# Handling Multiword Expressions in Causality Estimation

Shota Sasaki, Sho Takase, Naoya Inoue, Naoaki Okazaki, Kentaro Inui
Tohoku University
{sasaki.shota, naoya-i, okazaki, inui}@ecei.tohoku.ac.jp
takase.sho@lab.ntt.co.jp

**Abstract**

Previous studies on causality estimation mainly aquire causal event pairs from a large corpus based on lexico-syntactic patterns and coreference relations, and estimate causality by a statistical method. However, most of the previous studies assume event pairs can be represented by a pair of single words, therefore they cannot estimate multiword causality correctly (e.g."tired"-"give up") . In this paper, we create a list of multiword expressions and extend an existing method. Our evaluation demonstrates that the proper treatment of multiword expression events is effective and the proposed method outperforms the state-of-the-art causality estimation model.

## 1 Introduction

This paper addresses *causality estimation*, the task of estimating the strength of causality between two sentences. For example, consider the following two sentences:

(1) a. John was tired of the customer service.
b. John gave up using the product.

The task is to estimate that sentence (1a) is more causally related to sentence (1b) than non-causally related sentences such as "John opened a door.". Causality estimation is considered as an essential component of common sense reasoning.

A conventional approach to causality estimation is to construct a statistical model of causality relying on a large corpus in a semi-supervised manner. The main idea is two-fold: (i) collect causally related *word pairs* (e.g. *typhoon-die*) by exploiting the contextual proximity or discourse markers and (ii) apply them to a correlation measure (Chambers and Jurafsky, 2008; Luo et al., 2016) or a supervised classifier (Riaz and Girju, 2014; Granroth-Wilding and Clark, 2016).

A key limitation of the previous studies is that they model causality in terms of *word pairs*, not taking into account the causality represented by multiword expressions. For example, in example (1), a causality estimation model is expected to consider the causality between *tired* and *gave up* (i.e. *stop something*). However, the previous models consider only word pairs; therefore, it would improperly estimate the causality based on word pairs such as *tired-give* and *tired-up*. Because each individual word in multiword expressions might have a completely different meaning from the whole, it is crucial to solve this problem.

To address the above issue, this paper proposes a method that can estimate the causality between events represented by multiword expressions. Specifically, we obtained the list of multiword expressions from Wiktionary[1] to acquire the causality of multiword expressions from a corpus. Our experiments demonstrate that the proposed method outperforms the state-of-the-art method on Choice of Plausible Alternatives (COPA) (Roemmele et al., 2011), which can be regarded as a variant of causality estimation.

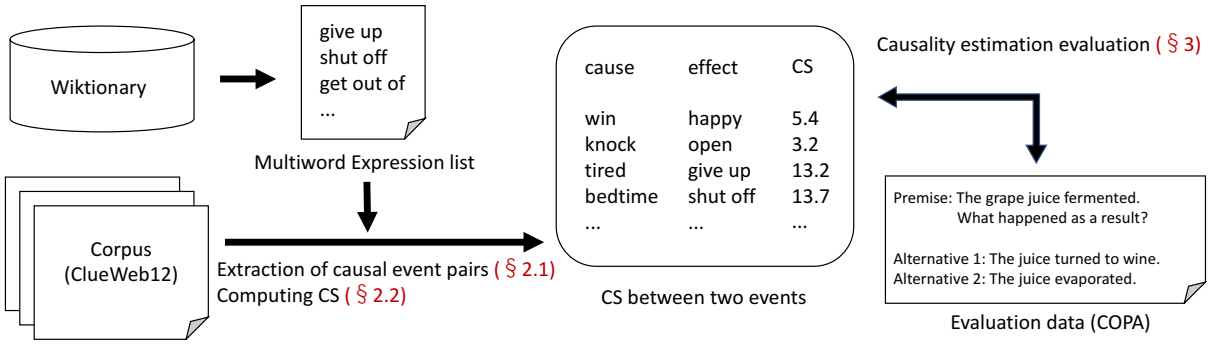


Figure 1: An overview of the proposed method.

## 2 Proposed Method

Following Luo et al. (2016), we extract a pair of causal events from a corpus by using causal markers (e.g. "B because A"), and model the strength of causality by using *Causal Strength* as a statistical measure. During the extraction of causal event pairs, the proposed method considers a multiword expression as a unit of an event as well as a single word. Figure 1 illustrates an overview of the proposed method.

### 2.1 Extraction of Causal Event Pairs

Considering that a template-matching approach is relatively successful in causality estimation (Luo et al., 2016), we first extract sentences matching a predefined template from a corpus. A template consists of a causal marker and two slots $A, B$, where $A, B$ indicates cause and effect, respectively (e.g. "B because A"). This study uses the list of templates provided by Luo et al. (2016).

Suppose that the following sentence is matched with a template "Because A, B":

(2) [Because] John was tired of the customer service$_A$, John gave up using the product$_B$.

From each slot, we extract nouns, verbs, adjectives and adverbs defined in WordNet (Fellbaum, 1998)[2]. In the sentence above, this yields {tired, customer, service} from A and {give, up, use, product} from B. Finally, we take the Cartesian product of these two sets, which yields the causal word pairs such as (tired$_c$, give$_e$), (tired$_c$, up$_e$) and (tired$_c$, use$_e$), to obtain all possible pairs of cause and effect words.

As exemplified above, most of the previous studies including Luo et al. (2016) assume that any events can be represented by a single word, measuring associations on word pairs. However, this assumption does not hold for causality represented by multiword expressions (e.g. "give up" in the above example). To identify events represented by multiword expressions correctly, we make a list of multiword expressions from Wiktionary[3], a publicly available dictionary edited by Wiktionary community members. In this study, we focus only on multiword predicates (MWPs), predicates consisting of multiple words (e.g., "give up"). To create the list of MWPs, we extracted 33,274 verb Wiktionary entries whose titles consist of two or three words. By using this list, we acquire a causal event pair represented by multiword expressions (e.g. (tired$_c$, give up$_e$)) in addition to single-word event pairs (e.g. (tired$_c$, give$_e$) and (tired$_c$, up$_e$)).

### 2.2 Causal Strength

After extracting causal event pairs, we estimate causality between events. In this paper, we use *Causal Strength* (henceforth, CS) proposed by Luo et al. (2016). Causal Strength is similar to pointwise mutual information (PMI) but more superior in modeling causality, combining two factors: the necessary factor

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

[2] During the extraction, all words are lemmatized and lowercased.

[3] https://en.wiktionary.org/wiki/Wiktionary:Main_Page

and the sufficient factor. Formally, for a causal event $i_c$ and a effect event $j_e$, the two factors are defined by the following equations:

$$\mathrm{CS}_{\mathrm{nec}}(i_c, j_e) = \frac{p(i_c|j_e)}{p^{\alpha}(i_c)} = \frac{p(i_c, i_e)}{p^{\alpha}(i_c)p(j_e)}, \tag{1}$$

$$\mathrm{CS}_{\mathrm{suf}}(i_c, j_e) = \frac{p(j_e|i_c)}{p^{\alpha}(j_e)} = \frac{p(i_c, i_e)}{p(i_c)p^{\alpha}(j_e)}, \tag{2}$$

where $\mathrm{CS}_{\mathrm{nec}}(i_c, j_e)$ is the necessary factor, $\mathrm{CS}_{\mathrm{suf}}(i_c, j_e)$ is the sufficient factor, and $\alpha$ is a hyper-parameter. We set $\alpha = 0.66$, the same value as Luo et al. (2016); **?**. By using $\mathrm{CS}_{\mathrm{nec}}(i_c, j_e)$ and $\mathrm{CS}_{\mathrm{suf}}(i_c, j_e)$, Causal Strength is defined as,

$$\mathrm{CS}(i_c, j_e) = \mathrm{CS}_{\mathrm{nec}}(i_c, j_e)^{\lambda}\mathrm{CS}_{\mathrm{suf}}(i_c, j_e)^{1-\lambda}, \tag{3}$$

where $\lambda$ is a hyper-parameter.

# 3 Experiment

To examine the necessity of proper treatment of multiword expressions, we compare the proposed method against existing causality estimation models, and conduct an ablation study.

## 3.1 Dataset

To extract causal event pairs, we used the ClueWeb12[4], a large-scale corpus consisting of 700 million documents crawled from the Web. We evaluated the proposed method on the task of Choice of Plausible Alternatives (COPA) (Roemmele et al., 2011), which is a widely-used benchmark of commonsense-reasoning models. Each COPA problem consists of a *premise* sentence and two *alternative* sentences as follows:

> **Premise:** The grape juice fermented. *What happened as a result?*
> **Alternative 1:** The juice turned to wine.
> **Alternative 2:** The juice evaporated.

The task is to choose the most plausible alternative as either the cause or effect of the given premise (e.g. Alternative 1 in the above example). For our evaluation, we used the publicly available COPA dataset[5], which consists of 500 development and 500 test problems.

## 3.2 Settings

We evaluate the proposed model against four existing baseline models: (i) "Random", a random baseline model, (ii) "PMI" (Roemmele et al., 2011), modeling the causality between two word pairs in terms of their co-occurrences within a particular window-size on Project Gutenberg corpus[6], (iii) "PMI-EX" (Gordon et al., 2011), the improved version of PMI using millions of personal stories extracted from the Weblogs and (iv) "CS w/o MWP" (Luo et al., 2016), the state-of-the-art system of COPA that achieved an accuracy of 70.2%.

We use Stanford CoreNLP (Manning et al., 2014) for POS tagging and lemmatization. The hyper-parameter $\lambda$ in the Causal Strength method is tuned from 0.0 to 1.0 in increments of 0.1 on the development set; see Table 1 for the actual values used in this experiment.

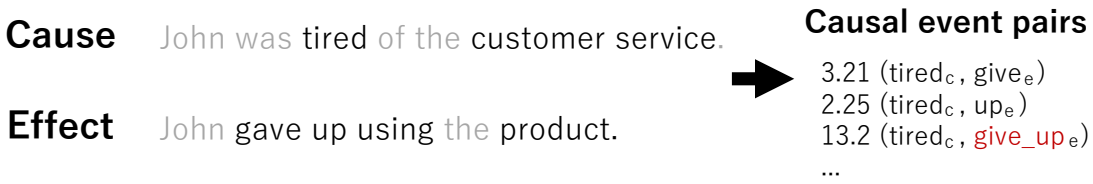


Figure 2: The outline of causality estimation between two sentences.

Table 1: Results of COPA evaluation.

| Method | Corpus | Accuracy (%) |
|---|---|---|
| Random | | 50.0 |
| PM I (Roemmele et al., 2011) | Project Gutenberg | 58.8 |
| PMI-EX (Gordon et al., 2011) | Personal stories | 65.4 |
| CS w/o MWP$_{\lambda=1.0}$ (Luo et al., 2016) | Causal Net | 70.2 |
| CS w/o MWP$_{\lambda=0.8}$ | ClueWeb12 | 69.9 |
| CS w/ MWP$_{\lambda=0.7}$ | ClueWeb12 | **71.2** |

### 3.3 Estimating Causality between Sentences

Let $S_c$ and $S_e$ be a sentence describing cause and effect, respectively. We first pre-process the sentences by lemmatization and removal of stop words[7], and then extract content words (nouns, verbs, adjectives and adverbs included in WordNet (Fellbaum, 1998))[8] from the sentences. As in the extraction of causal word pairs (see Sec. 2.1), the proposed method extracts a multiword expression as a single word if it is found in the list of multiword expressions. Let $W(S)$ be the words extracted from $S$ by this procedure. The causality score between $S_c$ and $S_e$ is then calculated as,

$$\mathrm{Score}(S_c, S_e) = \frac{1}{|W(S_c)||W(S_e)|} \sum_{w_i \in W(S_c)} \sum_{w_j \in W(S_e)} \mathrm{CS}(w_i, w_j). \quad (4)$$

For example, in Figure 2, $W(S_c) = \{\text{tired, customer, service}\}$ and $W(S_e) = \{\text{give, up, give\_up, use, product}\}$ hold; therefore, $\mathrm{Score}(S_c, S_e)$ is given by $(3.21 + 2.25 + 13.2 + ...)/(3 \cdot 5)$.

To solve a COPA problem, given a premise $P$ and two alternatives $A_1, A_2$, we identify the most plausible alternative as $A_i$ that maximizes $\mathrm{Score}(P, A_i)$ for effect questions; $\mathrm{Score}(A_i, P)$ for cause questions.

### 3.4 Results and Discussion

Table 1 shows the accuracy of the proposed model against the baselines on the 500 COPA test problems. The results indicate that the proposed method ("CS w/ MWP") outperformed the other existing models including CS, the state-of-the-art model of COPA (Luo et al., 2016) (by 1.0%). To see the effectiveness of the proper treatment of multiword-expression events, we also evaluated the proposed model without using the list of multiword expressions ("CS w/o MWP"). The results indicate that the proper treatment of multiword expression events significantly improves the accuracy of causality estimation (by 1.3%).

We manually analyzed how the proposed method improves the CS score on the COPA problems. The analysis revealed that the proper treatment of multiword expression events indeed rectifies the calculation of CS score in some COPA problems. For instance, consider the following problem:

[4] http://lemurproject.org/clueweb12/index.php
[5] http://people.ict.usc.edu/~gordon/copa.html
[6] http://www.gutenberg.org
[7] We used the list of stop words defined in Natural Language Toolkit: http://www.nltk.org/
[8] In addition, top-10 frequent words in a corpus are excluded (personal communication).

**Premise:** The father *shut off* the children's television. *What was the cause of this?*
**Alternative 1:** It was bedtime for the children.
**Alternative 2:** The children were watching cartoons.

In the premise, the multiword expression *shut off* represents an event of *to turn off*. However, each individual word, e.g. *shut* standing for *to close*, has a completely different meaning from the whole *shut off*. In this problem, the proposed model successfully estimates $\mathrm{CS}(\text{bedtime}_c, \text{shut off}_e)=13.7$ as opposed to $\mathrm{CS}(\text{bedtime}_c, \text{shut}_e)=1.88$. This indicates that the proposed model captures the causality represented by a multiword expression properly, i.e. the causality between "to shut off (electricity)" and "bedtime". Other such examples include (i) $\mathrm{CS}(\text{wait}_c, \text{take a seat}_e)=12.9$ (c.f. $\mathrm{CS}(\text{wait}_c, \text{take}_e)=3.28$, $\mathrm{CS}(\text{wait}_c, \text{seat}_e)=2.38$) and (ii) $\mathrm{CS}(\text{think}_c, \text{come up with}_e)=5.23$ (c.f. $\mathrm{CS}(\text{think}_c, \text{come}_e)=4.02$).

To evaluate how well the proposed system identifies multiword expressions, we randomly extracted 50 multiword expressions identified by the system from the development set. The analysis of these instances reveals that 18.0% (9/50) of them were incorrectly recognized, where one typical error is exemplified by *jog on* in *"I jogged on the treadmill."*: *jogged on* standing for *jogging* here, is incorrectly identified as the idiom *jog on* standing for "*to continue with one's pursuit*". To understand the potential effect of multiword expressions, we manually crafted the list of multiword expressions that are needed for solving the COPA test questions. The maximum accuracy[9] of the oracle system using the manually-crafted list was 71.0%, which suggests that the proposed method achieved almost equal score to the oracle score.

To gain further insights, we analyzed the remaining 82.0% (41/50) of correctly recognized multiword expressions. It reveals that proper causality estimation often requires the system to expand an event unit to another word as well as to recognize a multiword expression; for instance, when the causality between *"The stain came out of the shirt."* and *"I bleached the shirt"* is estimated, *stain come out*, rather than *come out*, is more appropriate as an event unit. Extending an event unit beyond a multiword expression would impose a severe data sparseness problem, which is to be addressed in our future work.

## 4 Related Work

Previous studies proposed a wide variety of approaches to causality estimation (Do et al., 2011; Kozareva, 2012; Riaz and Girju, 2014). Do et al. (2011) employed statistical measures such as PMI and inverse document frequency (IDF) from a corpus to model causality between events. Kozareva (2012) applied a bootstrap algorithm to acquire causal event pairs. Riaz and Girju (2014) extracted training data from FrameNet (Baker et al., 1998) to learn a classifier for a causal relation. However, these studies assume that an event is representable by a single word. Chambers and Jurafsky (2008)'s *Narrative Schema* uses a predicate-argument structure as an event unit, but a predicate is restricted to a single word.

Roemmele et al. (2011) introduced a baseline model of COPA that uses PMI between words on English documents in Project Gutenberg. Gordon et al. (2011) improved the baseline model introduced by Roemmele et al. (2011) by using personal stories extracted from Weblogs instead of Project Gutenberg. Luo et al. (2016) refined PMI to capture causality between events more accurately. In this paper, we employed the statistical measure proposed by Luo et al. (2016) because they achieved the state-of-the-art performance on the COPA dataset.

## 5 Conclusion

In this paper, we created the list of multiword expressions from Wiktionary, and proposed a method to capture causality of multiword expressions by extending the existing causality estimation model. We demonstrated the effectiveness of using a multiword expression list, reporting a new state-of-the-art

[9]We report a maximum accuracy because the manually created MWP list is specific to the test set and hence prevents us from tuning the hyper-parameter in the development set.

performance on COPA. Our future work includes using a combination of a predicate and an object as a unit of a causal event.

## References


Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pp. 86–90. Association for Computational Linguistics.

Chambers, N. and D. Jurafsky (2008). Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pp. 789–797. Association for Computational Linguistics.

Do, Q., Y. S. Chan, and D. Roth (2011). Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 294–303. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database. Bradford Books.* MIT Press.

Gordon, A. S., C. A. Bejan, and K. Sagae (2011). Commonsense causal reasoning using millions of personal stories. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 1180–1185. AAAI Press.

Granroth-Wilding, M. and S. Clark (2016). What happens next? event prediction using a compositional neural network model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 2727–2733. AAAI Press.

Kozareva, Z. (2012). Cause-effect relation learning. In *Workshop Proceedings of TextGraphs-7: Graph-based Methods for Natural Language Processing*, pp. 39–43. Association for Computational Linguistics.

Luo, Z., Y. Sha, K. Q. Zhu, S. won Hwang, and Z. Wang (2016). Commonsense causal reasoning between short texts. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*, pp. 421–430. AAAI press.

Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60. Association for Computational Linguistics.

Riaz, M. and R. Girju (2014). Recognizing causality in verb-noun pairs via noun and verb semantics. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pp. 48–57. Association for Computational Linguistics.

Roemmele, M., C. A. Bejan, and A. S. Gordon (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.