Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 662–678
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities

Du-Seong Chang [a,*], Key-Sun Choi [b]

[a] *Spoken Language Research Team, KT, 17 Woomyeon-dong, Seocho-gu, Seoul 137-792, Republic of Korea*
[b] *Division of Computer Science, Korea Advanced Institute of Science and Technology, BOLA, KORTERM, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea*

Received 5 October 2004; accepted 11 April 2005
Available online 15 June 2005

Abstract

This work aims to extract possible causal relations that exist between noun phrases. Some causal relations are manifested by lexical patterns like causal verbs and their sub-categorization. We use lexical patterns as a filter to find causality candidates and we transfer the causality extraction problem to the binary classification. To solve the problem, we introduce probabilities for word pair and concept pair that could be part of causal noun phrase pairs. We also use the cue phrase probability that could be a causality pattern. These probabilities are learned from the raw corpus in an unsupervised manner. With this probabilistic model, we increase both precision and recall. Our causality extraction shows an *F*-score of 77.37%, which is an improvement of 21.14 percentage points over the baseline model. The long distance causal relation is extracted with the binary tree-styled cue phrase. We propose an incremental cue phrase learning method based on the cue phrase confidence score that was measured after each causal classifier learning step. A better recall of 15.37 percentage points is acquired after the cue phrase learning.

Keywords: Causality; Pattern learning; Word pair probability; Cue phrase probability; Unsupervised learning

* Corresponding author. Tel.: +82 50 2797 8797; fax: +82 2 526 5909.
E-mail addresses: dschang@kt.co.kr (D.-S. Chang), kschoi@cs.kaist.ac.kr (K.-S. Choi).

0306-4573/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2005.04.004

1. Introduction

Causality or *Causal relation* refers to "the relation between a cause and its effect or between regularly correlated events".[1] *Causality pattern* is defined as a word, a phrase or their comprising pattern, which connects one event to the other with causal relation. In this paper, we aim to extract possible causal relations that exist between noun phrases. The causality patterns are used for connecting the cause and effect noun phrases. *Cue phrase* is a lexical, phrasal or structural causal cue between the cause and effect noun phrases. Cue phrases that connect two noun phrases are known to be causal verbs (Girju, 2003; Girju & Moldovan, 2002). For example, in (1a), the verb "cause" is a cue phrase to connect the two noun phrases "the oral bacteria" and "gum disease". In (1b), the two noun phrases "early childhood sun exposure" and "skin cancer" are not directly connected by one verb. To determine such long-distance causal relation, we introduce tree structured cue phrases.

(1a) The oral bacteria that *cause* gum disease appear to be the culprit.
(1b) Early childhood sun exposure *is* particularly *important in the development of* skin cancer.

Some word pairs in noun phrase pairs indicate the causality of the noun phrase. The word pair "bacteria" and "disease" is an example of a causal word pair. When the noun phrase pair "oral bacteria" and "gum disease" are causally related, we can infer that the noun pair "bowel bacteria" and "bowel disease" may be causally related. Causal word pairs are learned from cause–effect noun phrase pairs. We define *word pair probability* as the probability of the word pair that is part of causal noun phrase pairs. The pair of concept classes "bacteria" and "disease" (in MeSH, [B03] and [C23.550.288][2]) also helps identify the causal relation between noun phrase pairs. *Concept pair probability* is defined as the probability of the concept pair that has causal relation. Cue phrases connecting two causal noun phrases are also considered to have connection probability. We define *cue phrase probability* as the probability of the cue phrase that connects causal noun phrase pairs.

We use cue phrases as a filter to find causality candidates and we transfer the causality extraction problem to the binary classification. With the above probabilities, we will propose a causality classifier based on the Naïve Bayes classifier. These probabilities are learned from the raw corpus in an unsupervised manner. In Section 2, selected works are compared for causality extraction. Our classification model will be explained in Section 3. If we have a causality classifier, a set of new cue phrases is acquired from the automatically generated causality-annotated corpus. Furthermore, the long distance causality recognition will be resolved with the binary tree-styled cue phrases and their incremental learning method. The cue phrase learning method will be explained in Section 4. Finally, an evaluation is given for the proposed models.

2. Related works

Previous works on causality analysis mainly used the pattern matching approach. The probabilistic classification approach is applied to related research domain like rhetorical structural analysis. This work uses the cue phrase filter and introduces a new classification model based on cue phrase and word pair probabilities. A learning method for cue phrase and classifier is also proposed.

[1] From Merriam-Webster's Online Dictionary.
[2] They represent [bacteria] and [disease]. These concept numbers follow the biomedical ontology of MeSH (Medical Subject Headings; US National Library of Medicine, 2004).

2.1. Causality recognition and inference

Causal relations are expressed in various forms in the literature: between the subject and object position of noun phrases as in (2a), between two sentences or phrases as in (2b), or in the intra-structure of a noun phrase as in (2c). Causal relation also exists between paragraphs that describe events.

(2a) Gum disease is caused by plaque.
(2b) Because gum disease is usually painless, you may not know you have it.
(2c) Disease-causing sticky film of bacteria.

In the above examples, each cause event is connected to its effect event by the causative verb (''cause''), causal connectives (''because''), or intra-NP structure. In this paper, we focus on the causal relation between noun phrases as in (2a).

The inference model on causality, including the Bayesian approach of Cooper and Herskovits (1991) and the conditional independency pattern searching approach of Spirtes, Glymour, and Scheines (1993), are not the focus of this paper. We concentrate on the search for causal relations that are explicitly represented on the raw corpus.

2.2. Causality analysis

Initial works on causality analysis used hand-made causality patterns to find causality (Joskowsicz, Ksiezyk, & Grishman, 1989; Kaplan & Berry-Rogghe, 1991). Low, Chan, Choi, Chin, and Lay (2001) used hand-made causal semantic templates and a concept term dictionary for the restricted domain. For the financial news domain, they reported finding causality related to Hong Kong stock movement with a precision of 76%. Khoo, Kornfit, Oddy, and Myaeng (1998, 2000) used the semi-automatic causality pattern learning on the syntactically analyzed corpus and introduced the cue phrase learning method. However, the causality pattern matching method has a limited performance since not all sentences selected by patterns guarantee causality. They reported that the precision of the causality identifier pattern was about 76%.

Girju and Moldovan (2002) used the inter-noun phrase causal relation to improve the question-answering performance. To extract inter-noun phrase causal relations, they used the cue phrase filter and the noun class ranking based on WordNet (Miller, 1995). We call this the ''dictionary-based ranking model''. With simple rules reflecting the meaning of head nouns, each noun phrase pair is classified into 5 ranked classes, which are called the ''noun class rank''. The ranking rules are summarized as follows: rank 1 is given if the head nouns' senses belong to causation classes like ''human action'', ''phenomena'', ''state'', ''psychological feature'' and ''event'' among the WordNet synsets; rank 2 if only the effect head nouns' senses belong to the causation class; rank 3 if the effect head noun is enumerated to the causation class[3]; rank 4 if some senses of the cause head nouns are ''causal agents''; and finally, rank 5 is given based on the number of senses and the frequency of verbs.

The examination regarding ranks 1–4 as causal relations shows a precision score of 65.5%. Their decision tree classifier learned on the causality-annotated corpus showed a precision of 73.91% (Girju, 2003). In their works, cue phrases were verb phrases automatically acquired from WordNet and the corpus. After they collected causal noun phrase pairs from WordNet gloss definitions, they found the corresponding causality patterns form of <noun phrase-1, verb/verb expression, noun phrase-2> from the corpus. Finally, 72 cue phrases (which are verbs) were selected to connect the noun phrase pairs from the corpus.

[3] Girju and Moldovan (2002) said that if the effect is represented by an enumeration of noun phrase and the head noun of at least one of them has all the senses in one of the causation class, than the others also refer to causality in that context. They classified this relationship as causation of rank 3.

Marcu and Echihabi (2002) used the inter-sentence word pair probability for discriminating the rhetorical relation between sentences. To distinguish the causal relation from other rhetorical relations, they used the sentence pairs connected with "Because of" and "Thus". With the Naïve Bayes classifier, sentence pairs were classified as either "causal" or not. The result showed an accuracy of 57% in inter-sentence causality extraction.

For the supervised learning of the causality classifier, a causality-annotated corpus is required. However, the construction of such a corpus would take much effort. The supervised method has the limitation of being scaled up. We learned the word pair and cue phrase probabilities through the unsupervised learning method (Chang & Choi, 2005). This kind of learning method has the benefit of using a large-sized text corpus as the training set. We used the word pair probability to fulfill the low precision of cue phrases. In cases where a dictionary or WordNet is used as the basis of causality, the unregistered words in the dictionaries hinder the search for correct causal relation. We solved this unknown word problem by using the word pair probability. If the dictionary or word-sense mapping module is not available, then the proposed model is more attractive.

The proposed classification and learning models with modifications will be described in detail, and some examinations on cue phrase learning and concept pair probability will be added. The cue phrases from previous works on causality analysis were simple verb patterns given by humans or semi-automatically learned. In this paper, we introduce a binary tree-styled cue phrase for considering the long distance causality. Then, we also try to automate the cue phrase acquisition.

3. Causality extraction model

3.1. Causality extraction system

Consider our proposed causality extraction system flow in Fig. 1. Causality candidates were extracted from the raw corpus through a noun phrase chunking with syntactic analysis. We used cue phrases as a

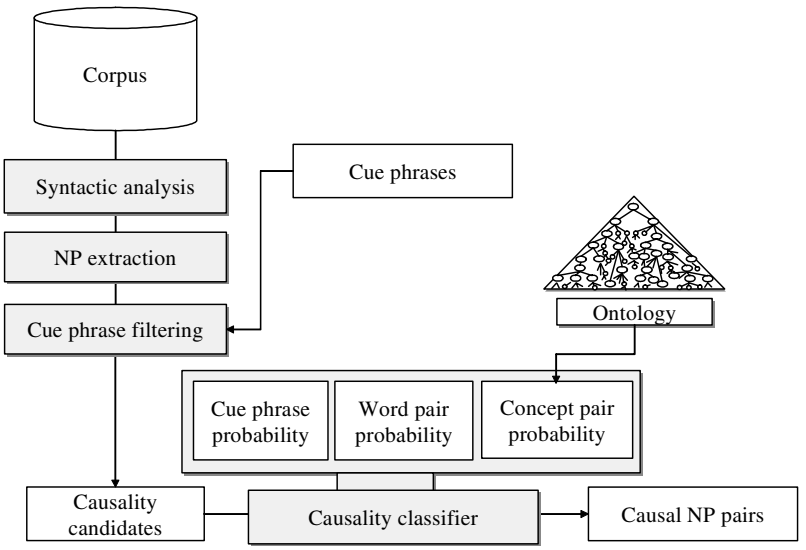

Fig. 1. Proposed causality extraction system.

filter to find causality candidates. The causality analysis problem was redefined as a classification problem to assign one of two causal classes, ''causal'' or ''non-causal'', to the causality candidates. Causal noun phrase pairs were selected by the cue phrase, word pair and concept pair probabilities.

3.2. Causality representation: cue phrases and causality candidates

Consider sentences (3a)–(3c). The two noun phrases are not directly connected by a verb phrase. When we limit the cue phrase to a verb or verb pattern, these long distance causal relations are hardly captured.

(3a) Radon *is* the nation's second-leading *cause of* lung cancer.
(3b) Sun exposure *is* particularly *important in the development of* skin cancer.
(3c) Skin cancer *had occurred because of* the sunburn.

A *Cue phrase* is re-defined as a verb-rooted syntactic tree, which connects one noun phrase to the other with causal relation. Fig. 2 shows syntactic trees for sentences (3a)–(3c) and their corresponding cue phrases which are included in boxes. These syntactic patterns normalize the syntactic variation such as the passive or the verbal chain. With a dependency tree-based cue phrase, we can consider the syntactic variation of the cue phrase. Here ''CNP'' and ''ENP'' refer to the cause and effect noun phrases, respectively.

The input of the classifier is a causality candidate that is filtered by cue phrases. The causality candidate is expressed by a ternary composed of a noun phrase pair and a cue phrase: <cause noun phrase candidate, cue phrase, effect noun phrase candidate>. To represent and match the syntactic tree pattern, the preorder

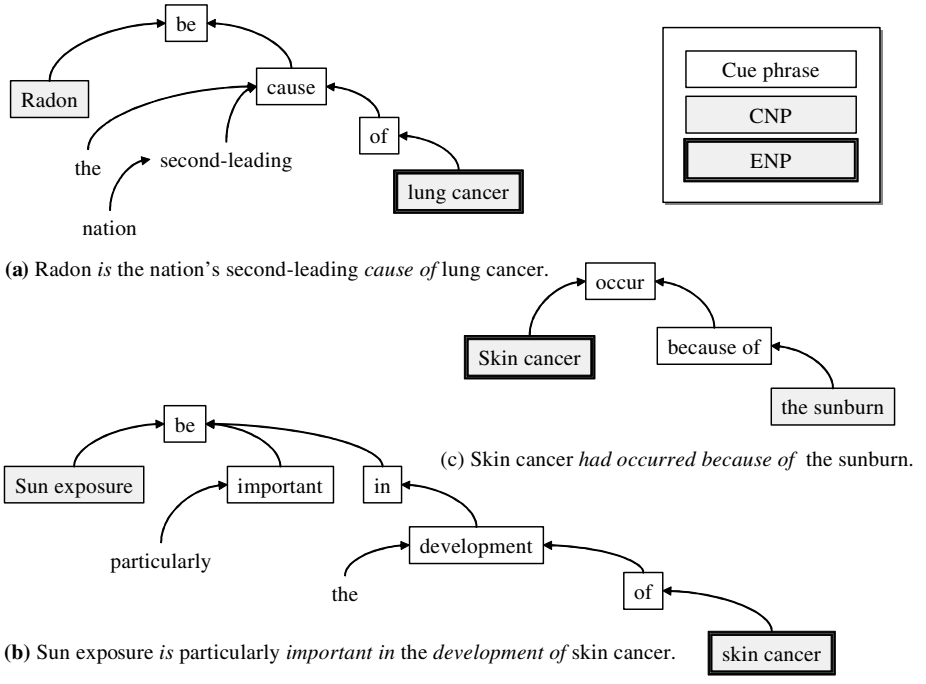

Fig. 2. Syntactic trees and their corresponding cue phrases.

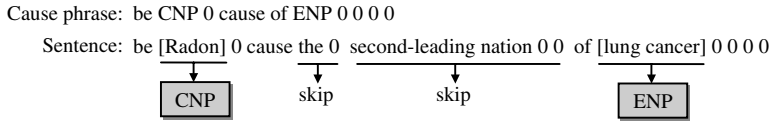

Fig. 3. Cue phrase matching process.

string expression (Luccio, Enriquez, Rieumont, & Pagli, 2001) was introduced. The preorder string of tree W in the root string l is defined as

$$W = l\ 0 \text{ (if } W \text{ is leaf node) or } l\ W_1, \ldots, W_h, 0 \text{ (if } W \text{ has subtrees, } W_1, \ldots, W_h)$$

For each node of the syntactic tree encountered, the corresponding label l is entered in the preorder string W. And for each return to the previous node a 0 is entered in W. With this expression, the cue phrase pattern matching time was reduced to the sublinear time.[4] Fig. 3 shows the preorder string of the sentence (3a) and its cue phrase matching sequence. The cue phrase matching algorithm is as follow: (1) Find the cue phrase root word (e.g. 'be') from the preordered string of the sentence. (2) Match the cue phrase and the preorder string. (3) When it mismatched, skip a subtree that is the same number of labels and 0's. (4) Noun phrases can be matched with the CNP and ENP slot. (5) The cue phrase matching success only if all slots and words of the cue phrase are matched. Causality candidates extracted from sentences (3a)–(3c) were as follows:

(4a) <"Radon", "be CNP 0 cause of ENP 0 0 0 0", "lung cancer">
(4b) <"sun exposure", "be CNP 0 important 0 in development of ENP 0 0 0 0 0", "skin cancer">
(4c) <"sunburn", "occur ENP 0 because of CNP 0 0 0 0", "skin cancer">

3.3. Naïve Bayes causality classifier

The causality classifier classified the causality candidate ternary (t_i) into "causality (c_1)" or "non-causality (c_0)". The class c^* of the ternary t_i was computed as shown in Formula (1):

$$c^* = \arg\max_{c_j} P(c_j \mid t_i) = \arg\max_{c_j} \frac{P(c_j)P(t_i \mid c_j)}{P(t_i)} \tag{1}$$

When we consider the cue phrase CP_{t_t}, the concept pair SP_{t_t}, and word pairs $LP_{t_t k}$ as in causal features of the ternary t_i, $P(t_i \mid c_j)$ in Formula (1) will be rewritten as Formula (2).[5] In Formula (2), $|t_i|$ represents the total number of word pairs in ternary t_i. We assumed that these features are all independent each other.

$$P(t_i \mid c_j) = P(CP_{t_i} \mid c_j)P(SP_{t_i} \mid c_j) \prod_{k=1}^{|t_i|} P(LP_{t_i k} \mid c_j) \tag{2}$$

In Formula (2), $P(CP_{t_i} \mid c_j)$, $P(SP_{t_i} \mid c_j)$, and $P(LP_{t_i k} \mid c_j)$ are the cue phrase probability, concept pair probability and the word pair probability, respectively, which were defined in Section 1. These probabilities can be learned from the causality-annotated ternary set. However, the construction of the causality-annotated

[4] Luccio et al. (2001) said, "Given two rooted ordered trees, T and P of n and m vertices, respectively, all occurrences of P as a subtree of T can be determined in $O(m + \log n)$ time after pre-processing T in $\Theta(n)$ time."

[5] In the evaluation, the concept pair probability is not fully used for some implementational reasons which are summarized in Section 5.2.

ternary set consumes time and effort. In this paper, we used a raw corpus rather than causality-annotated corpora. To make it possible, the EM (Expectation–Maximization) procedure was used with the Naïve Bayes classifier. In the next section, we summarize the unsupervised causal classifier learning method in Chang and Choi (2005).

3.4. Classifier learning with EM

The Naïve Bayes classifier is bootstrapped from the initial classifier. The training data is the causality candidate ternary set that was filtered by cue phrases. There are three training stages. In the initialization stage, we build an initial Naïve Bayes classifier from an initial classifier. As an initial classifier, we use the noun class rank described in Section 2.2. It is a dictionary-based classifier and does not need the extra training sequence. After the whole training corpus is classified with the initial classifier, highly ranked ternaries are selected as the initial causality-annotated set. From this selected annotation data, the parameters of the initial Naïve Bayes classifier are estimated. The initial Naïve Bayes classifier is from the automatically causality-annotated NP pairs that is a noisy data. To counter this problem, we use only the high causality-ranked small set initially, and use the raw corpus together on the EM learning sequence.

The second training stage is called the Expectation step. The whole training corpus, including the annotated part, is classified with the current classifier. The final training stage is called Maximization step. From the newly classified data, parameters are re-estimated. Parameters trained in EM are prior probability $P(c_j)$, cue phrase probability $P(CP_{t_i}|c_j)$, word probability $P(LP_{t_i k}|c_j)$, and concept probability $P(SP_{t_i}|c_j)$. The parameters are estimated with Laplace smoothing method (Laplace, 1814, 1995) for word pairs unseen in the training data. The Expectation step and Maximization step are repeated while the classifier parameters improve.

3.5. Causality classification model

The trained classifier can be combined with the noun class rank probability and the cue phrase confidence score. The noun class rank probability, $P(c_j|\text{rank}_{t_i})$, is defined as the probability of the causal class for the given noun class rank of the ternary. The cue phrase confidence score, $P(c_j|CP_{t_i})$, is defined as the probability of the causal class for the given cue phrase. These probabilities are learned from the automatically annotated corpus.

We propose some classification models. The classification model $CP + LP$ uses the cue phrase probability and the word pair probability as shown in (3). In this model, the noun class rank is also used as a back-off model. If the cue phrase probability and the word pair probability, $P(c_j|t_i)$, cannot decide on an evident causal class, the noun class rank probability is used. To do this, a discrimination value, $\text{Dist}(t_i)$, is introduced as shown in (4). The threshold h is a constant.

$$P_{CP+LP}(c_j|t_i) = \begin{cases} P(c_j|t_i) & \text{if } \text{Dist}(t_i) > h \\ P(c_j|\text{rank}_{t_i}) & \text{otherwise} \end{cases} \tag{3}$$

$$\text{Dist}(t_i) = \left| \frac{\log P(c_0|t_i) - \log P(c_1|t_i)}{\log P(c_0|t_i) + \log P(c_1|t_i)} \right| \tag{4}$$

The classification model $CP + LP + NC$ uses the cue phrase probability and the word pair probability combined with the noun class rank probability as shown in (5). The sum of weights w_{lp} and w_{nc} must be 1.

$$P_{CP+LP+NC}(c_j|t_i) = w_{lp} \times P(c_j|t_i) + w_{nc} \times P(c_j|\text{rank}_{t_i}) \tag{5}$$

The cue phrase confidence score is also learned from the automatically annotated corpus. The classification model $CP + LP + NC + CPC$ uses the cue phrase probability and the word pair probability com-

bined with the noun class rank probability and the cue phrase confidence score as shown in (6). The sum of weights w_{lp}, w_{nc} and w_{cpc} must be 1.

$$P_{CP+LP+NC+CPC}(c_j \mid t_i) = w_{lp} \times P(c_j \mid t_i) + w_{nc} \times P(c_j \mid \text{rank}_{t_i}) + w_{cpc} \times P(c_j \mid CP_{t_i}) \quad (6)$$

4. Cue phrase learning

The causality classifier was bootstrapped from an initial classifier with the raw corpus. The causality classifier requires a pre-defined set of cue phrase. Cue phrases were also learned from the set of causal noun phrase pairs and the open set of web pages. When we decide that two noun phrases are causally connected, the path connecting these two noun phrases on the syntactic tree is possibly a cue phrase. Fig. 4 shows the cue phrase learning flow that is composed of three steps: the initial noun phrase pair selection, the cue phrase candidate extraction, and the cue phrase selection.

From the causal noun phrase pairs, we selected initial noun phrase pairs such as "sun exposure" and "skin cancer". Sentences that contain initial noun phrase pairs were extracted from the web pages. Cue phrase candidates were generated from these sentences after the noun phrase was replaced with the noun phrase slot. The extracted cue phrase candidates were sorted with the confidence score that was calculated by the causality classifier. Then the cue phrase candidates with low confidence scores were removed. These new cue phrases took part in the causality classifier.

4.1. Initial noun phrase pair selection

The initial noun phrase pairs were extracted from 5 million news articles which are a part of TREC (Text REtrieval Conference; Harman, 1992) corpus. The corpus was automatically annotated with the causality classifier in Section 3. If a noun phrase pair ep_i appeared in k times in the corpus, the total causality $PC(ep_i)$

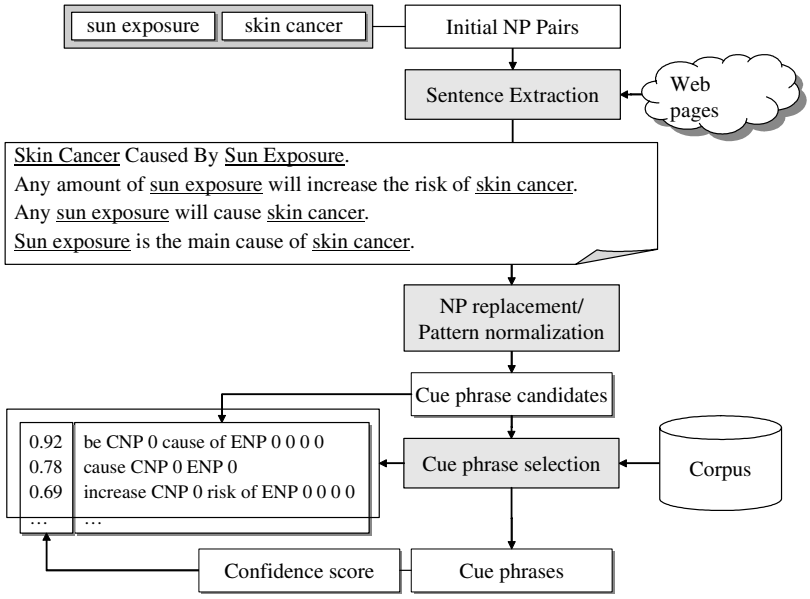

Fig. 4. Cue phrase learning process.

Table 1
The initial noun phrase pairs (in part)

Total causality	Noun class	Noun phrase pair
1.89063	1	a disease, blindness
1.26203	1	events, the accident
1.2305	1	smoking, cancer
1.22994	1	the chaos, injury
5.17184	2	the HIV virus, AIDS
2.4252	2	gases, global warming
2.38468	2	residential areas, noise

of a noun phrase pair ep_i is computed with the causal probabilities of ternaries $\{t_{ik}\}$ containing the noun phrase pair ep_i. The total causality of a noun phrase pair is high at the condition that the noun phrase pair appeared frequently and with a high causal probability.

$$PC(ep_i) = \frac{1}{2}\sum_k (P(c_1 \mid t_{ik}) - P(c_0 \mid t_{ik})) \tag{7}$$

The set of initial noun phrase pairs was selected based on the *total causality* of the given noun phrase pair. From the 200 000 causal noun phrase pair candidates in the corpus, we selected 584 causal noun phrase pairs (0.3%) under the following conditions:

(1) The noun class rank of the noun phrase pair is higher than 4.
(2) The total causality of the noun phrase pair is over than 1.

Table 1 shows a part of the automatically selected initial noun phrase pairs.

4.2. Cue phrase candidate extraction

Sentences that contain initial noun phrase pairs were gathered from web pages. For each event pairs, we collected 100 sentences in maximum. After syntactic analysis, the keyword noun phrase of the syntactic tree was replaced with the noun phrase slot, CNP and ENP. From the normalized syntactic tree, the syntactic patterns including the two noun phrase slots were selected as cue phrase candidates. From 15 000 sentences, we found 1142 syntactic patterns as cue phrase candidates.

4.3. Cue phrase selection

Cue phrase candidates were sorted with the confidence score. The confidence score of the cue phrase candidates could be learned with the causal classifier. Fig. 5 shows the cue phrase selection flow.

To obtain the confidence score of the cue phrase candidates, the new classifier that uses the original cue phrases and all new cue phrase candidates were learned. The classifier learning sequence is the same as that in Section 3.4. The initial classifier for the cue phrase learning is the noun class rank and the word pair probability. For every step of new classifier learning, the cue phrase confidence score was updated. After removing the cue phrase candidates with low confidence scores, the cue phrases were settled. Cue phrases with low confidence score were omitted and the new pattern was added to the cue phrases. These new cue phrases took part in the causality classifier. Table 2 shows a part of these cue phrases.

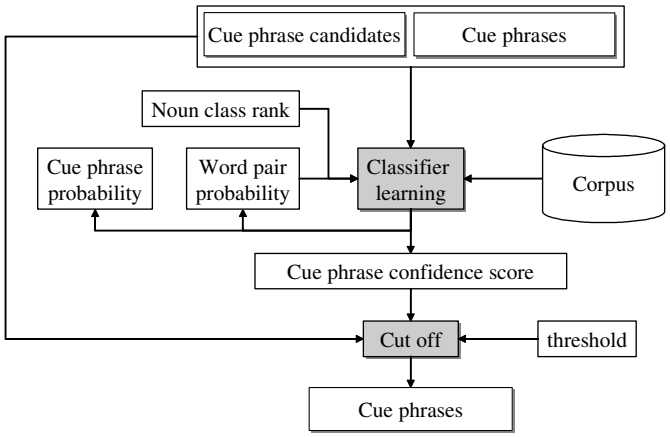

Fig. 5. Cue phrase selection flow.

Table 2
Cue phrases (in part)

be <ENP> 0 result of <CNP> 0 0 0 0
cause <CNP> 0 <ENP> 0 0 0
cause <CNP> 0 form of <ENP> 0 0 0 0
be <CNP> 0 virus cause <ENP> 0 0 0 0
be <CNP> 0 one of cause of <ENP> 0 0 0 0 0 0
be <CNP> 0 cause of <ENP> 0 0 0 0
cause kind of <CNP> 0 0 0 <ENP> 0 0
be <ENP> 0 due to <CNP> 0 0 0
give <CNP> 0 birth 0 to <ENP> 0 0 0
be cause of <ENP> 0 0 0 <CNP> 0 0
be <ENP> 0 act of <CNP> 0 0 0 0

4.4. The incremental cue phrase learning

The causality classifier proposed in Section 3.1 is bootstrapped from an initial classifier and the raw corpus. It uses pre-defined cue phrases. Cue phrases are learned with the established classifier and the causality-annotated corpus as shown in Figs. 4 and 5. After new cue phrases are added, new classifier is re-estimated and new cue phrase candidates are selected again from the annotation of this new classifier. Fig. 6 show these two learning system together.

The cue phrase learning sequence requires an existing causality classifier. The first Naïve Bayes causality classifier is bootstrapped from the dictionary-based classifier that uses only 72 causal verbs defined in Girju and Moldovan (2002). From the initial causality classifier and the initial cue phrase set, cue phrases and the causality classifier are incrementally learned by turns. The incremental cue phrase learning process can be repeated in a fixed number or while new patterns with high confidence scores are discovered.

When we added the candidate to the patterns, we added only some candidates that have high confidence scores. In the first loop, 19 cue phrases were selected from the 1142 cue phrase candidates. Finally, 81 cue phrases were added by the incremental cue phrase learning method.

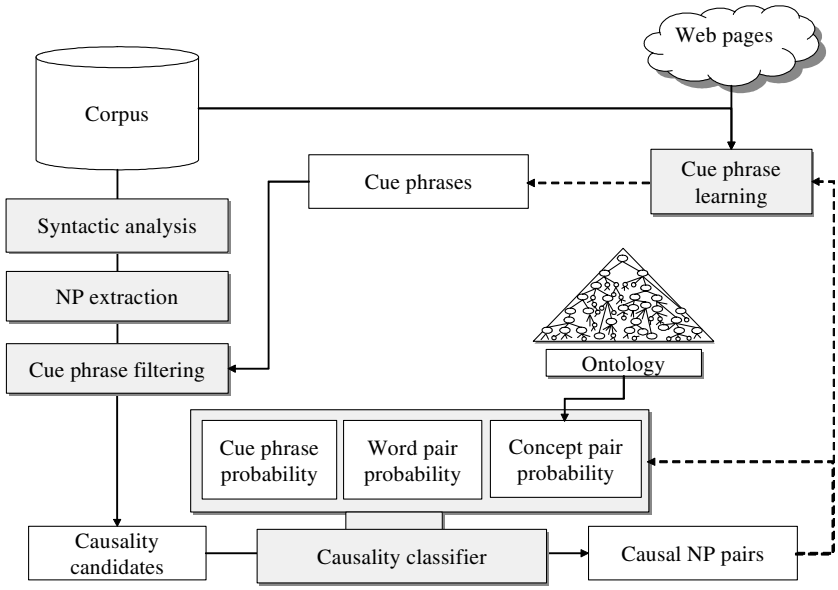

Fig. 6. The causality classifier and cue phrase learning system.

5. Evaluation

5.1. Training and test set

A part of the TREC corpus was used for causality extraction. The training corpus was composed of 5 million sentence-sized articles from the LA TIMES (1989–1990) and Wall Street Journal (1987–1990). We used two test sets, which were selected from different domains. The first one was from Wall Street Journal articles. The other was from the Medline medical encyclopedia of A.D.A.M. Inc. All sentences in the test sets included the word "cancer". The first one, which we called as *cTREC*, came from the general domain. The other, which we called as *cADAM*, came from the medical domain. Table 3 shows the specification of the corpus. All corpora were syntactically analyzed by a dependency parser (Tapanainen & Jarvinen, 1997).

Two human annotators manually classified the test sets; one was the first author and the other the medical domain expert. They had a 72.8% agreement on the results. A gold standard was made after discussions between the annotators.

Table 3
The specification of the corpus

		# of words	# of sentences	# of ternary filtered
Training set	WSJ	44.1M	1.9M	91192
	LA TIMES	72.1M	3.1M	117430
	Total	116.2M	5.0M	208 622
Test set	*cTREC*	40476	1379	205
	cADAM	23323	1147	166
	Total	63799	2526	371

Table 4
Causality extraction result

Classification model	Test set	Precision	Recall	*F*-value
NC	*cTREC*	76.47	57.07	65.36
	cADAM	71.43	30.12	42.37
	Total	74.89	45.01	56.23
LP with No EM	*cTREC*	74.74	69.27	71.90
	cADAM	72.73	43.37	54.34
	Total	74.05	57.68	64.85
CP + LP	*cTREC*	81.18	67.32	73.60
	cADAM	78.33	56.63	65.73
	Total	80.00	62.53	70.20
CP + LP + NC	*cTREC*	81.87	68.29	74.47
	cADAM	75.42	53.61	62.68
	Total	79.24	61.73	69.39
CP + LP + NC + CPC	*cTREC*	81.18	67.32	73.60
	cADAM	79.84	59.64	68.28
	Total	80.61	63.88	71.28

5.2. Evaluation

The cue phrase probability (*CP*) and the word pair probability (*LP*) were trained on the training set. As an initial classifier, the noun class rank was used. For the parameter initialization, all ternaries were ranked with noun classes and highly ranked ternaries were selected as a causality-annotated set. As a result, ternaries ranked by 1–3 were annotated to "causal" (c_1), and parts of ternaries ranked by 5 were annotated to "non-causal" (c_0). The initial causality-annotated ternaries were 18% of the training set. The causality classifier was bootstrapped from the causality-annotated set and the unlabeled training set.

Table 4 shows the evaluation result on the test sets. The classification model *NC* follows the model of Girju and Moldovan (2002), which uses the cue phrase filter and the noun class rank. The classification model *LP with No EM* follows the classification model of Marcu and Echihabi (2002), which uses the word pair probability without EM process.

The last three models are the proposed models. For the classification model *CP + LP*, we assigned 0 to the value of the threshold *h*. And for the noun class (*NC*) weight w_{nc} and the cue phrase confidence score (*CPC*) weight, 0.1 was assigned. For the evaluation, we used the pre-defined cue phrases based on the 72 causal verbs of Girju and Moldovan (2002).

Contribution of the cue phrase probability and the word pair probability. The proposed model *CP + LP* showed an *F*-value of 70.20%, which was an improvement of 13.97 percentage points from the baseline model (*NC*). In all the proposed models, the causality extraction performance was increased. We can say that the cue phrase probability and the word pair probability are useful for causality extraction.

Contribution of the noun class on domains. For the general domain test set (*cTREC*), the result of the combined with the noun class (*CP + LP + NC*) improved the performance in both the recall and the precision from the non-combined (*CP + LP*). However, for the medical domain test set (*cADAM*), the performance was decreased by 3.05 percentage points. This was caused by unknown words in the medical domain test set. Terminologies and pronouns in the specific domain included more unknown words than in the

general domain. For the baseline model *NC*, the unknown words in *cADAM* decreased the performance by 15.1% in precision and by 11.1% in the recall. We can say that the noun class is useful in the general but not in the specific domain.

Contribution of the cue phrase confidence score. The classification model combined with the cue phrase confidence score (*CP* + *LP* + *NC* + *CPC*) did not show any significant improvement over the non-combined model (*CP* + *LP* + *NC*). This is because the cue phrase probability and the cue phrase confidence score shared the same information space.

Robustness of the proposed model. In the proposed model (*CP* + *LP*), 37.5% of the unknown word-causing errors in the baseline system (*NC*) were correctly classified. The proposed model did not refer the word sense. It only referred the word pair frequency in the corpus. We can say that the proposed model is free from unknown words.

High performance of unsupervised learning. The proposed models are learned in an unsupervised manner. They do not require the pre-annotated data. Nevertheless, the performance is relatively high.

Table 5
Causality extraction with concept pair probabilities

Classification model	Test set	Precision	Recall	*F*-value
CP + *LP* + *SP*	*cTREC*	83.23	67.80	74.73
	cADAM	77.88	53.01	63.08
	Total	81.07	61.19	69.74
CP + *LP* + *SP* + *NC* + *CPC*	*cTREC*	84.13	68.39	75.44
	cADAM	79.67	59.04	67.82
	Total	81.93	63.55	71.58

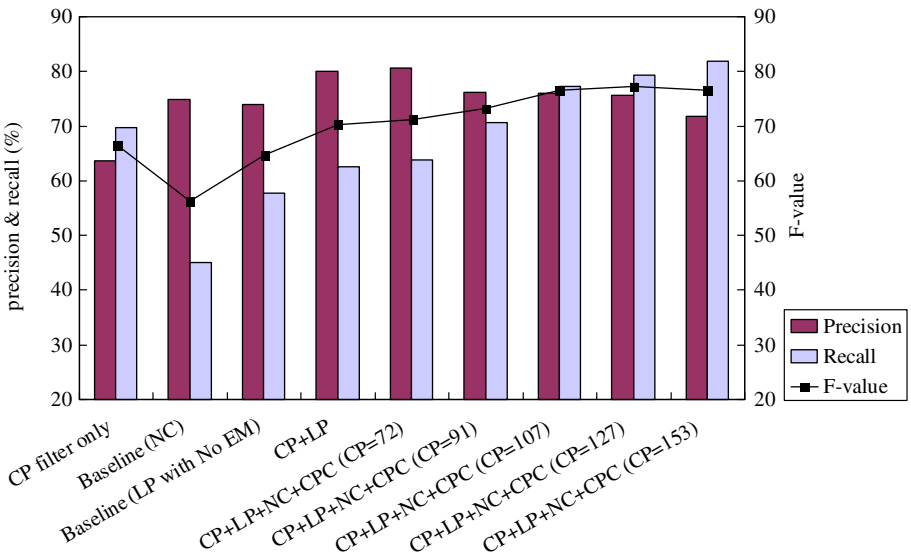

Fig. 7. Causality extraction after cue phrase learning.

Table 6
Causality extraction after cue phrase learning

Classification model	Cue phrases	Precision	Recall	*F*-value
CP + *LP* + *NC* + *CPC* (CP = 91)	Initial 72 causal verbs	80.81	81.63	81.22
	Trained cue phrases	46.81	28.57	35.48
	Total	76.16	70.62	73.29
CP + *LP* + *NC* + *CPC* (CP = 127)	Initial 72 causal verbs	79.48	82.99	81.20
	Trained cue phrases	60.98	64.94	62.89
	Total	75.58	79.25	77.37

Table 7
Causality extraction result on the test set *wTREC*

Classification model	Precision	Recall	*F*-value
NC	80.13	45.79	58.33
CP + *LP*	84.38	50.47	63.16
CP + *LP* + *NC*	87.69	53.27	66.28
CP + *LP* + *NC* + *CPC* (CP = 72)	86.36	53.27	65.90
CP + *LP* + *NC* + *CPC* (CP = 91)	86.49	59.81	70.72
CP + *LP* + *NC* + *CPC* (CP = 127)	86.84	61.68	72.13

Contribution of concept pair probability. We proposed the usage of concept pair probability, which is based on the concept class of the noun phrases. To find the concept class of each noun phrase, we can use a dictionary like WordNet. However, this is not the simple way since there are many noun phrases that represent two or more concept classes. To bypass the word sense disambiguation, we consider the head word pair of the noun phrases as the concept pair. Table 5 shows the performance evaluation of the concept pair probability (*SP*). In the table, they show meaningless improvements with the concept pair probability. It indicates that the headword pair was not enough and the word sense disambiguation problem had to be solved for causality extraction.

Contribution of the incremental cue phrase learning method. The effect of incremental cue phrase learning was evaluated in Fig. 7. In the first step, we added 19 patterns to the cue phrases. In the next step, 16 were added. Finally, we had a total of 153 cue phrases. From the *F*-value curve, the performance was maximized in the case where the number of cue phrases was 127. With small scarification of the precision, we can get greater recall value through the incremental cue phrase learning method. As a result, the total performance (*F*-value) was increased by 6.09 percentage points. About 17% of the causal noun phrase pairs were extracted by cue phrases that were added by the incremental learning process. Table 6 shows the performance contrast between the original 72 causal verbs and the trained tree-styled cue phrases.

Applicability of the proposed causality extraction and learning method. Terminologies of test corpus might affect the performance of the causality extraction. Table 7 shows the causality extraction performance on the new test set named *wTREC*. All sentences in *wTREC* are from general domain articles and include the word "war". *wTREC* contains 1827 sentences and 107 causalities. Although the recall is comparatively low, *F*-value is elevated with the causality classifier and the cue phrases learning sequences. The final result showed an *F*-value of 72.13% on the new test set. This result says that the proposed causality extraction and learning method is applicable to all sentences.

6. Conclusion

An improved approach was introduced in this paper for causality extraction. Previous works on causality extraction mainly used the lexical pattern matching and WordNet. We use lexical patterns as a filter to find causality candidates and we transfer the causality extraction problem to the binary classification. With this approach, we managed to combine possible classification features and introduce any kind of learning method.

The bootstrapping method was found useful for learning the Naïve Bayes causality classifier on the raw corpus. Empirical results suggested feasible features for the causality extraction. The cue phrase and the word pair probabilities are two of them, and noun class rank showed good performance in such domain that dictionary works well. The main advantage of the proposed causality extraction model over that of Girju (2003) is the robustness. The proposed model empirically shows high performance without dictionary-based feature. The benefit of the binary tree-styled cue phrase expression is its ability to match the long distance causality. With this cue phrase expression and the incremental cue phrase learning method, we automate the cue phrase learning sequence. In summary, we proposed the improved methods on causality extraction and cue phrase learning. The results of evaluations were promising.

The proposed causality extraction was used for the causal question answering. The causal question answering is available on the web; the causal browsing that uses the proposed system can be accessed there as well (Chang, 2003). We found direct/indirect causal relations with the proposed causality extraction. Fig. 8 shows the causal network for the term "protein". It was automatically generated from 2000 document-

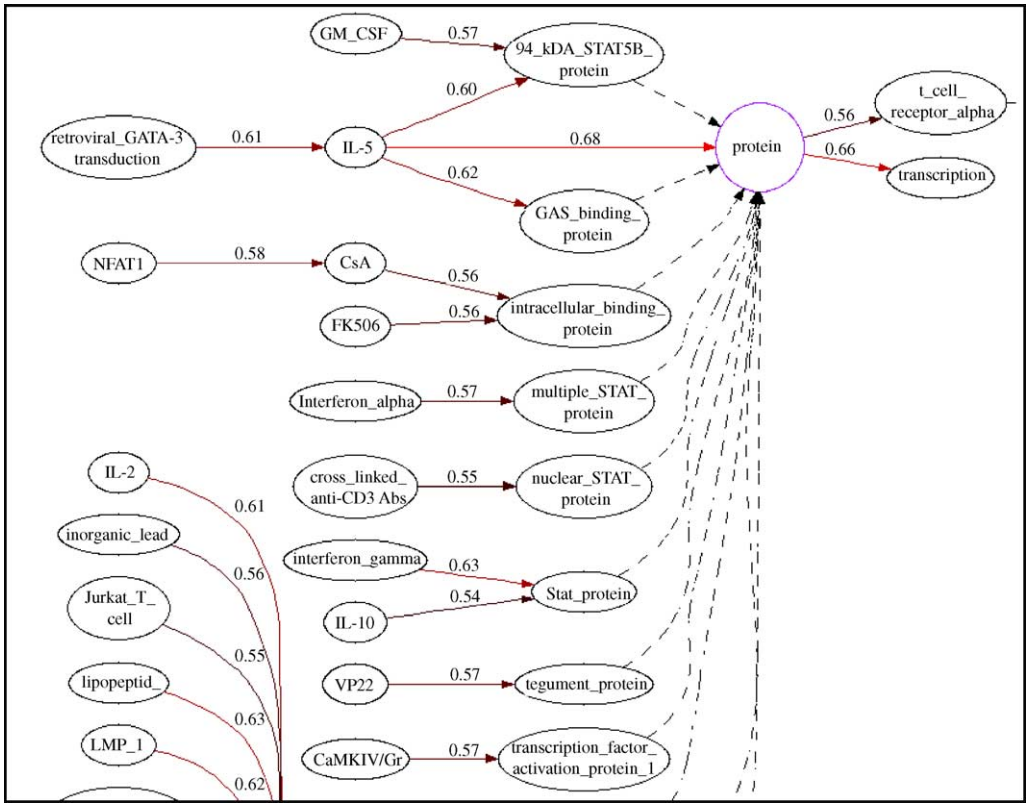

Fig. 8. Causal network for the term "protein" (in part).

sized biological domain paper abstracts (Tsujii, 2003). In the figure, the dotted line shows the hypernym relation. The solid line shows the causal relation in which the numeric label on each edge represents the causal probability described in documents.[6] The causality between "IL-5" and "protein" is originated from the sentence (5a). In the sentence (5b) and (5c), "IL-5" also causes two noun phrases that have common hypernym, "protein". These three causality and two hypernym relations are represented in the figure. The cue phrases "induce CNP 0 ENP 0 0" and "activate CNP 0 ENP 0 0" worked as filter.

(5a) IL-5 induced two proteins that bound to the gamma-activating sequence.
(5b) The 94 kDa STAT5B protein was activated by both IL-5 and GM-CSF.
(5c) We found that IL-5 induces two GAS-binding proteins in eosinophils.

The focus of this paper is restricted within the inter-NP causality. In the medical domain, especially for medical encyclopedia corpus, the cause and effect are rarely in one sentence. We are now expanding the search space to the inter-sentences causality. The preliminary trial on Korean inter-sentence causality extraction showed a little chance to be improved (Chang & Choi, 2005).

Acknowledgment

This research was supported in part by KISTEP Strategic National R&D Program under brain science program and by KOSEF under contract for bank of language resources.

References

Chang, D. S., & Choi, K. S. (2005). Causal relation extraction using cue phrase and lexical pair probabilities. In *Natural language processing—IJCNLP 2004*, *Lecture notes in computer science*, Vol. 3248, pp. 61–70.

Chang, D. S. (2003). Causality question answering system. KORTERM/KAIST. Available from http://gensum.kaist.ac.kr/~dschang/ENC/CQA.html.

Cooper, G., & Herskovits, E. (1991). A Bayesian method for constructing Bayesian belief networks from database. In *Proceedings of the conference on uncertainty in AI*, pp. 86–94.

Girju, R. (2003). Automatic detection of causal relations for question answering. In *The 41st annual meeting of the association for computational linguistics, workshop on multilingual summarization and question answering—Machine learning and beyond*, Sapporo, Japan.

Girju, R., & Moldovan, D. (2002). Mining answers for causation questions. In *AAAI symposium on mining answers from texts and knowledge bases*.

Harman, D. (1992). Overview of the first Text REtrieval Conference (TREC-1). In *The first Text REtrieval Conference (TREC-1)*, pp. 1–20.

Joskowsicz, L., Ksiezyk, T., & Grishman, R. (1989). Deep domain models for discourse analysis. In *The annual AI systems in government conference*, pp. 195–200.

Kaplan, R. M., & Berry-Rogghe, G. (1991). Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition, 3*(3), 317–337.

Khoo, C. S. G., Chan, S., & Niu, Y. (2000). Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pp. 336–344.

Khoo, C. S. G., Kornfit, J., Oddy, R. N., & Myaeng, S. H. (1998). Automatic extraction of cause–effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing, 13*(4), 177–186.

Laplace, Pierre Simon marquis de. (1995). *Philosophical essay on probabilities*. New York: Springer-Verlag.

Laplace, Pierre Simon marquis de. (1814). *Essai philosophique sur les probabilities*. Paris: Mme. Ve. Courcier.

[6] The directed acyclic graph in Fig. 8 is meaningful on the condition that: (1) It is selected from one document or the same domain with the same topic. (2) It is independent from events that are not represented in the graph.

Low, B. T., Chan, K., Choi, L. L., Chin, M. Y., & Lay, S. L. (2001). Semantic expectation-based causation knowledge extraction: A study on Hong Kong stock movement analysis. In *Lecture notes in computer science: Advances in knowledge discovery and data mining: 5th Pacific-Asia conference, PAKDD 2001 Hong Kong, China, April 2001, proceedings*. Heidelberg: Springer-Verlag.

Luccio, F., Enriquez, A. M., Rieumont, P. O., & Pagli, L. (2001). Exact rooted subtree matching in sublinear time, TR-01-14, University of Pisa.

Marcu, D., & Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th annual meeting of the association for computational linguistics conference*, Philadelphia, PA, pp. 368–375.

Miller, G. (1995). WordNet: A lexical database. *Communications of the ACM, 38*(11), 39–41.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. In *Springer lecture notes in statistics*. New York: Springer-Verlag.

Tapanainen, P., & Jarvinen, T. (1997). A non-projective dependency parser. In *Proceedings of the 5th conference on applied natural language processing*, pp. 64–71.

Tsujii, J. (2003) *Genia 3.01* from the GENIA project home page http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA.

US National Library of Medicine (2004). Medical subject headings from National Institute of Health, http://www.nlm.nih.gov/mesh.

Du-Seong Chang is a senior researcher at the spoken language research team, KT since 1993, where his interests include text mining, discourse system, embedded spoken system, and the realization of the multi modal system. He is PhD candidate at the division of computer science, KAIST. He received his MS in computer science from KAIST (1993).

Key-Sun Choi is a professor at the division of computer science, KAIST since 1988, where his interests include computational terminology and lexicography, Korean language engineering, and cognitive science. He is the director of KORTERM and the secretary of ISO/TC37/SC4. He received his PhD in computer science from KAIST (1986), worked for NEC (1987–1988) and NHK (2002) in Tokyo, and studies in CSLI, Stanford University (1997).