Extracting Causal Knowledge Using Clue Phrases and Syntactic Patterns

Hiroki Sakaji[1], Satoshi Sekine[2], and Shigeru Masuyama[1]

[1] Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi 441-8580, Japan
`sakaji@smlab.tutkie.tut.ac.jp, masuyama@tutkie.tut.ac.jp`
[2] NewYork University, 715 Broadway, 7th floor, New York, NY 10003 USA
`sekine@cs.nyu.edu`

Abstract. This paper proposes a method to extract causal knowledge (cause and effect relations) using clue phrases and syntactic patterns from Japanese newspaper articles concerning economic trends. For example, a sentence fragment "World economy recession due to the subprime loan crisis ..." contains causal knowledge in which "World economy recession" is an effect phrase and "the subprime loan crisis" is its cause phrase. These relations are found by clue phrases, such as "ため (*tame*: because)" and "により (*niyori*: due to)". We, first, investigated newspaper corpus by annotating causal knowledge and clue phrases. We found that some specific syntactic patterns are useful to improve accuracy to extract causal knowledge. Finally, we developed our system using the clue phrases and the syntactic patterns and showed the evaluation results on a large corpus.

1 Introduction

A large amount of machine-readable textual documents including Web pages and newspaper articles are now available. We can find a lot of valuable information for many real applications in the documents by text mining. One of such information is "causal knowledge". We expect causal knowledge in the economic domain would be useful to forecast economic trends and prevent loss of business opportunities. For example, if we get causal knowledge where *cause* is "the Year 2000 problem" and *effect* is "the decline of the sales of hotels", we infer that problems like "the Year 2000 problem" may cause the decline of the sales of hotels. (Note: Hotels heavily rely on IT technologies. However, hotel managers are very busy handling daily business, and the hotel industry was not ready for such a problem until the last minutes.) Nevertheless, extracting such knowledge costs prohibitively high and time-consuming. Thus, there are some attempts to extract causal knowledge from textual documents automatically [1–4].

We propose a method to extract causal knowledge from Japanese newspaper articles concerning economic trends using syntactic patterns. For example, "World economy recession due to the subprime loan crisis ...", here, "World economy recession" is an effect phrase, "the subprime loan crisis" is its cause phrase and “due to” is a clue phrase, respectively. These cause and effect relations are explicitly expressed by the clue phrase “により (*niyori*: due to)”. However, if we just use the clue phrase to extract causal knowledge, it extracts a lot of noises. By investigating the corpus, we found that there are some syntactic patterns in cause-effect phrases and these patterns are useful to improve the accuracy of the results. Furthermore, causal knowledge is sometimes expressed in more than one sentence. Our method can also handle causal knowledge stridden over two consecutive sentences. Finally, we developed our system to extract causal knowledge using the clue phrases and the syntactic patterns and evaluated on a large corpus.

2 Related Work

A lot of work has been done on causal information extraction from a large corpus. Inui et al. proposed a method for acquiring causal relations (*cause*, *effect*, *precond* and *means*) from a complex sentence containing a Japanese resultative connective “ため (*tame*: because)” [1]. The Japanese resultative connective “ため (*tame*: because)” is a strong clue for causal information. In their research, clue phrases other than “ため (*tame*: because)” are not used, therefore, their method can not extract causal relations expressed by other clue phrases. In contrast, our method can extract many causal relations expressed by 36 clue phrases.

Khoo et al. proposed a method for extracting cause-effect information from newspaper articles by applying patterns made manually[2]. Furthermore, they proposed a method for extracting causal knowledge from a medical database by applying graphical patterns[3]. However, in their research, both *cause* and *effect* need to be contained together in the same sentence. Thus, these methods are not able to extract causal knowledge stridden over two sentences, while, our method can extract causal knowledge in such a case.

Chang et al. proposed a method for extracting causal relations that exist between noun phrases using clue phrases and word pair probabilities[5]. This probability was defined as the probability of a causal noun phrase pair. They used a bootstrapping method for learning Naive Bayes causality classifier. Girju proposed a method for automatic detection and extraction of causal relations based on clue phrases[4]. In their paper, the causal relation is expressed by a pair of noun phrases. Girju used WordNet as semantic constraints for selecting candidate pairs. Hence, her method can not extract unknown phrases that are not in WordNet. In contrast, our method deals with causal knowledge expressed by not only noun phrases but also verb phrases and sentences. In addition, our method can extract unknown phrases.

Sakai et al. proposed a method for extracting cause information from Japanese financial articles concerning business performance[6]. Their method only extracts cause phrases. On the other hand, our method can also extract effect phrases.

Table 1. List of tags

Tag	Description	Examples
CLUE	clue phrase	ため (*tame*: because), から (*kara*: from)
CAUSE_VP	verb cause phrase	株式市場が下落した (*kabushiki sijyou ga gerakushita*: stock market declined)
CAUSE_NP	noun cause phrase	景気の回復 (*keiki no kaihuku*: recovery of economy)
EFFECT	effect phrase	世界不況 (*sekai hukyou*: world economy recession)
EFFECT_SBJ	subject of effect phrase	農産物価格は、(*nousanbutsu kakaku ha*: agricultural price)
EFFECT_PRED	predicate of effect phrase	下落した (*geraku shita*: fell off)
INV	word which changes verb/noun of cause phrase	の (*no*: of), こと (*koto*: thing)

3 Investigation of Clue Phrases

We investigate 300 newspaper articles concerning economic trends using clues for extracting appropriate causal knowledge. We expect that newspaper articles concerning economic trends contain a large amount of causal knowledge. Newspaper articles concerning economic trends are acquired by Sakaji's[7] method from the Nikkei newspaper published from 1990 to 2005.

3.1 Tagging rules

Newspaper articles concerning economic trends are annotated with the following tags (in Table 1) to investigate how causal knowledge and the clue phrases are expressed. A human annotates tags to causal knowledge that is expressed explicitly by a fact and its account in a sentence or two adjacent sentences. We have two notes about the definition. In Japanese, the subject and the predicate can be expressed separately in a sentence. In such a case, the subject of an effect phrase is annotated with "EFFECT_SBJ" and the predicate of an effect phrase is annotated with "EFFECT_PRED". Otherwise, the effect phrase is annotated with "EFFECT".

We found that specific clue phrases are used depending on the type of the cause phrase. For example, clue phrase "ため (*tame*: because)" is used only when the cause phrase is a verb phrase. However, if the cause is a predicate, a special word, such as "の (*no*: of)" or "こと (*koto*: thing)" is used to convert from a verb to a noun. We annotated "INV" tag to such words.

3.2 Results of Tagging

Inui et al. also reported an investigation of clue phrases[8] on the social domain of a general newspaper, the Mainichi newspaper. Table 2 shows the number of

Table 2. The number of investigated articles, the total number of clue phrases, the set of clue phrases

	Num. of articles	Total num. of clue phrases	Set of clue phrases
inui's investigation	750	219	34
our investigation	300	695	154

Table 3. The number of phrases used as clue phrases(Frequency), the total number of phrases(Sum), the ratio of phrases used as clue phrases(Rate)

Clue phrase	Frequency	Sum	Rate
で (*de*: by)	155	2385	0.065
による (*ni yoru*: by)	48	244	0.197
で、(*de*: by)	48	340	0.141
から (*kara*: from)	46	646	0.071
ため、(*tame*: because)	30	55	0.545
を背景に (*wo haikei ni*: behind)	24	26	0.923
から、(*kara*: from)	17	46	0.367

investigated articles, the total number of clue phrases and the set of clue phrases of the investigation by Inui's and ours. It shows that our investigation found more clue phrases per article than Inui's. We believe that the articles concerning economic trends have more causal knowledge than that of the social domain, though minor differences of definitions may cause some differences. Clue phrases with the seven highest frequencies found by our investigation are shown in Table 3.

In Table 3, clue phrases such as "で (*de*: by)", "による (*ni yoru*: by)" and "から (*kara*: from)" frequently appear in articles. However, these clue phrases' ratio between the frequency that the phrase is used as a clue phrase compared with the total frequency of the phrase in the corpus is very low. Therefore, it is not appropriate to use them in our automatic causal knowledge extraction system. We define the following score using the frequency and the ratio, in order to accurately extract causal knowledge.

$$Score(t_i) = \log(TF(t_i)) \times R(t_i) \tag{1}$$

Here, $TF(t_i)$ is frequency that phrase t_i is used as a clue phrase. $R(t_i)$ is a ratio that phrase t_i is used as a clue phrase. Clue phrases with the ten highest *score* are shown in Table 4.

4 Extraction of Causal Knowledge

From our investigation, we found that most of the cause phrases appear before clue phrases in a sentence. On the other hand, effect phrases appear in various

Table 4. Score of clue phrases

Clue phrase	Frequency	Sum	Rate	Score
を背景に (*wo haikei ni*: behind)	24	26	0.923	1.274
を背景に、(*wo haikei ni*: behind)	10	10	1.000	1.000
を受け、(*wo uke*: under)	12	14	0.857	0.925
を挙げる (*wo ageru*: quote)	10	12	0.833	0.833
ため、(*tame*: because)	30	55	0.545	0.806
に伴う (*ni tomonau*: with)	14	23	0.609	0.698
に伴い、(*ni tomonau*: with)	6	7	0.857	0.667
を反映して (*wo haneishite*: reflect)	6	7	0.857	0.667
に加え、(*ni kuwae*: besides)	9	13	0.692	0.661

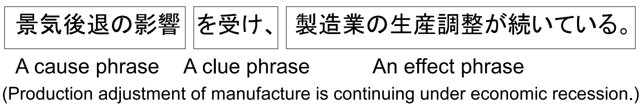

Fig. 2. An example of Pattern B

locations in a sentence. Furthermore, we found that clue phrases have four types of syntactic roles in indicating causal knowledge. Considering these observations we propose a method based on syntactic information. That is, we categorize patterns of cause-effect phrases by the location of the phrases and syntactic roles of clue phrases into four patterns as follows. These four patterns cover 86% of the cause-effect phrases in the corpus that we examined.

Pattern A: both a predicate and a subject are effect phrases in a sentence (see Fig. 1). The clue phrase has a role of connecting a predicate of an effect phrase and a subject of an effect phrase with a cause phrase.

Pattern B: an effect phrase appears after a cause phrase in a sentence (see Fig. 2). The clue phrase has a role of connecting an effect phrase with a cause phrase.

Pattern C: an effect phrase is the sentence just before a sentence including a clue phrase (see Fig. 3). The clue phrase has a role of indicating a previous sentence that is a cause phrase.

Pattern D: an effect phrase appears before a cause phrase in a sentence (see Fig. 4). The clue phrase has a role that is modified by an effect phrase and a cause phrase.

Patterns A, B and D extract causal knowledge from one sentence. On the other hand, Pattern D extracts causal knowledge from two adjacent sentences.

In the following three subsections, we will explain how we identify the patterns and how to extract cause-effect phrases using syntactic information.

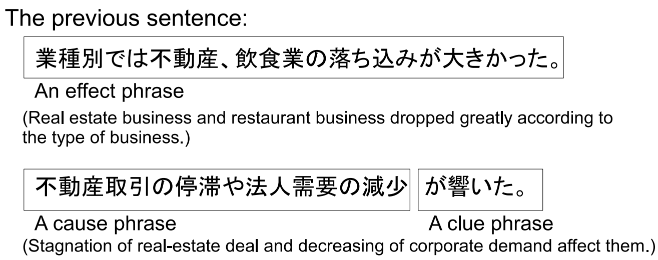

Fig. 3. An example of Pattern C

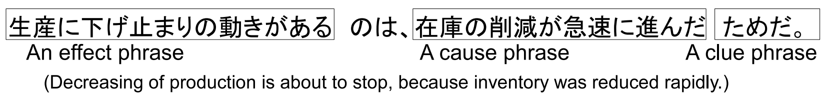

Fig. 4. An example of Pattern D

4.1 Parser

The sentences that include clue phrases are parsed by a Japanese dependency analyzer, Cabocha[3]. For example,

サブプライムローンの危機により、世界不況が起こった。
(World economy recession was caused by the subprime loan crisis.)

The parsed sentence is shown in Fig. 5. Here, a *bunsetu* is a basic block in Japanese composed of several words. The sequence of *bunsetu*s in a sentence are expressed as $(b_1, b_2, ..., b_n)$, a subscript of b is a *bunsetsu*'s number. *Bunsetu*s are assigned consecutive numbers, beginning from one, in ascending order from the beginning of a sentence. For example, in Fig. 5, “サブプライムローンの (*sabupuraimuro-n no*: the subprime loan)” has a number 1.

4.2 Pattern Identification

Here, we define the *core bunsetu* as the *bunsetu* that is rearmost *bunsetu*s composing a clue phrase. In addition, we define the *base point bunsetu* as the *bunsetu* modified by the core *bunsetu* (see Fig. 5). Our method searches causal knowledge and identifies an appropriate syntactic pattern.

Step 1: Search a sentence including clue phrases.
Step 2: If the clue phrase includes a punctuation “。” or a punctuation “。” appears after the clue phrase, go to Step 4. Otherwise, go to Step 3.

[3] http://chasen.org/~taku/software/cabocha/

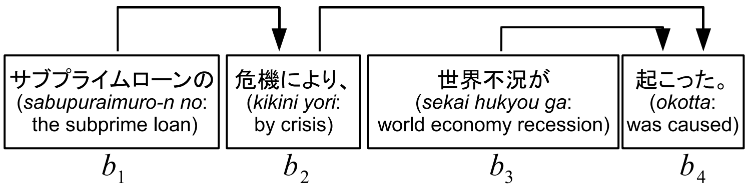

Fig. 5. An example of a parsed sentence

Step 3: If the base point *bunsetu* is a verb phrase and a *bunsetu* that modifies the base point *bunsetu* includes a particle, Pattern A is chosen. Otherwise, Pattern B is chosen. Go to step 5.
Step 4: If a *bunsetu* that modifies a core *bunsetu* includes a particle, Pattern D is chosen. Otherwise, Pattern C is chosen.
Step 5: Stop.

4.3 Cause Effect Phrase Extraction

Each pattern consists of a "cause phrase extraction process" and an "effect phrase extraction process". Each of these processes has a string variable CAN (an abbreviation of a candidate) and a numeric variable M. CAN is assigned "null" and M is assigned 0 as an initial value, respectively. First, we will describe the cause extraction process for all patterns, then we will describe the effect extraction process.

Cause Extraction

Patterns A and B

Step 1: Search a core *bunsetu* in the parsed sentence. The core *bunsetu*'s number is assigned to M.
Step 2: If CAN is "null", CAN is assigned b_{M-1} except for a particle, else connect b_{M-1} with CAN, and assign it to CAN. M is assigned $M-1$.
Step 3: Repeat Step 2 until $M-1$ becomes 0 or b_{M-1} modifies a phrase that has a number larger than the core *bunsetu*'s number.
Step 4: Acquire CAN as a cause phrase. □

Patterns C and D

Step 1: Search a core *bunsetu* in the parsed sentence. The core *bunsetu*'s number is assigned to M.
Step 2: If CAN is "null", CAN is assigned b_{M-1} except for a particle, else connect b_{M-1} with CAN, and assign it to CAN. M is assigned $M-1$.
Step 3: Repeat Step 2 until $M-1$ becomes 0 or b_{M-1} modifies the core *bunsetu* and b_{M-1} includes a syndetic particle(*kakari joshi*) or a conjunctive particle(*setsuzoku joshi*).
Step 4: Acquire CAN as a cause phrase. □

Effect Extraction

<u>Pattern A</u>

The procedure to extract the predicate of an effect phrase:

Step 1: Search a base point *bunsetu* from the parsed sentence and assign it to *CAN*. The base point *bunsetu*'s number is assigned to M.
Step 2: Connect b_{M-1} with *CAN*, and assign it to *CAN*. M is assigned $M-1$.
Step 3: Repeat Step 2 until $M-1$ becomes 0 or b_{M-1} becomes a core *bunsetu*.
Step 4: Acquire *CAN* as a predicate of an effect phrase. □

The procedure to extract the subject of an effect phrase:

Step 1: Search a *bunsetu* that modifies a base point *bunsetu* and includes a syndetic particle(*kakari joshi*) or a case particle(*kaku joshi*) from the parsed sentence. Assign this *bunsetu* to *CAN*. The *bunsetu*'s number is assigned to M.
Step 2: If b_{M-1} modifies b_M, connect b_{M-1} with *CAN*, and assign it to *CAN*, else go to Step 4. M is assigned $M-1$.
Step 3: Repeat Step 2 until $M-1$ becomes 0.
Step 4: Acquire *CAN* as a subject of an effect phrase. □

Fig. 1 illustrates an example of extracting causal knowledge by Pattern A. When extracting a predicate of an effect phrase, first, *CAN* is assigned a base point *bunsetu* "ある (*aru*: is)". M is assigned a base point *bunsetu*'s number 9. Next, connect b_8 with *CAN*, and assign "状況にある (*jyoukyou ni aru*: is in a situation)" to *CAN*. M is assigned 8. Repeat this process, until b_{M-1} becomes a core *bunsetu* "反映し、(*han ei shi*: reflecting)". Then, *CAN* "依然厳しい状況にある (*izen kibishii jyoukyou ni aru*: is still in a difficult situation)" is extracted as a predicate of an effect phrase.

When extracting a subject of an effect phrase, first, *CAN* is assigned a *bunsetu* "収益は、(*syuueiki ha*: benefits)" that modifies a base point *bunsetu* "ある (*aru*: is)" and includes a syndetic particle(*kakari joshi*) "は (*ha*)". M is assigned this *bunsetu*'s number 2. Next, connect b_1 with *CAN* "収益は、(*syuueiki ha*: benefits)", and assign "売り上げと収益は、(*uriage to syuueki ha*: sales and benefits)" to *CAN*. Because b_1 modifies b_2. M is assigned 1. Then, $M-1$ becomes 0. Therefore, *CAN* "売り上げと収益は、(*uriage to syuueki ha*: sales and benefits)" is extracted as a subject of an effect phrase.

When extracting a cause phrase, first, M is assigned a core *bunsetu*'s number 5. Assign b_4 except for particles "停滞 (*teitai*: stagnation)" to *CAN* because *CAN* is "null". M is assigned 3. Next, connect b_3 "需要の (*jyuyou no*: demand)" with *CAN* "停滞 (*teitai*: stagnation)", and assign "需要の停滞 (*jyuyou no teitai*: stagnation of demand)" to *CAN*. M is assigned 2. *CAN* "需要の停滞 (*jyuyou no teitai*: stagnation of demand)" is acquired as a cause phrase, because b_2 modifies "ある (*aru*: is)" that has a number larger than the core *bunsetu*'s number.

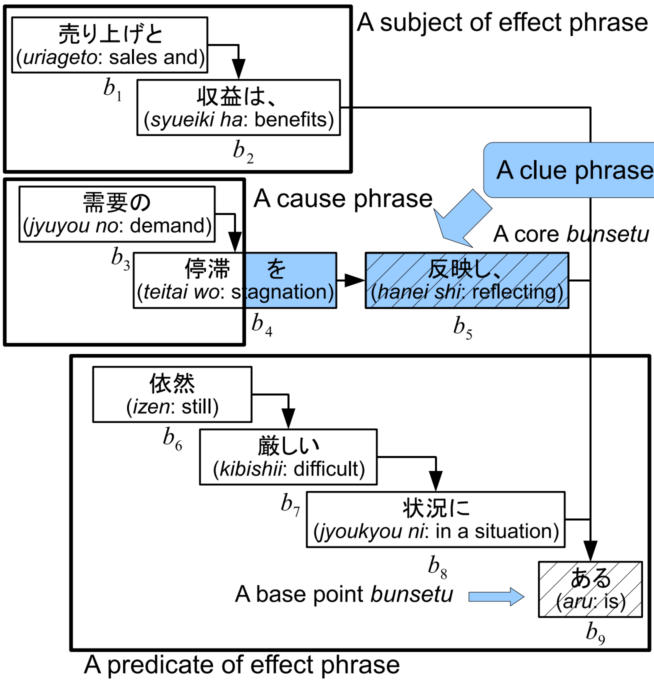

Fig. 1. An example of Pattern A

Pattern B

Pattern B extracts effect phrases in a manner similar to Pattern A's extraction of predicates of effect phrases, except the base point *bunsetu* is a noun phrase. If a base point *bunsetu* is a noun phrase, effect phrases are extracted in a manner described below:

Step 1: Search a base point *bunsetu* from the parsed sentence and assign it to *CAN*. The base point *bunsetu*'s number is assigned to M.
Step 2: If b_{M+1} is modified by a *bunsetu* that has a number smaller than a core *bunsetu*'s number, connect *CAN* with b_{M+1}, and assign it to *CAN*, else go to Step 4. M is assigned $M + 1$.
Step 3: Repeat Step 2 until b_M becomes the rearmost *bunsetu* in the sentence.
Step 4: Acquire *CAN* except for a terminal particle as an effect phrase. □

Pattern C

The sentence just before the sentence including a clue phrase is extracted as an effect phrase.

Pattern D

The effect phrases are extracted in a manner described below:

Table 5. Extraction results

	Recall	Precision	F-Measure
Cause phrase	0.917	0.757	0.829
Effect phrase	0.809	0.526	0.638
Both	0.770	0.449	0.567

Step 1: Search a *bunsetu* that modifies a core *bunsetu* and includes a syndetic particle(*kakari joshi*). Assign this *bunsetu* except for a particle and punctuation to *CAN*. The *bunsetu*'s number is assigned to M.
Step 2: If b_{M-1} modifies b_M, connect b_{M-1} with *CAN*, and assign it to *CAN*, else go to Step 4. M is assigned $M-1$.
Step 3: Repeat Step 2 until $M-1$ becomes 0.
Step 4: Acquire *CAN* as an effect phrase. □

5 Evaluation

In this section, we evaluate our method. A set of 200 new articles concerning economic trends are used for the evaluation. The clue phrases we used are the highest scored 40 clue phrases in Section 3.2 except for "で、(*de*: by)", "ほか、(*hoka*: except)", "ため (*tame*: because)" and "による (*ni yoru*: by)". Then we run the system.

First, the system output is evaluated automatically by matching the output with the human made answers. However, minor differences are sometimes acceptable as correct, a human evaluates again when there is a difference. The total evaluation results for cause and effect are presented in Table 5. Table 6 shows the accuracy for each clue phrases with more than 5 frequency. Recall is the percentage of correctly extracted phrases out of the phrases found by the annotator. Precision is the percentage of the correctly extracted phrases out of all phrases found by the system. F-Measure is a combination of recall and precision equally weighted, and is calculated by the following formula 2.

$$F\text{-}Measure = 2 * precision * recall/(precision + recall) \qquad (2)$$

F-Measure attains a high value only when both recall and precision are high.

As shown in Table 6, for clue phrases "に加え、(*ni kuwae*: besides)", "の効果が (*no kouka ga*: effect)" and "の影響が (*no eikyou ga*: influenced by)" no effect phrases are extracted. The F-measure for the cause and effect phrases are 0.829 and 0.639, respectively. We believe that the result of cause phrase extraction is good for a causal knowledge extraction task. Therefore, we consider that cause phrase extraction is sufficient for a real-life application. However, the result of effect phrase extraction is not satisfactory.

Table 6. Precision for each cause and effect phrases

Clue phrase	num. of phrases	Precision	
		Cause phrase	Effect phrase
を背景に (*wo haikei ni*: behind)(B)	14	0.923	0.923
ため、(*tame*: because)(A)	6	0.500	0.200
ため、(*tame*: because)(B)	24	0.708	0.664
に伴う (*ni tomonau*: with)(B)	20	0.750	0.500
に加え、(*ni kuwae*: besides)(B)	10	0.400	0
ためだ。(*tameda*: because)(C)	9	0.889	0.889
ためだ。(*tameda*: because)(D)	6	1.000	0.833
により (*ni yori*: by)(A)	9	0.778	0.667
により (*ni yori*: by)(B)	11	0.455	0.273
の効果が (*no kouka ga*: effect)(B)	8	0.750	0
の影響が (*no eikyou ga*: influenced by)(B)	15	0.733	0
により、(*ni yori*: by)(B)	14	0.929	0.786
によって (*ni yotte*: by)(B)	10	0.500	0.500
から、(*kara*: from)(A)	12	0.833	0.667
から、(*kara*: from)(B)	30	0.700	0.567

Table 7. The number of false positives and false negatives

	num. of errors	num. of false positives	num. of false negatives
Cause phrase	71	51	20
Effect phrase	139	103	36

6 Error Analysis

We found two kinds of errors, "false positive" and "false negative". False positive is the one where the extracted phrase is not a causal phrase. False negative is the one where the causal phrase is not extracted. We investigated the number of false positives and false negatives, shown in Table 7. We can see that the number of errors of cause phrases are twice the number of errors of effect phrases. Also, the number of false positives are more than the number of false negatives in both cause and effect phrase extraction. Therefore, we consider that a method for determining the presence or absence of causal knowledge is necessary.

Next, we investigated precision of each pattern. In the investigation, we exclude false positive examples in order to examine performance of syntactic patterns in detail. The results are shown in Table 8. A number of Pattern B's faults was the most highest in the four patterns. Most of Pattern B's faults were the parser errors. The rest of them were algorithmic design errors. We expect that errors are decreased by improving an algorithm's conditional expression.

Table 8. Each precision of pattern

	num. of phrases	Precision	
		Cause phrase	Effect phrase
Pattern A	38	0.978	0.816
Pattern B	113	0.901	0.761
Pattern C	13	0.867	0.917
Pattern D	6	1.000	1.000

7 Conclusion

We proposed a method that extracts causal knowledge using clue phrases and syntactic patterns from newspaper articles concerning economic trends. We, first, investigated our newspaper corpus by annotating causal knowledge and clue phrases. Then, we found that some specific syntactic patterns are useful to improve accuracy to extract causal knowledge. Finally, we developed our system to extract causal knowledge using the clue phrases and the syntactic patterns and show the evaluation results on a large corpus.

References

1. Inui, T., Inui, K., Matsumoto, Y.: Acquiring causal knowledge from text using the connective marker *tame*. Journal of Information Processing Society of Japan **45**(3) (2004) 919–933
2. Khoo, C.S., Kornfilt, J., Oddy, R.N., Myaeng, S.H.: Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. Literary and Linguistic Computing **13**(4) (1998) 177–186
3. Khoo, C.S., Chan, S., Niu, Y.: Extracting causal knowledge from a medical database using graphical patterns. In: Proceedings of the 38th ACL. (2000) 336–343
4. Girju, R.: Automatic detection of causal relations for question answering. In: In ACL Workshop on Multilingual Summarization and Question Answering. (2003) 76–83
5. Chang, D.S., Choi, K.S.: Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. Information Processing and Management **42**(3) (2006) 662–678
6. Sakai, H., Masuyama, S.: Cause information extraction from financial articles concerning business performance, ieice trans. IEICE Trans. Information and Systems **E91-D**(4) (2008) 959–968
7. Sakaji, H., Sakai, H., Masuyama, S.: Automatic extraction of basis expressions that indicate economic trends. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). (2008) 977–984
8. Inui, T., Okumura, M.: Investigating the characteristics of causal relations in japanese text. In: The 43rd Annual Meeting of the Association for Computational Linguistics, Workshop on Frontiers in Corpus Annotation II: Pie in the Sky. (2005)