Dialog System for Open-Ended Conversation Using Web Documents

Masahiro Shibata
Faculty of Information Science and Electrical Engineering, Kyushu University, Japan
E-mail: shibata@lang.is.kyushu-u.ac.jp

Tomomi Nishiguchi
Toshiba Corporation Digital Media Network Company, Japan
E-mail: tomomi.nishiguchi@toshiba.co.jp

Yoichi Tomiura
Faculty of Information Science and Electrical Engineering, Kyushu University, Japan
E-mail: tom@is.kyushu-u.ac.jp

Keywords: dialog system, web corpus, surface cohesion, semantic coherence

Received: October 12, 2008

We have developed a new type of open-ended dialog system that generates proper responses to users' utterances using the abundant documents available on the World Wide Web as sources. Existing knowledge-based dialog systems provide meaningful information to users, but they are unsuitable for open-ended input. The system Eliza, while it can handle open-ended input, gives no meaningful information. Our system lies between the above two dialog systems; it can converse on various topics and gives meaningful information related to the user's utterances. The system selects an appropriate sentence as a response from documents gathered through the Web on the basis of surface cohesion and shallow semantic coherence. We developed a trial system to converse about movies and experimentally found that the proposed method generated appropriate responses at a rate of 66%.

Povzetek: Razvit je sistem dialoga z uporabo spletnih strani.

1 Introduction

We have developed a new type of open-ended dialog system that generates proper responses to users' utterances using the abundant documents available on the World Wide Web as sources.

Many practical knowledge-based dialog systems, such as telephone weather forecast [10] and online air travel planning systems [11] , assume that users can form lingual expressions and make their requests clearly enough to be recognized. Under these conditions, such systems can determine a user's intention using methods like pattern-matching, because the user's aim is definite and the possible input is restricted; therefore, the systems can provide correct answers from prepared databases. This type of dialog system works well for specialized tasks, but it is unsuitable for open-ended input.

Chat systems for open-ended conversations also exist. Eliza [9] is a psychotherapeutic counselling system that converts user's utterances into interrogative sentences or makes non-substantive responses, such as “Really?”, “I see”, or “What happened?”, to avoid answering the user clearly. Consequently, Eliza provides no meaningful information to users.

Previous studies have chiefly examined such dialog systems, that is, those that give users correct information in a narrow domain using structured knowledge and those that give users natural (but valueless) responses in an unlimited domain. We have developed a new type of dialog system that lies between these two types [7] . Our system converses naturally in a broad domain and gives users information related to their interests. To generate the system's utterances, an appropriate response to a user's utterance is selected from documents gathered through the Web. The appropriateness of each selected response is estimated by its level of surface cohesion and shallow semantic coherence. Surface cohesion, which depends on centering theory, is the naturalness of the dialog on the surface. Semantic coherence, defined according to the degree of predictability of the response with words in the user's utterance and the informational value of the words in that response, is the relativity of the user's and system's utterances.

Most question answering (QA) systems, such as those described in Refs. [1] and [5] , also extract the best answer in response to a user's natural language question from a large corpus on the basis of its similarity or relativity to the input question. However, QA systems differ from our system on the following points. First, QA systems mostly deal with questions that have correct answers, while our system generates responses to declarative sentences or interrogative sentences that do not necessarily have specific correct answers. Second, most QA systems depend on a named entity extraction method [2] and aim to extract the appropriate words or phrases for the answers. In contrast, our system depends on a sentence extraction method that attempts to make appropriate responses to a user's utterances. Therefore, we must consider the naturalness of the dialog as well as its relativity to user's utterances in our system.

Here we report on our trial system for conversing about movies. Our dialog method does not use any domain-specific knowledge, but rather the domain of the corpus. Although the trial system is limited to movie-related topics, our dialog system can converse on various topics by gathering web pages about those topics.

We now describe the method used for selecting a proper response from candidates and experimentally evaluate our trial system.

2 Example dialog

We assume that user's and system's utterances are made alternately. Below is an actual sample dialog between the system (*S*) and a user (*U*).

S_0: 映画の話をしましょう。(Let's talk about movies.)

U_1: 亡国のイージスを見ました。(I saw “Aegis.”)

S_1: へぇー。 (Ahh.)

U_2: 原作は「終戦のローレライ」で知られる福井晴敏さんの小説だそうですね。(I heard the original story of this movie was written by Harutoshi Fukui, who is famous for the novel “Lorelei, the Witch of the Pacific Ocean.”)

S_2: 日本推理作家協会賞、日本冒険小説協会大賞、大藪春彦賞をトリプル受賞した、58 万部を超えるベストセラー小説です。(It is a bestseller, selling over 580,000 copies, which won the Mystery Writers of Japan, Inc. Award, the Japanese Adventure Fiction Association Award, and the Haruhiko Oyabu Award.)

U_3: うらかぜがミサイルを迎撃するシーンの緊迫感にドキドキしました。(The tense scene where the Urakaze intercepted a missile was exciting.)

S_3: 亡国のイージスは、防衛庁も協力して撮影したようですが、当初は作品の内容から協力するかどうかもめたそうで、あの石破茂の英断で協力することになったようです。(“Aegis” was shot with the cooperation of the Japan Defence Agency. At first, thinking about the story, they were hesitant to cooperate, but Shigeru Ishiba made the final decision to cooperate.)

U_4: ヨンファ役の中井貴一がはまり役でした。(Kiichi Nakai was a natural for the role of Yeonhwa.)

S_4: 揺れのない強い意志の仙石を真田広之は好演してましたし、クールなヨンファを中井貴一が良く演じ、役者たちはそれぞれ見せていた。(The main actors played their roles well. Hiroyuki Sanada was good as Sengoku, who had a firm and strong will, and Kiichi Nakai was also good in the role of the cool Yeonhwa.)

S_0 is the fixed phrase to start the conversation. Our system searches the corpus for the most appropriate sentence for responding to the user's utterance and uses it as the system's utterance. S_2 - S_4 are the system's utterances generated using selected sentences from the corpus. If the system fails to select an appropriate sentence, it generates a simple response, such as S_1, to continue the conversation.

3 Method of proposed system

Our system uses a corpus of sentences gathered from web documents. When a user inputs an utterance, the system searches the corpus for the most appropriate sentence for responding to the user's utterance based on surface cohesion and shallow semantic coherence. Each user's utterance and each candidate sentence in the corpus is only analysed morphologically[1] because we could not prepare enough reliable data, such as case frames, to analyse sentences syntactically in open-ended conversation.

In this section, we describe the construction of the corpus from web documents and explain how surface cohesion and semantic coherence are determined by our system. We then explain how the system selects the most appropriate sentence from the corpus as its response.

3.1 Constructing the utterance candidate corpus

For the source of candidates for the system's utterances, we collected documents about movies to make an utterance candidate corpus for the current trials. We gathered web pages using the web search API developed by Yahoo! Developer Network[2] with the keywords “α 映画(eiga).” α is any expression that is to be the main theme of a dialog. We selected a movie title as α in our trial system. “映画(eiga)” means “movie” in Japanese. This word was used to exclude pages irrelevant to movies from the search. However, the web pages returned by the search engine included documents without α as the main theme. Therefore we then prepared a simple title filter to remove such pages in the following manner. Because the title tag of an html page often expresses the main theme of the page, we kept web pages that included α in the title tag, i.e., “<title>$\cdots\alpha\cdots$ </title>,” and removed those that did not. A set of utterance candidates, $\Gamma(\alpha)$, consists of all the sentences extracted from the web pages that passed through the above filter. Furthermore, we added information about the sentence number and the web page URL to each sentence in $\Gamma(\alpha)$. The utterance candidate corpus is $\bigcup_{\alpha}\Gamma(\alpha)$. In our trial system, this consisted of 2,580,602 sentences extracted from 44,643 documents.

[1] We use ChaSen (http://chasen.naist.jp/hiki/ChaSen) for analysing Japanese morphology.

[2] http://developer.yahoo.co.jp

3.2 Surface cohesion

Our system selects utterance candidates that maintain surface cohesion of the dialog. Centering theory [4] deals with the transition of the central concern of the dialog and has been applied to Japanese [8] . On the basis of that application, we regard the centralness of discourse entities in a Japanese sentence to be ranked as follows:

1. zero pronoun,[3]
2. noun phrase with postposition "は(wa),"
3. noun phrase with postposition "が(ga)," and
4. noun phrase with postposition "を(wo)."

We call these ranks centralness ranks. In Japanese, the case of a noun phrase is determined by the postposition appended to it. "が(ga)" indicates the subject case, "を(wo)" indicates the object case, and "は(wa)" indicates the topic of a sentence. The system requires a case frame dictionary to decide precisely whether an utterance has a zero pronoun or not. Therefore, we apply this simple rule: if a sentence does not have a noun phrase with postposition "は(wa)" or "が(ga)," we assume that a zero pronoun is the subject case. Furthermore, it is difficult to decide the antecedent of a pronoun without a case frame dictionary. We regard the antecedent as the noun phrase (including zero pronouns) that has the highest centralness rank in the previous utterance.

In this paper, we call the noun phrase with the highest centralness rank in a user's utterance (*U*) the focus. The focus (*f*) of a user's utterance *U* is decided on the basis of simplified centering theory:

- if *U* has a noun phrase *NP* with postposition "は(wa)," *f* is *NP*,
- if *U* has a noun phrase *NP* with postposition "が(ga)" and does not have a noun phrase with postposition "は (wa)," *f* is *NP*, and
- if *U* has neither a noun phrase with postposition "は(wa)" nor a noun phrase with postposition "が(ga)," there is a zero pronoun and *f* is its antecedent. Centering theory says this antecedent is the entity that has the highest centralness rank in the system's utterance just before *U*. Therefore, *f* is the focus of the system's previous utterance.

If the system's utterance *S* just after *U* includes a topic, that is the same as the focus of *U*, then the topic transition between *U* and S is natural. Moreover, it is also natural in many cases, where *S* includes a topic that is the same as the main theme of the dialog.

3.3 Semantic coherence

We define the semantic coherence between utterances using content words (nouns, verbs, and adjectives) is defined as:

$$r(w, w') = \log P(w' \mid w) - \log P_D(w'), \qquad (1)$$

where *w* is a content word in the previous *U*, and *w*' is a content word in a candidate for *S*. $P(w' \mid w)$ is the probability that a sentence includes *w′* when its preceding sentence includes *w*. $P_D(w')$ is defined as $df(w')/|D|$, where $df(w')$ is the number of web documents including *w*' in the corpus and *D* is all the web documents in the corpus. $-\log P(w' \mid w)$ refers to the conditional information of *w*' the occurrence of a sentence including just after a sentence including *w*. It represents the predictability from *w* to *w*'. When *w*' can be easily predictable from *w*, it becomes low. On the other hand, $-\log P_D(w')$ represents the information of *w*'. When *w*' only appears in a few documents, it becomes high. Thus, $r(w, w')$ is high if *w*' is easily predictable from *w* and if *w*' only appears in certain documents. In our trial system, $P(w' \mid w)$ is determined by maximum likelihood estimation using the utterance candidate corpus.

$r(w, w')$ is $-\infty$ when $P(w' \mid w)$ is zero. We will define the semantic coherence between utterances in Section 3.4.3 so that the existence of *w*' in a candidate sentence does not affect it when $r(w, w')$ is $-\infty$. $P(w' \mid w)$ is not confident when the frequency of *w* in the corpus is low. We also regard $r(w, w')$ as $-\infty$ when the frequency of *w* is lower than the threshold *θ*. In the trial system, *θ* is set to 5. The selection of *θ* will be investigated in future work.

3.4 Generating system's utterances

Our system generates utterances in the following manner.

3.4.1 Generating system's utterances

We assume that a conversation with our system is relatively short and that the main theme does not change throughout the dialog. Moreover, we assume that the main theme of the dialog has the highest centralness rank in a user's first utterance. Thus, when starting a conversation, our system selects the noun phrase with the highest centralness rank in the user's first utterance[4] and sets the main dialog theme from this noun phrase. The system's utterance candidates are collected by their main themes, as described in Section 3.1. When α is the main dialog theme, there should be more chances to select appropriate sentences for the system's response from $\Gamma(\alpha)$ than from $\Gamma(\alpha')$; $\alpha \neq \alpha'$. Therefore, we restrict the system's utterance candidates to $\Gamma(\alpha)$.

[3] Case elements are often omitted in Japanese. These invisible case elements are called zero pronouns.

[4] We assume that there are no zero pronouns in this sentence.

3.4.2 Selecting sentences including the focus from Γ(α)

Let U be a user's utterance, S be a candidate for the system's utterance just after U, and f be the focus of U. As described in Section 3.2, the topic transition between U and S is natural when S includes the same topic as f, and thus system tries to select such candidates.

We assume that the topic of a sentence is a word or a phrase with a high centralness rank. Therefore, our system selects sentences that have a zero pronoun with an f antecedent, or f with postposition "は(wa)," "が(ga)," or "を(wo)" from $\Gamma(\alpha)$. In actuality, considering the accuracy of an anaphoric analysis, the system regards the following type of sentences as sentences having a zero pronoun with an f antecedent: a series of at the most m (=2 in the trial system) sentences judged to have a zero pronoun in accordance with the method described in Section 3.2 just after a sentence having f with postposition "は(wa)."

3.4.3 Filtering and ranking by semantic coherence

The candidates that successfully pass through the step outlined in Section 3.4.2 have surface cohesion. Our system selects the sentence that has the highest semantic coherence from among these candidates as the system's final candidate.

The semantic coherence between content words is defined in Section 3.3. Let U be a user's utterance, S a candidate for the system's utterance just after U, $CW(U;f)$ a set of all the content words in U except for the focus f, and $CW(S;f)$ a set of all the content words in S except for f. We define the semantic coherence between U and S basically as the sum of the semantic coherences between the content words in U and S. However, not all content words in S have high semantic coherence with content words in U, even when S is an appropriate response for U. Therefore we restrict the sum to, at the most, K highest values.[5] In addition, we have to take into account the possibility that $r(w,w')$ is $-\infty$. We, then, define the semantic coherence, $R(U,S;f)$, between U and S with the focus f as

$$R(U,S;f) = \underset{(w,w')\in CW(U;f)\times CW(S;f)}{FSUM(K)} r(w,w'), \qquad (2)$$

where $FSUM(K)_{x\in X}\, g(x)$ is the sum of, at the most, K highest finite values of $g(x)$ ($x \in X$) when there is one or more finite values of $g(x)$ ($x \in X$) and is $-\infty$ when there are no finite values of $g(x)$ ($x \in X$).[6] In calculating semantic coherence, we use content words but not the focus f because f has already been used in the filter described in Section 3.4.2.

Before ranking by semantic coherence, the system removes candidates that have fewer content words than K. Such candidates tend to have higher values of R when all candidates have negative values of R because $r(w,w')$ can be a negative finite value. However, such candidates tend to be meaningless as responses.

It is also possible that the semantic coherence between U and every candidate that pass through the filter described in Section 3.4.2 is low. In such case, we use the semantic coherence threshold.[7] If no candidates that pass through the filter have a higher semantic coherence than the threshold, our trial system makes the judgment that there is no candidate that has sufficient semantic coherence.

3.4.4 Selecting sentences for system's utterances with the main theme as focus

As mentioned in Section 3.2, the topic transition between a user's utterance (U) and the system's utterance (S) just after U is natural if S includes the same topic as the main theme. In a case where no candidate has a semantic coherence (R) higher than the threshold described in Section 3.4.3, the system tries to generate an utterance including the same topic as the main theme α. That is, our system selects sentences that have a zero pronoun with an α antecedent, or α with postposition "は(wa)," "が(ga)," or "を(wo) from $\Gamma(\alpha)$ and executes the selection described in Section 3.4.3. If no candidate has an R higher than the threshold, our system generates the fixed utterance "へえー(Ahh)" to continue the conversation.

3.4.5 Generating system's utterances

Our system basically outputs a sentence selected by the processes described in Sections 3.4.3 and 3.4.4 without change. However, if the system's utterance candidate was selected on the basis of the focus (f) of the user's utterance and f has no modifier in the candidate sentence, we can remove the noun phrase f in the candidate sentence to make it a zero pronoun sentence because its antecedent can be identified according to centering theory. This removal strengthens surface cohesion.

4 Examination

To evaluate the performance of our dialog system, we investigated the naturalness of the system's utterances given as responses to utterances made by some users who conversed with our trial system.

[5] In accordance with the results of a preliminary experiment, we set K as 3 in our system.

[6] For instance, suppose that $g(1)=5$, $g(2)=-1$, $g(3)=2$, and $g(4)=g(5)=-\infty$. In this situation,
$FSUM(2)_{x\in\{1,2,3,4\}}\, g(x) = g(1)+g(3) = 7$,
$FSUM(2)_{x\in\{3,4,5\}}\, g(x) = g(3) = 2$,
$FSUM(2)_{x\in\{4,5\}}\, g(x) = -\infty$, and $FSUM(2)_{x\in\{\}}\, g(x) = -\infty$.

[7] In accordance with the results of a preliminary experiment, the threshold was set to -2.5.

Table 1 : Results of human evaluation of the naturalness of the system's utterances generated by the proposed method.

User ID	(1)	(2)	(3)	(4)	(5)	(6)	Total (Ratio)
Level-3	14	34	60	42	36	25	211 (29%)
Level-2	48	28	39	59	58	44	276 (38%)
Level-1	64	50	21	20	35	56	246 (34%)

Table 2 : Evaluation results of the naturalness of the system's utterances for each movie. (JP refers to a Japanese movie and KR refers to a Korean movie in α.)

Movie α	Level-3	Level-2	Level-1	Generation failed	|Γ(α)|
Densha Otoko (JP)	17	37	30	2	28,236
Charlie and the Chocolate Factory	25	45	16	3	21,282
Finding Neverland	24	29	30	5	18,274
Howl's Moving Castle (JP)	29	33	26	2	16,874
Aegis (JP)	31	35	13	10	13,072
Chicago	25	29	26	5	8,030
Windstruck (KR)	21	18	16	34	7,514
Bridget Jones's Diary	12	22	35	17	5,635
Giant	6	2	42	39	3,514
Deep Blue	21	26	12	30	2,662

4.1 Experimental methodology

We selected 10 movies, each title (α) having many utterance candidates ($\Gamma(\alpha)$), and asked 6 participants to watch them all.

We then asked them to converse with our dialog system, with the movie titles as the main dialog themes. We call a user's utterance and the system's response to it an utterance pair. Every person had three conversations for each α , with one conversation consisting of five utterance pairs, not including S_0. Generally speaking, when a conversation becomes too long, the main theme may change; therefore, to avoid changes in the main theme, we limited each conversation to five utterance pairs. In this trial of conversations with our system, the 6 participants had 180 conversations, resulting in 900 utterance pairs.

We next asked the participants to grade the naturalness of the system's utterance in each utterance pair into one of three levels:

level-3: system's utterance is natural as a response to the user's utterance,
level-2: system's utterance is acceptable as a response to the user's utterance,
level-1: system's utterance is unnatural as a response to the user's utterance.

We investigated the performance of our dialog system using these human evaluations.

4.2 Experimental results

Of the 900 user's utterances, inquiries requiring a specific correct answer were made 20 times. Getting correct answers to such inquiries, however, is not the aim of our system, because such answer can be obtained using other dialog methods such as knowledge-based systems. Thus we excluded those utterance pairs in which the user's utterance was such an inquiry, and evaluated the performance of the system using the remaining 880 utterance pairs.

The system's utterance in 147 utterance pairs was “へえー(Ahh).” That is, our system failed to generate an appropriate response to 147 user's utterances (17% of the 880 utterance pairs) by the proposed method.

Table 1 shows the results of the human evaluation of the 733 system's utterances that were selected from the utterance candidate corpus and generated as the system's responses. Each row corresponds to the naturalness levels and each column corresponds to each user (1) - (6) and the total utterances at each level. Two hundred and eleven system's utterances (29% of the 733 system's utterances) were natural; 276 system's utterances (38%) were acceptable; and 246 system's utterances (34%) were unnatural. If utterances evaluated as level-3 and level-2 (“natural” and “acceptable”) are regarded as appropriate responses to user's utterances, then our system succeeded in generating appropriate responses 66% of the time.

Table 2 shows the evaluation results for each movie. We can see that the number of failures in selecting system's utterances (i.e. the number of generated “へえー(Ahh)” responses) tends to be small when $|\Gamma(\alpha)|$ (the number of sentences in $\Gamma(\alpha)$) is large, as we expected. In contrast, the relation between $|\Gamma(\alpha)|$ and the number of natural or acceptable responses is not clear. For instance, “Densha Otoko” has a large number (28,236) of candidate sentences in the candidate utterance corpus; however, the system generated 30 unnatural utterances. Conversely, “Aegis” has a relatively small number (13,072) of candidate sentences, although the system generated only 13 unnatural utterances.

By investigating the system's utterances evaluated as level-1, we found that insufficient filtering during construction of the utterance candidate corpus may have caused the generation of the system's unnatural responses. For example, web pages gathered with the keyword “映画 ジャイアンツ” included not only documents about the movie “Giant,” but also a lot of documents about the Japanese professional baseball team, Yomiuri Giants.[8] For some movie titles, the simple title filter we prepared was still insufficient for identifying the main theme of all gathered web documents. We must reconsider the filtering function when constructing the utterance candidate corpus.

5 Discussion: semantic coherence

This paper proposes a dialog strategy in which the system selects a sentence appropriate as the response to a user's utterance from the abundant available documents and generates it as the system's utterance. The semantic coherence described in Sections 3.3 and 3.4.3 is only a tentative definition: how to best define semantic coherence remains a matter for debate. In this section, we consider the reformation of semantic coherence in our study.

As Equation 1, we defined the semantic coherence between content words *w* and *w'* as the sum of the predictability ($\log P(w'|w)$) and the information of word *w'* ($-\log P_D(w')$). As Equation 2, we defined the semantic coherence between a user's utterance *U* and a candidate for the system's utterance *S* as the sum of the semantic coherences between content words in *U* and *S*. However, with regard to the appropriateness for the system's utterance, it is sufficient if at least a part of the content words in *S* can be predicted easily and a part includes a relatively large amount of information. A content word *w'* in *S* does not have to have high predictability and a large amount of information simultaneously. Furthermore, the statistical and information-related theoretic meaning of $R(U,S;f)$, defined as the sum of the *K* highest $r(w,w')$, is not clear. Therefore, we redefine a new semantic coherence ($R'(U,S;f)$) which is satisfied with the following properties.

(a) $R'(U,S;f)$ is high when at least a part of the content words in *S* is associated strongly with *U*.

(b) $R'(U,S;f)$ is high when at least a part of the content words in *S* contains a large amount of information.

We suppose $P(\{w'_1, w'_2, \cdots, w'_k\}|U)$ to be the probability of a sentence occurring which contains whole *w'₁, w'₂*, ... , and *w'ₖ* after *U*. (This is sufficient even if the sentence contains other words.) When $\log P(\{w'_1, w'_2, \cdots, w'_k\}|U)$ is close to zero, the content words w'_1, w'_2, ... , and w'_k are associated strongly with *U*. In contrast, when this value is smaller, w'_1, w'_2, ... , and w'_k become harder to be associated with *U*. We assume that w'_i occurs independently of other w'_j. Based on this assumption, $P(\{w'_1, w'_2, \cdots, w'_k\}|U)$ can be approximated as follows:

$$P(\{w'_1, w'_2, \cdots, w'_k\}|U) \cong \prod_{j=1}^{k} P(w'_j|U), \qquad (3)$$

where $P(w'_j|U)$ is the probability that a sentence containing w'_j occurs after *U*. When $P(w'_j|U)$ is higher, w'_j is easily predictable from *U*.

When $P(w'_j|U)$ is zero, w'_j cannot be predicted from *U*. We can assume that $P(w'_j|U)$ may depend on the combination of words in *U*, but huge quantities of training data are required to calculate the reliable estimations of the probabilities based on this assumption. Therefore, we suppose the following approximation:

$$P(w'_j|U) \cong \max_{w \in U} P(w'_j|w), \qquad (4)$$

where $P(w'_j|w)$ is the probability that a sentence containing w'_j occurs after a sentence containing *w*, which is same as the definition given in Equation 1. This equation can be interpreted to mean that the occurrence of w'_j does not depend on the combination of some words, but rather only word *w* in *U*.

We substitute Equation 4 into Equation 3 and get the following equation:

$$\log P(\{w'_1, w'_2, \cdots, w'_k\}|U) \cong \sum_{j=1}^{k} \log \max_{w \in U} P(w'_j|w).$$

As Assumption (a), all content words in *S* are not always predictive easily from *U* and the number of content words are different among candidate sentences. Therefore, we regard *L* content words as words contributing to the predictability as the same as $R(U,S;f)$. In addition, considering the case that *S* has fewer than *L* content words or the case that *S* has fewer than *L* content words which are predictable from *U* (i.e. $P(w'_j|U) > 0$), we define $R'_1(U,S;f)$, the predictability from *U* to *S*, as follows:

$$R'_1(U,S;f) = \underset{w' \in CW(S;f)}{FSUM(L)} \log \max_{w \in U} P(w'|w_i).$$

Unlike the definition of Equation 2 in Section 3.4.3, we consider the predictability from *U*'s focus *f* to content word *w'* in system's utterance candidate *S*.

The information of w'_j can be evaluated by $-\log P_D(w'_j)$, in the same was as in Equation 1 in Section 3.3. Therefore, considering Assumption (b), we

[8] Both the movie title “Giant” and the baseball team “Giants” have the same spelling, “ジャイアンツ,” in Japanese.

regard the information that S contains as the sum of, at most, M highest $-\log P_D(w'_j)$, as follows:

$$R'_2(U,S;f) = \underset{w' \in CW(S;f)}{FSUM(M)}(-\log P_D(w')).$$

Obviously, future work must consider how to set the most appropriate L and M.

Finally, we can redefine the semantic coherence $R'(U,S;f)$ as the sum of $R'_1(U,S;f)$ and $R'_2(U,S;f)$ as follows:

$$R'(U,S;f) = R'_1(U,S;f) + R'_2(S;f).$$

As mentioned above, some reformations are being considered for semantic coherence, which needs to be more properly defined.

6 Discussion: semantic coherence

In this paper, we selected movies as dialog themes and developed a system for having idle conversations about those movies. However, the use of our system only in idle conversations does not fully show the system's usefulness. For example, one useful area of application of our method is the following. Suppose that a user is interested in something and wants to obtain some information about it; a situation where a user wants to consult on a matter but only has a vague idea about the topic and cannot think of proper keywords for successful searching on the Web. In these situations, our system may be able to provide some clues for searching. In the course of having a conversation with our system, the user has the chance to learn some useful terms or keywords from the system's utterances. Once the user gets these keywords, they can pursue their interest in the theme using information retrieval techniques such as keyword search or the QA method and obtain more detailed information. We may not be able to use our system for information retrieval alone, but it has the potential to attain more flexible dialog for information provision through combined use with other information retrieval techniques.

Other than movies, there are various other dialog domains around which we can develop our system in a similar manner, such as books, food, and baseball. For such domains, we can construct an utterance candidate corpus by the ad-hoc method described in this paper of gathering web documents, and we can relatively effortlessly select the most likely main themes. In contrast, there are also dialog domains for which it is difficult to determine the most appropriate main themes to use to construct the corpus. As for the step of searching for appropriate sentences from the corpus, there are also dialogs in which candidate documents may not be sufficiently narrowed down in terms of only the main theme, such as was shown above by the dialog about “Giant.” Therefore, instead of dividing of corpus sentences into $\Gamma(\alpha)$ by the main theme α, we would estimate the central topics of each document previously gathered by web crawlers and insert them into each document as document keywords. This can be accomplished by existing techniques of automatic keyword extraction [6] [3] . In addition, we can put several keywords into a document in place of a main theme. If a document has several keywords, the content of the document can then represented by the combination of these keywords, which leads to a solution of synonym problem like that encountered with the word “Giant.”

Further, to improve the system generality, we believe that we should abolish the main theme from our dialog strategy. Even in dialogs between humans, the main dialog theme is not often set before starting the conversation. The themes of dialogs are fixed, and also changed, over the course of conversation. The central topics of utterances (focuses) seem to decide the topic of the dialog. Therefore, we should preserve the focuses of some previous utterances and make set F of these focuses and the focus of the current user's utterance. Then, in the selection step of the system's utterance, we match F and keyword set K of each document and narrow down the sentence selection area to those documents matched with F. This will enable a narrowing down of the search area without setting the dialog main theme in advance.

As for matching F and K, we can simply check, for example, whether F and K have more than r common components or not. Of course, we can also suppose other ways of matching. Automatic keyword extraction generally uses a value representing the validity of the keyword (e.g., TF-IDF value). Therefore, each word in K seems to have its evaluated value as its weight. On the other hand, the words in F can be assigned weight depending on either their centralness or weight, reflected in the fact that the topics of utterances are gradually forgotten over time.

7 Conclusion

We explained a method of generating a natural response to a user's utterance in an open-ended conversation by retrieving an appropriate sentence from documents on the Web. Furthermore, we investigated the performance of our trial system using this method by having it actually converse with people. Our system could generate a natural response to a user's utterance 66% of the time. Finally, we discussed the redefinition of semantic coherence and instruction of document keywords as an extension of our method to better apply our system to other dialog domains.

Our trial system is only capable of making a idle conversation about movies. However, our approach of selecting the proper system's utterance from the corpus has potential to be usefulness in a number of engineering applications. For example, by combining other search techniques, our method could be used in an information retrieval system that converses naturally instead of functioning as a conversational Web search engine. Moreover, if the utterance candidate corpus includes sentences extracted from blog pages about a certain product, we can expect that a person considering whether to purchase the product or not will be able to have a useful conversation with our dialog system.

To realize these applications, we must consider some extensions of our method, such as those discussed in Sections 5 and 6. Furthermore, future work must investigate the usefulness of our method with respect to the practical use of information provision.

References

[1] T. Akiba, K.Itou, and A. Fujii (2004). Question answering using "common sense" and utility maximization principle. *In Working Notes of 4th NTCIR Workshop*, pp. 297-303.

[2] A. Bhole. (2007). Extracting named entities and relating them over time based on Wikipedia. *Informatica*, vol. 30, no. 4, pp. 463-468.

[3] J. Dobša and B. D. Bašić (2007). Approximate representation of textual documents in the concept space . *Informatica*, vol. 31, no. 1, pp. 21-22.

[4] B. Grosz, A. Joshi, and S. Weinstein (1995). Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, vol. 21, no. 2, pp. 203-226.

[5] S. Lee and G. G. Lee (2003). A question answering system for Japanese. *In Proceedings of the 3rd NTCIR Workshop*, pp. 31-38.

[6] S. Sato, H. Hayashi, N. Maki, and M. Inoguchi (2007). Development of automatic accumulated newspaper articles on disasters. *In Proceedings of 2nd International Conference on Urban Disaster Reduction*, (CD-ROM).

[7] M. Shibata, Y. Tomiura, H. Matsumoto, T. Nishiguchi, K. Yukino, and A. Hino (2006). Dialog system for new idea generation support. *In Proceedings of Computer Processing of Oriental Languages, 21st International Conference, ICCPOL2006*, pp. 490-497.

[8] M. Walker, M. Iida, and S. Cote (1994). Japanese discourse and the process of centering. *Computational Linguistics*, vol. 20, no. 2, pp. 193-233.

[9] J. Weizenbaum (1966). Eliza – a computer program for the study of natural language communication between man and machine. *Communication of the ACM*, vol. 9, no. 1, pp. 36-45.

[10] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hanzen, and L. Hetherington (2000). Jupiter: A telephone-based conversational interface for weather information. *IEEE Trans. SAP*, vol. 8, no. 1, pp. 100-112.

[11] V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goddeau, J. Glass, and E. Brill (1994). Pegasus: A Spoken language interface for online air travel planning. *Speech Communication*, vol. 15, pp. 331-340.