Why-Question Answering using Intra- and Inter-Sentential Causal Relations

Jong-Hoon Oh* **Kentaro Torisawa†** **Chikara Hashimoto ‡** **Motoki Sano§**
Stijn De Saeger¶ **Kiyonori Ohtake‖**

Information Analysis Laboratory
Universal Communication Research Institute
National Institute of Information and Communications Technology (NICT)
{*rovellia,†torisawa,‡ch,§msano,¶stijn,‖kiyonori.ohtake}@nict.go.jp

Abstract

In this paper, we explore the utility of *intra-* and *inter-sentential causal relations* between terms or clauses as evidence for answering why-questions. To the best of our knowledge, this is the first work that uses both intra- and inter-sentential causal relations for why-QA. We also propose a method for assessing the appropriateness of causal relations as answers to a given question using the semantic orientation of *excitation* proposed by Hashimoto et al. (2012). By applying these ideas to Japanese why-QA, we improved precision by 4.4% against all the questions in our test set over the current state-of-the-art system for Japanese why-QA. In addition, unlike the state-of-the-art system, our system could achieve very high precision (83.2%) for 25% of all the questions in the test set by restricting its output to the confident answers only.

1 Introduction

"Why-question answering" (why-QA) is a task to retrieve answers from a given text archive for a why-question, such as "Why are tsunamis generated?" The answers are usually text fragments consisting of one or more sentences. Although much research exists on this task (Girju, 2003; Higashinaka and Isozaki, 2008; Verberne et al., 2008; Verberne et al., 2011; Oh et al., 2012), its performance remains much lower than that of the state-of-the-art factoid QA systems, such as IBM's Watson (Ferrucci et al., 2010).

In this work, we propose a quite *straightforward* but *novel* approach for such difficult why-QA task. Consider the sentence **A1** in Table 1, which represents the causal relation between the cause, "the ocean's water mass ..., waves are generated," and its effect, "Tsunamis ... are generated." This is a good answer to the question, "Why are tsunamis generated?", since the effect part is more or less equivalent to the (propositional) content of the question. Our method finds text fragments that include such causal relations with an effect part that resembles a given question and provides them as answers.

A1	[Tsunamis that can cause large coastal inundation are generated]$_{effect}$ <u>because</u> [the ocean's water mass is displaced and, much like throwing a stone into a pond, waves are generated.]$_{cause}$
A2	[Earthquake causes seismic waves which set up the water in motion with a large force.]$_{cause}$ <u>This causes</u> [a tsunami.]$_{effect}$
A3	[Tsunamis]$_{effect}$ <u>are caused by</u> [the sudden displacement of huge volumes of water.]$_{cause}$
A4	[Tsunamis weaken as they pass through forests]$_{effect}$ <u>because</u> [the hydraulic resistance of the trees diminish their energy.]$_{cause}$
A5	[Automakers in Japan suspended production for an array of vehicles]$_{effect}$ <u>because</u> [the magnitude 9 earthquake and tsunami hit their country on Friday, March 11, 2011.]$_{cause}$

Table 1: Examples of intra/inter-sentential causal relations. Cause and effect parts of each causal relation, marked with [..]$_{cause}$ and [..]$_{effect}$, are connected by the underlined cue phrases for causality, such as *because*, *this causes*, and *are caused by*.

Since this idea looks quite intuitive, many people would probably consider it as a solution to why-QA. However, to our surprise, we could not find any previous work on why-QA that took this approach. Some methods utilized the causal relations between terms as evidence for finding answers (i.e., matching a cause term with an answer text and its effect term with a question) (Girju, 2003; Higashinaka and Isozaki, 2008). Other approaches utilized such clue terms for causality as "because" as evidence for finding answers (Murata et al., 2007). However, these algorithms did not check whether an answer candidate, i.e., a text fragment that may be provided as an answer, explicitly contains a complex causal relation sen-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1733–1743,
Sofia, Bulgaria, August 4-9 2013.

tence with the effect part that resembles a question. For example, **A5** in Table 1 is an incorrect answer to "Why are tsunamis generated?", but these previous approaches would probably choose it as a proper answer due to "because" and "earthquake" (i.e., a cause of tsunamis). At least in our experimental setting, our approach outperformed these simpler causality-based QA systems.

Perhaps this approach was previously deemed infeasible due to two non-trivial technical challenges. The first challenge is to accurately identify a wide range of causal relations like those in Table 1 in answer candidates. To meet this challenge, we developed a sequence labeling method that identifies not only *intra-sentential causal relations*, i.e., the causal relations between two terms/phrases/clauses expressed in a single sentence (e.g., **A1** in Table 1), but also the *inter-sentential causal relations*, which are the causal relations between two terms/phrases/clauses expressed in two adjacent sentences (e.g., **A2**) in a given text fragment.

The second challenge is assessing the appropriateness of each identified causal relation as an answer to a given question. This is important since the causal relations identified in the answer candidates may have nothing to do with a given question. In this case, we have to reject these causal relations because they are inappropriate as an answer to the question. When a single answer candidate contains many causal relations, we also have to select the appropriate ones. Consider the causal relations in **A1–A4**. Those in **A1–A3** are appropriate answers to "Why are tsunamis generated?", but not the one in **A4**. To assess the appropriateness, the system must recognize *textual entailment*, i.e., "tsunamis (are) generated" in the question is entailed by all "tsunamis are generated" in **A1**, "cause a tsunami" in **A2** and "tsunamis are caused" in **A3** but not by "tsunamis weaken" in **A4**. This quite difficult task is currently being studied by many researchers in the RTE field (Androutsopoulos and Malakasiotis, 2010; Dagan et al., 2010; Shima et al., 2011; Bentivogli et al., 2011). To meet this challenge, we developed a relatively simple method that can be seen as a lightweight approximation for this difficult RTE task, using excitation polarities (Hashimoto et al., 2012).

Through our experiments on Japanese why-QA, we show that a combination of the above methods can improve why-QA accuracy. In addition, our proposed method can be successfully combined with other approaches to why-QA and can contribute to higher accuracy. As a final result, we improved the precision by 4.4% against all the questions in our test set over the current state-of-the-art system of Japanese why-QA (Oh et al., 2012). The difference in the performance became much larger when we only compared the highly confident answers of each system. When we made our system provide only its confident answers according to their confidence score given by our system, the precision of these confident answers was 83.2% for 25% of all the questions in our test set. In the same setting, the precision of the state-of-the-art system (Oh et al., 2012) was only 62.4%.

2 Related Work

Although there were many previous works on the acquisition of intra- and inter-sentential causal relations from texts (Khoo et al., 2000; Girju, 2003; Inui and Okumura, 2005; Chang and Choi, 2006; Torisawa, 2006; Blanco et al., 2008; De Saeger et al., 2009; De Saeger et al., 2011; Riaz and Girju, 2010; Do et al., 2011; Radinsky et al., 2012), their application to why-QA was limited to causal relations between terms (Girju, 2003; Higashinaka and Isozaki, 2008).

As previous attempts to improve why-QA performance, such semantic knowledge as WordNet synsets (Verberne et al., 2011), semantic word classes (Oh et al., 2012), sentiment analysis (Oh et al., 2012), and causal relations between terms (Girju, 2003; Higashinaka and Isozaki, 2008) has been used. These previous studies took basically bag-of-words approaches and used the semantic knowledge to identify certain semantic associations using terms and n-grams. On the other hand, our method explicitly identifies intra- and inter-sentential causal relations between terms/phrases/clauses that have complex structures and uses the identified relations to answer a why-question. In other words, our method considers more complex linguistic structures than those used in the previous studies. Note that our method can complement the previous approaches. Through our experiments, we showed that it is possible to achieve a higher precision by combining our proposed method with bag-of-words approaches considering semantic word classes and sentiment analysis in our previous work (Oh et al.,

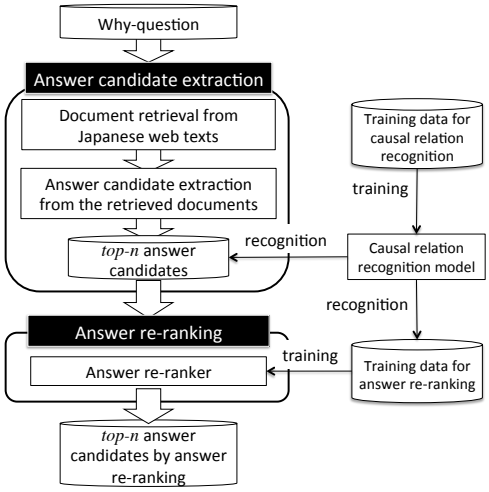

Figure 1: System architecture

2012).

3 System Architecture

We first describe the system architecture of our QA system before describing our proposed method. It is composed of two components: answer candidate extraction and answer re-ranking (Fig. 1). This architecture is basically the same as that used in our previous work (Oh et al., 2012). We extended our previous work by introducing causal relations recognized from answer candidates to the answer re-ranking. The features used in our previous work are very different from those in this work, and we found that combining both improves accuracy.

Answer candidate extraction: In our previous work, we implemented the method of Murata et al. (2007) for our answer candidate extractor. We retrieved documents from Japanese web texts using *Boolean AND* and *OR* queries generated from the content words in why-questions. Then we extracted passages of five sentences from these retrieved documents and ranked them with the ranking function proposed by Murata et al. (2007). This method ranks a passage higher when it contains more query terms that are closer to each other in the passage. We used a set of clue terms, including the Japanese counterparts of *cause* and *reason*, as query terms for the ranking. The top ranked passages are regarded as *answer candidates* in the answer re-ranking. See Murata et al. (2007) for more details.

Answer re-ranking: Re-ranking the answer candidates is done by a supervised classifier (SVMs) (Vapnik, 1995). In our previous work, we employed three types of features for training the re-ranker: morphosyntactic features (n-grams of morphemes and syntactic dependency chains), semantic word class features (semantic word classes obtained by automatic word clustering (Kazama and Torisawa, 2008)) and sentiment polarity features (word and phrase polarities). Here, we used semantic word classes and sentiment polarities for identifying such semantic associations between a why-question and its answer as "if a disease's name appears in a question, then answers that include nutrient names are more likely to be correct" by semantic word classes and "if something undesirable happens, the reason is often also something undesirable" by sentiment polarities. In this work, we propose causal relation features generated from intra- and inter-sentential causal relations in answer candidates and use them along with the features proposed in our previous work for training our re-ranker.

4 Causal Relations for Why-QA

We describe causal relation recognition in Section 4.1 and describe the features (of our re-ranker) generated from causal relations in Section 4.2.

4.1 Causal Relation Recognition

We restrict causal relations to those expressed by such cue phrases for causality as (the Japanese counterparts of) *because* and *as a result* like in the previous work (Khoo et al., 2000; Inui and Okumura, 2005) and recognize them in the following two steps: extracting causal relation candidates and recognizing causal relations from these candidates.

4.1.1 Extracting Causal Relation Candidates

We identify cue phrases for causality in answer candidates using the regular expressions in Table 2. Then, for each identified cue phrase, we extract three sentences as a causal relation candidate, where one contains the cue phrase and the other two are the previous and next sentences in the answer candidates. When there is more than one cue phrase in an answer candidate, we use all of them for extracting the causal relation candidates, assuming that each of the cue phrases is linked to different causal relations. We call a cue phrase used for extracting a causal relation candidate a *c-marker* (*causality marker*) of the candidate to distinguish it from the other cue phrases in the same causal relation candidate.

Regular expressions	Examples
(D|の)? ため P?	ため (for), のため (for), そのため (as a result), のために (for)
ので	ので (since or because of)
こと (から|で)	ことから (from the fact that), ことで (by the fact that)
(から|ため) C	からだ (because), ためだ (It is because)
D? RCT (P|C)+	理由は (the reason is), 原因だ (is the cause),この理由から (from this reason)

Table 2: Regular expressions for identifying cue phrases for causality. D, P and C represent demonstratives (e.g., この (this) and その (that)), postpositions (including case markers such as が (nominative), の (genitive)), and copula (e.g., です (is) and である (is)) in Japanese, respectively. RCT, which represents Japanese terms meaning *reason*, *cause*, or *thanks to*, is defined as follows: RCT = {理由 (reason), 原因 (cause), 要因 (cause), 引き金 (cause), おかげ (thanks to), せい (thanks to), わけ (reason) }.

4.1.2 Recognizing Causal Relations

Next, we recognize the spans of the cause and effect parts of a causal relation linked to a c-marker. We regard this task as a sequence labeling problem and use Conditional Random Fields (CRFs) (Lafferty et al., 2001) as a machine learning framework. In our task, CRFs take three sentences of a causal relation candidate as input and generate their cause-effect annotations with a set of possible cause-effect IOB labels, including Begin-Cause (B-C), Inside-Cause (I-C), Begin-Effect (B-E), Inside-Effect (I-E), and Outside (O). Fig 2 shows an example of such sequence labeling. Although this example is about sequential labeling shown on English sentences for ease of explanation, it was actually done on Japanese sentences.

We used the three types of feature sets in Table 3 for training the CRFs, where j is in the range of $i-4 \leq j \leq i+4$ for current position i in a causal relation candidate.

Type	Features
Morphological feature	$m_j, m_j^{j+1}, pos_j, pos_j^{j+1}$
Syntactic feature	$s_j, s_j^{j+1}, b_j, b_j^{j+1}$
C-marker feature	$(m_j, cm), (m_j^{j+1}, cm)$
	$(s_j, cm), (s_j^{j+1}, cm)$

Table 3: Features for training CRFs, where $x_j^{j+1} = x_j x_{j+1}$

Morphological features: m_j and pos_j in Table 3 represent the j^{th} morpheme and the POS tag.

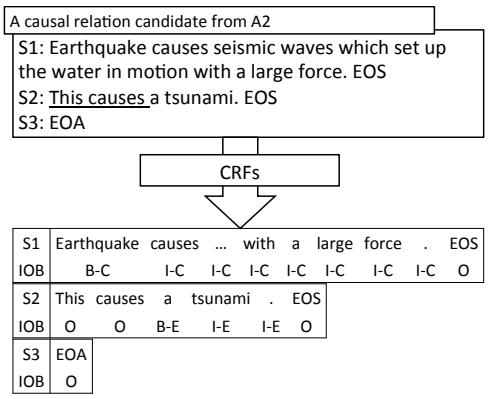

Figure 2: Recognizing causal relations by sequence labeling: Underlined text *This causes* represents a c-marker, and EOS and EOA represent *end-of-sentence* and *end-of-answer candidates*.

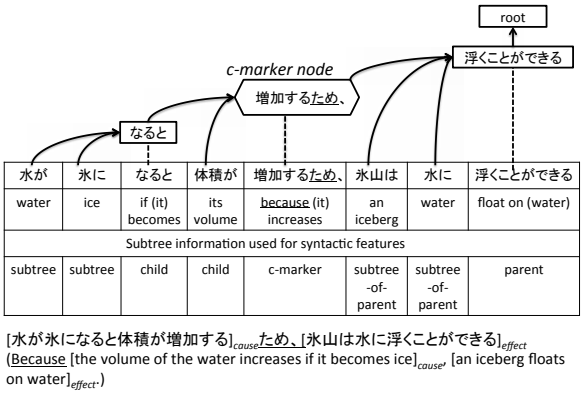

Figure 3: Example of syntactic information related to a c-marker used for syntactic features

We use JUMAN[1], a Japanese morphological analyzer, for generating our morphological features.

Syntactic features: The span of the causal relations in a given causal relation candidate strongly depends on the c-marker in the candidate. Especially for intra-sentential causal relations, their cause and effect parts often appear in the subtrees of the c-marker's node or those of the c-marker's parent node in a syntactic dependency tree structure. Fig. 3 shows an example that follows this observation, where the c-marker node is represented in a hexagon and the other nodes are in a rectangle. Note that each node in Fig. 3 is a word phrase (called *a bunsetsu*), which is the smallest unit of syntactic analysis in Japanese. A *bunsetsu* is a syntactic constituent composed of a content word and several function words such as postpositions and case markers. Syntactic dependency is represented by an arrow in Fig. 3. For example, there is syntactic dependency from word phrase 水が

[1] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

(water) to なると (if (it) becomes), i.e., 水が $\xrightarrow{dep}$ なると. We encode this subtree information into s_j, which is the syntactic information of a word phrase to which the j^{th} morpheme belongs. s_j only has one of six values: 1) the c-marker's node (*c-marker*), 2) the c-marker's child node (*child*), 3) the c-marker's parent node (*parent*), 4) in the c-marker's subtree but not the c-marker's child node (*subtree*), 5) in the subtree of the c-marker's parent node but not the c-marker's node (*subtree-of-parent*) and 6) the others (*others*). b_j is the word phrase information of the j^{th} morpheme (m_j) that represents whether m_j is in the beginning or inside a word phrase. For generating our syntactic features, we use KNP[2], a Japanese syntactic dependency parser.

C-marker features: As our c-marker features, we use a pair composed of c-marker cm and one of the following: m_j, m_j^{j+1}, s_j, or s_j^{j+1}.

4.2 Causal Relation Features

We use terms, partial trees (in a syntactic dependency tree structure), and the semantic orientation of excitation (Hashimoto et al., 2012) to assess the appropriateness of each causal relation obtained by our causal relation recognizer as an answer to a given question. Finding answers with term matching and partial tree matching has been used in the literature of question answering (Girju, 2003; Narayanan and Harabagiu, 2004; Moschitti et al., 2007; Higashinaka and Isozaki, 2008; Verberne et al., 2008; Surdeanu et al., 2011; Verberne et al., 2011; Oh et al., 2012), while that with the excitation polarity is proposed in this work.

We use three types of features. Each feature type expresses the causal relations in an answer candidate that are determined to be appropriate as answers to a given question by term matching (tf_1–tf_4), partial tree matching (pf_1–pf_4) and excitation polarity matching (ef_1–ef_4). We call these causal relations used for generating our causal relation features *candidates of an appropriate causal relation* in this section. Note that if one answer candidate has more than one candidate of an appropriate causal relation found by one matching method, we generated features for each appropriate candidate and merged all of them for the answer candidate.

[2] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

Type	Description
tf_1	word n-grams of causal relations
tf_2	word class version of tf_1
tf_3	indicator for the existence of candidates of an appropriate causal relation identified by term matching in an answer candidate
tf_4	number of matched terms in candidates of an appropriate causal relation
pf_1	syntactic dependency n-grams (n dependency chain) of causal relations
pf_2	word class version of pf_1
pf_3	indicator for the existence of candidates of an appropriate causal relation identified by partial tree matching in an answer candidate
pf_4	number of matched partial trees in candidates of an appropriate causal relation
ef_1	types of noun-polarity pairs shared by causal relations and the question
ef_2	ef_1 coupled with each noun's word class
ef_3	indicator for the existence of candidates of an appropriate causal relation identified by excitation polarity matching in an answer candidate
ef_4	number of noun-polarity pairs shared by the question and the candidates of an appropriate causal relation

Table 4: Causal relation features: n in n-grams is $n = \{2, 3\}$ and n-grams in an effect part are distinguished from those in a cause part.

4.2.1 Term Matching

Our term matching method judges that a causal relation is a candidate of an appropriate causal relation if its effect part contains at least one content word (nouns, verbs, and adjectives) in the question. For example, all the causal relations of **A1–A4** in Table 1 are candidates of an appropriate causal relation to the question, "Why is a tsunami generated?", by term matching with question term *tsunami*.

tf_1–tf_4 are generated from candidates of an appropriate causal relation identified by term matching. The n-grams of tf_1 and tf_2 are restricted to those containing at least one content word in a question. We distinguish this matched word from the other words by replacing it with QW, a special symbol representing a word in the question. For example, word 3-gram "this/cause/QW" is extracted from *This causes tsunamis* in **A2** for "Why is a tsunami generated?" Further, we create a word class version of word n-grams by converting the words in these word n-grams into their corresponding word class using the semantic word classes (500 classes for 5.5 million nouns) from our previous work (Oh et al., 2012). These word classes were created by applying the automatic word clustering method of Kazama and Torisawa (2008) to 600 million Japanese web pages. For example, the word class version of word 3-gram

"this/cause/QW" is "this/cause/QW,WC$_{tsunami}$", where WC$_{tsunami}$ represents the word class of a *tsunami*. tf_3 is a binary feature that indicates the existence of candidates of an appropriate causal relation identified by term matching in an answer candidate. tf_4 represents the degree of the relevance of the candidates of an appropriate causal relation measured by the number of matched terms: one, two, and more than two.

4.2.2 Partial Tree Matching

Our partial tree matching method judges a causal relation as a candidate of an appropriate causal relation if its effect part contains at least one partial tree in a question, where the partial tree covers more than one content word. For example, only the causal relation **A1** among **A1–A4** is a candidate of an appropriate causal relation for question "Why are tsunamis generated?" by partial tree matching because only its effect part contains partial tree "tsunamis $\xrightarrow{dep}$ (are) generated" of the question.

pf_1–pf_4 are generated from candidates of an appropriate causal relation identified by the partial tree matching. The syntactic dependency n-grams in pf_1 and pf_2 are restricted to those that contain at least one content word in a question. We distinguish this matched content word from the other content words in the n-gram by converting it to QW, which represents a content word in the question. For example, syntactic dependency 2-gram "QW $\xrightarrow{dep}$ cause" and its word class version "QW,WC$_{tsunami}$ $\xrightarrow{dep}$ cause" are extracted from *Tsunamis that can cause* in **A1**. pf_3 is a binary feature that indicates whether an answer candidate contains candidates of an appropriate causal relation identified by partial tree matching. pf_4 represents the degree of the relevance of the candidate of an appropriate causal relation measured by the number of matched partial trees: one, two, and more than two.

4.2.3 Excitation Polarity Matching

Hashimoto et al. (2012) proposed a semantic orientation called *excitation polarities*. It classifies predicates with their argument position (called *templates*) into *excitatory*, *inhibitory* and *neutral*. In the following, we denote a template as "[*argument position,predicate*]." According to Hashimoto's definition, excitatory templates imply that the function, effect, purpose, or the role of an entity filling an argument position in the templates is activated/enhanced. On the contrary, inhibitory templates imply that the effect, purpose or the role of an entity is deactivated/suppressed. Neutral templates are those that neither activate nor suppress the function of an argument.

We assume that the *meanings* of a text can be *roughly* captured by checking whether each noun in the text is *activated* or *suppressed* in the sense of the excitation polarity framework, where the activation and suppression of each entity (or noun) can be detected by looking at the excitation polarities of the templates that are filled by the entity. For instance, effect part "tsunamis that can cause large coastal inundation are generated" of **A1** roughly means that "tsunamis" are *activated* and "inundation" is (or can be) activated. This activation/suppression configuration of the nouns is *consistent* with sentence "tsunamis are caused" in which "tsunamis" are activated. This consistency suggests that **A1** is a good answer to question "Why are tsunamis caused?", although the "tsunamis" are modified by different predicates; "cause" and "generate." On the other hand, effect part "tsunamis weaken as they pass through forests" of **A4** implies that "tsunamis" are suppressed. This suggests that **A4** is not a good answer to "Why are tsunamis caused?" Note that the consistency checking between activation/suppression configurations of nouns[3] in texts can be seen as a rough but lightweight approximation of the recognition of textual entailments or paraphrases.

Following the definition of excitation polarity in Hashimoto et al. (2012), we manually classified templates[4] to each polarity type and obtained 8,464 excitatory templates, such as [が, 増える] ([subject, increase]) and [が, 向上する] ([subject, improve]), 2,262 inhibitory templates, such as [を, 防ぐ] ([object, prevent]) and [が, 死ぬ] ([subject, die]), and 7,230 neutral templates such as [を, 考える] ([object, consider]). With these templates, we obtain activation/suppression configurations (including neutral) for the nouns in the causal relations in the answer candidates and ques-

[3] Because the activation/suppression configurations of nouns come from an excitation polarity of templates, "[*argument position,predicate*]," the semantics of verbs in the templates are implicitly considered in this consistency checking.

[4] Varga et al. (2013) has used the same templates as ours, except they restricted their excitation/inhibitory templates to those whose polarity is consistent with that given by the automatic acquisition method of Hashimoto et al. (2012).

tions.

Next, we assume that a causal relation is appropriate as an answer to a question if the effect part of the causal relation and the question share at least one common noun with the same polarity. More detailed information concerning the configurations of all the nouns in all the candidates of an *appropriate* causal relation (including their cause parts) and the question are encoded into our feature set ef_1–ef_4 in Table 4 and the final judgment is done by our re-ranker.

For generating ef_1 and ef_2, we classified all the nouns coupled with activation/suppression/neutral polarities in a causal relation into three types: SAME (the question contains the same noun with the same polarity), DiffPOL (the question contains the same noun with different polarity), and OTHER (the others). ef_1 indicates whether each type of noun-polarity pair exists in a causal relation. Note that the types for the effect and cause parts are represented in distinct features. ef_2 is the same as ef_1 except that the types are augmented with the word classes of the corresponding nouns. In other words, ef_2 indicates whether each type of noun-polarity pair exists in the causal relation for each word class. ef_3 indicates the existence of candidates of an appropriate causal relation identified by this matching scheme, and ef_4 represents the number of noun-polarity pairs shared by the question and the candidates of an appropriate causal relations (one, two, and more than two).

5 Experiments

We experimented with causal relation recognition and why-QA with our causal relation features.

5.1 Data Set for Why-Question Answering

For our experiments, we used the same why-QA data set as the one used in our previous work (Oh et al., 2012). This why-QA data set is composed of 850 Japanese why-questions and their top-20 answer candidates obtained by answer candidate extraction from 600 million Japanese web pages. Three annotators checked the top-20 answer candidates of these 850 questions and the final judgment was made by their majority vote. Their inter-rater agreement by Fleiss' kappa reported in Oh et al. (2012) was substantial (κ = 0.634). Among the 850 questions, 250 why-questions were extracted from the Japanese version of *Yahoo! Answers*, and another 250 were created by annotators. In our previous work, we evaluated the system with these 500 questions and their answer candidates as training and test data in 10-fold cross-validation. The other 350 why-questions were manually built from passages describing the causes or reasons of events/phenomena. These questions and their answer candidates were used as additional training data for testing subsamples in each fold during the 10-fold cross-validation. In our why-QA experiments, we evaluated our why-QA system with the same settings.

5.2 Data Set for Causal Relation Recognition

We built a data set composed of manually annotated causal relations for evaluating our causal relation recognition. As source data for this data set, we used the same 10-fold data that we used for evaluating our why-QA (500 questions and their answer candidates). We extracted the causal relation candidates from the answer candidates in each fold, and then our annotator (not an author) manually marked the span of the cause and effect parts of a causal relation for each causal relation candidate, keeping in mind that the causal relation must be expressed in terms of a c-marker in a given causal relation candidate. Finally, we had a data set made of 16,051 causal relation candidates, 8,117 of which had a true causal relation; the number of intra- and inter-sentential causal relations were 7,120 and 997, respectively.

Note that this data set can be partitioned into ten folds by using the 10-fold partition of its source data. We performed 10-fold cross validation to evaluate our causal relation recognition with this 10-fold data.

5.3 Causal Relation Recognition

We used CRF++[5] for training our causal relation recognizer. In our evaluation, we judged a system's output as correct if both spans of the cause and effect parts overlapped those in the gold standard. Evaluation was done by precision, recall, and F_1.

	Precision	Recall	F_1
BASELINE	41.9	61.0	49.7
INTRA-SENT	84.5	75.4	79.7
INTER-SENT	80.2	52.6	63.6
ALL	83.8	71.1	77.0

Table 5: Results of causal relation recognition (%)

Table 5 shows the result. BASELINE represents

[5] http://code.google.com/p/crfpp/

the result for our baseline system that recognizes a causal relation by simply taking the two phrases adjacent to a c-marker (i.e., before and after) as cause and effect parts of the causal relation. We assumed that the system had an oracle for judging correctly whether each phrase is a cause part or an effect part. In other words, we judged that a causal relation recognized by BASELINE is correct if both cause and effect parts in the gold standard are adjacent to a c-marker. INTRA-SENT and INTER-SENT represent the results for intra- and inter-sentential causal relations and ALL represents the result for the both causal relations by our method. From these results, we confirmed that our method recognized both intra- and inter-sentential causal relations with over 80% precision, and it significantly outperformed our baseline system in both precision and recall rates.

	Precision	Recall	F_1
ALL-"MORPH"	80.8	66.4	72.9
ALL-"SYNTACTIC"	82.9	67.0	74.1
ALL-"C-MARKER"	76.3	51.4	61.4
ALL	**83.8**	**71.1**	**77.0**

Table 6: Ablation test results for causal relation recognition (%)

We also investigated the contribution of the three types of features used in our causal relation recognition to the performance. We evaluated the performance when we removed one of the three types of features (ALL-"MORPH", ALL-"SYNTACTIC" and ALL-"C-MARKER") and compared the results in these settings with the one when all the feature sets were used (ALL). Table 6 shows the result. We confirmed that all the feature sets improved the performance, and we got the best performance when using all of them. We used the causal relations obtained from the 10-fold cross validation for our why-QA experiments.

5.4 Why-Question Answering

We performed why-QA experiments to confirm the effectiveness of intra- and inter-sentential causal relations in a why-QA task. In this experiment, we compared five systems: four baseline systems (MURATA, OURCF, OH and OH+PREVCF) and our proposed method (PROPOSED).

MURATA corresponds to our answer candidate extraction.

OURCF uses a re-ranker trained with only our causal relation features.

OH, which represents our previous work (Oh et al., 2012), has a re-ranker trained with morphosyntactic, semantic word class, and sentiment polarity features.

OH+PREVCF is a system with a re-ranker trained with the features used in OH and with the causal relation feature proposed in Higashinaka and Isozaki (2008). The causal relation feature includes an indicator that determines whether the causal relations between two terms appear in a question-answer pair; cause in an answer and its effect in a question. We acquired the causal relation instances (between terms) from 600 million Japanese web pages using the method of De Saeger et al. (2009) and exploited the top-100,000 causal relation instances in this system.

PROPOSED has a re-ranker trained with our causal relation features as well as the three types of features proposed in Oh et al. (2012). Comparison between OH and PROPOSED reveals the contribution of our causal relation features to why-QA.

We used TinySVM[6] with a linear kernel for training the re-rankers in OURCF, OH, OH+PREVCF and PROPOSED. Evaluation was done by P@1 (Precision of the top-answer) and Mean Average Precision (MAP); they are the same measures used in Oh et al. (2012). P@1 measures how many questions have a correct top-answer candidate. MAP measures the overall quality of the top-20 answer candidates. As mentioned in Section 5.1, we used 10-fold cross-validation with the same setting as the one used in Oh et al. (2012) for our experiments.

	P@1	MAP
MURATA	22.2	27.0
OURCF	27.8	31.4
OH	37.4	39.1
OH+PREVCF	37.4	38.9
PROPOSED	**41.8**	**41.0**

Table 7: Why-QA results (%)

Table 7 shows the evaluation results. Our proposed method outperformed the other four systems and improved P@1 by 4.4% over OH, which is the-state-of-the-art system for Japanese why-

[6] http://chasen.org/∼taku/software/TinySVM/

QA. OURCF showed the performance improvement over MURATA. Although this suggests the effectiveness of our causal relation features, the overall performance of OURCF was lower than that of OH. OH+PREVCF outperformed neither OH nor PROPOSED. This suggests that our approach is more effective than previous causality-based approaches (Girju, 2003; Higashinaka and Isozaki, 2008), at least in our setting.

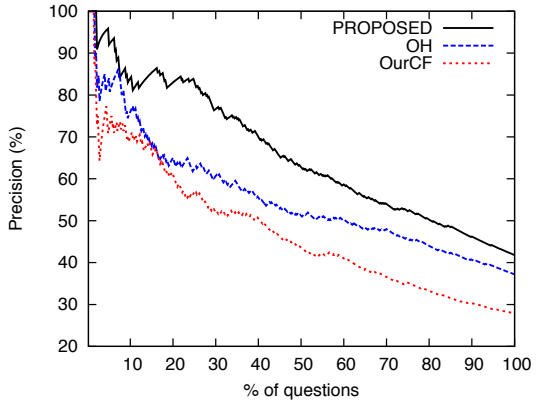

Figure 4: Effect of causal relation features on the top-answers

We also compared *confident* answers of OURCF, OH, and PROPOSED by making each system provide only the k confident top-answers (for k questions) selected by their SVM scores given by each system's re-ranker. This reduces the number of questions that can be answered by a system, but the top-answers become more reliable as k decreases. Fig. 4 shows this result, where the x axis represents the percentage of questions (against all the questions in our test set) whose top-answers are given by each system, and the y axis represents the precision of the top-answers at a certain point on the x axis. When both systems provided top-answers for 25% of all the questions in our test set, our method achieved 83.2% precision, which is much higher than OH's (62.4%). This experiment confirmed that our causal relation features were also effective in improving the quality of the highly confident answers.

However, the high precision by our method was bound to confident answers for a small number of questions, and the difference in the precision between OH and PROPOSED in Fig. 4 became smaller as we considered more answers with lower confidence. We think that one of the reasons is the relatively small coverage of the excitation polarity lexicon, a core resource in our excitation polarity matching. We are planning to enlarge the lexicon to deal with this problem.

Next, we investigated the contribution of the intra- and inter-sentential causal relations to the performance of our method. We used only one of the two types of causal relations for generating causal relation features (INTRA-SENT and INTER-SENT) for training our re-ranker and compared the results in these settings with the one when both were used (ALL (PROPOSED)). Table 8 shows the result. Both intra- and inter-sentential causal relations contributed to the performance improvement.

	P@1	MAP
INTER-SENT	39.0	39.7
INTRA-SENT	40.4	40.5
ALL (PROPOSED)	**41.8**	**41.0**

Table 8: Results with/without intra- and inter-sentential causal relations (%)

We also investigated the contributions of the three types of causal relation features by ablation tests (Table 9). When we do not use the features by excitation polarity matching (ALL-$\{ef_1$–$ef_4\}$), the performance is the worst. This implies that the contribution of excitation polarity matching exceeds the other two.

	P@1	MAP
ALL-$\{tf_1$–$tf_4\}$	40.8	40.7
ALL-$\{pf_1$–$pf_4\}$	41.0	40.9
ALL-$\{ef_1$–$ef_4\}$	39.6	40.5
ALL (PROPOSED)	**41.8**	**41.0**

Table 9: Ablation test results for why-QA (%)

6 Conclusion

In this paper, we explored the utility of intra- and inter-sentential causal relations for ranking answer candidates to why-questions. We also proposed a method for assessing the appropriateness of causal relations as answers to a given question using the semantic orientation of excitation. Through experiments, we confirmed that these ideas are effective for improving why-QA, and our proposed method achieved 41.8% P@1, which is 4.4% improvement over the current state-of-the-art system of Japanese why-QA. We also showed that our system achieved 83.2% precision for its confident answers, when it only provided its confident answers for 25% of all the questions in our test set.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research (JAIR)*, 38(1):135–187.

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa Dang, and Danilo Giampiccolo. 2011. The seventh pascal recognizing textual entailment challenge. In *Proceedings of TAC*.

E. Blanco, N. Castell, and Dan I. Moldovan. 2008. Causal relation extraction. In *Proceedings of LREC'08*.

Du-Seong Chang and Key-Sun Choi. 2006. Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. *Information Processing and Management*, 42(3):662–678.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 16(1):1–17.

Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata. 2009. Large scale relation acquisition using class dependent patterns. In *Proceedings of ICDM '09*, pages 764–769.

Stijn De Saeger, Kentaro Torisawa, Masaaki Tsuchida, Jun'ichi Kazama, Chikara Hashimoto, Ichiro Yamada, Jong Hoon Oh, István Varga, and Yulan Yan. 2011. Relation acquisition using word classes and partial patterns. In *Proceedings of EMNLP '11*, pages 825–835.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of EMNLP '11*, pages 294–303.

David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. 2010. Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pages 76–83.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of EMNLP-CoNLL '12*.

Ryuichiro Higashinaka and Hideki Isozaki. 2008. Corpus-based question answering for why-questions. In *Proceedings of IJCNLP '08*, pages 418–425.

Takashi Inui and Manabu Okumura. 2005. Investigating the characteristics of causal relations in Japanese text. In *In Annual Meeting of the Association for Computational Linguistics (ACL) Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of ACL-08: HLT*, pages 407–415.

Christopher S. G. Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of ACL '00*, pages 336–343.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML '01*, pages 282–289.

Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. 2007. Exploiting syntactic and shallow semantic kernels for question answer classification. In *Proceedings of ACL '07*, pages 776–783.

Masaki Murata, Sachiyo Tsukawaki, Toshiyuki Kanamaru, Qing Ma, and Hitoshi Isahara. 2007. A system for answering non-factoid Japanese questions by using passage retrieval weighted based on type of answer. In *Proceedings of NTCIR-6*.

Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of COLING '04*, pages 693–701.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Takuya Kawada, Stijn De Saeger, Jun'ichi Kazama, and Yiou Wang. 2012. Why question answering using sentiment analysis and word classes. In *Proceedings of EMNLP-CoNLL '12*, pages 368–378.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of WWW '12*, pages 909–918.

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *ICSC '10*, pages 361–368.

Hideki Shima, Hiroshi Kanayama, Cheng wei Lee, Chuan jie Lin, Teruko Mitamura, Yusuke Miyao, Shuming Shi, and Koichi Takeda. 2011. Overview of NTCIR-9 RITE: Recognizing Inference in TExt. In *Proceedings of NTCIR-9*.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2011. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37(2):351–383.

Kentaro Torisawa. 2006. Acquiring inference rules with temporal constraints by using japanese coordinated sentences and noun-verb co-occurrences. In *Proceedings of HLT-NAACL '06*, pages 57–64.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Istvan Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of ACL '13*.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2008. Using syntactic information for improving why-question answering. In *Proceedings of COLING '08*, pages 953–960.

Suzan Verberne, Lou Boves, and Wessel Kraaij. 2011. Bringing why-qa to web search. In *Proceedings of ECIR '11*, pages 491–496.