What Kinds and Amounts of Causal Knowledge Can Be Acquired from Text by Using Connective Markers as Clues?

INUI Takashi, INUI Kentaro, and MATSUMOTO Yuji

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5, Takayama, Ikoma, 630-0192, Japan
{takash-i,inui,matsu}@is.aist-nara.ac.jp

Abstract. This paper reports the results of our ongoing research into the automatic acquisition of causal knowledge. We created a new typology for expressing the causal relations — *cause*, *effect*, *precond(ition)* and *means* — based mainly on the volitionality of the related events. From our experiments using the Japanese resultative connective *"tame"*, we achieved 80% recall with over 95% precision for the *cause*, *precond* and *means* relations, and 30% recall with 90% precision for the *effect* relation. The results indicate that over 27,000 instances of causal relations can be acquired from one year of Japanese newspaper articles.

1 Introduction

In many fields including psychology and philosophy, the general notion of causality has been a research subject since the age of ancient Greek philosophy. From the early stages of research into artificial intelligence, many researchers have been concerned with common-sense knowledge, particularly cause-effect knowledge, as a source of intelligence. Relating to this field, ways of designing and using a knowledge base of causality information to realize natural language understanding have also been actively studied [14, 3]. For example, knowledge about the preconditions and effects of actions is commonly used for discourse understanding based on plan recognition. Figure 1-(a) gives a typical example of this sort of knowledge about actions, which consists of precondition and effect slots of an action labeled by the header.

This knowledge-intensive approach to language understanding results in a bottleneck due to the prohibitively high cost of building and managing a comprehensive knowledge base. Despite the considerable efforts put into the CYC [8] and OpenMind [16] projects, it is still unclear how feasible it is to try to build such a knowledge base manually. Very recently, on the other hand, several research groups have reported on attempts to automatically extract causal knowledge from a huge body of electronic documents [1, 7, 2, 13]. While these corpus-based approaches to the acquisition of causal knowledge have considerable potential, they are still at a very preliminary stage in the sense that it is not yet clear what kinds and what amount of causal knowledge they might extract,

(a) plan operator

dry-laundry-in-the-sun($actor, $laundry)	
precondition:	weather(sunny)
effect:	get-dry($laundry)
decomposition:	hang($actor, $laundry)

(b) causal relations

precond(⟨it is sunny⟩, ⟨dry the laundry in the sun⟩)
effect(⟨dry the laundry in the sun⟩, ⟨the laundry gets dry⟩)
means(⟨hang laundry⟩, ⟨dry the laundry in the sun⟩)

Fig. 1. The example of plan operator and causal relations

how accurate the process could be, and how effectively extracted knowledge could be used for language understanding.

Motivated by this background, we are reporting the early results of our approach to automatic acquisition of causal knowledge from a document collection. In this work, we consider the use of resultative connective markers such as "because" or "so" as linguistic clues for knowledge acquisition. For example, given the following sentences (1), we may be able to acquire the causal knowledge given in Figure 1-(a), which can be decomposed into two finer-grained causal relations as given in Figure 1-(b):

(1) a. Because it was a sunny day today, the laundry dried well.
b. It was not sunny today, so John couldn't dry the laundry in the sun.

The idea of using these sorts of connective markers to acquire causal knowledge is not novel in itself. In this paper, however, we address the following subset of the above-mentioned unexplored issues, focusing on knowledge acquisition from Japanese texts:

- What classification typology should be given to causal relations that can be acquired using clues provided by connective markers (in Section 5),
- How accurately can acquired relation instances be classified (in Section 6 and Section 7), and
- How many relation instances can be acquired from currently available document collections (in Section 7).

2 Causal knowledge

We regard causal knowledge instances as binominal relations such as in Figure 1-(b). The headings indicate causal relations and arguments indicate related events held in causal relation with each other. Given text segments like (1), the process of acquiring causal knowledge would form two independent phases: argument identification and causal relation estimation.

Table 1. Typology of causal relations

Causal relations	Meaning	Examples of linguistic tests
cause(SOA_1, SOA_2)	SOA_1 causes SOA_2	SOA_2 happened as a result of the fact that SOA_1 happened.
effect(Act_1, SOA_2)	SOA_2 is the effect of Act_1	SOA_2 happens as a result of the execution of Act_1.
precond(SOA_1, Act_2)	SOA_1 is a precondition of Act_2	Act_2 cannot be done unless SOA_1 holds/happens. If SOA_1 holds/happens, one will often execute Act_2.
means(Act_1, Act_2) (same subjects)	Act_1 is a means of executing Act_2	Someone executes Act_1 in order to execute Act_2. If someone executes Act_1, then she can execute Act_2.

2.1 A typology of causal relations

One of the main goals of discourse understanding is the recognition of the intention behind each volitional action appearing in a given discourse. In intention recognition, therefore, it is important to distinguish volitional actions (e.g., the action of "drying laundry") from all the other sorts of non-volitional states of affairs (e.g., the event of "laundry drying"). For convenience, in this paper, we refer to the former simply as actions (Act) and the latter as states of affairs (SOA) except where a more precise specification is needed. We need to classify causal relations with respect to the volitionality of their arguments.

Given the distinction between actions and SOAs, the causal knowledge base needed for intention recognition can be considered as consisting of:

- the causal relation between SOAs,
- the precondition relation between SOAs and actions,
- the effect relation between actions and SOAs, and
- the means relation between actions.

These relations should not be confused. For example, the confusion between precondition and effect may lead to a fatally wrong inference — hanging laundry causes it to become dry, but never causes a sunny day.

Based on the distinction between these relations, we have created a typology of causal relations as summarized in Table 1. In the table, Act_i denotes a volitional action and SOA_i denotes a non-volitional state of affairs. The first column of the table gives the necessary condition for each relation class. For example, *effect*(Act_1, SOA_2) denotes that, if the *effect* relation holds between two arguments, the first argument must be a volitional action and the second must be a non-volitional state of affairs. On the other hand, it is not easy to provide rigorously sufficient conditions for each relation class. To avoid addressing unnecessary philosophical issues, we provide each relation class with a set of

linguistic tests that loosely specify the sufficient condition. Several examples of the linguistic tests we use are also presented in Table 1.

2.2 Arguments of causal relations

Our proposed collection of causal relations should constitute of a higher level of abstraction than mere rhetorical relations. When a causal relation is estimated from text, we must therefore abstract away subjective information including tense, aspect and modality of the arguments. Similarly, it is desirable that some propositional elements of arguments are also abstracted to conceptual categories, like Asia → LOCATION_NAME. Thus, for the acquisition of causal knowledge, we also need to automatize this abstraction process. In this paper, however, we focus on the relation estimation problem. We process the arguments as follows (see the example (2)):

- Maintaining all propositional information, and
- Discarding all subjective and modal information.

(2) I am familiar with Asia because I traveled around Asia.
→ *effect*(⟨travel around Asia⟩, ⟨be familiar with Asia⟩)

Representation of Arguments We represent arguments of causal relation instances by natural language expressions such as Figure 1-(b) and (2) instead of by any formal semantic representation language for the following two reasons. First, it has proven difficult to design a formal language that can fully represent the diverse meanings of natural language expressions. Second, as discussed in [6], there has been a shift towards viewing natural language as the best means for knowledge representation. In fact, for example, all the knowledge in the Open Mind Commonsense knowledge base is represented by English sentences [16], and Liu et al. [9] reported that it could be successfully used for textual affect sensing.

3 The source of knowledge

3.1 Causal relations and connective markers

Let us consider the following examples, from which one can obtain several observations about the potential sources of causal knowledge.

(3) a. The laundry dried well today because it was sunny.
b. The laundry dried well, though it was not sunny.
c. If it was sunny, the laundry could dry well.
d. The laundry dried well because of the sunny weather.
→ e. *cause*(⟨it is sunny⟩, ⟨laundry dries well⟩)

Table 2. Frequency distribution of connective markers

ga	(but)	131,164	*kara*	(because)	10,209
tame	(because)	76,087	*node*	(because)	9,994
to	(if/when)	56,549	*nara*	(if)	7,598
(re-)ba	(if)	48,606	*tara*	(if)	6,027
nagara	(while)	13,796	*noni*	(but)	2,917

Table 3. Frequency distribution of *tame* in the intra-sentential contexts

	Types of *tame* phrase	Freq	Examples
a	adverbial verb phrase	42,577	*hare-ta-tame sentakumono-ga yoku kawai-ta.* sunny-PAST-*tame* laundry-NOM well dry-PAST The laundry dried well because it was sunny.
b	other types	33,510	*kore-ha ryokousya-no-tame-no kansouki-desu.* this-TOPIC tourist-*tame*-GEN tumble dryer-COPULA This is a tumble drier for the tourist.

(4) a. Mary used a tumble dryer because she had to dry the laundry quickly.
b. Mary could have dried the laundry quickly if she had used a tumble dryer.
c. Mary used a tumble dryer to dry the laundry quickly.
d. Mary could have dried the laundry more quickly with a tumble dryer.
→ e. *means*(⟨use a tumble dryer⟩, ⟨dry laundry quickly⟩)

First, causal knowledge can be acquired from sentences with various connective markers. (3e) is a *cause* relation instance that is acquired from subordinate constructions with various connective markers as in (3a) – (3d). Likewise, the other classes of relations are also acquired from sentences with various connective markers as in (4). The use of several markers is advantageous for improving the recall of the acquired knowledge.

Second, it is also interesting to see that the source of knowledge could be extended to sentences with an adverbial minor clause or even a prepositional phrase as exemplified by (3d), (4c) and (4d). Note, however, that the acquisition of causal relation instances from such incomplete clues may require additional effort to infer elliptical constituents. To acquire a *means* relation instance (4e) from (4d), for example, one might need the capability to paraphrase the prepositional phrase "with a tumble dryer" to a subordinate clause, say, "if she had used a tumble dryer".

Third, different kinds of instances can be acquired with the same connective marker. For example, the type of knowledge acquired is a *cause* relation from sentence (3a), but with a *means* relation from (4a). Thus, one needs to create a computational model that is able to classify the samples according to the causal relation implicit in each sentence. This is the issue we address in the following sections.

3.2 Japanese connective markers

The discussion of English in Section 3.1 applies equally to Japanese. One could acquire the same causal relation instances from sentences with various connective markers such as *tame* (because, in order to), *ga* (but) and *(re-)ba* (if) . On the other hand, different kinds of causal relation instances could be acquired from the same connective marker.

Table 2 shows the frequency distribution of connective markers in the collection of Nihon Keizai Shimbun newspaper articles from 1990. Observing this distribution, we selected *tame* as our target for exploration because (1) the word *tame* is used relatively frequently in our corpus, and (2) the word *tame* is typically used to express causal relations more explicitly than other markers.

Next, Table 3 shows the frequency distribution of the intra-sentential contexts in which *tame* appears in the same newspaper article corpus. The word *tame* is most frequently used as an adverbial connective marker accompanying a verb phrase that constitutes an adverbial subordinate clause (see Table 3-(a)). Hereafter, sentences including such clauses will be referred to as *tame*-complex sentences. We were pleased to observe this tendency because, as argued above, the acquisition from complex sentences with adverbial subordinate clauses is expected to be easier than from sentences with other types of clues such as nominal phrases (see Table 3-(b)). Based on this preliminary survey, we restrict our attention to the *tame*-complex sentences.

4 Related Work

There have been several studies aiming at the acquisition of causal knowledge from text. Garcia [1] used verbs as causal indicators for causal knowledge acquisition in French. Khoo et al. [7] acquired causal knowledge with manually created syntactic patterns specifically for the MEDLINE text database. Girju et al. [2] and Satou et al. [13] tried to acquire causal knowledge by using connective markers in the same way as we do. However, the classification of causal relations that we described in this paper is not taken into consideration in their methods.

It is important to note that our typology of causal relations is not just a simple subset of common rhetorical relations as proposed in Rhetorical Structure Theory [10]. For example, (3) shows that a *cause* relation instance could be acquired not only from a REASON rhetorical relation (exemplified by (3a)), but also from CONTRAST and CONDITION relations ((3b) and (3c), respectively). A collection of causal relations should be considered as representing knowledge of a higher level of abstraction rather than as a collection of rhetorical relations. In other words, causal relation instances are knowledge that is needed to explain why rhetorical relations are coherent. For example, it is because you know the causal relation (3e) that you can understand (3a) to be coherent but (5) to be incoherent.

(5) *The laundry dried well today though it was sunny.

Table 4. Distribution of causal relations held by *tame*-complex sentences in $\mathcal{S}_1$

SC denotes the subordinate clause and MC denotes the matrix clause. Act_s and SOA_s denote an event referred to by the SC, and Act_m and SOA_m denote an event referred to by the MC.

class	SC	MC	frequency	Most frequent relation and its ratio	
A	SOA	SOA	229	*cause*(SOA_s,SOA_m)	0.96 (220/229)
B	Act	SOA	161	*effect*(Act_s,SOA_m)	0.93 (149/161)
C	SOA	Act	225	*precond*(SOA_s,Act_m)	0.90 (202/225)
D	Act	Act	379	*means*(Act_m,Act_s)	0.85 (323/379)
		total	994		0.90 (894/994)

5 Causal relations in tame-complex sentences

Before moving into the classification of *tame*-complex sentences, in this section we describe the causal relations implicit in *tame*-complex sentences. We examined their distribution as follows:

Step 1. First, we took random samples from a newspaper article corpus of 1000 sentences that were automatically categorized into *tame*-complex sentences. Removing interrogative sentences and sentences from which a subordinate-matrix clause pair was not properly extracted due to preprocessing (morphological analyzer) errors, we had 994 remaining sentences. We refer to this set of sentences as $\mathcal{S}_1$.

Step 2. Next, we manually divided the 994 sentences composing $\mathcal{S}_1$ into four classes depending on the combination of volitionality of the subordinate and matrix clauses. The frequency distribution of the four classes (A – D) is shown in the left-hand side of Table 4.

Step 3. We then examined the distribution of the causal relations we could acquire from the samples of each class using the linguistic tests examplified in Table 1 [1].

The right-hand side of Table 4 shows the most abundant relation and its ratio for each class A – D. For example, given a *tame*-complex sentence, if the subordinate clause refers to a volitional action and the matrix clause refers to a non-volitional SOA (namely, class B), they are likely to hold a relation *effect*(Act_s, SOA_m) with a probability of 0.93 (149/161).

The following are examples of cases where the most abundant relation holds.

[1] The clausal volitionality and the causal relations were judged using the linguistic test. To estimate reliability of judgements, two subjects majoring in computational linguistics are currently annotating the texts with both volitionality and causal relations. We calculated κ statistical measure with 200 previously annotated samples. The κ value was 0.93 for the volitionality, 0.88 for causal relations.

(6) *tai-de* *manguroubu-wo* *hakaisi-ta-tame* *daisuigai-ga* *hasseisi-ta.*
in Thailand mangrove-ACC destroy-PAST-*tame* flooding-NOM occur-PAST
Serious flooding occurred because mangrove swamps were destroyed in Thailand.

Act_s: (someone) destroy mangrove swamps in Thailand
SOA_m: serious flooding occur

→ *effect*(⟨destroy mangrove swamps in Thailand⟩, ⟨serious flooding occur⟩)

(7) *pekin-eno* *kippu-wo* *kau-tame* *kippuuriba-ni* *i-tta.*
for Beijing ticket-ACC buy-*tame* to ticket office go-PAST
(I) went to the ticket office in order to buy a ticket for Beijing.

Act_s: (I) buy a ticket for Beijing
Act_m: (I) go to the ticket office

→ *means*(⟨go to the ticket office⟩, ⟨buy a ticket for Beijing⟩)

The distribution shown in Table 4 is quite suggestive. As far as *tame*-complex sentences are concerned, if one can determine the value of the volitionality of the subordinate and matrix clauses, one can classify *tame*-complex sentences into the four relations — *cause*, *effect*, *precond* and *means* — with precision of 85% or more. Motivated by this observation, in the next section we first address the issue of automatic estimation of clausal volitionality before moving onto the issue of automatic classification of causal relations.

6 Estimation of volitionality

In this section, we present our approach to estimating clausal volitionality.

6.1 Preliminary analysis

In our previous work, we found that clausal volitionality depends mostly on the verb of the clause. That is, if certain clauses contain the same verb, the volitionality values of these clauses will also tend to be the same. Nevertheless, there are some counterexamples. For example, both the subordinate clause of (8a) and the matrix clause of (8b) contain the same verb *kakudaisuru* (expand), however, (8a) refers to the volitional action and (8b) refers to the non-volitional SOA.

(8) a. *seisannouryoku-wo* *kakudaisuru-tame* *setubitousisuru.*
production ability-ACC expand-*tame* make plant investment
(A company) will make plant investments to expand production ability.

b. *kanrihi-ga* *sakugensi-ta-tame* *eigyourieki-ga* *kakudaisi-ta.*
cost-NOM reduce-PAST-*tame* profit-NOM expand-PAST
Business profit expanded as a result of management costs being reduced.

Table 5. Feature set used for volitionality estimation

V:Verb C:Case M:Modality

class — descriptions
V EDR — Four features indicating the verb class given by the EDR concept dictionary [18]: (1) true if the verb is "movement" or "action", false otherwise; (2) true if the verb is "state", "change" or "phenomenon", false otherwise; (3) true if both (1) and (2) are true; (4) true if neither of the above is true
V ALT-J/E — A set of binary features indicating the verb class given by the dictionary incorporated in the ALT-J/E translation system [5, 4]: "state","continuous situation", "momentary situation", "intransitive", "transitive", "auxiliary", "potential", "spontaneous", "causative", "passivizable", "indirect-passivizable"
V Goi-Taikei — Verbal semantic attributes in Goi-Taikei [4].
C Marker — "*ga* (subject)", "*wo* (object)"
C Element — The concept of case element descrived in Goi-Taikei [4].
M Tense — "*-ru* (present)" form or "*-ta* (past)" form.
M Aspect — "*-teiru* (-ing)" form or not.
M Voice — "*-reru* (passive)" form or not. "*-seru* (causative)" form or not.
M Potential — "*-dekiru* (can)" form or not.
M Negative — "*-nai* (not)" form or not.
Subject — Whether or not the subject is a human or an organization.

Table 6. Ratio of volitionality of each clause

		frequency	
		Act / SOA	total
$\mathcal{S}_1$	Subordinate clause	539 / 455	994
	Matrix clause	603 / 391	994
$\mathcal{S}_2$	Subordinate clause	613 / 372	985
	Matrix clause	650 / 335	985

Still, there will be factors in addition to the verb that help determine clausal volitionality. As a result of analyzing *tame*-complex sentences in $\mathcal{S}_1$, we found the following new characteristics of volitionality:

- The volitionality value of a clause tends to be a non-volitional SOA when the subject is not a person or an organization.
- The volitionality value of a clause tends to change depending on whether it appears as a subordinate clause or a matrix clause.
- The volitionality value of a clause tends to change based on modality, such as tense.

6.2 Estimation of Volitionality by SVMs

We investigated experimentally how accurately the volitionality value (volitional action or non-volitional SOA) of each clause can be estimated by using Support Vector Machines (SVMs) [17] – an accurate binary classification algorithm.

Experimental Conditions Table 5 shows the features we used to represent the sentences. While almost all features can be automatically extracted, it is not so easy to extract the "Subject" feature. Because subject phrases usually do not appear overtly in Japanese complex sentences. In this experiment, we implemented a simple subject feature extractor with about 60% precision.

We used all the sentences in $\mathcal{S}_1$ as training samples and another new *tame*-complex sentence set $\mathcal{S}_2$ as test samples. The set $\mathcal{S}_2$ includes 985 *tame*-complex sentences sampled from newspaper articles issued in a different year than $\mathcal{S}_1$. The frequency distribution of clausal volitionality of both $\mathcal{S}_1$ and $\mathcal{S}_2$ are shown in Table 6.

In addition to the characteristics of clausal volitionality mentioned before, we found little evidence of a correlation between the volitionality values of matrix and subordinate clauses. So, in this experiment, we created a separate classifier for each clause. We used the quadratic polynomial kernel as a kernel function.

Results The accuracy is 0.885 for the subordinate clauses, and 0.888 for the matrix clauses. The baseline accuracy is 0.853. Here, the baseline denotes the accuracy achieved by applying a simple classification strategy where (a) if the verb of the input clause appeared in the training set, the clause was classified by a majority vote, and (b) if the voting was even or the verb was not included in the training set, the clause was classified as volitional action by default.

Our results obtained through SVMs outperforms the baseline accuracy.

Next, we introduced a reliability metric to obtain a higher accuracy. When the reliability of estimating the volitionality value is known, the accuracy of automatic classification of causal relations can be improved by removing samples where the reliability of estimating the volitionality value is low. For the estimation of reliability, we used the absolute values of the discriminate function (the distances from the hyperplane) output by the SVMs. We set up the reliability threshold value α, and then assumed that a judgment would only be decided for a sample when the reliability was greater than α. By varying α, we obtained the coverage-accuracy curves of Figure 2 [2] . These results confirm that the problem of clausal volitionality estimation is solvable with very high confidence.

7 Automatic classification of causal relations

We investigated how accurately we could classify the causal relation instances contained in *tame*-complex sentences. For this purpose, we again used SVMs as the classifier.

7.1 Experimental conditions

We set up four classes — *cause*, *effect*, *precond* and *means*. The features we used to represent the sentences are as follows:

[2] Coverage = # of samples output by the model / # of samples. Accuracy = # of samples correctly output by the model / # of samples output by the model.

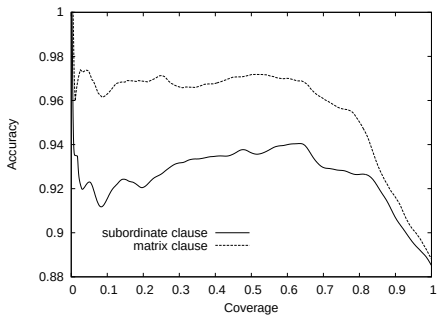

Fig. 2. Coverage-accuracy curves (clausal volitionality estimation)

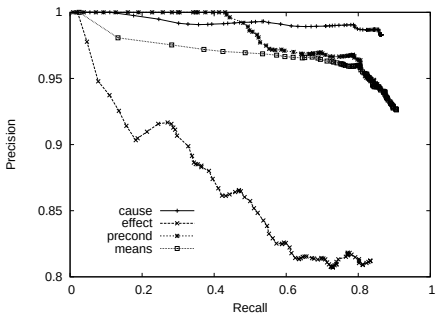

Fig. 3. Recall-precision curves (causal relation classification)

i. All the features shown in Table 5,
ii. The volitionality value estimated by the technique described in the previous section, and
iii. Whether the subjects of the two clauses in the sentence are the same.

The third, subjects agreement feature can be automatically extracted by using the technique described in Nakaiwa et al. [12] with a high level of precision. However, in this experiment, we were unable to implement this method. Instead, a simple rule-based extractor was used.

The data are the same as those in Section 6.2. We used the sentences in $\mathcal{S}_1$ as training samples and $\mathcal{S}_2$ as test samples. We first estimated the volitionality value and its reliability using all the data. Then, we removed about 20% of the samples by applying to the reliability metric.

The one-versus-rest method was used so that we could apply SVMs to multiple classifications. When the discriminate function value acquired from two or more classifiers with this technique was positive, one classifier with the maximum function value was ultimately selected.

7.2 Results

We refer to the maximum discriminate function value obtained through the one-versus-rest method as s_1, and to the second highest one as s_2. We then obtained the results shown in Table 8 and Figure 3 [3] through the same procedure as described for reliability in Section 6.2, where the classification reliability was

[3] For each relation R:
Recall = # of samples correctly classified as R / # of samples holding the target relation R,
Precision = # of samples correctly classified as R / # of samples output as being R.

Table 7. Distribution of causal relations held by *tame*-complex sentences in $\mathcal{S}_2$

class	Most frequent relation and its ratio	
A	*cause*(SOA_s,SOA_m)	0.98 (193/196)
B	*effect*(Act_s,SOA_m)	0.78 (108/139)
C	*precond*(SOA_s,Act_m)	0.94 (166/176)
D	*means*(Act_m,Act_s)	0.79 (375/474)
		0.85 (842/985)

Table 8. Accuracy of causal relation classification

	3-point averaged precision			
	cause	*effect*	*precond*	*means*
With volitionality	0.992	0.859	0.989	0.984
Upper bound	0.996	0.882	0.993	0.988
Without volitionality	0.769	0.588	0.943	0.722

defined as $s_1 + (s_1 - s_2)$. The 3-point averaged precision in Table 8 represents the summary of the recall-precision curves. This value is the 3-point average of precision where the 3 points are recall = 0.25, 0.50, 0.75.

The first row of Table 8 shows that our causal relation classifier performed with high precision. All relations excluding *effect* relation class achieved over 0.95. The second row shows the current upper bound of causal classification. These are the results in the case that classifiers were trained with the feature information for the two primitive features, the subject feature and the subjects agreement feature, by using a human judge instead of our simple feature extractor in an effort to avoid machine-induced errors in input data. The third row shows the results in the case that classifiers were trained without the volitionality values. It is clear that clausal volitionality plays an important role in classifying causal relations.

7.3 Discussion

Let us estimate the amount of knowledge one can acquire from *tame*-complex sentences in a collection of one year of newspaper articles with approximately 1,500,000 sentences in total.

Suppose that we want to acquire causal relations with a precision of, say, 99% for *cause* relation, 95% for *precond* and *means* relations, and 90% for *effect* relation. First, it can be seen from Figure 3 that we achieved 79% recall (REC) for the *cause* relation, 30% for *effect*, 82% for *precond*, and 83% for *means*. Second, assume that the frequency ratios (FR) of these relations to all the *tame*-complex sentences are as given in Table 7. In this case, for example, the frequency ratio of the *cause* relation class was 193/1000 = 19%. From these, it can be seen

that we achieved 64% recall: $0.19^{cause}_{\mathrm{FR}} \times 0.79^{cause}_{\mathrm{REC}} + 0.11^{effect}_{\mathrm{FR}} \times 0.30^{effect}_{\mathrm{REC}} + 0.17^{precond}_{\mathrm{FR}} \times 0.82^{precond}_{\mathrm{REC}} + 0.38^{means}_{\mathrm{FR}} \times 0.83^{means}_{\mathrm{REC}} = 0.64$.

Finally, since we collected about 42,500 *tame*-complex sentences from one year of newspaper articles (see Table 3), we expect to acquire over 27,000 instances of causal relations ($\simeq 42,500 \times 0.64$). This number accounts for 1.8% of all the sentences (1,500,000 sentences), and is not small in comparison to number of causal instances included in the Open Mind Commonsense knowledge base [15] and Marcu's results [11].

8 Conclusion

Through our approach to acquiring causal knowledge from text, we made the following findings:

- If one can determine the volitionality of the subordinate and matrix clauses of a Japanese *tame*-complex sentence, the causal relation can be classified as *cause*, *effect*, *precond* or *means* with a precision of over 85% on average (Table 4 and Table 7).
- By using SVMs, we achieved 80% recall with over 95% precision for the *cause*, *precond* and *means* relations, and 30% recall with 90% precision for the *effect* relation (Figure 3).
- The classification results indicate that over 27,000 instances of causal relations can be acquired from one year of Japanese newspaper articles.

In future work, we will extend the connective markers covered to include frequent connective markers such as *ga* (but) and *re-ba* (if). More importantly, what we have discussed in this paper is not specific to Japanese, so we want to investigate application to English connectives as well. We also plan to design a computational model for applying the acquired knowledge to natural language understanding and discourse understanding.

Acknowledgements

We would like to express our special thanks to the creators of Nihongo-Goi-Taikei and several of the dictionaries used in the ALT-J/E translation system at NTT Communication Science Laboratories, and the EDR electronic dictionaries produced by Japan Electronic Dictionary Research Institute. We would also like to thank Nihon Keizai Shimbun, Inc. for allowing us to use their newspaper articles. We are grateful to the reviewers for their suggestive comments, Taku Kudo for providing us with his dependency analyzer and SVM tools, and Eric Nichols and Campbell Hore for proofreading.

References

1. D. Garcia. COATIS, an NLP system to locate expressions of actions connected by causality links. In *Proc. of the 10th European Knowledge Acquisition Workshop*, pages 347–352, 1997.

2. R. Girju and D. Moldovan. Mining answers for causation questions. In *Proc. the AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 2002.
3. J. R. Hobbs, M. Stickel, D. Appelt, and P. Martion. Interpretation as abduction. *Artificial Intelligence*, 63:69–142, 1993.
4. S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. *Goi-Taikei - A Japanese Lexicon*. Iwanami Shoten, 1997.
5. S. Ikehara, S. Shirai, A. Yokoo, and H. Nakaiwa. Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–. In *Third Machine Translation Summit: MT Summit III*, pages 101–106, Washington DC, 1991.
6. L. M. Iwanska and S. C. Shapiro. *Natural Language Processing and Knowledge Representation - Language for Knowledge and Knowledge for Language*. The MIT Press, 2000.
7. C. S. G. Khoo, S. Chan, and Y. Niu. Extracting causal knowledge from a medical database using graphical patterns. In *Proc. of the 38th. Annual Meeting of the Association for Computational Linguistics (ACL2000)*, pages 336–343, 2000.
8. D. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11), 1995.
9. H. Liu, H. Lieberman, and T. Selker. A model of textual affect sensing using real-world knowledge. In *Proc. of the International Conference on Intelligent User Interfaces*, pages 125–132, 2003.
10. W. C. Mann and S. A. Thompson. Rhetorical structure theory: A theory of text organization. In *USC Information Sciences Institute, Technical Report ISI/RS-87-190*, 1987.
11. D. Marcu. An unsupervised approach to recognizing discourse relations. In *Proc. of the 40th. Annual Meeting of the Association for Computational Linguistics (ACL2002)*, pages 368–375, 2002.
12. H. Nakaiwa and S. Ikehara. Intrasentential resolution of japanese zero pronouns in a machine translation system using semantic and pragmatic constraints. In *Proc. of the 6th TMI*, pages 96–105, 1995.
13. H. Satou, K. Kasahara, and K. Matsuzawa. Retrieval of simplified causal knowledge in text and its application. In *Proc. of The IEICE,Thought and Language*, 1998. (in Japanese).
14. R. Schank and R. Abelson. *Scripts Plans Goals and Understanding*. Lawrence Erlbaum Associates, 1977.
15. P. Singh. The public acquisition of commonsense knowledge. In *Proc. of AAAI Spring Symposium on Acquiring Linguistic Knowledge for Information Access*, 2002.
16. D. G. Stork. Character and document research in the open mind initiative. In *Proc. of Int. Conf. on Document Analysis and Recognition*, pages 1–12, 1999.
17. V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
18. T. Yokoi. The edr electronic dictionary. *Communications of the ACM*, 38(11):42–44, 1995.