Proceedings of the IASTED International Conference
Artificial Intelligence and Soft Computing (ASC 2011)
June 22 - 24, 2011 Crete, Greece

EVALUATION OF UTTERANCES BASED ON CAUSAL KNOWLEDGE RETRIEVED FROM BLOGS

Motoyasu Fujita, Rafal Rzepka and Kenji Araki
Graduate School of Information Science and Technology
Hokkaido University
Kita-ku, Kita 14, Nishi 4
Sapporo, Hokkaido, Japan
email: {fuji_riv, kabura, araki}@media.eng.hokudai.ac.jp

ABSTRACT
In this paper, we describe the effectiveness of utterance generation using causal knowledge for a dialogue system. Recently, there has been a variety of research on non-task-oriented dialogue systems; however, an effective approach has not yet been developed. One of the most important reasons for this is that non-task-oriented dialogue systems lack common sense knowledge, which is not in their databases. As the first step towards solving this problem, we concentrated on causal knowledge containing reasons and effects, which can provide unwritten meanings for utterance understanding and generating modules. In this paper we investigated how an utterance generated with knowledge related to user input can improve an existing conversational system. Experiment results show that utterance generation using causal knowledge can improve a conversational system.

KEY WORDS
Natural Language Processing, Causal Knowledge, Utterance Generation

1. Introduction

Research on non-task-oriented dialogue systems, often called chatbots, is not very common because it is difficult for such systems to predict the users intentions using prior knowledge. To resolve this problem, we started with the automatic addition of causal knowledge. Because such knowledge includes cause and effect relationships, we presumed that a conversational system should use this to guess the relationship between the users input and the system's world knowledge that might be used for an elaborative response, which is proven to be better than a simple one [4]. The following utterances are an example of a dialogue using cause-effect knowledge.

- **User:** It is warm today and the weather is good.
- **System:** You can air your sheets on the balcony.

In the above example, the system responded to the relationship of warm and weather is good. Such a response is difficult for previous systems if such rule is not described. In addition, it is difficult to describe all possible events beforehand. However, systems can more easily respond to such inputs by using causal knowledge. We presumed that causal knowledge helps to understand the user's intentions. In this paper, we confirm the validity of utterance generation using causal knowledge from blogs. First, we will show how we extracted causal knowledge from a blog corpus to create a large database of causes and effects. Second, we will introduce methods of how the system generates utterances using the corpus. Finally, we will compare our system with another state-of-the-art dialogue system.

2. Related Work

An example of a non-task-oriented conversational system described in natural language processing literature is Modalin, developed by Higuchi et al[1]. It is a free-topic keyword-based conversational system for Japanese that automatically extracts sets of words related to a conversation topic from Web resources, which was proved to outperform classic ELIZA-like [2] dialogue systems and be easy to combine with other algorithms [3]. After the search engine results extraction process, Modalin generates an utterance, adds modality, and verifies the semantic reliability of the generated phrase. Over 80% of the extracted word associations were evaluated as being correct, and adding modality improved the system significantly. However, in the case of a system that uses templates to generate an utterance, the manual preparation of templates is laborious and causes problems as noun and verb associations are filled in randomly. Several studies have described extracting causes and effects. Inui et al [5] proposed an algorithm for automatic acquisition of causal knowledge from document collection using a Japanese connective word, *tame* ("because"; hereafter, *italics* indicate words in Japanese). By using machine-learning techniques, they achieved 80% recall with over 95% precision for the causes, precondition, and means. For the effects, 30% recall with 90% precision was obtained. However, they admit that their instances are difficult to use in reasoning. Because newspapers are used as a source, the topic range of extracted knowledge is narrow and lacks commonsensical entries. Sakaji et al. [6] have extracted causal knowledge using 36 clue phrases and

DOI: 10.2316/P.2011.716-043

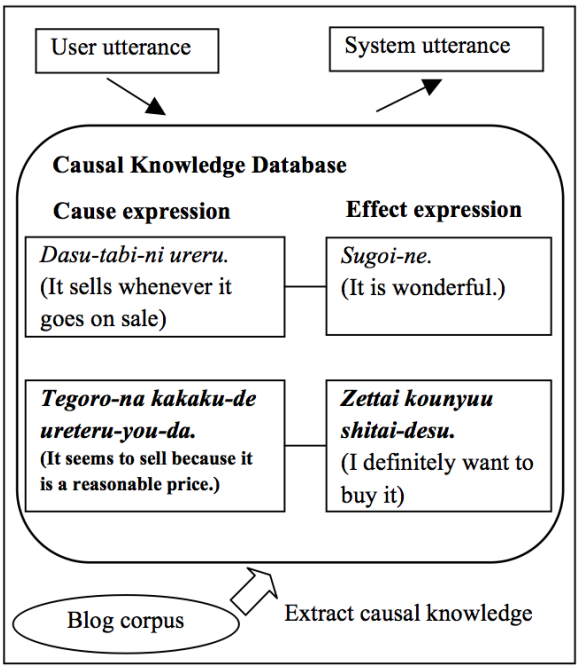

Figure 1. Outline of system

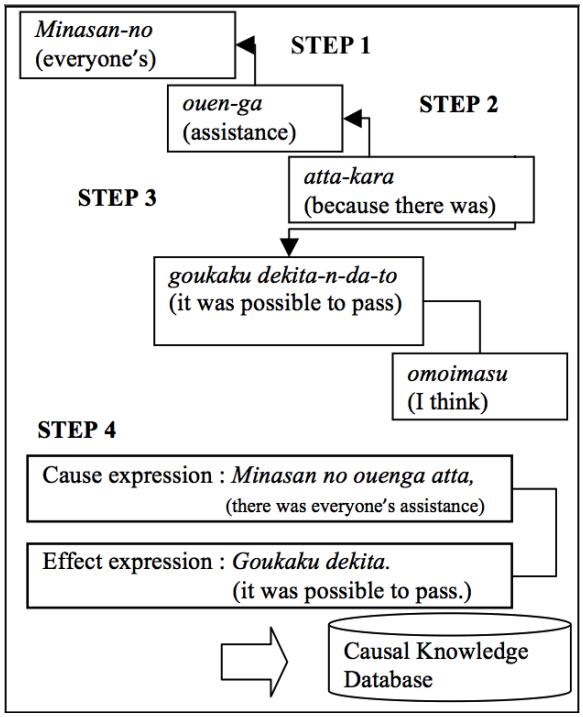

Figure 2. Causal knowledge extraction process

syntactic patterns from Japanese newspaper articles concerning economic trends. They achieved 92% recall with 76% precision for causes, and 81% recall with 53% precision for effects. Their study utilizes various clue phrases and is easy to apply in a reasoning system, but the clue phrases are difficult to use in a dialogue system and the problem of narrow topic coverage remains.

3. Outline of System

Figure 1 shows an outline of our system, which generates utterances using causal knowledge. The system generates utterances by extracting information from a Causal Knowledge Database that we created. In the following sections, we will first explain our extraction of causal knowledge from a blog corpus [7]. Second, we will explain our utterance generating method. Third, we will explain how we used the extracted causal knowledge in our preliminary experiment. Finally, we will describe our evaluation experiment and results.

4. Extracting Causal Knowledge

To examine the validity of causal knowledge used in dialogue, we used a blog corpus to create a database of causes and effects.

4.1 Extraction Process

In this subsection, we explain our method of causal knowledge extraction using blogs. For this purpose we used a blog corpus created from the Ameba service for Japanese bloggers [8]. The corpus is written in colloquial Japanese and contains about 350 million sentences. We presumed that utterance generation using causal knowledge extracted from blog entries will improve the naturalness of the grammar more than utterance generation using the inflexible templates of Modalin. We confirmed that our hypothesis was correct; this will be described in detail later. We acquired causal knowledge based on dependency relations in a similar manner to Sakaji et al[6]. We used the same dependency analyzer, CaboCha[9].

We used three clue words to extract causal knowledge: *kara*, *tame* and *node* (grammatically, they indicate conditional function). These words are confirmed to be suitable for the extraction of causal knowledge[5]. The extraction process is shown in Figure 2, and is performed in the following way.

- **Step 1:** Search for sentences which include clue words in the blog corpus.
- **Step 2:** Extract cause expressions from sentences which include a clue word.
- **Step 3:** Extract effect expression from sentences which include a clue word.
- **Step 4:** Make an Cause-Effect entry for the Causal Knowledge Database.

The cause expression and effect expression are paired and stored in the Causal Knowledge Database.

In the next section we will explain utterance generation using the database created in Step 4.

5. Utterance Generation

5.1 Outline of Utterance Generation

In this section, we will explain the utterance generation process using the example sentence "it seems to sell because it was a reasonable price", and Figure 3 to illustrate.

(a) **Extracting important words from a user utterance:** The first step for generating utterances is to extract important words from the user's input. Important words are independent words: nouns, verbs and adjectives, but adverbial nouns are excluded. We assume that these words include important semantic information for extracting causal knowledge. In the above example (Figure 3) they are "price", "sell" and "reasonable".

(b) **Extracting important dependency relations from a user utterance:** Important dependency relations are a combination of previously extracted nouns and important words in a dependency relation to this noun. In the above example they are "reasonable - price" and "sell - price".

(c) **Using Causal Knowledge Database:** The next step is to search the Causal Knowledge Database for important dependency relations extracted from user's utterance. In this example, a sentence from the database "It seems to sell because it is a reasonable price" contain the same important dependency relations as in the user's input.

(d) **Collecting candidates:** The sentence obtained in c) is followed by another sentence which is included in the causal relationship, and this is saved as an utterance candidate. For instance, "I definitely want to buy it."

5.2 Preliminary Experiment

We defined the method described in subsection 5.1 as a prototype system, and compared it with Modalin in order to examine its performance. Modalin can perform conversations with users in a non-topic-constrained manner, i.e. the topic can be set freely by the user. It generates responses towards user's utterances in the following way:

(a) Extracting keywords from user utterance

(b) Extracting word associations from the Web

(c) Generating sentence proposition using word associations

(d) Adding modality to the sentence proposition

Table 1. The result of preliminary experiment

	A	B	C	D
Prototype system	3.16	3.13	2.32	2.08
Modalin	1.77	1.48	1.89	1.98

We performed experiments using both systems in order to examine the naturalness and effectiveness of the utterances using the Causal Knowledge Database. Before the experiment, we prepared 20 sentences, which were inputted during a conversational experiment. Modalin produced 20 final sentences and the prototype system produced 15 randomly chosen candidates for each input as utterances. Three human users evaluated both systems utterances, which were mixed for a fair experiment. The purpose was to check whether there were any candidates that score lower than Modalin's output. The users were asked the following questions in order to evaluate:

(A) Was the utterance grammatically natural?

(B) Was the utterance semantically natural?

(C) Was the vocabulary rich?

(D) Did you get an impression that the system followed the user's intentions?

The answers for these questions were given on a 5-point semantic differential Likert scale. The results are shown in Table 1. In the preliminary experiment, the prototype system scored highest in all questions. However, the algorithm was not capable of choosing one utterance candidate, therefore it was not yet possible to use it as a dialogue system.

5.3 Calculation of Similarity

In order to select the best candidate, we decided to calculate the similarity between the user's utterance and the causal part of cause-effect pairs stored in the database. Because standard approaches as [11] are difficult to implement for Japanese, we used the technique of Bag of Words, referring to Shimohata [10]. The calculation of similarity is shown in (1).

- **Ai:** Agreement between important words of the user's utterance and candidate sentence
- **Iu:** Number of important words within the user's utterance
- **Ic:** Number of important dependencies within the candidate sentence
- **Ad:** Agreement between the dependency relations of the user's utterance and candidate sentence
- **Du:** Number of dependency relations within the user's utterance

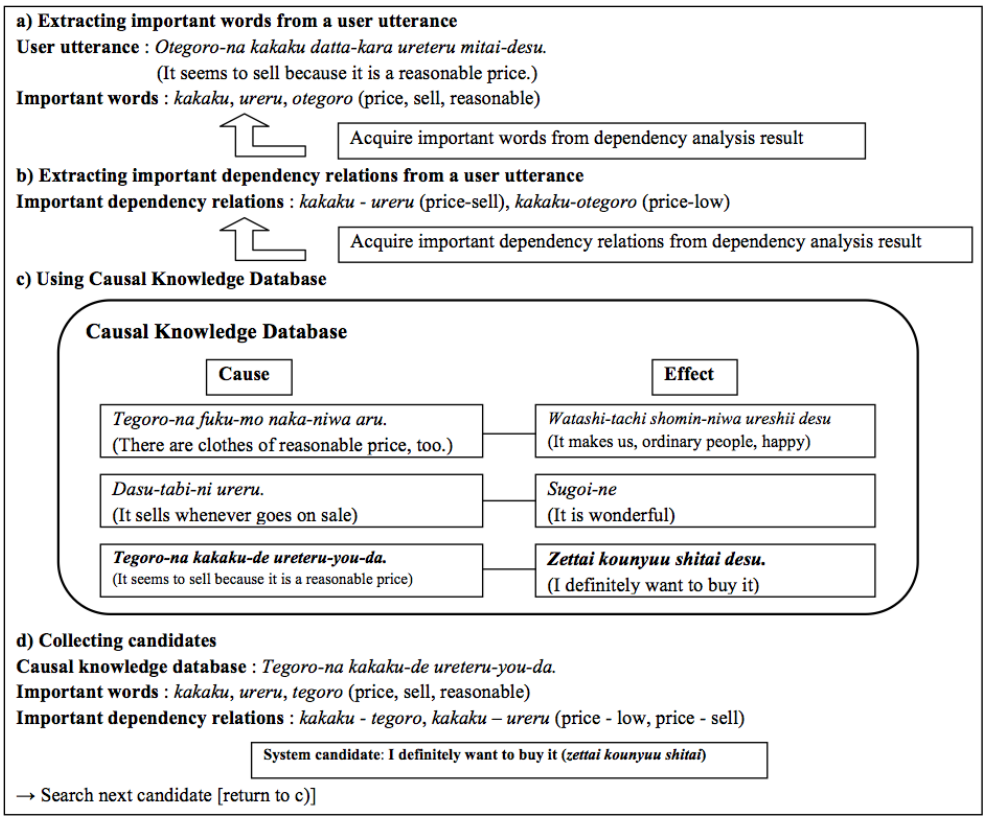

Figure 3. Utterance generation process explained with an example

$$\text{(1)} \qquad \text{Sim} = \log\left(\frac{\left(\frac{Ai}{Iu} + \frac{Ai}{Ic}\right)}{Wi} + \frac{\frac{Ad}{Du} + \frac{Ad}{Dc}}{Wr}\right)$$

- **Dc:** Number of dependency relations within the candidate sentence
- **Wi:** Weight of important words ($|Iu - Ai| > 3 \rightarrow 3$), ($|Iu - Ai| \leq 2 \rightarrow 2$)
- **Wr:** Weight of dependency relations ($|Du - Ad| > 3 \rightarrow 2$) ($|Du - Ad| \leq 2 \rightarrow 1$)

The weights were set experimentally. Using the above calculation method, the system was able to select one sentence from the list of candidates and output it as an utterance. This allowed us to create a dialogue system, which we named ”Causalin”.

6. Evaluation Experiment and Result

In this section, we describe an evaluation experiment we performed in order to confirm the performance of Causalin.

Table 2. Results of evaluation experiment

System	A	B	C	D	E	F
Causalin(ran)	3.24	3.10	3.26	2.86	2.74	2.12
Causalin(sim)	3.77	3.70	3.33	3.19	3.24	2.54
Modalin	3.09	2.95	2.79	2.54	2.46	2.24

In this experiment, we used three utterance generation systems and 50 new utterance sets.

6.1 Detail of Experiment

The three systems used for evaluation were the prototype system, Causalin and Modalin. However, the prototype system generates several utterance candidates, so there was a need to select one at random. The prototype system was named Causalin(ran), and Causalin using similarity was named Causalin(sim). We performed an utterance generation experiment with these three systems and 50 new utterances. Eight participants took part in experiment. Six of them were PhD students and two were company employees. The previous question set was extended to include the original set of questions used by Higuchi et al.[1]:

(A) Was the utterance grammatically natural?

Table 3. Significance differences between Causalin(ran) and Modalin

	Question	A	B	C	D	E	F
Causalin(ran) and Modalin	P value	0.2254	0.4576	>0.0001	0.0006	0.0010	0.4281
	significant on 5% level?	No	No	Yes	Yes	Yes	No
	significant on 1% level?	No	No	Yes	Yes	Yes	No

Table 4. Significance differences between Causalin(sim) and Modalin

	Question	A	B	C	D	E	F
Causalin(sim) and Modalin	P value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0048
	significant on 5% level?	Yes	Yes	Yes	Yes	Yes	Yes
	significant on 1% level?	Yes	Yes	Yes	Yes	Yes	Yes

(B) Was the utterance semantically natural?

(C) Was the vocabulary rich?

(D) Did you get an impression that the system possesses any knowledge?

(E) Did you get an impression that the system was human-like?

(F) Did you get an impression that the system followed the user's intentions?

The answers to these questions were given on a 5-point semantic differential scale. After completing the above questionnaire, evaluators answered a final question, Which system do you find most interesting?. The results are shown in Table 2, Table 5 and Figure 4. Further, Table 3 shows significant differences between Causalin(ran) and Modalin, and Table 4 between Causalin(sim) and Modalin. The evaluation is explained in detail in the next section.

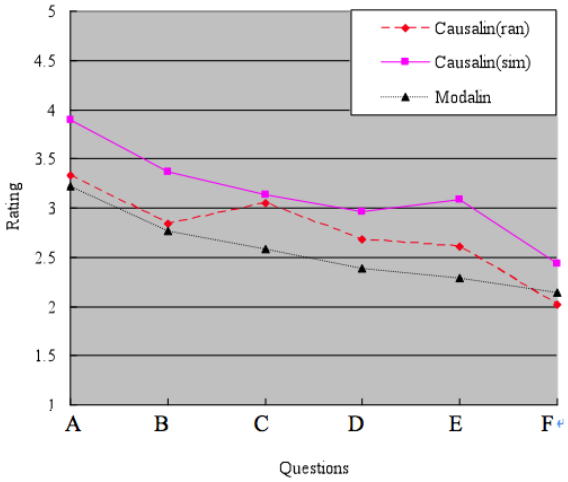

Figure 4. Rating comparison of all three systems

6.2 Experiment Results

In this section, we describe the results of the evaluation experiment. For Question A, Causalin(sim) received an average score of 3.77, while Modalin received 3.09. The statistical difference was 0.68 points. For Question B, Causalin(sim) obtained an average score of 3.70 against Modalin with 2.95. The difference was 0.75 points. For Question C, Causalin(sim) obtained an average score of 3.33, and Modalin 2.79. The difference was 0.54 points. For Question D, Causalin(sim) received an average score of 3.19 and Modalin 2.54, with a difference of 0.65 points. For Question E, Causalin(sim) outperformed Modalin with 3.24 against 2.46. The difference was 0.78 points. For Question F, Causalin(sim) scored 2.54, and Modalin 2.24. The difference was 0.30 points. For all questions, Causalin(sim) acquired the highest average score, and statistical significance was confirmed on a 1% level. Causalin(ran) and Modalin appeared not to be significant on a 5% level for Questions A, B and F. These results do not prove the effectiveness of utterance generation using a blog corpus, however no user selected Modalin in the final question for overall evaluation. Therefore, where the sentence generation method is concerned, we have proved that Causalin's technique of using blog-extracted sentences improves the user's impression in comparison to the template method of Modalin. The above results clearly showed that utterance generation using causal knowledge can expand a dialogue system and improve the impression of rich vocabulary and knowledge. In addition, such a system was proven to be more human-like than one that does not use any kind of causal reasoning. We also confirmed that using blogs for utterance generation can resolve problems with sentences that are unnatural both syntactically and semantically.

Table 5. The result of final question: Which system do you find most interesting?

System	Final Question (people)
Causalin(ran)	2
Causalin(sim)	6
Modalin	0

7. Conclusion and Future Work

In this research, we investigated the effectiveness of using causal knowledge for generating utterances. First, we automatically collected causal knowledge from a vast blog corpus. Second, we designed an utterance generation process using causes and effects. Third, we confirmed the performance of a dialogue system that utilizes causal knowledge namely, that the system gives the user an impression that it can reason about why things happen and what may happen next. The results showed that our approach also improved the user's impression of the system's vocabulary and knowledge. For the next step, we are planning to normalize the causal knowledge database in order to be utilized by other systems. After cleaning up the unnatural entries, we are going to create an algorithm for transforming causes and effects into forms that can be used by the Japanese version [13] of ConceptNet [14] or a causal relations network proposed by Sato [15]. We also plan to use the database for learning linguisitc patterns not only for explicit but also for implicit causal relations as proposed by Girju [12].

References

[1] S. Higuchi, R. Rzepka, K. Araki, *A Casual Conversation System Using Modality and Word Associations Retrieved from the Web*, in Proceedings of The 2008 Conference on Empirical Methods on Natural Language Processing (EMNLP08), Honolulu, USA, 2008, pp. 382-390.

[2] J. Weizenbaum, *ELIZA – A computer program for the study of natural language communication between man and machine*, Commun. ACM, vol.9, no.1, pp.36-45, 1966.

[3] R. Rzepka, S. Higuchi, M. Ptaszynski, P. Dybala and K. Araki, *When Your Users Are Not Serious – Using Web-based Associations, Affect and Humor for Generating Appropriate Utterances for Inappropriate Input*, Transactions of the Japanese Society for AI, 25(1), pp.114-121, 2010.

[4] R. Tokuhisa and R. Terashima, *An Analysis Of 'Distinctive' Utterances In Non-task-oriented Conversational Dialogue*, Transactions of the Japanese Society for Artificial Intelligence, 22(4), 2007, pp. 425-435.

[5] T. Inui, K. Inui, and Y. Matsumoto, *Acquiring Causal Knowledge from Text Using the Connective Marker tame*, Journal of Information Processing Society of Japan, 45(3), 2004, pp. 919-933.

[6] H. Sakaji, S. Sekine, and S. Masuyama, *Extracting Causal Knowledge Using Clue Phrases and Syntactic Patterns*, PAKM 2008. LNCS (LNAI), vol.5345, 2008.

[7] J. Maciejewski, M. Ptaszynski and P. Dybala, *Developing a Large Scale Corpus for Natural Language Processing and Emotion Research in Japanese*, International Workshop on Modern Science and Technology 2010 (IWMST 2010), September 4-5, 2010, Kitami Institute of Technology, Kitami, Japan.

[8] Ameba Blog, www.ameblo.jp

[9] T. Kudo and Y. Matsumoto, *Japanese Dependency Analysis Using Cascaded Chunking*, Journal of Information Processing Society of Japan, 43(6), 2002, pp. 1834-1842.

[10] M. Shimohata, E. Sumita and Y. Matsumoto, *A Method for Retrieving a Similar Sentence and Its Application to Speech Translation*, Journal of Natural Language Processing 11(4), 2004, pp. 105-126.

[11] V. Hatzivassiloglou, Judith L. Klavans, E. Eskin, *Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning*, Proceedings of the 23rd annual international ACM SIGIR conference on Research and Development in Information Retrieval, Athens, Greece, July 24-28, 2000, pp. 224-231

[12] Girju, R, *Automatic detection of causal relations for question answering*, Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, 2003, vol. 12, pp. 76-83

[13] T. Roberts, R. Rzepka and K. Araki, *A Japanese Natural Language Toolset Implementation for ConceptNet*, Proceedings of Commonsense Knowledge in the AAAI 2010 Fall Symposium (Technical Report FS-10-02), pp. 88-89, Arlington, USA, November, 2010.

[14] C. Havasi, R. Speer and J. Alonso, *ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge*. In Proceedings of Recent Advances in Natural Languages Processing, 2007, pp. 277-293.

[15] Sato, T., Horita, *Assessing the plausibility of inference based on automated construction of causal networks using web-mining*, Sociotechnica, 2006, pp. 66.74