Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

A Semi-Supervised Learning Approach to Why-Question Answering

Jong-Hoon Oh[*] **Kentaro Torisawa**[†] **Chikara Hashimoto**[‡] **Ryu Iida**[§]
Masahiro Tanaka[¶] **Julien Kloetzer**[‖]
National Institute of Information and Communications Technology (NICT), Kyoto, 619-0289, Japan
{[*]rovellia,[†]torisawa,[‡]ch,[§]ryu.iida,[¶]mtnk,[‖]julien}@nict.go.jp

Abstract

We propose a semi-supervised learning method for improving *why-question answering* (why-QA). The key of our method is to generate training data (question-answer pairs) from causal relations in texts such as "[Tsunamis are generated]$_{effect}$ because [the ocean's water mass is displaced by an earthquake]$_{cause}$." A naive method for the generation would be to make a question-answer pair by simply converting the effect part of the causal relations into a why-question, like "Why are tsunamis generated?" from the above example, and using the source text of the causal relations as an answer. However, in our preliminary experiments, this naive method actually failed to improve the why-QA performance. The main reason was that the machine-generated questions were often incomprehensible like "Why does (it) happen?", and that the system suffered from overfitting to the results of our automatic causality recognizer. Hence, we developed a novel method that effectively filters out incomprehensible questions and retrieves from texts answers that are likely to be paraphrases of a given causal relation. Through a series of experiments, we showed that our approach significantly improved the precision of the top answer by 8% over the current state-of-the-art system for Japanese why-QA.

1 Introduction

Why-question answering (why-QA) aims to retrieve answers to such why-questions as "Why are tsunamis generated?" from a given text archive (Girju 2003; Higashinaka and Isozaki 2008; Verberne et al. 2008; 2010; 2011; Oh et al. 2012; 2013). Recent works (Verberne et al. 2010; 2011; Oh et al. 2012; 2013) have taken a machine-learning approach with a supervised classifier for answer ranking and successfully improved the why-QA performance. For the given pairs of a question and an answer candidate passage, the classifier arranges them into *correct* pairs and *incorrect* pairs, or gives a score indicating the likelihood that the pair is correct, which is used for ranking the answer candidates.

In this paper, we propose a semi-supervised learning method that exploits automatically generated training data for improving the supervised classifier for answer ranking in our baseline why-QA system, which is actually an implementation of our previous work (Oh et al. 2013). The key of our proposed method is its use of causal relations in texts for generating training data (question-answer pairs). For this purpose, we use the causality recognition method of our previous work (Oh et al. 2013), which uses some pre-defined cue words such as "because." The **CR** shown below is an example of the recognized causal relations, where its cause and effect parts are marked with [..]$_{cause}$ and [..]$_{effect}$. Although our target language is Japanese, for ease of explanation, we use English examples in this paper.

A1	..... Sudden vertical **displacements** of the seafloor by **earthquakes**, disturb the **ocean's** surface, **displace water**, and generate destructive tsunami waves. This is basically how a tsunami is generated. .....
A2	..... It is crucial to provide immediate warnings after a potential tsunami generation because tsunamis are often unpredictable. .....

Table 1: Answer passages to "Why are tsunamis generated?" The vocabulary overlap between answer passages and the expected answer from the **CR**$_{cause}$ is represented in bold.

CR: [Tsunamis are generated]$_{effect}$ because [the ocean's water mass is displaced by an earthquake.]$_{cause}$

Note that the cause part (**CR**$_{cause}$) can be regarded as an expected answer to question "Why are tsunamis generated?," which can be automatically generated from the effect part (**CR**$_{effect}$). One might think that such a question-answer pair derived from the source texts of those causal relations can be used directly as a training instance for why-QA. However, in our preliminary experiments, this naive method failed to improve our baseline why-QA system. The main reason was that it suffered from overfitting to the results of our automatic causality recognizer, which can only recognize a limited variation of causal relations expressed by a limited set of cue words such as "because" and that the machine-generated questions were often incomprehensible: "Why does (it) happen?" Hence we developed a novel method that retrieves answers to a question from web texts to avoid the overfitting problem and that effectively filters out incomprehensible questions.

First, we explain our solution to the overfitting problem. We use our baseline why-QA system for retrieving answers to a question from web texts in the following way. Assume that our baseline why-QA system returned the passages in Table 1 as answers to "Why are tsunamis generated?," which is a question from the $\mathbf{CR}_{effect}$. **A1** (a correct answer to the question) has a large vocabulary overlap with the expected answer, $\mathbf{CR}_{cause}$, but **A2** (an incorrect one) does not. From this observation, we established the hypothesis shown below for identifying correct answers:

Hypothesis: Given such a causal relation as **CR** and a question generated from its effect part, the correct answer passage to the question (in the results of our baseline why-QA system) has and indeed must have a large vocabulary overlap with the cause part of the causal relation.

Note that **A1** can be regarded as another expression of **CR** or its *paraphrase*. An important point is that there is no explicit clue such as "because" for recognizing the causal relation in **A1** and our automatic causality recognition method actually fails to detect it. Nevertheless, to improve the why-QA performance, it is important to enable why-QA systems to recognize a wide range of such *indirectly expressed* causal relations as candidates for answers to why-questions. If we give as a positive training instance a pair of question "why are tsunamis generated?" and its answer **A1**, identified by the hypothesis, to our why-QA system, it should help the system to deal with a wider range of indirectly expressed causality and lead to a higher performance. On the other hand, if we use the naive method and give as a positive training instance the pair of the same question and **CR**, which is a *directly expressed* causality by "because" so that it is recognizable by our causality recognition method, it may lead to overfitting of the why-QA system toward the pattern of causality expressions recognizable by the causality recognition method. Of course, the simple hypothesis actually has a negative side effect that gives false negative labels to a correct answer, which has a small vocabulary overlap with an expected answer expressed in automatically recognized causal relations. Despite such a problem, we show that our method using the hypothesis leads to a significant performance improvement over a state-of-the-art why-QA system (Oh et al. 2013).

We also found that many *incomprehensible* questions were generated from causal relations by the naive method. Since these questions often included realized anaphors, such as pronouns, or zero anaphors, which are the omitted arguments of predicates and frequently occur in Japanese (Iida et al. 2007), they could not be interpreted without contexts in their original causal relation texts. For instance, incomprehensible question "Why was (he) awarded the Nobel Prize?" was generated from causal relation **CR'**, where "(he)" represents a zero anaphor.

CR': [Because Dr. Prusiner discovered prions]$_{cause}$, [(he) was awarded the Nobel Prize.]$_{effect}$

Actually, new training data obtained from such incomprehensible questions degraded the performance of our baseline why-QA system. Hence we developed a novel method of judging the *comprehensibility of questions* using a supervised classifier with subset-tree kernels (Moschitti 2006) and used only the questions judged as comprehensible by the classifier for our semi-supervised learning. We empirically validated its effectiveness through our experiments.

The contributions of this work are summarized as follows.

- To the best of our knowledge, this work is the first attempt to take a semi-supervised learning approach for why-QA.
- We propose a novel method for a non-trivial task, to generate training data (question-answer pairs) for why-QA with causal relations in texts.
- Our method generates comprehensible questions from causal relations and then retrieves from web texts their answers, which are likely to be paraphrases of a given causal relation, by using our baseline why-QA system and vocabulary overlap between the answers and the causal relations. These paraphrases of a given causal relation allow why-QA systems to learn with a wide range of causality expressions and to recognize such causality expressions as candidates for answers to why-questions.
- We showed that our proposed method improved the precision of the top answers by 8% against the current state-of-the-art system of Japanese why-QA (Oh et al. 2013).

2 System Architecture

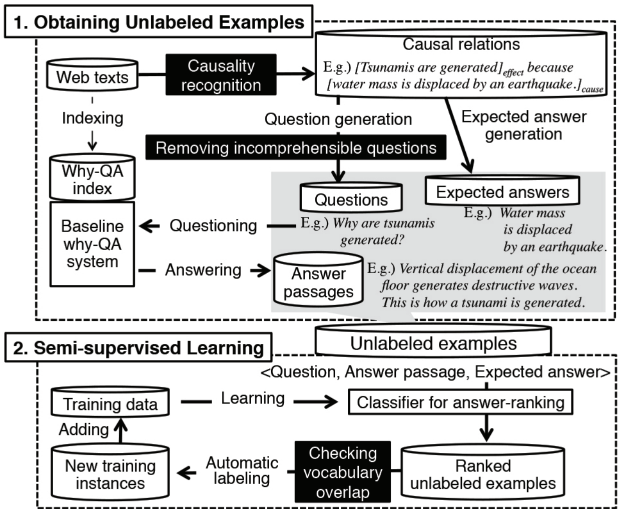

1. Obtaining Unlabeled Examples
Web texts
Causality recognition
Causal relations
E.g.) [Tsunamis are generated]effect because [water mass is displaced by an earthquake.]cause
Indexing
Question generation
Expected answer generation
Removing incomprehensible questions
Why-QA index
Questions
Expected answers
Baseline why-QA system
Questioning
E.g.) Why are tsunamis generated?
E.g.) Water mass is displaced by an earthquake.
Answering
Answer passages
E.g.) Vertical displacement of the ocean floor generates destructive waves. This is how a tsunami is generated.
2. Semi-supervised Learning
Unlabeled examples
<Question, Answer passage, Expected answer>
Training data
Learning
Classifier for answer-ranking
Adding
New training instances
Automatic labeling
Checking vocabulary overlap
Ranked unlabeled examples

Figure 1: System architecture

Figure 1 shows the overall architecture of our proposed method. We automatically extracted causal relations from Japanese web texts using the causality recognition method of our previous work (Oh et al. 2013). We restricted our causal relation to one whose cause and effect parts were connected by some pre-defined cue phrases[1] and identified the boundaries of the cause and effect parts using sequential labeling with conditional random fields (CRFs) (Lafferty, McCallum, and Pereira 2001). Precision, recall, and F-score of this method reported in Oh et al. (2013) are 83.8%, 71.1%,

[1] We used Oh et al. (2013)'s cue phrases that included those indicating *for*, *as a result*, *from the fact that*, *because*, *the reason is*, and *is the cause*.

and 77.0%, respectively (see Oh et al. (2013) for more details). Using this method, we recognized about 656 million causal relations from two billion web texts.

Next, we generated why-questions and their expected answers from these causal relations. After removing the incomprehensible questions from the generation result, the remaining questions were given to our baseline why-QA system (described in Section 3) as input to get the top-n answer passages for each question; we set n to 20 in this work. Then we made unlabeled examples with a set of triples: a why-question, its expected answer, and an answer passage to the why-question. We detail this procedure in Section 4.

Our semi-supervised learning iteratively generates new training instances from the classifier's results on the unlabeled examples and trains a classifier for answer ranking using the training data enlarged with the new training instances (see Section 5 for details).

3 Baseline Why-QA System

The baseline why-QA system, which is our implementation of Oh et al. (2013), is composed of *answer retrieval* and *answer ranking*. We use it to obtain unlabeled examples to be labeled in our semi-supervised learning algorithm for improving the classifier in the answer ranking.

3.1 Answer retrieval

This module provides a set of passages for a why-question retrieved from two billion Japanese web texts. To make answer retrieval from such large-scale texts efficient, we restrict our retrieved passages to those composed of the seven consecutive sentences and containing at least one cue phrase for recognizing causal relations used in Oh et al. (2013). We extracted about 4.2 billion passages that satisfy these conditions from the two billion web texts and indexed them for keyword search using a Lucene search engine[2].

We used two types of queries generated from a given why-question for answer retrieval. Let a set of nouns and a set of verbs/adjectives in a why-question be $N = \{n_1, \cdots, n_j\}$ and $VA = \{va_1, \cdots, va_k\}$, respectively. From our observation that a correct answer is likely to contain all the nouns in a why-question, we generated two Boolean queries: "n_1 AND $\cdots$ AND n_j" and "n_1 AND $\cdots$ AND n_j AND (va_1 OR $\cdots$ OR va_k)." Then, for each of the queries, we retrieved the top-1200 passages that were ranked by a *tf-idf* scoring model implemented in the Lucene search engine. We combined the resulting passages of the two queries and passed them to the next step, which is answer ranking.

3.2 Answer ranking

A supervised classifier (SVMs (Vapnik 1995)) was used to rank these retrieved passages. We used the same features employed in our previous work (Oh et al. 2013) for training the classifier: morpho-syntactic 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)), sentiment polarity features (word and phrase polarities), and causal relation features (n-grams, semantic word classes, and excitation polarities (Hashimoto et al. 2012) in causal relations whose effect part is matched with a question). All the passages passed from the answer retrieval were ranked in descending order by their scores given by the classifier, and then the top-20 answer passages in the ranked list were chosen as output of the baseline why-QA system.

Note that the classifier in answer ranking uses as features the causal relations recognized by our automatic causality recognizer, which is an implementation of Oh et al. (2013)'s method. As mentioned in the Introduction, if we give the classifier many question-answer pairs (as new training instances) directly generated from such automatically recognized causal relations and their original texts, it might lead to such undesirable side effects that the resulting classifier would put excessive weight on the causal relation features. In our preliminary experiments, we actually observed such side effects. Hence, rather than using the source text of causal relations as answers, we retrieved from texts answers that tended to be paraphrases of the given causal relations using our baseline why-QA system.

4 Unlabeled Examples

4.1 Generating comprehensible questions

From causal relations acquired by using our causality recognizer, we generated why-questions by simply converting the effect part of the causal relations to their why-question form using a small set of hand-coded rules: e.g., "Tsunamis are generated" is converted into "Why are tsunamis generated?" We found that many incomprehensible questions such as "Why was (he) awarded the Nobel Prize?" were generated mainly due to pronouns and zero anaphors, which frequently appear in Japanese (Iida et al. 2007). Especially, the topics and subjects of the predicates appearing at the beginning of sentences were often missing in the effect part of the causal relations automatically extracted from sentences like the **CR'** shown in the Introduction. Hence we developed a novel method of identifying comprehensible questions using the following procedures: 1) filtering out questions containing any pronoun and 2) filtering out questions, in which any essential argument of the predicates, such as subject and object, is missing. To the best of our knowledge, this work is the first attempt to judge the *comprehensibility of machine-generated questions* for question-answering. For the second filtering, we used subset-tree kernels[3] implemented in SVM-Light (Joachims 1999; Moschitti 2006) and trained the kernels using the combinations[4] of trees and vectors shown below.

- All the subset-trees of a phrase structure tree (e.g., [NP [N *tsunamis*] P])
- All the subset-trees of a phrase structure tree where their nouns are replaced with their corresponding word class (the same 500 word classes for five million words as those in Oh et al. (2013)) (e.g., [NP [N $\mathrm{W}_{disaster}$] P], where

[2] http://lucene.apache.org

[3] http://disi.unitn.it/moschitti/Tree-Kernel.htm

[4] We used "-t 5 -C + -S 1 -d 1" as the parameter of the training.

"tsunamis" is replaced with W$_{disaster}$ representing the word class involving "disaster")

- Vectors expressing morpheme and POS tag n-grams

For this training, we hand-created 9,500 comprehensible why-questions and used them as positive examples. We also automatically generated 16,094 negative examples by deleting one or two among the topic, the subject, and the object of the predicates (i.e., topic, subject, and object words followed by Japanese postpositions '*-ha*' (topic marker), '*-ga*' (subject case marker), and '*-wo*' (object case marker), respectively) in the comprehensible questions. These examples were parsed by a Japanese dependency parser, J.DepP[5] (Yoshinaga and Kitsuregawa 2009), and their dependency trees were converted into phrase structure trees to generate subset-trees. This conversion was done by simply introducing NP (if the head word is a noun), VP (if the head word is a verb or an adjective), and OP (otherwise) to a parent node of a word phrase (*bunsetsu*[6]) in the dependency tree. For the training data, we randomly selected 90% of the positive examples and then selected negative examples automatically generated from the randomly selected positive ones. We used the remainder as the test data. In this setting, the subset-tree kernels achieved 84.5% precision and 88.9% recall. In our why-QA experiments in Section 6 we used a 90% precision setting to identify the comprehensible questions (i.e., SVM scores higher than 0.4).

4.2 Obtaining unlabeled examples

If a question is judged comprehensible, we extracted its expected answer from the cause part of the causal relation from which the question was generated. We obtained about 3.8 million pairs of questions and expected answers from 656 million causal relations. Then we retrieved the top-20 answer passages for each of these questions using our baseline why-QA system. Finally, we made unlabeled examples, $U = \{(q, e, p_j)\}$, where q is a question, e is an expected answer to q, and p_j is a passage in the top-20 answer passages $\{p_1, \cdots, p_{20}\}$ to question q. For example, let **A1** in Table 1 be a retrieved answer passage to the question, "Why are tsunamis generated?" Then we can make an unlabeled example in Table 2 with the answer passage as well as the question and its expected answer from the **CR**.

5 Semi-supervised Learning

The pseudo-code of our semi-supervised learning algorithm is given in Figure 2. The algorithm is an iterative procedure, which takes, as input, unlabeled examples U generated by the procedure in Section 4, manually labeled training data L^0, and maximum iteration number l. At the i-th iteration, the procedure 1) trains classifier c^i using existing training data L^i ($c^i := Learn(L^i)$), 2) automatically assigns labels to some unlabeled examples selected from U according to their classification results by c^i ($L^i_U := Label(c^i, U)$), and 3) enlarges the current training data with new training instances ($L^{i+1} := L^i \cup L^i_U$). The iteration terminates when either of the two stop conditions is satisfied: the procedure already iterates l times ($i = l$) or no more new training instances can be added to L^i_U ($L^i_U = \phi$).

q	Why are tsunamis generated?
e	The ocean's water mass is displaced by an earthquake.
p_j	..... Sudden vertical displacements of the seafloor by earthquakes, disturb the ocean's surface, displace water, and generate destructive tsunami waves. This is basically how a tsunami is generated. .....

Table 2: Example of unlabeled examples

```
**Input**: Unlabeled examples ($U$), initial training data ($L^0$), and maximum iteration number ($l$)
**Output**: A classifier ($c^n$)
$i = 0$
**repeat**
    $c^i := Learn(L^i)$
    $L^i_U := Label(c^i, U)$
    $L^{i+1} := L^i \cup L^i_U$
    $i = i + 1$
**until** stop condition is met
```

Figure 2: Our semi-supervised learning algorithm

5.1 Automatic labeling of unlabeled examples

This procedure, $Label(c^i, U)$ in Figure 2, labels unlabeled examples based on the hypothesis described in the Introduction and is composed of *candidate selection* and *automatic labeling*. Recall that unlabeled example $u = (q, e, p_j) \in U$ is represented as a triple of why-question q, expected answer e to q, and answer passage p_j that is returned by our baseline why-QA system with q as its input. Also recall that q and e were obtained from the same causal relation.

Candidate selection To select candidates for a new training instance, we took an *uncertainty sampling* approach (Lewis and Gale 1994) that has been used in active learning (Tong and Koller 2001; Zhu et al. 2008; Lughofer 2012; Settles 2012; Zhu and Ma 2012) and semi-supervised learning with multi-view classifiers (Cao, Li, and Lian 2003; Gu, Zhu, and Zhang 2009; Oh, Uchimoto, and Torisawa 2009; Oh et al. 2010). At the i-th iteration of the whole procedure, classifier c^i, which is actually SVMs, provides score value s for each question-passage pair (q, p_j). We represent this by $s = c^i(q, p_j)$. First, our procedure chooses the question-passage pair with the highest score s among the pairs for identical question q. Let the chosen pair for question q' be (q', p'). Then we check whether classifier c^i is *uncertain* about its prediction of (q', p'). We assume that c^i's prediction of (q', p') is uncertain if $s' = c^i(q', p')$ is close enough to the decision boundary for the SVMs, i.e., 0 (Tong

[5] http://www.tkl.iis.u-tokyo.ac.jp/∼ynaga/jdepp/

[6] A bunsetsu, which is a syntactic constituent composed of a content word and several function words such as postpositions and case markers, is the smallest phrase of syntactic analysis in Japanese.

and Koller 2001). More precisely, (q', p') can be selected as a candidate to be assigned a label and added to the existing training data, when all the following conditions hold where e' is the expected answer to q' such that q' and e' came from the same causal relation.

- $|c^i(q', p')| < \alpha$ ($\alpha > 0$), where α is a threshold value to determine the uncertainty of the classifier's prediction.
- Answer passage p' does not include the causal relations from which q' and e' were extracted.
- (q', p') does not appear in current labeled example L^i.

Note that the first condition is for uncertainty sampling and the second is posed to avoid giving a heavy bias to the causal relation features in training a classifier. The third is for preventing the same examples from being added to the training data more than once.

Automatic labeling Once (q', p') is selected as a candidate for automatic labeling, we compute the vocabulary overlap between answer passage p' and expected answer e' to question q'. In automatic labeling, we assume that p' is likely to be a correct answer, if the overlap between p' and e' is large. The overlap is estimated using function $VO(e', p')$ defined below, where $T(x)$ is a set of terms (i.e., content words including nouns, adjectives, and verbs) in x, and $S(p')$ is a set of subsequent two sentences extractable from answer passage p':

$$VO(e', p') = max_{s \in S(p')} \frac{|T(e') \cap T(s)|}{|T(e')|} \qquad (1)$$

Note that the terms in the subsequent two sentences in answer passage p' rather than those in whole answer passage p' are matched with those in expected answer e'. We use this scheme because of our observation that the terms in an expected answer are likely to coherently appear in subsequent two sentences in a correct answer passage. Using vocabulary overlap estimation $VO(e', p')$, the procedure decides the label for question-passage pair (q', p') as follows. Note that we pose additional conditions on the overlap between question q' and answer passage p' (i.e. $VO(q', p')$), based on the observation that the correct answer/passage should have a large vocabulary overlap with the question as well:

- If $VO(e', p') > \beta$ and $VO(q', p') > \beta$, then the label is "correct," where $0.5 < \beta < 1$.
- If $VO(e', p') < 1 - \beta$ or $VO(q', p') < 1 - \beta$, then the label is "incorrect."

Finally, all the automatically labeled examples are ranked by $|c^i(q', p')|$ in ascending order. The top K examples are selected from the ranked result and added to L^i_U. Intuitively, the automatically labeled examples, which current classifier c^i classified with lower confidence (i.e., of which current classifier c^i is more uncertain about its prediction) has higher priority to be selected as new training instances. Note that we have three parameters, α, β, and K. We determined their values through experiments using our development data.

6 Experiments

6.1 Data

Why-QA Data Set For the experiments, we used the same data set as the one used in our previous works (Oh et al. 2012; 2013). This data set is composed of 850 Japanese why-questions and their top-20 answer passages obtained from 600 million Japanese web texts by using the answer-retrieval method of Murata et al. (2007). Three annotators checked whether an answer passage in the question-passage pair is a correct answer to the question, and final judgments were made by majority vote[7].

For our experiments, we divided the data set (17,000 question-passage pairs) into training, development, and test data. For our training data, we first selected 7,000 question-passage pairs for 350 questions, which we used only as training data in our previous works (Oh et al. 2012; 2013) and randomly chose 8,000 question-passage pairs for 400 questions from the remainders of the first selection. We equally divided the 2,000 question-answer pairs for 100 questions, which is the remainder after the selection of the training data, into development and test data.

The training data was used as the initial training data (L^0 in Figure 2) for training the initial classifier (c^0 in Figure 2). The development data was used to determine several parameters in our semi-supervised learning including α (the threshold value to determine whether a classifier is uncertain about its classification result), β (the threshold value to determine the label of an unlabeled example in the automatic labeling), and K (the maximum number of automatically labeled examples added to the existing training data at each iteration). All the combinations of α, β, and K derived from $\alpha \in \{0.2, 0.3, 0.4\}$, $\beta \in \{0.6, 0.7, 0.8\}$, and $K \in \{150, 300, 450\}$ were tested over the development data, and $\alpha = 0.3$, $\beta = 0.7$, and $K = 150$, which showed the best performance, were used for our final experiments. We also set the maximum iteration number (l in Figure 2) to 40, where the performance converged in the test with these selected parameters on the development data.

Unlabeled examples (U in Figure 2) We randomly sampled 11 million causal relations (about 1/60 of all of the causal relations) from the 656 million causal relations that were automatically acquired from our two billion web texts by using our causality recognizer (Oh et al. 2013). We generated questions and their extracted expected answers from these selected causal relations and identified the comprehensible ones from these questions. We obtained 56,775 comprehensible questions and got the top-20 answer passages by using these questions as input for our baseline why-QA system. We denote as U_{SC} the unlabeled examples obtained using only these comprehensible questions.

We also prepared another type of unlabeled example to show the contribution of the comprehensible questions to our semi-supervised learning. We generated questions from the 11 million causal relations and used all of the randomly

[7] Their inter-rater agreement by Fleiss' kappa reported in Oh et al. (2012) was 0.634.

selected 100,000 questions without removing the incomprehensible ones for obtaining the unlabeled examples. We denote these unlabeled examples as U_{All}. Finally, we had unlabeled examples such that $|U_{SC}|$ = 514,674, $|U_{All}|$ = 1,548,998 and $|U_{SC} \cap U_{All}|$ = 17,844.

6.2 Evaluation with Why-QA Data Set

We conducted the experiments with the systems in the following different settings of answer ranking:

- MURATA represents Murata et al. (2007)'s unsupervised method that was used for creating the Why-QA Data Set.
- INIT represents our implementation of Oh et al. (2013), the state-of-the-art method for why-QA, that we used as the initial classifier in our semi-supervised learning.
- ATONCE uses a classifier obtained using all the automatically labeled examples (without the restriction of K=150) obtained at the first iteration of our semi-supervised learning algorithm as well as the initial training data as its training data. Comparison between ATONCE and our proposed method can show the effect of iterations in our semi-supervised learning.
- OURS(U_{All}) represents a variation of our method, where U_{All} is used as the unlabeled examples for our semi-supervised learning. Comparison between OURS(U_{All}) and our proposed method can show the contribution of comprehensible questions in our method.
- OURS(CR) is another variation of our method. In this setting, we used U_{SC} as unlabeled examples and allowed our semi-supervised learning algorithm to select question-passage pairs as a candidate for new training instances even when the answer passage contained the same causal relation as the one from which the question was extracted. Since such question-passage pairs can cause system's overfitting to the results of an automatic causality recognizer, OURS(CR) can show its effect in our semi-supervised learning.
- OURS(U_{SC}) represents our proposed method.
- UPPERBOUND represents a system that always locates all the n correct answers to a question in top-n ranks if they are in the test data. This indicates the upper-bound of performance in this experiment.

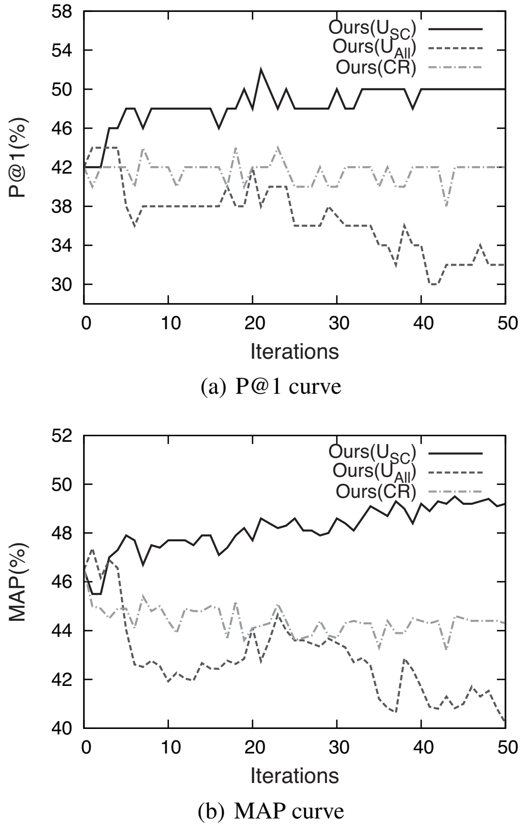
We used TinySVM[8] with a linear kernel for training a classifier in all the systems except for MURATA and UPPERBOUND. Evaluation was done by Precision of the top answer (P@1) and Mean Average Precision (MAP), both of which are the same measures used in Oh et al. (2013).

Table 3 shows the performance of the seven systems. R(P@1) and R(MAP) represent the relative performance of each system to that of UPPERBOUND. More precisely, R(P@1) and R(MAP) are P@1 and MAP evaluated using all the questions that have at least one correct answer in the test data.

None of ATONCE, OURS(U_{All}), and OURS(CR) outperformed INIT, which used only the manually created initial training data. On the contrary, our proposed method OURS(U_{SC}) consistently showed better performance in both P@1 and MAP than INIT. This implies that 1) the iterative process in our proposed method contributed to the performance improvement; 2) only the new training instances obtained by the comprehensible questions improved the performance; and 3) the new training instances generated directly from causal relations negatively affected answer ranking in our experimental setting. Further, we found that the R(P@1) of OURS(U_{SC}) is about 75.8% (=50.0/66.0). This indicates that answer ranking by our method can locate a correct answer to a why-question in the top rank with high precision if an answer-retrieval module can retrieve at least one correct answer to the question from web texts.

	P@1	MAP	R(P@1)	R(MAP)
MURATA	30.0	36.1	45.5	54.7
INIT	42.0	46.5	63.6	70.5
ATONCE	42.0	45.4	63.6	68.8
OURS(U_{All})	34.0	41.7	51.5	63.2
OURS(CR)	42.0	44.4	63.6	67.3
OURS(U_{SC})	**50.0**	**48.9**	**75.8**	**74.1**
UPPERBOUND	66.0	66.0	100	100

Table 3: Performance of systems



(a) P@1 curve

(b) MAP curve

Figure 3: P@1 and MAP curves

Figure 3 shows the P@1 and MAP curves of OURS(CR),

[8] http://chasen.org/~taku/software/TinySVM/

$Q1$	Why are emissions from diesel vehicles so toxic that plants wither and die?
A_{Q1}	... Compared with a gasoline engine, a diesel engine has higher fuel efficiency, lower CO2 emissions but higher NOx emissions. NOx causes air-pollution and is related to acid rain and the destruction of the ozone. *This also kills plants on the ground*, leading to subsequent huge environmental problems. ...
$Q2$	Why wasn't (the main donjon of) Edo Castle rebuilt after it was destroyed by the 1657 Meireki fire?
A_{Q2}	... The height of Edo Castle's main donjon was comparable with that of a 20-story building. ... However, after its destruction in the big fire in 1657, (the main donjon of Edo Castle) was not rebuilt. *Though government elders proposed rebuilding it, Daimyo Hoshina Masayuki rejected the proposal*, insisting that "we must priotize helping people recover from the damage and suffering caused by this big fire." ...

Table 4: Examples of answers (translated from Japanese) that only our proposed method could return as the top answer.

Ours(U_{All}) and Ours(U_{SC}) as the iterations advance toward 50 in our semi-supervised learning. The P@1 and MAP of Ours(U_{SC}) reached 50% and 49.2%, respectively, after 50 iterations. On the contrary, Ours(U_{All}) achieved the better performance in both P@1 and MAP than its initial stage only for the first several iterations. We believe that the noisy labeled examples introduced by *incomprehensible* questions badly affected the performance. Further Ours(CR) had the better performance than its initial stage at only the three iteration points and its MAP was consistently worse than that of its starting point. These results indicate that our solutions to the two problems in generating training data from causal relations (i.e. incomprehensible questions and system's overfitting to the results of an automatic causality recognizer explained in the Introduction) worked well, and all of our solutions contributed to the performance improvement.

Table 4 shows the top answers returned by our proposed method Ours(U_{SC}), where the other five systems except for UpperBound failed to return them as the top answer. The clause or sentence in italics in answer passages A_{Q1} and A_{Q2} represents certain causality related to the answers to each of questions $Q1$ and $Q2$. Note that the causality in the examples is expressed without such explicit clue words as "because." We believe that our proposed method provided training data with which our answer ranker could learn a pattern of such causality and, as a result, it returned those answers as the top answers.

6.3 Evaluation with answer retrieval in our baseline why-QA system

We performed another evaluation to show that Ours(U_{SC}), which is the answer ranker trained with our proposed method, works well under a different answer-retrieval setting from that used for creating the Why-QA Data Set (i.e., the answer-retrieval method of Murata et al. (2007) and 600 million web texts). For this purpose, we retrieved answer passages using all the questions in the development and test data of the Why-QA Data Set as input to the answer-retrieval module in our baseline why-QA system, which was described in Section 3.1. Then, to get the top-5 answer passages to each of the questions, we ranked the retrieved passages with Init and Ours(U_{SC}) in the previous experiment. Three annotators checked the top-5 answer passages returned by Init and Ours(U_{SC}), and their final judgement was given by majority voting[9]. The evaluation was done by P@1, P@3, and P@5. P@N measures how many questions have correct answers in the top-N answer passages returned by a system. Table 5 shows the result. Our proposed method outperformed Init in P@1, P@3, and P@5. This result shows that our proposed method consistently improved answer ranking in the different settings for answer retrieval in a why-QA system, at least in our experimental setting.

	P@1	P@3	P@5
Init	43.0	65.0	71.0
Ours(U_{SC})	**50.0**	**68.0**	**75.0**

Table 5: Performance of systems with answer retrieval in our baseline why-QA system

7 Conclusion

In this paper, we presented a novel approach to why-QA, which is semi-supervised learning that exploits automatically generated training data for why-QA. We propose a novel method for a non-trivial task, to generate training data (question-answer pairs) for why-QA using causal relations in texts. Our method generates comprehensible questions from causal relations and retrieves from web texts answers to the questions, which are likely to be paraphrases of a given causal relation, using our baseline why-QA system and vocabulary overlap between answers and causal relations. These paraphrases of a given causal relation in the retrieved answers allow why-QA systems to learn a wide range of causality expression patterns and to recognize such causality expressions as candidates for answers to why-questions. Through our experiments, we showed that our proposed method achieved 8% improvement in precision at the top answer over the current state-of-the-art system for Japanese why-QA, which was actually used as a starting point for our semi-supervised learning. In future work, we plan to extend our proposed method with event causality (Hashimoto et al. 2014; 2015), entailment/contradiction patterns (Kloetzer et al. 2013; 2015), and zero anaphora resolution (Iida et al. 2015).

[9] Their inter-rater agreement by Fleiss' kappa was 0.722

References

Cao, Y.; Li, H.; and Lian, L. 2003. Uncertainty reduction in collaborative bootstrapping: Measure and algorithm. In *Proceedings of ACL '03*, 327–334.

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

Gu, P.; Zhu, Q.; and Zhang, C. 2009. A multi-view approach to semi-supervised document classification with incremental naive bayes. *Computers & Mathematics with Applications* 57(6):1030–1036.

Hashimoto, C.; Torisawa, K.; Saeger, S. D.; Oh, J.-H.; and Kazama, J. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of EMNLP-CoNLL '12*, 619–630.

Hashimoto, C.; Torisawa, K.; Kloetzer, J.; Sano, M.; Varga, I.; Oh, J.; and Kidawara, Y. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of ACL '14*, 987–997.

Hashimoto, C.; Torisawa, K.; Kloetzer, J.; and Oh, J. 2015. Generating event causality hypotheses through semantic relations. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2396–2403.

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

Iida, R.; Komachi, M.; Inui, K.; and Matsumoto, Y. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the ACL Workshop: 'Linguistic Annotation Workshop'*, 132–139.

Iida, R.; Torisawa, K.; Hashimoto, C.; Oh, J.; and Kloetzer, J. 2015. Intra-sentential zero anaphora resolution using subject sharing recognition. In *Proceedings of EMNLP '15*, 2179–2189.

Joachims, T. 1999. Making large-scale SVM learning practical. In Schölkopf, B.; Burges, C.; and Smola, A., eds., *Advances in Kernel Methods - Support Vector Learning*. Cambridge, MA: MIT Press. chapter 11, 169–184.

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

Kloetzer, J.; Saeger, S. D.; Torisawa, K.; Hashimoto, C.; Oh, J.; Sano, M.; and Ohtake, K. 2013. Two-stage method for large-scale acquisition of contradiction pattern pairs using entailment. In *Proceedings of EMNLP '13*, 693–703.

Kloetzer, J.; Torisawa, K.; Hashimoto, C.; and Oh, J. 2015. Large-scale acquisition of entailment pattern pairs by exploiting transitivity. In *Proceedings of EMNLP '15*, 1649–1655.

Lafferty, J.; McCallum, A.; and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML '01*, 282–289.

Lewis, D. D., and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *Proceedings of SIGIR '94*, 3–12.

Lughofer, E. 2012. Hybrid active learning for reducing the annotation effort of operators in classification systems. *Pattern Recognition* 45(2):884–896.

Moschitti, A. 2006. Making tree kernels practical for natural language learning. In *Proceedings of EACL '06*, 113–120.

Murata, M.; Tsukawaki, S.; Kanamaru, T.; Ma, Q.; and Isahara, H. 2007. A system for answering non-factoid Japanese questions by using passage retrieval weighted based on type of answer. In *Proceedings of NTCIR-6*.

Oh, J.-H.; Yamada, I.; Torisawa, K.; and De Saeger, S. 2010. Co-star: A co-training style algorithm for hyponymy relation acquisition from structured and unstructured text. In *Proceedings of COLING '10*, 842–850.

Oh, J.-H.; Torisawa, K.; Hashimoto, C.; Kawada, T.; Saeger, S. D.; Kazama, J.; and Wang, Y. 2012. Why question answering using sentiment analysis and word classes. In *Proceedings of EMNLP-CoNLL '12*, 368–378.

Oh, J.-H.; Torisawa, K.; Hashimoto, C.; Sano, M.; Saeger, S. D.; and Ohtake, K. 2013. Why-question answering using intra- and inter-sentential causal relations. In *Proceedings of ACL '13*, 1733–1743.

Oh, J.-H.; Uchimoto, K.; and Torisawa, K. 2009. Bilingual co-training for monolingual hyponymy-relation acquisition. In *Proceedings of ACL-IJCNLP '09*, 432–440.

Settles, B. 2012. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Tong, S., and Koller, D. 2001. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research* 2:45–66.

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

Verberne, S.; Boves, L.; Oostdijk, N.; and Coppen, P.-A. 2008. Using syntactic information for improving why-question answering. In *Proceedings of COLING '08*, 953–960.

Verberne, S.; Boves, L.; Oostdijk, N.; and Coppen, P.-A. 2010. What is not in the bag of words for why-QA? *Computational Linguistics* 36(2):229–245.

Verberne, S.; van Halteren, H.; Theijssen, D.; Raaijmakers, S.; and Boves, L. 2011. Learning to rank for why-question answering. *Inf. Retr.* 14(2):107–132.

Yoshinaga, N., and Kitsuregawa, M. 2009. Polynomial to linear: Efficient classification with conjunctive features. In *Proceedings of EMNLP '09*, 1542–1551.

Zhu, J., and Ma, M. 2012. Uncertainty-based active learning with instability estimation for text classification. *ACM Trans. Speech Lang. Process.* 8(4):5:1–5:21.

Zhu, J.; Wang, H.; Yao, T.; and Tsou, B. K. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of COLING '08*, 1137–1144.