Weakly Supervised Multilingual Causality Extraction from Wikipedia

Chikara Hashimoto
Yahoo Japan Corporation
1-3 Kioicho, Chiyoda-ku, Tokyo 102-8282, Japan
chashimo@yahoo-corp.jp

Abstract

We present a method for extracting causality knowledge from Wikipedia, such as *Protectionism* → *Trade war*, where the cause and effect entities correspond to Wikipedia articles. Such causality knowledge is easy to verify by reading corresponding Wikipedia articles, to translate to multiple languages through Wikidata, and to connect to knowledge bases derived from Wikipedia. Our method exploits Wikipedia article sections that describe causality and the redundancy stemming from the multilinguality of Wikipedia. Experiments showed that our method achieved precision and recall above 98% and 64%, respectively. In particular, it could extract causalities whose cause and effect were written distantly in a Wikipedia article. We have released the code and data for further research.

1 Introduction

Much of the world consists of entities that causally depend on each other. Therefore, causality knowledge, e.g., *Protectionism* → *Trade war*,[1] is useful for many tasks such as why-QA (Oh et al., 2017), reading comprehension (Berant et al., 2014), and event prediction (Radinsky et al., 2012).

Although many methods have been proposed for causality extraction from text (Ning et al., 2018; Gao et al., 2018; Kruengkrai et al., 2017; Rehbein and Ruppenhofer, 2017; Dunietz et al., 2017; Hidey and McKeown, 2016; Zhao et al., 2016; Hashimoto et al., 2014, 2012; Do et al., 2011; Riaz and Girju, 2010; Abe et al., 2008), they have rarely addressed three issues that are important for constructing a causality knowledge base (CKB). First, we should be able to *verify* extracted causalities, so that the CKB can sustain the credibility of its information.[2] Second, it would be desirable to easily *translate* the CKB to multiple languages, to avoid duplicating the construction effort for different languages. Third, it would also be desirable to automatically *connect* the CKB to other knowledge bases (KBs), to bring together KB construction efforts in various communities and thus maximize their synergistic effect.

Therefore, we propose a method for extracting causalities from Wikipedia by using cause and effect entities that correspond to Wikipedia articles; for example, for *Tobacco* → *Lung cancer*, English Wikipedia has articles titled *Tobacco* and *Lung cancer*. Such causalities satisfy the above three desiderata. First, we can easily *verify* such causalities, because Wikipedia articles tend to credibly attest them; for example, the *Tobacco* article states that inhaling its smoke can cause *Lung cancer*.[3] In contrast, knowledge from other sources such as the web text tends to be difficult to verify, owing to a deluge of false information. Second, causalities extracted from Wikipedia can be *translated* trivially to multiple languages, because Wikidata (Vrandečić and Krötzsch, 2014), a free, multilingual KB, provides links among Wikipedia articles on the same subject in different languages. Third, it is easy to *connect* causalities whose cause and effect are the topics of Wikipedia articles to other KBs derived from Wikipedia, e.g., Freebase (Bollacker et al., 2008), DBpedia (Lehmann et al., 2009), Knowledge Graph (Singhal, 2012), YAGO2 (Hoffart et al., 2013), and Wikidata. In addition, Wikidata also provides links to external KBs, which connect the causalities to those KBs.

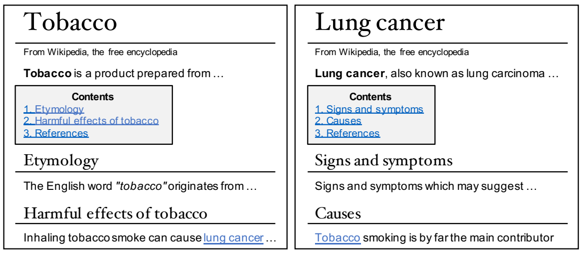
Because our task is to identify causal relations between entities that are the topics of Wikipedia articles, a simple solution would be to use relation extraction (RE) methods (Vashishth et al., 2018; Zhang et al., 2018, 2017; Zhou et al., 2016;

[1] In this paper, $A \rightarrow B$ denotes that A causes B.

[2] The verifiability in this paper means that one can verify extracted causalities with credible sources of information.

[3] Most of the Wikipedia contents mentioned in this paper were downloaded on January 7th, 2019.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2988–2999,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics



Tobacco

From Wikipedia, the free encyclopedia

Tobacco is a product prepared from …

Contents
1. Etymology
2. Harmful effects of tobacco
3. References

Etymology

The English word *"tobacco"* originates from …

Harmful effects of tobacco

Inhaling tobacco smoke can cause lung cancer …

Lung cancer

From Wikipedia, the free encyclopedia

Lung cancer, also known as lung carcinoma …

Contents
1. Signs and symptoms
2. Causes
3. References

Signs and symptoms

Signs and symptoms which may suggest …

Causes

Tobacco smoking is by far the main contributor

Figure 1: Sections that describe causality in Wikipedia: *Harmful effects of tobacco* and *Causes*. We can extract *Tobacco* → *Lung cancer* from these sections.

Lin et al., 2016; Wang et al., 2016; Cai et al., 2016; Shen and Huang, 2016). Most RE methods, however, identify relations between entities co-occurring in a sentence (with exceptions such as Quirk and Poon, 2017; Peng et al., 2017) and cannot deal with a large portion of our test data (67.3%, as we detail in §3.2), because the cause and effect entities in the data tend to appear distantly in a Wikipedia article. This is reasonable given that the subject of an encyclopedia article would not typically be repeated throughout the article, as the author may expect readers to know that the whole article is about the subject.

We aim to extract causalities from Wikipedia regardless of whether their cause and effect co-occur in a sentence by training a causality classifier from causality instances whose cause and effect may appear distantly. Here, we face two problems:

Lack of labeled data: For the learning of the causality classifier, there is no data marking causes and effects in Wikipedia articles, and manually creating such data would be laborious. A possible solution would be distant supervision (Mintz et al., 2009) using external KBs to mark relevant entities in articles. Because Wikipedia evolves rapidly, however, relying on such KBs would make the resulting CKB obsolete. Therefore, how can we instantly acquire such data directly from Wikipedia?

Lack of redundancy: An encyclopedia tends to avoid redundancy in text, which many knowledge extraction methods rely on to reliably estimate the probability of extracted knowledge (Christensen et al., 2011; Downey et al., 2005). Therefore, how can we reliably estimate the probability of a causality extracted from Wikipedia?

For the lack of labeled data, we can accurately extract hundreds of causalities by harnessing the property of Wikipedia that some articles have sections describing relevant causalities. For example, as shown in Fig. 1, the *Lung cancer* article has a section named *Causes* that mentions *Tobacco*, and the *Tobacco* article has a section named *Harmful effects of tobacco* that mentions *Lung cancer*. Accordingly, we can extract *Tobacco* → *Lung cancer* and then use it for the learning of the classifier. All the supervision we need is a handful of keywords for such sections, e.g., *causes* and *effects*.

For the lack of redundancy, we exploit another kind of redundancy stemming from the *multilinguality* of Wikipedia: the same subject may be described by articles in different languages; for example, *Tobacco* is described in 112 languages. We thus use a data source consisting of Wikipedia articles in nine languages: English (en), German (de), French (fr), Spanish (es), Italian (it), Portuguese (pt), Swedish (sv), Dutch (nl), and Polish (pl). This requires our method to be language independent. For these nine languages, however, the only required linguistic analysis is detection of word boundaries, i.e., white space. This simplicity has an advantage of allowing our method to parse Wikipedia quickly to keep the CKB up to date.

We evaluated our method by using the relation triples in Wikidata, which represents a causality by either a has cause or a has effect relation (§3). Our method achieved precision and recall above 98% and 64%, respectively, rivaling an *oracle* relation extractor that perfectly detected the causality between entities co-occurring in a sentence. We also confirmed that the multilingual redundancy of Wikipedia was effective: using more languages led to significantly better performances.

Our contributions are five-fold. (1) We proposed the three desiderata for CKBs: *verifiability*, *translatability*, and *connectivity*. (2) We presented the ideas of exploiting the *causality-describing sections* and *multilingual redundancy* of Wikipedia. (3) Based on these ideas, we proposed a weakly-supervised, multilingual causality extraction method. (4) We evaluated our method in relatively large-scale settings. (5) We have released the code and data from this study (§6).[4]

In this paper we focus on causality extraction. We will present how to construct the CKB from causalities extracted by our method in future.

In this study, we define a causality $A \rightarrow B$ according to Wikipedia and Wikidata. Specifically,

[4] The code and data from this study will be available at https://research-lab.yahoo.co.jp/people/chikara_hashimoto.html.

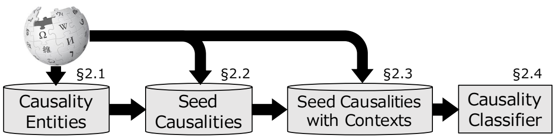


Figure 2: Overview of our method. Numbers indicate sections that describe corresponding components.

A → *B* if Wikipedia describes *A* as causing *B* either explicitly or implicitly, or if *A* (*B*) has the has effect (has cause) relation to *B* (*A*) in Wikidata.

2 Proposed Method

Our method learns a causality classifier that, given an entity pair (e_1, e_2), determines if $e_1 \rightarrow e_2$ holds. The entities are Wikidata identifiers (IDs) having at least one corresponding Wikipedia article. For example, Q1566 is the Wikidata ID for *Tobacco*, which is described in Wikipedia in 112 languages.

Figure 2 shows the learning process; the method identified entities that tend to participate in causality (§2.1), extracted seed causalities between such entities from causality-describing sections as illustrated in Fig. 1 (§2.2), extracted the contexts of the seed causalities from articles in multiple languages (§2.3), and finally learned the classifier from the multilingual contexts of the seeds (§2.4).

The resulting classifier thus takes (e_1, e_2) as input, examines the multilingual contexts of e_1 and e_2, and determines if $e_1 \rightarrow e_2$ holds.

The Python code that accompanies this paper is an implementation of our method (§6).

2.1 Causality Entity Extraction

Some entities, e.g., *Tobacco*, are more likely to participate in causality than others, e.g., *A (the alphabet letter)*. We call them *causality entities*. To accurately extract seed causalities in the next step, we first extracted causality entities by identifying articles that had causality-describing sections: *Tobacco* and *Lung cancer* in Fig. 1 were thus regarded as causality entities.

To identify causality-describing sections, we manually prepared a handful of keywords that tended to appear in the titles of such sections for the nine languages: en, de, fr, es, it, pt, sv, nl, and pl. This was the only supervision of our method. Specifically, we chose *Cause*, *Causes*, *Effect*, and *Effects* as such keywords for en and translated them to the other languages by reference to Wiktionary.[5] We have released these keywords (§6).

We then extracted causality entities from the Wikipedia dump of each language and kept only those entities that appeared in more than one language. The identify of an entity across languages could be confirmed by Wikidata IDs; e.g., *Tobacco* (en) and *Tabaco* (es) both correspond to Q1566.

2.2 Seed Causality Extraction

A seed causality is an entity pair (e_1, e_2) such that e_1 appears in a causality-describing section, whose title contains *Cause* or *Causes* (in the case of en), in the article corresponding to e_2; and such that e_2 appears in a causality-describing section, whose title contains *Effect* or *Effects*, in the article corresponding to e_1. For instance, *Tobacco* → *Lung cancer* in Fig. 1 is a seed causality, as it satisfies the above condition.

We extracted seed causalities in this way, kept only those that appeared in more than one language, and consequently acquired 879 seed causalities from the nine languages. We have also released the seed causalities (§6).

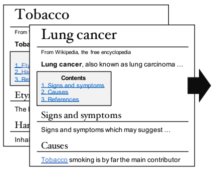
2.3 Seed Causality Context Extraction

As illustrated in Fig. 3, for each entity in a seed causality, we extracted its contexts from articles in multiple languages as features for the classifier.

The context window was up to 100 words to the left and right of each target entity. When a section title appeared in a context window, we shrank the window to the section title position so that the window would not cross the section boundary.

We restricted the articles from which contexts were extracted as follows. For $e_1 \rightarrow e_2$, the context of e_1 (e_2) was extracted from the article that corresponded to e_2 (e_1). For example, for *Tobacco* → *Lung cancer*, the context of *Tobacco* was extracted only from the *Lung cancer* article, and that of *Lung cancer* was extracted only from the *Tobacco* article. This reduced the processing time and helped extract only highly relevant contexts for a target causality. We extracted the contexts directly from the Wikipedia source texts with all markups kept intact, because those markups may have helped the classifier, and parsing the source texts would have slowed the process. In addition, we replaced e_1 and e_2 in the extracted contexts with the special symbols `__CAUSE__` and `__EFFECT__`.

[5] https://en.wiktionary.org/



English Wikipedia

```
there is sufficient
evidence that involuntary
smoking (exposure
to secondhand or
'environmental' __CAUSE__
smoke) causes __EFFECT__ in
humans.
```

French Wikipedia

```
la cause la plus
fréquente de
__EFFECT__ est
l'exposition
chronique à la fumée
de [[#tabagisme|
__CAUSE__ ]]
```

Spanish Wikipedia

```
algunos tipos de cáncer se
encuentran directamente
relacionados con el
__CAUSE__ , como es el
caso del [[__EFFECT__]] y
otros cánceres del tracto
respiratorio
```

Figure 3: Seed causality context extraction from Wikipedia articles of multiple languages.

We also removed line-break characters to extract contexts from across multiple lines.

For each seed causality, we concatenated all the contexts from multiple languages. For context extraction, we used only four languages —en, de, fr, and es— because they had the broadest coverage.

We also removed infoboxes from articles before context extraction, because we used Wikidata's triples for evaluation, and some of those have been transcribed into infoboxes.[6] This was for fairness rather than methodological considerations.

2.4 Learning Classifier

For learning of the causality classifier, we used the seed causalities as positive instances. In contrast, negative (non-causality) instances were entity pairs such that the article corresponding to one entity had a link to the article corresponding to the other entity. For example, the article on *Barack Obama* has a link to the article on *Hillary Clinton*, and hence, *Barack Obama* → *Hillary Clinton* could be used as a negative instance. In this way, we could obtain sensible negative instances, meaning that they were not totally random pairs but were semantically related to each other in some way. We obtained the same number (879) of negative instances as positive ones and similarly extracted their multilingual contexts; thus, we trained the classifier with 1,758 instances.

We used fastText (Joulin et al., 2016), a linear text classifier that averages word embeddings and makes a prediction based on the averaged embeddings, as the causality classifier because of its speed and accuracy. Among its hyperparameters, we tuned only the number of epochs and the learning rate, as detailed in §3.2. We did not use pretrained word embeddings for the classifier to minimize its dependence on external resources.

Although there would be more sophisticated approaches to modeling the multilingual contexts extracted from Wikipedia, we intentionally adopted the very simple modeling in this paper in order to show that even such a simple method could deliver good performances by using the multilingual contexts of Wikipedia. We will develop more sophisticated models in future.

3 Experiments

Using relation triples in Wikidata (§3.1), we evaluated our method by comparing it with various baselines (§3.2). We also measured the effectiveness of using multiple languages (§3.3).

We can summarize the results as follows. Our method (1) extracted causality with precision and recall above 98% and 64%, respectively; it (2) rivaled the performance of an oracle relation extractor that worked sentence-wise; and it (3) effectively used multiple languages.

3.1 Test Data

Our experiments were based on labeled data derived from Wikidata, which has various relation triples (e_1, rel, e_2), where e_1 and e_2 are entities ("item identifiers" in Wikidata terms, such as Q1566) between which the relation rel (a "property" in Wikidata terms, such as has cause) holds.

In short, we used triples whose relations were either has cause or has effect as positive (causality) instances and those with other relations as negative (non-causality) instances. Because we aimed to extract causalities that were easy to verify by reading individual Wikipedia articles, we restricted the triples to those whose component entities co-occurred in an article. Specifically, we used only triples such that the article corresponding to one entity had a link to the article corresponding to the other entity in at least one language. Consequently, we obtained 1,524 positive instances and the same number of negative instances, giving 3,048 instances in total.

The data was in no way easy to classify, because all the negative instances were not random pairs but had semantic relations that are as natural and common as causality. For example, in Wikidata, *World War I* (Q361) and the *Paris Peace Conference* (Q199820) show causality, as the former has

[6] https://www.wikidata.org/wiki/Wikidata:WikiProject_Infoboxes

the has effect relation with the latter. In contrast, *World War I* and the *German invasion of Belgium* (Q5551414) do not show causality, because they have a significant event relation between them.

The data accompany the paper (§6).

3.2 Performance of Proposed Method

Experimental Settings for Proposed Method

In this section, we denote our proposed method as PROP. In the experiments, it extracted multilingual contexts for each relation triple in the data, as described in §2.3. The data were used only for testing. Training was conducted using the automatically acquired seed causalities (§2).

Hyperparameter tuning for PROP was based on five-fold cross validation using the seed causalities so that we could maximize the F1 score on them. We only tuned the number of epochs and the learning rate of fastText; the former was chosen from 30, 50, 70, and 100, while the latter was chosen from 0.3, 0.5, 0.7, and 1.0. As a result, they were set to 100 and 1.0, respectively. The other fastText parameters were set to default values.

We used the default fastText threshold for classification. We conducted 10 runs and averaged the performance scores (accuracy, precision, recall, and F1) for evaluation.

Compared Methods

We evaluated the performance of PROP by comparing it with the following baseline methods described below: SECTION, INFOBOX, RELATED, and ORACLE RE.

SECTION This was simply the method of seed causality extraction described in §2.2, which extracted 879 causalities from Wikipedia. This was intended to evaluate two questions. The first was how well our idea of exploiting causality-describing sections in Wikipedia worked on its own. The second question was how much PROP improved the results, as it used the classifier trained on the output of the SECTION method.

INFOBOX This method extracted causality from the infoboxes in Wikipedia articles in the nine languages. Some infoboxes contain information on causality; for example, the article on *Candidiasis* has an infobox with the field *Causes*, whose value is *Candida*, which indicates *Candida* → *Candidiasis*. For the INFOBOX method to identify such causality-describing fields in infoboxes, we used the same set of keywords that PROP used to identify causality-describing sections. As discussed in §2.3, some Wikidata triples have been transcribed into infoboxes, which should give the INFOBOX method an advantage over PROP, as PROP was forbidden to use infoboxes.

RELATED This method regarded an entity pair whose entities were semantically related to each other as a causality, because such relatedness has been shown to imply causality (Do et al., 2011; Riaz and Girju, 2010). Specifically, this method used a semantic relatedness defined as $1-sr(a,b)$, where $sr(a,b)$ is a Wikipedia-link-based distance measure that was proposed by Milne and Witten (2008) and has been widely used (Lee et al., 2015). The measure is defined

$$sr(a,b) = \frac{log(max(|A|,|B|)) - log(|A \cap B|)}{log(|W|) - log(min(|A|,|B|))},$$

where a and b are the two articles (entities) of interest, A and B are the sets of all articles that link to a and b, respectively, and W is the entire set of Wikipedia articles. To classify entity pairs, we set a threshold for the relatedness score so that we could maximize F1 on the training data for PROP (§2.4). The threshold value ranged from 0.00 to 1.00 at intervals of 0.01. We used the link structure of en Wikipedia because it is the largest one.

ORACLE RE This method used an *oracle* relation extractor that, given two entities that co-occur in a sentence, never fails to determine whether they have a causal relation. For other pairs, whose component entities did not co-occur in a sentence, this method uniformly guessed that they had no causal relation. We used en Wikipedia for ORACLE RE because it has the broadest coverage. We segmented articles into sentences with spaCy,[7] which accurately recognizes sentence boundaries through dependency parsing. This method would show the upper-bound performance for our task of typical RE methods that work sentence-wise. This was an ambitious baseline given the performances of state-of-the-art RE methods: Vashishth et al. (2018) achieved a top-100 precision of 84% on the New York Time corpus (Riedel et al., 2010); Wang et al. (2016) achieved an F1 score of 88% on SemEval-2010 Task 8 (Hendrickx et al., 2010).

	Acc	Prec	Rec	F1
Section	50.56	100.00	1.12	2.21
Infobox	53.71	100.00	7.41	13.81
Related	68.86	66.23	76.97	71.20
Oracle RE	75.89	100.00	51.77	68.22
Prop	81.45	98.28	64.02	**77.53**

Table 1: Performance results of the compared methods.

Results

Table 1 lists the accuracy (Acc), precision (Prec), recall (Rec), and F1 of the compared methods.

The Section method achieved 100% precision, which indicates that our technique of exploiting causality-describing sections in Wikipedia could accurately extract causalities. As the method's recall indicates, however, it covered only a small portion of our target causalities.

The Infobox method also achieved 100% precision, though its coverage was also quite limited.

The Related method exhibited the highest recall, but the precision was unacceptably low for the subsequent manual labor that would be required to construct the CKB.

The Oracle RE method achieved 100% precision by design. Its recall was rather low because 67.3% of the entity pairs in the data (§3.1) consisted of entities that did NOT co-occur in a sentence. This means that most RE methods that work sentence-wise will miss a large portion of causalities, regardless of their accuracy.

Finally, Prop achieved the best F1 score, though its recall still had room for improvement. The fact that Prop outperformed the baselines, especially Oracle RE, clearly shows the effectiveness of our method.

3.3 Language Ablation Test

We also examined the effect of using more languages for extracting contexts by comparing the performance of the following variations of Prop: Prop, Prop$_{en.de.fr}$, Prop$_{en.de.es}$, Prop$_{en.fr.es}$, Prop$_{en}$, Prop$_{en.de}$, Prop$_{en.fr}$, and Prop$_{en.es}$. Unlike Prop, which used en, de, fr, and es for extracting multilingual contexts, the others used only the languages indicated by the subscript language symbols. For example, Prop$_{en.de}$ used only en and de Wikipedia articles.

For each variation of our method, we conducted 10 runs and averaged the resulting performance scores for evaluation.

	Acc	Prec	Rec	F1
Prop$_{en.de.fr.es}$	81.45	98.28	64.02	**77.53**
Prop$_{en.de.fr}$	80.54	98.16	62.24	76.18*
Prop$_{en.de.es}$	80.82	98.65	62.49	76.51*
Prop$_{en.fr.es}$	79.20	98.14	59.52	74.10*
Prop$_{en.de}$	79.92	99.21	60.32	75.26$^\diamond$
Prop$_{en.fr}$	78.39	98.32	57.76	72.77$^\diamond$
Prop$_{en.es}$	77.91	98.91	56.45	71.88$^\diamond$
Prop$_{en}$	76.43	98.91	53.45	69.40

Table 2: Results of language ablation tests. Prop is denoted as Prop$_{en.de.fr.es}$ for clarity. $*$ and $\diamond$ indicate statistically significant differences from the performances of Prop$_{en.de.fr.es}$ and Prop$_{en}$, respectively (paired t-test: $p < 0.01$).

Table 2 summarizes the results of the ablation tests. The upper half of the table indicates that removing one language from Prop tended to degrade its performance, while the lower half indicates that adding one language to Prop$_{en}$ tended to improve its performance. From these results, we conclude that using multilingual contexts is effective for causality extraction from Wikipedia.

3.4 Analysis

We first examine cases in which Prop succeeded and then analyze its error cases.

Causalities whose cause and effect entities co-occur in a sentence tend to have phrases indicating the causal relation between them, which helps identify the relation. For example, for *Adipsia* → *Hypernatremia*, there is the following sentence:

(1) *Adipsia* may be seen in conditions such as diabetes insipidus and may **result in** *hypernatremia*.

For causalities whose cause and effect do not co-occur in a sentence, it is likely that their causal relation are only indicated by multi-line texts or the structure of Wikipedia article. For example, for *Hormone therapy* → *Cancer pain*, there is the following list item in the *Cause* section of the article of *Cancer pain*:

(2) *hormone therapy*, which sometimes **causes** pain flares;

Although *Cancer pain* is not written explicitly in the item, we can guess that the *pain flare* refers to *Cancer pain*, because this list item is part of the contents of the *Cause* section of the article of *Cancer pain*. Prop can identify such a causality,

[7] https://spacy.io/

because pieces of Wikipedia source text that represent this kind of structure can be included in the window of the multilingual contexts and because all contexts are extracted from the articles that correspond to either the cause or the effect.

Regarding PROP's error analysis, we focus here on its false-negative errors, as it achieved high precision while its recall had room for improvement. Such errors were mostly due to the lack of evidence of causality; in some cases, even though the cause and effect entities both appear in an article, it does not indicate their causal relation. For example, for *Psoriasis* → *Woronoff's ring*, the cause is only mentioned as a list item in the *See also* section of the effect entity's article in en Wikipedia.

Other false-negatives included entity pairs for which not only causality but other relations hold. For example, for *International University Sports Federation* → *Universiade*, both the has effect and the organizer relations hold. These tend to be cases in which the Wikipedia article describes the cause as *organizing* (instead of *causing*) the effect.

4 Discussion

4.1 Three Desiderata for CKB

In §1 we discussed the desiderata for the CKB: verifiability, translatability, and connectivity. In this section we consider how well the causalities extracted by PROP satisfied the desiderata.

Verifiability

We examined 100 samples from the causalities in the data (§3.1) that PROP correctly classified as causalities (i.e., true positives) to measure their verifiability. In other words, we examined how many of them consisted of cause and effect entities between which we could easily identify causality by reading their individual Wikipedia articles. Causalities that were not described in en Wikipedia were regarded as unverifiable, because we assumed that our target users could understand only English. Of the 100 samples, 19 were not written in en Wikipedia.

As a result, 62.0% of the samples were verifiable. If we ignore those 19 samples that were not written in en Wikipedia, 76.5% of the samples were verifiable. For example, for *Onchocerca volvulus* → *Onchocerciasis*, the article on *Onchocerca volvulus* has the following sentence:

(3) *Onchocerca volvulus* is a nematode that **causes** *onchocerciasis*.

We thus conclude that causalities extracted by PROP tend to be verifiable.

Translatability

We next examined the number of languages to which each true-positive causality was translated, by using Wikidata's links among Wikipedia articles on the same subject in different languages. We targeted each of the nine languages. We regarded each causality as translated to a language if its cause and effect were both translated.

As a result, 74.9% of the causalities were translated to more than one language, which indicates that the causalities extracted by PROP tend to exhibit a high degree of translatability. Furthermore, 16.5% were translated to all nine languages, including, e.g., *Chemotherapy* → *Vomiting* (*Q974135* → *Q127076*) and *Treaty of Versailles* → *World War II* (*Q8736* → *Q362*).

Connectivity

We also examined how many of the true positives were connected to external KBs. We first listed all external IDs,[8] which Wikidata uses to identify external KBs such as Freebase, IMDb, and ISBN-13, with a lookup function in Wikidata,[9] resulting in 3,695 external IDs. We then made a table to map each Wikidata ID to the external IDs to which it is connected. With the table we counted the number of true-positive causalities whose cause and effect entities shared at least one external ID.

Consequently, 72.3% of the true-positives were connected to external KBs. For example, *Diabetes mellitus* → *Cataract* was connected to 19 KBs such as MeSH,[10] Freebase, and BabelNet (Navigli and Ponzetto, 2012). We thus conclude that causalities extracted by PROP indicate a high degree of connectivity.

4.2 Independence of Languages

We designed PROP to be language independent so that we could exploit multilingual redundancy. Unfortunately, unintended language dependence might arise as we use more languages. We are currently aware that we will need tokenizers if we use languages for which word boundaries are not as explicit as in the nine languages here, e.g., Chinese and Japanese. It will thus be a future challenge to

[8] https://www.wikidata.org/wiki/Wikidata:Identifiers

[9] https://www.wikidata.org/wiki/Special:ListProperties

[10] https://www.ncbi.nlm.nih.gov/mesh/

use many more languages while keeping PROP as language independent as possible.

4.3 Independence of External KBs

We also designed PROP not to rely on external KBs so that we could easily keep the CKB up to date with Wikipedia. If we relaxed this design policy, we would be able to use the triples in Wikidata as additional training data for our classifier. We thus plan to explore this direction of research.

Another external resource that is useful for PROP would be pre-trained word vectors (Bojanowski et al., 2017). We evaluated PROP$_{en+}$, which is the same as PROP$_{en}$ except that it used pre-trained word vectors.[11] The pre-trained word vectors slightly improved the performance; PROP$_{en+}$ achieved a F1 of 69.51, while PROP$_{en}$'s F1 was 69.40.

5 Related Work

5.1 Causality Extraction

Causality extraction methods can be classified with regard to what constitutes cause and effect: noun phrases, verb phrases, or clauses. The noun-phrase type, e.g., *global warming* → *malaria epidemic*, has mostly been addressed by RE methods, as we discuss in §5.2. The verb-phrase type, e.g., *get fired* → *live on unemployment insurance*, has been extracted by various methods (Ning et al., 2018; Gao et al., 2018; Rehbein and Ruppenhofer, 2017; Kruengkrai et al., 2017; Hashimoto et al., 2015, 2014, 2012; Do et al., 2011; Riaz and Girju, 2010; Abe et al., 2008). The clause type, e.g., *I hid the car key* → *She's mad*, has also been studied (Dunietz et al., 2017). Other types include causal embeddings (Sharp et al., 2016), which can be used for causal question answering (Oh et al., 2017). We focused here on the noun-phrase type because noun phrases can be components of verb phrases and clauses, and hence, our work may also contribute to the extraction of other types.

Another standpoint of classifying causality extraction is the information source, e.g., newspapers (Khoo et al., 1998), the web (Kruengkrai et al., 2017), parallel corpora (Hidey and McKeown, 2016),[12] images (Gao et al., 2018), and videos (Fire and Zhu, 2016). We used Wikipedia articles in multiple languages because they tend to be more credible than other sources, and because they allowed us to exploit multilingual redundancy.

Wikipedia articles in multiple languages also provide causalities that tend to satisfy the three desiderata discussed in §1, which is the novel perspective that we proposed and that many previous studies lacked. In addition, we proposed a novel method that is better suited for extracting such causalities than previous methods were.

5.2 Relation Extraction

In SemEval-2007 Task4 (Girju et al., 2007) and SemEval-2010 Task 8 (Hendrickx et al., 2010), the target relations included "Cause-Effect"; our study is also relevant to methods for those tasks (Lee et al., 2019; Wang et al., 2016; Shen and Huang, 2016; Cai et al., 2016; Xu et al., 2016). Other methods based on RE also addressed causality (Kim and Myaeng, 2016; De Saeger et al., 2011, 2009; Schoenmackers et al., 2010).

Although our method can be regarded as a relation extractor, it is different from the above methods because it is particularly tailored to causality extraction from Wikipedia. For this task, it is important to extract relation instances whose component entities do not co-occur in a sentence.

More recent studies have addressed *inter*-sentential RE for specialized domains (Noriega-Atala et al., 2018; Quirk and Poon, 2017; Peng et al., 2017), and Mandya et al. (2018) constructed a large-scale dataset for this task. Hence, we plan to incorporate these approaches into our method.

5.3 Knowledge Extraction from Wikipedia

Knowledge extraction from Wikipedia in general is also relevant to our study. Although there have been studies on extracting class concepts (Pasca, 2018), trivia (Tsurel et al., 2017), taxonomies (Flati et al., 2014), infobox contents (Wang et al., 2013), and various semantic relations (Wu and Weld, 2010), among other things, causality extraction from Wikipedia has rarely been addressed as far as we are aware. Hidey and McKeown (2016) addressed the extraction of linguistic markers indicating causality from Wikipedia, but those markers were not causalities *per se*.

5.4 Temporal Relation Extraction

Researchers have noticed that causality extraction and temporal relation extraction (Mani et al.,

[11] We used two million word vectors trained with subword information on Common Crawl available at `https://fasttext.cc/docs/en/english-vectors.html`.

[12] Precisely, the task of Hidey and McKeown (2016) is identifying linguistic cues that indicate causality.

2006) share some properties and can complement each other (Ning et al., 2018; Mirza and Tonelli, 2016; Bethard and Martin, 2008). In the future, we will also explore the possibility of exploiting temporal relation extraction methods for our task.

6 Accompanying Code and Data

We hereby release the Python code to implement our method, including modules for generating relevant data, with step-by-step instructions. We also release the following data sets: (a) the keywords for identifying causality-describing sections (§2.1), (b) the seed causalities (§2.2), and (c) the test data (§3.1). The code and data will be available at `https://research-lab.yahoo.co.jp/people/chikara_hashimoto.html`.

Note that 98.1% of the seed causalities are not included in Wikidata as causality, and they can increase Wikidata causalities by about 19.8%.

7 Conclusion

We have proposed a weakly supervised method for extracting causality from Wikipedia articles in multiple languages. The causalities extracted by our method tend to be easy to *verify*, *translate* to multiple languages, and *connect* to external KBs. Our key idea is to exploit the *causality-describing sections* and *multilingual redundancy* of Wikipedia. Our method achieved precision and recall above 98% and 64%, respectively, and it could even extract causalities whose cause and effect entities did not co-occur in a sentence.

References

Shuya Abe, Kentaro Inui, and Yuji Matsumoto. 2008. Two-phased event relation acquisition: Coupling the relation-oriented and argument-oriented approaches. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 1–8.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510.

Steven Bethard and James H. Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (ACL-HLT)*, pages 177–180.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 756–765.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the Sixth International Conference on Knowledge Capture*, pages 113–120.

Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata. 2009. Large scale relation acquisition using class dependent patterns. In *The Ninth IEEE International Conference on Data Mining (ICDM)*, 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 the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 825–835.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 294–303.

Doug Downey, Oren Etzioni, and Stephen Soderland. 2005. A probabilistic model of redundancy in information extraction. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 1034–1041.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. Automatically tagging constructions of causation and their slot-fillers. *Transactions of the Association for Computational Linguistics (TACL)*, 5:117–133.

Amy Fire and Song-Chun Zhu. 2016. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST) - Special Issue on Causal Discovery and Inference*, 7(2):23:1–23:22.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2014. Two is bigger (and better) than one: the wikipedia bitaxonomy project. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 945–955.

Qiaozi Gao, Shaohua Yang, Joyce Chai, and Lucy Vanderwende. 2018. What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 934–945.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 13–18.

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 the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 619–630.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong-Hoon Oh. 2015. Generating event causality hypotheses through semantic relations. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2396–2403.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 987–997.

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 *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 33–38.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1424–1433.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artif. Intell.*, 194:28–61.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Christopher S.G. Khoo, Jaklin Kornfilt, Sung Hyon Myaeng, and Robert N. Oddy. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13:177–186.

Jinho Kim and Sung-Hyon Myaeng. 2016. Discovering relations to augment a web-scale knowledge base constructed from the web. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics (WIMS)*, pages 16:1–16:12.

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. 2017. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 3466–3473.

Joohong Lee, Sangwoo Seo, and Yong Suk Choi. 2019. Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing.

Joonseok Lee, Ariel Fuxman, Bo Zhao, and Yuanhua Lv. 2015. Leveraging knowledge bases for contextual entity exploration. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1949–1958.

Jens Lehmann, Chris Bizer, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2124–2133.

Angrosh Mandya, Danushka Bollegala, Frans Coenen, and Katie Atkinson. 2018. A dataset for inter-sentence relation extraction using distant supervision. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 753–760.

David Milne and Ian H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 25–30.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 1003–1011.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 64–75.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational (ACL)*, pages 2278–2288.

Enrique Noriega-Atala, Paul Douglas Hein, Shraddha Satish Thumsi, Zechy Wong, Xia Wang, and Clayton Thomas Morrison. 2018. Inter-sentence relation extraction for associating biological context with events in biomedical texts. In *IEEE International Conference on Data Mining Workshops (ICDM)*, pages 722–731.

Jong-Hoon Oh, Kentaro Torisawa, Canasai Kruengkrai, Ryu Iida, and Julien Kloetzer. 2017. Multi-column convolutional neural networks with causality-attention for why-question answering. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 415–424.

Marius Pasca. 2018. Finding needles in an encyclopedic haystack: Detecting classes among wikipedia articles. In *Proceedings of the 2018 World Wide Web Conference (WWW)*, pages 1267–1276.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1171–1182.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 2012 World Wide Web Conference (WWW)*, pages 909–918.

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

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Proceedings of the 4th IEEE International Conference on Semantic Computing (ICSC)*, pages 361–368.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, ECML PKDD'10, pages 148–163.

Stefan Schoenmackers, Jesse Davis, Oren Etzioni, and Daniel Weld. 2010. Learning first-order horn clauses from web text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1088–1098.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 138–148.

Yatian Shen and Xuanjing Huang. 2016. Attention-based convolutional neural network for semantic relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2526–2536.

Amit Singhal. 2012. Introducing the knowledge graph: Things, not strings. Corporate blog.

David Tsurel, Dan Pelleg, Ido Guy, and Dafna Shahaf. 2017. Fun facts: Automatic trivia fact extraction from wikipedia. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 345–354.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. Reside: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1257–1266.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57:78–85.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1298–1307.

Zhigang Wang, Zhixing Li, Juanzi Li, Jie Tang, and Jeff Z. Pan. 2013. Transfer learning based cross-lingual knowledge extraction for wikipedia. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 641–650.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 118–127.

Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. 2016. Improved relation classification by deep recurrent neural networks with data augmentation. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 1461–1470.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2205–2215.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 35–45.

Sendong Zhao, Ting Liu, Sicheng Zhao, Yiheng Chen, and Jian-Yun Nie. 2016. Event causality extraction based on connectives analysis. *Neurocomputing*, 173(P3):1943–1950.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 207–212.