As a result, many people get information via web rather than the conventional mass media. On the other hand, the amount of information on the web is so huge that we often encounter the difficulty in finding information we want. Keyword search is the most widely-used means for the web information access. However, this style is not necessarily the best for information demands of all users who do not have definite goals or just want to know what would be interesting. To cope with user's vague information demands is an important mission for interactive spoken dialogue systems. Moreover, supporting user's information collection in a small-talk style is one of the new directions of spoken dialogue systems.

Existing spoken dialogue systems can be classified into two types (T.Kawahara, 2009): those using relational databases (RDB) such as the Airline Travel Information System (ATIS) (D.A.Dahl, 1994), and those using information retrieval techniques based on statistical document matching (T.Misu and T.Kawahara, 2010). The first scheme can achieve a well-defined task by using a structural database, but this scheme cannot be applied to the web information in which the structure and task are not well defined. The second scheme has been studied to handle large-scale texts such as web, but most of the conventional systems adopt a "bag-of-words" model, and naive statistical matching often generates irrelevant responses which have nothing to do with the user's requests. Our proposed scheme solves this problem by using information extraction based on semantic parsing from web texts, without constructing an RDB. We adopt the predicate-argument (P-A) structure generated by a parser as a baseline, but every P-A structure is not useful for information extraction and retrieval(Y.Kiyota et al., 2002; M.O.Dzikovska et al., 2003; S.Harabagiu et al., 2005). In fact, the useful information structure is dependent on domains. Conventionally, the templates for information extraction were hand-crafted (R.Grishman, 2003), but this heuristic process is so costly that it cannot be applied to a variety of domains on the web. In this paper, therefore, we pro-

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 59–66,
Portland, Oregon, June 17-18, 2011.

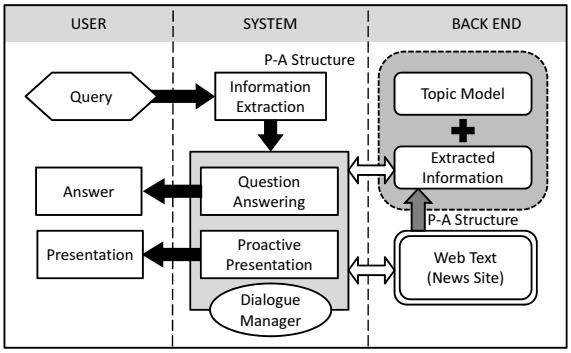

Figure 1: *System overview.*

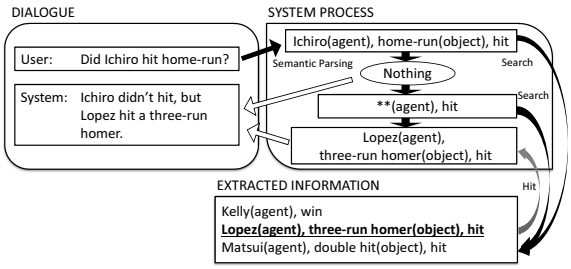

Figure 2: *Example of information extraction and dialogue.*

pose a filtering method of predicate-argument (P-A) patterns generated by the parser, in order to automatically define the domain-dependent useful information structure.

We also address flexible matching based on the P-A structure, because the exact matching often fails and does not generate any outputs. In order to retrieve most relevant information, we define similarity measures of predicates and arguments, which are also learned from a domain corpus.

In this paper, the proposed scheme is applied to a domain of baseball news, and implemented as a spoken dialogue system which can reply to the user's question as well as make proactive information presentation using a news website. An overview of this system is described in Section 2, and the template filtering method is presented in Section 3. Then, system response generation based on flexible matching is explained in Section 4. Finally, an evaluation of the system is presented in Section 5.

2 System Overview

2.1 Architecture

The architecture of the proposed spoken dialogue system is depicted in Fig. 1. First, information extraction is conducted by parsing web texts in advance. A user's query is also parsed to extract the same information structure, and the system matches the extracted information against the web information. According to the matching result, the system either answers the user's question or makes proactive presentation of information which should be most relevant to the user's request.

If the system finds some information which completely matches the user's query, the system makes a response using the corresponding web text. When the system cannot find exact information, it searches for some information which matches partially. For example, in Fig. 2, when a user asked "Did Ichiro hit a home-run?", the system cannot find exact information "[Ichiro (agent), home-run (object), hit]", but finds "[Lopez (agent), three-run homer (object), hit]" which is partially matched and most relevant. This information is used to generate a relevant response that the user would expect.

In the conventional RDB-based dialogue scheme, the system hardly makes relevant responses if it finds no matched entries, thus usually replies "There is no matched entries". In the conventional question-answering scheme, the same situation often happens. Occasionally, a set of close-matched answers may be found by statistical matching, but the found answers may not be relevant to the user's query. In the proposed scheme, we guarantee that the answer is at least partially matched to the user's query in terms of the information structure.

2.2 Information Extraction based on P-A Structure

We use the predicate argument (P-A) structure to define the information structure from web texts. The P-A structure represents a sentence with a predicate, arguments and their semantic cases, as shown in the previous examples. There are some required semantic cases depending on the type of the predicate (verb), and also arbitrary semantic cases like time, place, and other modifications. This structure is a classic concept in natural language processing, but recently, automatic semantic parsing has reached a practical level thanks to corpus-based learning tech-

niques (D.Kawahara and S.Kurohashi, 2006) and has been used for several large-scale tasks (D.Shen and M.Lapata, 2007; R.Wang and Y.Zhang, 2009; D.Wu and P.Fung, 2009). We use KNP[1] as a syntactic and semantic parser.

3 Extraction of Domain-Dependent P-A Templates

The P-A structure automatically generated by the semantic parser provides useful information structure as a baseline. However, every P-A pair is not meaningful in information navigation; actually, only a fraction of the patterns are useful, and they are domain-dependent. 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)]". We propose a method to automatically extract these useful patterns given a domain corpus. We assume each article in the newspaper corpus/websites is annotated with a domain such as sports-baseball and economy-stock.

The method is to filter P-A structure patterns (=templates) based on some statistical measure which accounts for the domain. The filtering process is also expected to eliminate inappropriate patterns caused by parsing errors. Moreover, in spoken dialogue systems, errors in automatic speech recognition (ASR) may result in erroneous matching. By eliminating irrelevant patterns, we expect robust information extraction for spoken input.

Specifically, the following two significance measures are investigated in this work.

3.1 TF-IDF Measure

First, we use the TF-IDF measure to evaluate importance of word w_i in a particular domain or topic t.

$$tfidf(w_i, t) = P(w_i|t) \log \frac{C(d)}{C(d : w_i \in d)} \quad (1)$$

The TF term is the occurrence probability of word w_i, defined as:

$$P(w_i|t) \approx \frac{C(w_i, t) + \alpha}{\sum_j \big(C(w_j, t) + \alpha\big)} \quad (2)$$

[1] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

where $C(w_i, t)$ is the occurrence count of word w_i in the domain t in the corpus, and α is a smoothing factor given by the Dirichlet process prior. The IDF term is the inverse log probability of documents containing word w_i:

$$\frac{C(d)}{C(d : w_i \in d)} \approx \frac{C(d) + \beta}{C(d : w_i \in d) + \beta} \quad (3)$$

where $C(d)$ is the number of documents (=newspaper articles) in the corpus and $C(d : w_i \in d)$ is the number of documents which contain w_i. β is a smoothing factor given by the Dirichlet process prior. We estimate α and β by a likelihood function using the training corpus. We compute the TF-IDF value for a predicate and each argument, and then compute their geometric mean to define the evaluation measure for a P-A template.

3.2 Naive Bayes (NB) Model

The second measure is based on the Naive Bayes model.

$$P(t|w_i) = \frac{C(w_i, t) + D_t\gamma}{C(w_i) + \gamma} \quad (4)$$

Here, γ is a smoothing factor and D_t is a normalization coefficient of the corpus size of the domain t.

$$D_t = \frac{\sum_j C(w_j, t)}{\sum_k C(w_k)}. \quad (5)$$

The evaluation measure for a P-A pattern is obtained by taking a geometric mean of the component words.

3.3 Clustering of Named Entities

The statistical learning often falls in the data sparseness problem, especially for proper nouns, for example, name of persons. Moreover, there may be mismatch in the set of named entities between the training corpus and the test phase. For robust estimation, we introduce classes for named entities (name of persons, organizations, places). Note that unifying all named entities in the corpus before computing the evaluation measure would weaken the significance of these entities. Thus, we compute statistics for every proper noun before clustering, and sum up values for the class afterwards. For example, the score for "[[person](agent), hit]" is computed as a sum over all persons of this pattern.

Table 1: Evaluation of template filtering.

model	feature	Precision	Recall	F
Baseline	-	0.444	1	0.615
TF-IDF	Predicate	0.587	0.840	0.691
	Argument	0.658	0.730	0.692
	P + A	0.513	0.843	0.638
NB	Predicate	0.601	0.879	0.714
	Argument	0.661	0.794	0.722
	P + A	**0.878**	0.726	**0.795**

3.4 Evaluation of Template Filtering

We performed an experimental evaluation to compare the effectiveness of the two significance measures (TF-IDF and Naive Bayes (NB)) in the Japanese professional baseball domain. The models are trained with the Mainichi Newspaper corpus 2008. The clustering of named entities is applied to both methods. The P-A templates having larger significance scores are selected. We determined a threshold for selecting templates using a development set which was held out from the test set by 10%. The test set was made from Mainichi newspaper's website which talks about games played between April 21-23, 2010. Manual annotation was made on typical predicates and semantic cases which can be used for question answering and proactive presentation. The filtering was performed on the test set by matching the patterns defined by each measure, and evaluated against the annotated answers in terms of recall, precision and F-measure (F). Table 1 lists the result for the two measures using predicate-only, argument-only, and both of them.

In this result, using both predicates and arguments in the Naive Bayes (NB) model performs the best. Compared with the baseline without any filtering, the proposed methods significantly improved precision with some degradation of recall. This property is important in realizing informative response generation robust against ASR and parsing errors. Among the selected templates, we can find typical and important patterns like "have a win", "come into pitch", and "make it consecutive wins". Most of recall errors are infrequent patterns, and majority of precision errors are those patterns that are frequently observed but not useful for presentation.

4 Presentation of Relevant Information

When the system fails to find exact information that matches the user's query, or the user does not speak for a while, the system tries to make proactive information presentation. It is based on the partially matched entries of the current or latest query. The fall-back is similar to collaborative response generation in the conventional spoken dialogue systems (D.Sadek, 1999), but it is intended for proactive information presentation using general documents.

4.1 Response generation based on partial matching

For preference among multiple components in the P-A pattern of the user query, we make use of the significance measure defined in Section 3. Specifically, we relax (=ignore) the component of the least significance score, then search for relevant information. If any entry is not still matched, we relax the next less significant component. If multiple entries are found with this matching, we need to select the most relevant entry. Thus, we introduce two scores of relevance. The relevance measure is defined in different manners for predicates (=verbs) and arguments (=nouns). The measure for arguments is defined based on the co-occurrence statistics in the corpus. The measure for predicate is defined based on distributional analysis of arguments.

4.2 Relevance measure of arguments

The relevance of argument words (=nouns) w_i and w_j is defined as

$$sim_{arg}(w_i, w_j) = \frac{\{C(w_i, w_j)\}^2}{C(w_i) \times C(w_j)}. \qquad (6)$$

Here, w_i is in the original query, and relaxed (ignored) in the partial matching, and w_j of the best relevance score is retrieved for response generation. In the example of Fig. 2, w_i is "Ichiro" and w_j is "Lopez".

4.3 Relevance measure of predicates

Distributional analysis (Z.Harris, 1951; Lin, 1998) has been used to define similarity of words, assuming that similar words have similar contexts. In this paper, we use the distribution of arguments which have a modification relation to predicates (Fig. 3)

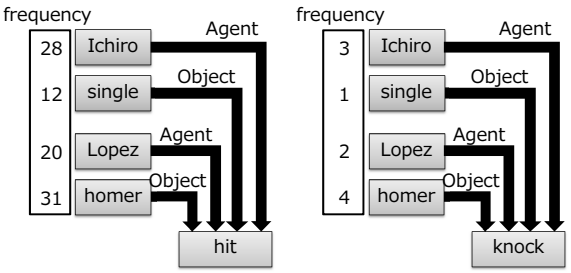

Figure 3: *Distribution analysis of P-A structure.*

(T.Shibata et al., 2008; P.Pantel et al., 2009). The relevance of predicate words w_{pre_i} and w_{pre_j} is defined as a cosine distance of occurrence vectors of the modifying arguments(J.Mitchell and M.Lapata, 2008; S.Thater et al., 2010). Here, argument entries are distinguished by their semantic cases such as Agent and Object, as shown in Fig 3. As the distribution of arguments is sparse and its reliable estimation is difficult, we introduce smoothing by using another distributional analysis of arguments, which is similar to the one in the previous section.

4.4 Bag-of-Words (BOW) Model

If no entry is matched with all possible partial matching, we resort to the naive "bag-of-words" (BOW) model, in which a sentence is represented with a vector of word occurrence and matching is done based on this vector. This method is widely used for document retrieval. We count only content words. In this method, we make use of the significance score for preference of the words when multiple candidates are matched for a short query.

The overall matching strategy of the proposed scheme is summarized in Fig. 4.

4.5 Selection of Relevant Information from Sentence

Answer or information presentation is generated based on the matched sentence in a newspaper article. As a sentence is often complex or made of multiple predicates, simple presentation of the sentence would be redundant or even irrelevant. Therefore, we select the portion of the matched P-A structure, to generate a concise response relevant to the user's query. For example, when a sentence "Ichiro hit a three-run homer in the seventh inning and Mariners won the game" is matched by the pattern

1. **Exact Matching** of P-A templates.
2. **Partial Matching** using significance measure for query relaxation and relevance score for candidate selection.
3. Back-off to **"Bag-of-Words" (BOW) model** with significance measure for disambiguation.

Figure 4: *Strategy for flexible matching in steps.*

"[Ichiro(agent), hit]", we select the former portion of the sentence which exactly answers the user's query, and generate a response "Ichiro hit a three-run homer in the seventh inning."

5 System Evaluation

We have implemented a spoken dialogue system based on the significance measure (Naive Bayes model) and the relevance measures, which were learned using the Mainichi Newspaper corpus of ten years (2000-2009). For evaluation of the system, we prepared 201 questions from news articles (September 19-26, 2010) seen at the website of Mainichi Newspaper[2]. Correct answers to the test queries were annotated manually. Evaluation was done with the text input as well as speech input. A word N-gram language model for ASR dedicated to the domain was trained using the relevant newspaper article corpus. The word error rate was approximately 24%.

The system responses for the test queries are categorized into one of the following four: correct answer only ("Correct"), case which includes the correct answer but also other redundant answers ("Ambiguous"), incorrect answer ("Incorrect"), and ("No Answer"). The ambiguous cases occur when multiple sentences or predicates are matched. We also calculate recall, precision and F-measure by counting individual answers separately even when multiple answers are output. The results based on these evaluation measures are summarized in Table 2 and Table 3 for text input and speech input.

In the tables, the proposed method is broken down into three phases as shown in Fig. 4: exact matching of P-A structure (Section 3), incorporation of the partial matching (Section 4.1), and back-off to the "bag-of-words" (BOW) model (Section 4.4). For comparison, we also tested the BOW model and

[2]http://www.mainichi.jp

Table 2: Evaluation of system response.

Input	Model	Correct	Ambiguous	Incorrect	No Answer
Text	Exact	29.9%	0.5%	1.5%	68.1%
	Exact+Partial	66.2%	5.0%	20.3%	8.5%
	Exact+Partial+BOW	69.7%	5.0%	25.3%	0.0%
	(cf) Bag-of-words (BOW)	46.8%	13.9%	39.3%	0.0%
	(cf) Sequence-of-words (SOW)	54.2%	11.4%	34.3%	0.0%
Speech (ASR)	Exact	19.4%	1.0%	0.5%	79.1%
	Exact+Partial	57.2%	6.0%	18.9%	17.9%
	Exact+Partial+BOW	64.1%	6.5%	28.9%	0.0%
	(cf) Bag-of-words (BOW)	39.8%	9.4%	48.8%	0.0%
	(cf) Sequence-of-words (SOW)	46.3%	10.4%	43.3%	0.0%

Table 3: Accuracy of system response.

Input	Model	Precision	Recall	F
Text	Exact	93.8%	30.3%	45.8%
	Exact+Partial	72.5%	71.1%	71.8%
	Exact+Partial+BOW	70.1%	74.6%	72.3%
	(cf) Bag-of-words (BOW)	49.8%	60.7%	54.7%
	(cf) Sequence-of-words (SOW)	55.2%	65.6%	60.0%
Speech (ASR)	Exact	89.1%	20.4%	33.2%
	Exact+Partial	65.8%	63.2%	64.5%
	Exact+Partial+BOW	61.7%	70.6%	65.9%
	(cf) Bag-of-words (BOW)	42.9%	49.3%	45.9%
	(cf) Sequence-of-words (SOW)	48.3%	56.7%	52.2%

"sequence-of-words" (SOW) model, which consider the sequence order in the BOW model. The exact matching assumes strong constraint of P-A patterns, so the generated answers are almost correct, but no answers are generated very often. By incorporating the partial matching and BOW model, the system can output more relevant answers. Compared with the BOW model, the proposed method achieves much higher ratio or precision of correct answers. F-measure is also higher by 17% absolute.

A similar tendency is observed for speech input, although the overall accuracy is degraded because of the ASR errors. However, degradation is relatively small considering the word accuracy of 76%. The partial matching works effectively even if the exact matching fails due to ASR errors. Moreover, the back-off to the BOW model is effective in ASR input.

The proposed method generates concise responses by selecting the relevant portion as described in Section 4.5, while the BOW method often generates long responses which includes many redundant portions. This property is particularly important in the speech interface.

We show a dialogue example in Fig. 5 which is in Japanese and translated to English for reference (=Italic).

6 Domain Portability

In the proposed scheme, we construct a domain-dependent system in a pre-determined domain. It is straightforward to port the system to other domains just by preparing a domain corpus. This system uses a newspaper corpus for training domain-dependent statistical models. Newspaper articles are classified according to domain categories such as sports/baseball and business/finance. Therefore, it is straightforward to select relevant articles for a particular domain. In the proposed scheme, we do not need any further annotations and all statistical models are trained in an unsupervised manner.

User:	Hanshin ha katta? *(Did Hanshin win the game?)*
System:	Hanshin ha nanakai ni Kanemoto, Hiyama no daida kousei de katta yo. *(Hanshin won the game by Kanemoto and Hyama's pinch-hiting in the seventh inning.)*
User:	Noumi ha touban shita? *(Did Noumi come in to pitch?)*
System	Noumi ga nanakai mushitten de kachitousyu ni natta yo. *(Noumi was the winning pitcher, by shutting out for seven innings.)*
User:	Sanshin wo tori mashita ka? *(Did he get strikeout? ("he" is omitted.))*
System:	Noumi ha nishi ichi-nirui deha Abe wo naikaku suraida- de sanshin ni totta yo. *(Noumi struck out Abe on a slider with two out two-on jam.)*

Figure 5: *Dialogue example (original in Japanese, translated to English).*

7 Conclusions

We have presented a new scheme of spoken dialogue systems which can talk about web texts in an interactive manner. The information extraction technique is adopted to conduct question answering as well as proactive information presentation. Filtering based on a statistical significant measure is introduced to automatically select useful templates in a given domain. Relevance measures are also defined for predicate and argument in order to retrieve relevant entries when the exact matching does not succeed. In experimental evaluations, we have demonstrated that the filtering works effectively and the system generates more relevant responses than the conventional method.

Ongoing works include application to other domains to demonstrate generality of the scheme.

References

D.A.Dahl. 1994. Expanding the scope of the ATIS task: The ATIS-3 corpus. In *Proc. ARPA Human Language Technology Workshop*, pages 43–48.

D.Kawahara and S.Kurohashi. 2006. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proc. HLT-NAACL*, pages 176–183.

D.Sadek. 1999. Design consideration on dialogue systems: From theory to technology - the case of Artimis -. In *Proc. ESCA workshop on Interactive Dialogue in Multi-Modal Systems*, pages 173–187.

D.Shen and M.Lapata. 2007. Using semantic roles to improve question answering. In *Proc. EMNLP-CoNLL*, pages 12–21.

D.Wu and P.Fung. 2009. Can semantic role labeling improve SMT? In *Proc. EAMT*, pages 218–225.

J.Mitchell and M.Lapata. 2008. Vector-based models of semantic composition. In *Proc. ACL-HLT*, pages 236–244.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. ACL and COLING*, pages 768–774.

M.O.Dzikovska, M.D.Swift, and J.F.Allen. 2003. Integrating linguistic and domain knowledge for spoken dialogue systems in multiple domains. In *Proc. of IJCAI-03 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

P.Pantel, E.Crestan, A.Borkovsky, A.-M.Popescu, and V.Vayas. 2009. Web-scale distributional similarity and entity set expansion. In *Proc. EMNLP*, pages 938–947.

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

R.Pieraccini, E.Tzoukermann, Z.Gorelov, J-L.Gauvain, E.Levin, C.-H Lee, and J.G.Wilpon. 1992. A speech understanding system based on statistical representation of semantics. In *Proc. IEEE-ICASSP*, volume 1, pages 193–196.

R.Wang and Y.Zhang. 2009. Recognizing textual relatedness with predicate-argument structure. In *Proc. EMNLP*, pages 784–792.

S.Harabagiu, A.Hickl, J.Lehmann, and D.Moldovan. 2005. Experiments with interactive question-answering. In *Proc. ACL*, pages 205–214.

S.Thater, H.Fürstenau, and M.Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proc. ACL*, pages 948–957.

T.Kawahara. 2009. New perspectives on spoken language understanding: Does machine need to fully understand speech? In *Proc. IEEE-ASRU*, pages 46–50.

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

T.Shibata, M.Odani, J.Harashima, T.Oonishi, and S.Kurohashi. 2008. Syngraph: A flexible matching method based on synonymous expression extraction from an ordinary dictionary and a web corpus. In *Proc. IJCNLP*, pages 787–792.

Y.Kiyota, S.Kurohashi, and F.Kido. 2002. ”dialog navigator” : A question answering system based on large text knowledge base. In *Proc. COLING*, pages 460–466.

Z.Harris. 1951. *Structural Linguistics*. University of Chicago Press.