A Crowdsourcing Approach for Annotating Causal Relation Instances in Wikipedia

Kazuaki Hanawa, Akira Sasaki, Naoaki Okazaki, Kentaro Inui
Tohoku University
{hanawa, aki-s, okazaki, inui}@ecei.tohoku.ac.jp

Abstract

This paper presents a crowdsourcing approach for annotating causal relation instances to Wikipedia. Because an annotation task cannot be decomposed into multiple-choice problems, we integrate a crowdsourcing service and brat, a popular on-line annotation tool, to provide an easy-to-use interface and quality control for annotation work. We design simple micro-tasks that involve annotating textual spans with causal relations. We issued the micro-tasks to crowd workers, and collected 95,008 annotations of causal relation instances among 8,745 summary sentences in 1,494 Wikipedia articles. The annotated corpus not only provides supervision data for automatic recognition of causal relation instances, but also reveals valuable facts for improving the annotation process of this task.

1 Introduction

Commonsense knowledge such as entities and events, and their causal relationships, are indispensable in various natural language processing (NLP) applications, including question answering (Oh et al., 2013; Oh et al., 2016; Sharp et al., 2016), hypothesis generation (Radinsky et al., 2012; Hashimoto et al., 2015), stance detection (Sasaki et al., 2016), and literature curation for systems biology (Pyysalo et al., 2015; Rinaldi et al., 2016).

In many previous researches, corpora for acquiring causal relations were built by annotating two text spans (e.g., entities) and their relations in the text (Doddington et al., 2004; Hendrickx et al., 2010; Pyysalo et al., 2015; Rinaldi et al., 2016; Dunietz et al., 2017; Rehbein and Ruppenhofer, 2017). However, this approach is extremely work intensive. It involves choosing a target domain, designing an ontology (semantic classes) of entities, building a corpus for named entity recognition, designing an annotation guideline for relations, and annotating the relations between entities. Building such a corpus also requires the annotation efforts of experts. For these reasons, this approach is almost non-scalable to various domains or genres of text although the knowledge of the causal relations is highly target-specific.

This paper presents an approach for harnessing causal relation instances to Wikipedia articles via crowdsourcing. Wikipedia is the central infrastructure for knowledge curation, as exemplified by Freebase (Bollacker et al., 2008) and Wikification (Mihalcea and Csomai, 2007). Therefore, we base Wikipedia articles for building a corpus with causal relation instances. This work represents a first step toward organizing the causal knowledge in Wikipedia articles covering various topics.

Recently, researchers have recognized the value of crowdsourcing services in constructing wide-ranging language resources at low cost (Brew et al., 2010; Finin et al., 2010; Gormley et al., 2010; Jha et al., 2010; Fort et al., 2011; Kawahara et al., 2014; Lawson et al., 2010; Hovy et al., 2014; Takase et al., 2016). Unfortunately, causal relations cannot be directly annotated by crowdsourcing. For this purpose, non-expert workers on crowdsourcing services require a clear and simple micro-task. A crowdsourcing service only provides a standardized interface for workers. The micro-tasks on this interface

31st Pacific Asia Conference on Language, Information and Computation (PACLIC 31), pages 336–345
Cebu City, Philippines, November 16-18, 2017

are often limited to multiple choice questions or free descriptions.

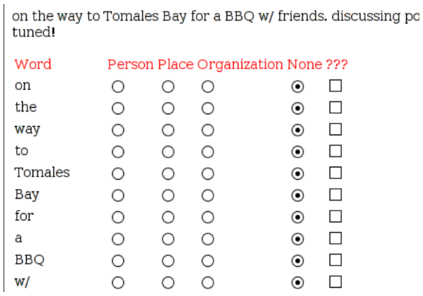
This study also explores the potential of crowdsourcing for collecting annotations about causal relation instances. To this end, we tailor a simple micro-task in which crowd workers annotate textual spans with causal relations to the title of a Wikipedia article. We also develop an annotation system that cooperates with a crowdsourcing service. By virtue of the widely used annotation tool brat[1] (Stenetorp et al., 2012), the system is easy to use and extendible to other annotation tasks.

We collected 95,008 annotations of causal relation instances for 8,745 summary sentences[2] in 1,494 Wikipedia articles. By analyzing the annotation results, we provide valuable hints for improving the annotation process in terms of the number of crowd workers necessary for an article, the number of agreements necessary for improving the quality of causal relation instances, syntactic profiles of annotated spans (e.g., noun and verb phrases), and common confusions of annotations.

The annotation results are also useful for mining expressions inverting polarity of causality (promotion and suppression) and provide supervision data for automatic extraction of causal relation instances from Wikipedia articles. We have released the annotation system, annotated corpus, and the automatic extraction tool on a dedicated website[3]. Although the corpus was built for Japanese Wikipedia articles, we here use English translations for illustrative purposes.

2 Related work

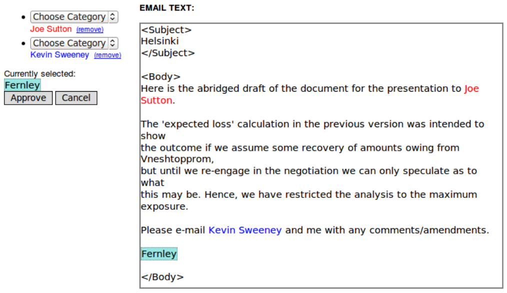
NLP researchers have built corpora for various NLP tasks through crowdsourcing. These tasks include part-of-speech tagging (Hovy et al., 2014), PP attachment (Jha et al., 2010), named entity recognition (Finin et al., 2010; Lawson et al., 2010), sentiment classification (Brew et al., 2010), relation extraction (Gormley et al., 2010), semantic modeling of relation patterns (Takase et al., 2016), and discourse parsing (Kawahara et al., 2014). In most of these tasks, the micro-tasks are designed as multiple-choice problems. For example, Brew et al. (2010) annotated sentiment polarity in a micro-task where workers labeled an article as positive, negative, or irrelevant. When the target task cannot be broken into micro-tasks of multiple-choice problems, a special approach is needed. Labeling of text spans falls into this category.



on the way to Tomales Bay for a BBQ w/ friends. discussing pc
tuned!

Word	Person	Place	Organization	None	???
on	○	○	○	◉	□
the	○	○	○	◉	□
way	○	○	○	◉	□
to	○	○	○	◉	□
Tomales	○	○	○	◉	□
Bay	○	○	○	◉	□
for	○	○	○	◉	□
a	○	○	○	◉	□
BBQ	○	○	○	◉	□
w/	○	○	○	◉	□

Figure 1: Named entity annotation by the multiple-choice method (Finin et al., 2010).



EMAIL TEXT:

```
<Subject>
Helsinki
</Subject>

<Body>
Here is the abridged draft of the document for the presentation to Joe
Sutton.

The 'expected loss' calculation in the previous version was intended to
show
the outcome if we assume some recovery of amounts owing from
Vneshtopprom,
but until we re-engage in the negotiation we can only speculate as to
what
this may be. Hence, we have restricted the analysis to the maximum
exposure.

Please e-mail Kevin Sweeney and me with any comments/amendments.

Fernley

</Body>
```

Figure 2: A custom interface for annotating named entities via crowdsourcing (Lawson et al., 2010).

Notwithstanding, corpora with span annotations built by crowdsourcing have been reported in several studies. Finin et al. (2010) annotated the boundaries and semantic classes of named entities by converting the annotation task into a micro-task of multiple-choice problems. They applied the standard interface of Amazon Mechanical Turk (see Figure 1). In this interface, the worker selected a label (PERSON, PLACE, ORGANIZATION, or NONE) from a row of radio buttons placed beside every word in a sentence. This interface not only reduces the readability of the sentence but also requires many selections of radio buttons. The closest work to ours is Lawson et al. (2010). They implemented a custom interface in which workers selected arbitrary spans of text and attached a label to each span (see Fig-

[1] http://brat.nlplab.org/

[2] The lead paragraph of a Wikipedia article containing a quick summary of the most important points of the article.

[3] http://www.cl.ecei.tohoku.ac.jp/

ure 2). However, their interface is specific to named entity recognition, and is not generalizable to other annotation tasks. In addition, their annotation tool has not been released to the public.

In contrast, we combine a crowdsourcing service with brat, a popular open-source annotation tool, to provide an easy-to-use interface and quality control for the annotation work. This approach is not limited to causal relations but can be adapted to any brat-supported tasks (e.g., part-of-speech tagging and information extraction). We also present a quality control mechanism that is applicable to any crowdsourcing services accepting free text for a micro-task.

Several studies have dedicated to identify causal relations mentioned in text. For instance, Dunietz et al. (2017) present the version 2.0 of Bank of Effects and Causes Stated Explicitly (BECauSE). The corpus includes annotations of causes and effects as well as seven semantic relations that are frequently associated with causation. Rehbein and Ruppenhofer (2017) use the similar annotation scheme for building a German corpus with some changes in the label set and the scope of causality. Built on top of well-established lingustic theories, these studies focus more on "causal language" (expressions of causation) than real-world causation. In contrast, our ultimate goal is acquisition of real-world causal knowledge by exploting Wikipedia as an encyclopedia. We thus design a curation process with crowdworkers involved in, focussing on how humans 'read' Wikipedia articles for causal knowledge.

3 Annotating promotion/suppression relations in Wikipedia articles

3.1 Labels of causal relations

This study annotates promotion/suppression relations (Hashimoto et al., 2012; Fluck et al., 2015) in Wikipedia articles. Here, "*X* promotes *Y*" means that *Y* is activated when *X* is activated. Analogously, "*X* suppresses *Y*" means that *Y* is inactivated when *X* is activated.

Many corpora for acquiring relational knowledge are created by annotating two entities and the relation between the pair of entities in a sentence (Doddington et al., 2004). However, this approach is too difficult for crowd workers because it requires locating the entities and considering the promotion/suppression relations for all possible pairs of entities. Moreover, to create a valuable corpus, it is important to annotate the promotion/suppression relations involving the article title (a variable *T*, hereafter), because the article is naturally intended to provide knowledge about *T*. Therefore, we force *T* to participate in an argument of a promotion/suppression relation. In other words, the annotation task is accomplished by labeling PRO ("*T* promotes *Y*"), SUP ("*T* suppresses *Y*"), PRO_BY ("*X* promotes *T*"), or SUP_BY ("*X* suppresses *T*") for text spans (denoted by *Y* for PRO and SUP, and denoted by *X* for PRO_BY and SUP_BY) in the article.

We randomly selected 1,494 articles belonging to nine categories and to the subcategories/sub-subcategories: "Social issues", "Disasters", "Diseases and disorders", "Innovation", "Policy", "Finance", "Energy technology", "Biomolecules" and "Nutrients". It is hoped that articles in these categories contain many promotion/suppression relations.

3.2 Annotation policy

The units to be annotated must also be defined in the annotation design. In this research, we examined two kinds of units: noun phrases and verb phrases. However, neither of these units were satisfactory for annotating promotion/suppression relations.

For example, consider the following sentences in the Wikipedia article "Nyctalopia"[4].

> Nyctalopia, also called night-blindness, is a condition making it difficult to see in relatively low light. Nyctalopia may exist from birth, or be caused by injury or severe malnutrition.

Among these sentences, we seek an instance of ⟨SUP, nyctalopia, see in relatively low light⟩. However, when we limited the annotation unit to noun phrases, we could not annotate the phrase "see in relatively low light". Similarly, when we limited the annotation unit to verb phrases, we failed to obtain ⟨PRO_BY, nyctalopia, injury⟩.

Furthermore, whether adopting noun phrases or verb phrases, the segmentation problem of

[4] `https://en.wikipedia.org/wiki/Nyctalopia`

noun/verb phrases remained. For example, both of "severe malnutrition" and "malnutrition" can be interpreted as causes of nyctalopia. When multiple overlapping spans are plausible, we need a criterion that prioritizes one span over the others. However, such a criterion is difficult to define. Instead of defining strict guidelines for annotation spans, we collect multiple annotations within an article and explore the best set of guidelines for crowd workers. A side product of this approach is the varying degree of confidence for each span in the corpus. Thus, this corpus provides useful hints for further improving the annotation process for causal relations.

3.3 Using brat in crowdsourcing

Quality control is a major concern in language resources built by crowdsourcing. In most crowdsourcing services, the quality of an annotation and the worker can be judged by inserting test questions with the correct annotations provided by the task designer.

Although test questions and verifications are essential for quality control, they are inapplicable to the annotation policy described in Section 3.2, because they measure annotation quality by exact match. In contrast, our approach allows multiple spans for arguments of causal relations. If such annotations are judged by exact match, almost all of them will be assessed as incorrect. Therefore, we incorporate the verification process of the test questions in brat, and feedback the annotation quality of a worker to the crowdsourcing service.

Figure 3 is an overview of the proposed system. The annotation procedure is described below:

1. Workers click the link to the modified version of brat (for working on an external website) from a virtual task in the crowdsourcing service.

2. Workers perform the annotation tasks on brat.

3. When a worker complete the set of micro-tasks, we measure the worker's performance against the test questions hidden in the set. As the performance measure, we adopted the character-level F1 score between the worker's annotations and the gold standard.

4. The worker is requested to return to the crowdsourcing service and enter the password issued on brat. If the F1 score of a worker's annotation exceed a specified threshold (0.3), the worker is issued a correct password and could claim rewards. If the F1 score is below this threshold, the worker is issued an incorrect password (with no rewards).

	PRO	SUP	PRO_BY	SUP_BY
Exact	0.192	0.192	0.132	0.197
Partial	0.448	0.325	0.379	0.380
Character	0.332	0.282	0.309	0.317

Table 1: Inter-annotator agreement of each relation (micro F1 score)

4 Annotation results

Using the system described in the previous section and the Yahoo! crowdsourcing service[5], we collected ten annotations per article. We offered an independent task for each promotion/suppression relation PRO, SUP, PRO_BY and SUP_BY. Besides simplifying the annotation task, this assignment ensures that a worker collecting the annotation results is unaware of other relations. Figure 4 shows an example of "Leukemia" annotations. As shown in this fugure, each span has a varying degree of confidence. Most annotators judged that leukemia causes "abnormal white blood cells", followed by "high numbers of abnormal white blood cells". In addition, we observe a nested structure in which leukemia promotes "abnormal white blood cells" but also suppresses the subsequence "white blood cells". Nested structures in annotations can reveal patterns (e.g., "abnormal X") that reverse the polarity of causal relations, for example, from promotion to suppression (see Section 4.5).

4.1 Inter-annotator agreement

How is the quality of the causal relation corpus constructed in this study? Table 1 shows the average inter-annotator agreements for each relation. The agreement between two annotations was measured by the F1 scores of the exact match, partial match and character-level match. The agreement of annotations for an article was obtained by micro-

[5] https://crowdsourcing.yahoo.co.jp/

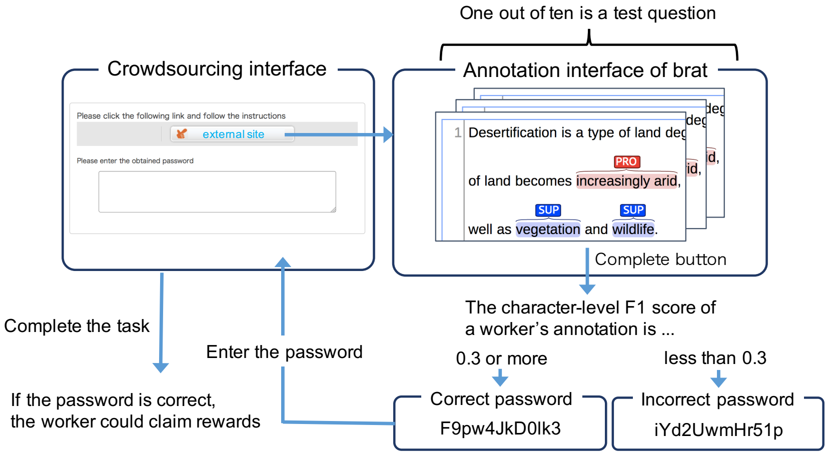
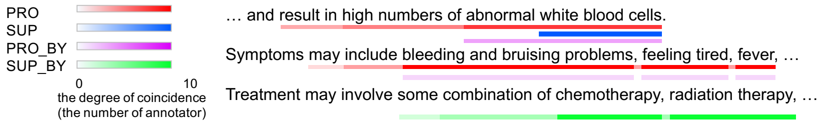

One out of ten is a test question
Crowdsourcing interface
Please click the following link and follow the instructions
external site
Please enter the obtained password
Annotation interface of brat
1 Desertification is a type of land deg
PRO
of land becomes increasingly arid,
SUP
SUP
well as vegetation and wildlife.
Complete button
Complete the task
Enter the password
The character-level F1 score of
a worker's annotation is ...
0.3 or more
less than 0.3
If the password is correct,
the worker could claim rewards
Correct password
F9pw4JkD0lk3
Incorrect password
iYd2UwmHr51p

Figure 3: Overview of the annotation system integrating Yahoo! crowdsourcing and brat



… and result in high numbers of abnormal white blood cells.

Symptoms may include bleeding and bruising problems, feeling tired, fever, …

Treatment may involve some combination of chemotherapy, radiation therapy, …

Figure 4: An excerpt of annotation results for "Leukemia" Wikipedia articles. The color at the bottom of the text indicates the relation label, and the color intensity indicates the degree of agreement between the workers.

PRO annotations	7,624
SUP annotations	2,923
PRO_BY annotations	5,387
SUP_BY annotations	1,127

Table 2: Number of annotations in the data created by 2-match aggregation.

averaging the agreements of all ($_{10}C_2 = 45$) pairs of workers. The exact match F1 score regards two annotations as matched when the start and the end of the segments are the same. The partial match F1 score regards two annotations as matched when they have an overlapping region. Although the inter-annotator agreements reported in Table 1 appear low, the results are reasonable considering the difficulty of the task.

4.2 Recommended number of annotations

The consistency of the annotations can be improved by adopting only spans with n or more exactly matched annotations. We call this treatment *n-match aggregation*. Figure 5 shows the micro-averages of the agreements between the raw annotations and those obtained by n-match aggregation. As shown in the figure, the highest consistency was achieved in 2-match aggregation. In other words, spans should be aggregated when two or more annotations are exactly matched. Therefore, the data created by 2-match aggregation were used in subsequent experiments. Table 2 shows the number of spans for each relation in the dataset.

Can we reduce the number of annotators per article without degrading the annotation quality? In this experiment, we extracted $_{10}C_m$ combinations of m annotations and calculated the micro-average of the agreements between the gold standard data (reference annotations used in the check questions) and n-match aggregations. The F1 score for each m and n is presented in Figure 6.

As shown in this figure, increasing the number of annotators improves the result; the more annotators (m) we use, the higher agreement we obtain from the n-match aggregations. Interestingly, the 2-match aggregation obtains high agreement in five annotations ($m = 5, n = 2$). Considering the tradeoff between number of annotations and cost, five annotations per article may be sufficient to achieve a satisfactory cost–performance balance.

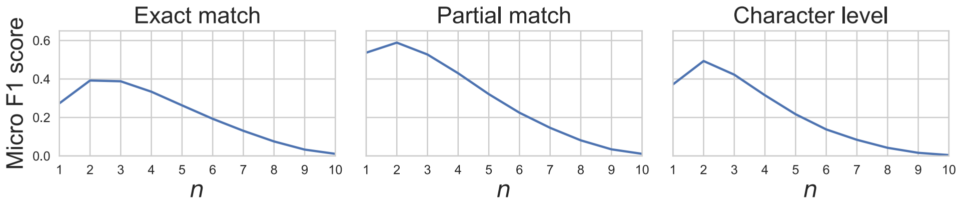


Figure 5: Agreement between the raw annotations and those obtained by n-match aggregation.

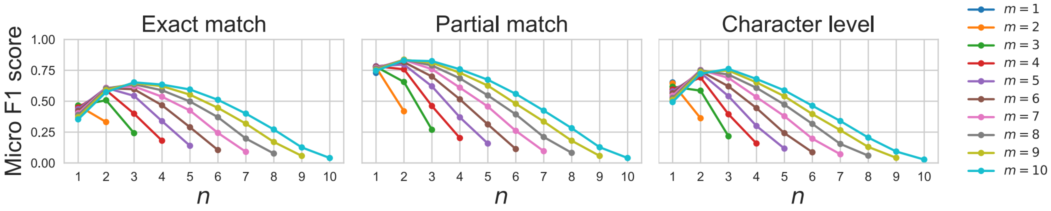


Figure 6: Agreement between the gold standard data and n-match aggregations from m annotations.

Part-of-speech	PRO	SUP	PRO_BY	SUP_BY	Average
Noun	85.99	97.66	90.06	90.76	90.17
Verb	9.53	0.60	4.24	4.79	5.76
Auxiliary verb	1.52	0	1.30	1.22	1.09
Adjective	0.59	0.04	0.45	0.22	0.41
Mark	2.15	1.61	0.35	2.56	2.27
Particle	0.19	0.04	0.40	0.45	0.27
Adverb	0.03	0.04	0.02	0	0.02
Prefix	0	0	0.04	0	0.01

Table 3: Percentage of part-of-speeches of head words of annotated spans.

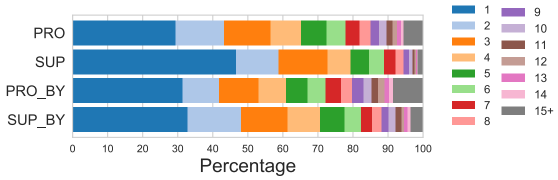


Figure 7: Distribution of word numbers in an annotated span.

4.3 Improving the annotation guidelines

In Section 3.2, we explained the conflict between defining noun phrases and verb phrases as the units of annotation spans. To which part-of-speech did the crowd workers tend to annotate causal relations? The ratios of part-of-speeches labeled during the annotations are listed in Table 3. Here, we focused on the part-of-speech of the last word of the annotated phrases[6]. As shown in Table 3, noun phrases constitute approximately 90.2% of the annotated spans, distantly followed by verb phrases (5.7%).

Further investigations revealed that noun phrases can be annotated within the verb phrases annotated by the workers. For example, when a worker annotates the verb phrase "increases the risk" with PRO, the noun phrase "the risk" can be annotated with the same relation. To estimate the number of such instances, we manually analyzed 300 randomly selected verb phrases from the annotations. We found that 53.0% of verb phrases were re-annotatable as noun phrases. Therefore, it may be sufficient to limit annotation spans to noun phrases in the guideline.

Figures 7 and 8 depict the distributions of the numbers of words and bunsetsu chunks[7], respectively, in an annotated span. Naturally, we observe that shorter phrases occupy most of the annotations. Unfortunately, the length of the spans to be annotated cannot be clarified. Therefore, determining the noun phrases prior to an annotation work may be unreasonable, but allowing crowd workers to choose their segment boundaries might be necessary.

[6] In Japanese, both noun and verb phrases are head final.

[7] The smallest meaningful sequence consisting of content word(s) attached with function word(s).

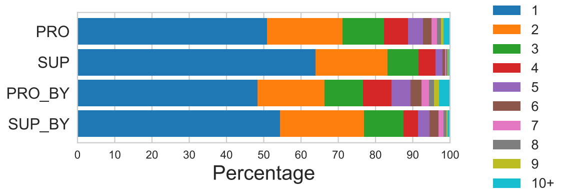


Figure 8: Distribution of numbers of bunsetsu chunks in an annotated span.

4.4 Annotation confusions

How do the crowd workers make erroneous annotations, and what relations are likely to be confused? To answer these questions, we analyzed the tendency of annotation confusions by comparing the data created by the 2-match aggregation and individual annotation data. Here, we define that a confusion occurs when the label assigned to a span differs from that allocated in the 2-match aggregation.

We analyze annotation confusions by the following method. Suppose that causal labels PRO, SUP, PRO_BY, and SUP_BY are annotated 4000, 3000, 2000, and 1000 times in a corpus, respectively. In other words, the annotation ratios of the PRO, SUP, PRO_BY, and SUP_BY labels are 40%, 30%, 20%, and 10%, respectively. In addition, suppose that some spans that should be labeled PRO are incorrectly labeled with SUP, PRO_BY, and SUP_BY. Let the number of incorrect labels be 200, 300, and 100, respectively.

Assuming that labeling errors follow the same probability distribution as the individual labels, the expectation of incorrectly labeling PRO as SUP is given by

$$(200 + 300 + 100) \times \frac{30}{30 + 20 + 10} = 300. \quad (1)$$

The peculiarity of the confusions PRO → SUP can be measured by the deviation between the number of incorrect annotations and the above expectation. Here, we adopt a modified chi-squared test (in which the numerator is not squared):

$$\frac{\text{observation} - \text{expectation}}{\text{expectation}} = \frac{\text{observation}}{\text{expectation}} - 1. \quad (2)$$

For PRO → SUP confusions, Equation 2 gives $200/300 - 1 = -0.333$, indicating that the annotation errors were 33.3% fewer than expected.

		incorrect			
		PRO	SUP	PRO_BY	SUP_BY
correct	PRO	-	-0.510	0.425	0.019
	SUP	-0.612	-	-0.405	1.037
	PRO_BY	0.556	-0.198	-	-0.567
	SUP_BY	-0.222	0.969	-0.670	-

Table 4: Deviations (ratios) from expected numbers of annotation errors.

Table 4 shows the results of Equation 2 for all kinds of confusions. According to this table, few workers confused the polarity of a causal relation, e.g., PRO and SUP. Most of the confusions were caused by the direction of a causal relation such as PRO and PRO_BY. Such confusions might be reduced by decomposing the annotation task into two steps. In the first step, workers could annotate the polarity of a causal relation regardless of its direction (i.e., by equating PRO and PRO_BY). The direction of the causal relations, e.g., PRO and PRO_BY, could then be classified in the next step.

4.5 Nested structure of promotion and suppression

Figure 4 shows an interesting example of nested spans of promotion and suppression. To examine patterns that reverse the polarity of causal relations, we extract regions containing overlapping regions of PRO and SUP annotations. The overlapping regions can be divided into four types: (PRO = SUP) the regions of PRO and SUP are identical; (PRO $\supset$ SUP) the region of PRO contains that of SUP; (PRO $\subset$ SUP) the region of SUP contains that of PRO; and (OTHER) the regions of PRO and SUP have overlaps but no inclusion relation.

Table 5 gives an example and lists the number of instances of each type. The majority of nested spans occur when PRO completely contains SUP (PRO $\supset$ SUP). This means that suppression relations are often described by polarity inversion patterns such as "decrease in *X*", "prevent *X*", and "reject *X*". In cases of two polarity inversion patterns, e.g., "fail to prevent unintentional results", we find opposite inversions in which SUP completely contains PRO (PRO $\subset$ SUP). The type PRO = SUP simply denotes annotation errors (confusions between PRO and SUP).

Table 6 lists some polarity inversion patterns mined by this analysis. Some of these patterns

Type	Number	Example		
		A part of a sentence	Pro	Sup
Pro = Sup	154	paralysis of the limb occurs	paralysis of the limb	paralysis of the limb
Pro $\supset$ Sup	1,850	exhibits a decrease in platelets	a decrease in platelets	platelets
Pro $\subset$ Sup	54	fail to prevent unintentional results	unintentional results	prevent unintentional results
OTHER	85	can control smoke caused by fire	control smoke	smoke caused by fire

Table 5: Examples and numbers of Pro and Sup overlaps.

Japanese	English	Number
X障害	X disorder	53
Xの低下	decline in X	25
X異常	X abnormality	12
X減少	decrease in X	9
Xを阻害	inhibition of X	7
Xの治療	treatment of X	7
Xが障害される	X is impaired	6
抗X	anti-X	6
Xを防ぐ	prevent X	5
Xの制御	control of X	5
X被害	X damage	4
X汚染	X pollution	4
Xを拒否	reject X	3
Xの代替	alternative to X	3

Table 6: Examples of polarity inversion patterns.

are easily crafted by humans, e.g., "decline in X" and "prevent X". However, this analysis also mines novel patterns within noun phrases, e.g., "X damage (health damage)" and within words, e.g., "anti-X (antidepressant)".

4.6 Automatic recognition of causal relations

How do the data created in this research contribute to acquiring causal relation instances from Wikipedia articles? We formalize this task as a sequential labeling problem of predicting labels of promotion/suppression for words in a sentence. We use the data built by 2-match aggregation as a training data. Because the dataset includes spans with multiple relation labels (as explained in Section 4.5), we build a model for each relation.

The sequential labeling was performed by a one-layer bi-directional LSTM. The dimension of the input word vectors and the hidden layer was 300. In addition, word vectors were initialized with ones trained on Japanese Wikipedia articles. The IOB2 notation was applied to the causal relations, such as B-Pro, I-Pro, B-Sup, I-Sup. All occurrences of the title phrase in the article text were replaced with __TITLE__. With this replacement, the model can learn the textual clues between the title phrase and an argument of a relation. We also deleted expressions in parentheses, which often describe pronunciations.

The F1 scores of Pro, Sup, Pro_by and Sup_by were 0.365, 0.282, 0.315 and 0.167, respectively. Although the F1 scores are relatively low, the prediction performance is reasonable because the F1 score of the annotator agreement was approximately 0.5.

5 Conclusion

We presented a crowdsourcing-based approach for annotating causal relation instances to Wikipedia articles. For this purpose, we designed a simple microtask in which crowd workers annotated textual spans having causal relations with the title of a Wikipedia article. To provide an easy-to-use interface with sufficient quality control, we integrated the crowdsourcing service with the brat interface. The annotated corpus not only provides supervision data for automatic recognition of causal relation instances, but also reveals valuable facts for improving the annotation process.

In the imminent future, we consider refining the annotation process as suggested in Section 4, and increase the size and variety of the corpus. We will also extend target articles to other languages (e.g., English) because the approach in this paper is not language-specific. In addition to the intrinsic evaluation (automatic recognition of causal relations), we plan to extrinsically evaluate the corpus; for example, by applying the model trained on the corpus to a downstream task such as question answering and stance detection

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15H01702 and 15H05318 and JST CREST.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. of SIGMOD 2008*, pages 1247–1250.

Anthony Brew, Derek Greene, and Pádraig Cunningham. 2010. Using crowdsourcing and active learning to track sentiment in online media. In *Proc. of ECAI 2010*, pages 145–150.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Proc. of LREC 2004*, pages 837–840.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proc. of the 11th Linguistic Annotation Workshop*, pages 95–104.

Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88.

Juliane Fluck, Sumit Madan, Tilia Renate Ellendorff, Theo Mevissen, Simon Clematide, Adrian van der Lek, and Fabio Rinaldi. 2015. Track 4 overview: Extraction of causal network information in biological expression language (BEL). In *Proc. of the Fifth BioCreative Challenge Evaluation Workshop*, pages 333–346.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Matthew R. Gormley, Adam Gerber, Mary Harper, and Mark Dredze. 2010. Non-expert correction of automatically generated relation annotations. In *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 204–207.

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 *Proc. of EMNLP-CoNLL 2012*, pages 619–630.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong-Hoon Oh. 2015. Generating event causality hypotheses through semantic relations. In *Proc. of AAAI 2015*, pages 2396–2403.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proc. of the 5th International Workshop on Semantic Evaluation*, pages 33–38.

Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proc. of ACL 2014*, pages 377–382.

Mukund Jha, Jacob Andreas, Kapil Thadani, Sara Rosenthal, and Kathleen McKeown. 2010. Corpus creation for new genres: A crowdsourced approach to PP attachment. In *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 13–20.

Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. 2014. Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proc. of COLING 2014*, pages 269–278.

Nolan Lawson, Kevin Eustice, Mike Perkowitz, and Meliha Yetisgen-Yildiz. 2010. Annotating large email datasets for named entity recognition with mechanical turk. In *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 71–79.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *Proc. of CIKM 2007*, pages 233–242.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra- and intersentential causal relations. In *Proc. of ACL 2013*, pages 1733–1743.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. 2016. A semi-supervised learning approach to why-question answering. In *Proc. of AAAI-16*, pages 3022–3029.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, and Sophia Ananiadou. 2015. Overview of the cancer genetics and pathway curation tasks of BioNLP shared task 2013. *BMC Bioinformatics*, 16(10):S2.

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

Ines Rehbein and Josef Ruppenhofer. 2017. Catching the common cause: Extraction and annotation of causal relations and their participants. In *Proc. of the 11th Linguistic Annotation Workshop*, pages 105–114.

Fabio Rinaldi, Tilia Renate Ellendorff, Sumit Madan, Simon Clematide, Adrian van der Lek, Theo Mevissen, and Juliane Fluck. 2016. BioCreative V track 4: a shared task for the extraction of causal network information using the Biological Expression Language. *Database: The Journal of Biological Databases and Curation*, page baw067.

Akira Sasaki, Junta Mizuno, Naoaki Okazaki, and Kentaro Inui. 2016. Stance classification by recognizing related events about targets. In *Proc. of WI 2016*, pages 582–587.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision. In *Proc. of EMNLP 2016*, pages 138–148.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proc. of EACL 2012 (demonstrations)*, pages 102–107.

Sho Takase, Naoaki Okazaki, and Kentaro Inui. 2016. Composing distributed representations of relational patterns. In *Proc. of ACL 2016*, pages 2276–2286.