SeRI: A Dataset for Sub-event Relation Inference from an Encyclopedia

Tao Ge[1,2(✉)], Lei Cui[2], Baobao Chang[1], Zhifang Sui[1], Furu Wei[2], and Ming Zhou[2]

[1] School of EECS, Peking University, Beijing, China
{chbb,szf}@pku.edu.cn
[2] Microsoft Research Asia, Beijing, China
{tage,lecu,fuwei,mingzhou}@microsoft.com

Abstract. Mining sub-event relations of major events is an important research problem, which is useful for building event taxonomy, event knowledge base construction, and natural language understanding. To advance the study of this problem, this paper presents a novel dataset called SeRI (**S**ub-**e**vent **R**elation **I**nference). SeRI includes 3,917 event articles from English Wikipedia and the annotations of their sub-events. It can be used for training or evaluating a model that mines sub-event relation from encyclopedia-style texts. Based on this dataset, we formally define the task of sub-event relation inference from an encyclopedia, propose an experimental setting and evaluation metrics and evaluate some baseline approaches' performance on this dataset.

1 Introduction

Event relation inference is an important research topic because event relations are not only indispensable for constructing event knowledge bases but also helpful for natural language understanding and knowledge inference. As one of the most important event relations, sub-event relation knowledge has been proved to be useful in many applications [2,6,14,19–23]. For example, the sub-event knowledge *sub*(*Battle of the Atlantic*, *World War II*) can greatly help textual entailment (e.g. (1)) and knowledge inference (e.g., (2)) task.

(1) He died in the *Battle of the Atlantic*. → He died in *World War II*.
(2) time(*World War II*, 1939-1945) → time(*Battle of the Atlantic*, years during 1939-1945)

Despite the significance of general sub-event knowledge, there is little work on mining the sub-event knowledge from the web, which may be due largely to a lack of datasets. To advance the research of this problem, this paper presents a dataset called SeRI (**S**ub-**e**vent **R**elation **I**nference). SeRI contains 3,917 event articles[1] from English Wikipedia, and the annotations for their sub-events, which

[1] Event articles refer to the articles that describe a major event in Wikipedia, like Fig. 1.

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11109, pp. 268–277, 2018.
https://doi.org/10.1007/978-3-319-99501-4_23


Battle of the Atlantic

From Wikipedia, the free encyclopedia

For the Atlantic naval campaign of World War I, see Atlantic U-boat campaign of World War I.

The **Battle of the Atlantic** was the longest continuous military campaign[6][7] in World War II, running from 1939 to the defeat of Germany in 1945. At its core was the Allied naval blockade of Germany, announced the day after the declaration of war, and Germany's subsequent counter-blockade. It was at its height from mid-1940 through to the end of 1943. The Battle of the Atlantic pitted U-boats and other warships of the *Kriegsmarine* (German navy) and aircraft of the *Luftwaffe* (German Air Force) against the Royal Canadian Navy, Royal Navy, the United States Navy, and Allied merchant shipping. The convoys, coming mainly from North America and predominantly going to the United Kingdom and the Soviet Union, were protected for the most part by the British and Canadian navies and air forces. These forces were aided by ships and aircraft of the United States from September 13, 1941.[8] The Germans were joined by submarines of the Italian Royal Navy (*Regia Marina*) after their Axis ally Italy entered the war on June 10, 1940.

As an island nation, the United Kingdom was highly dependent on imported goods. Britain required more than a million tons of imported material per week in order to be able to survive and fight. In essence, the Battle of the Atlantic was a tonnage war: the Allied struggle to supply Britain and the Axis attempt to stem the flow of merchant shipping

Officers on the bridge of an escorting British destroyer stand watch for enemy submarines, October 1941

Fig. 1. The event article of *Battle of the Atlantic* event in Wikipedia. The red rectangle provides with the annotation information and the underlined phrases are mentions of other events, linking to the corresponding event articles. (Color figure online)

can be used for training and evaluating a model for mining sub-event relations from an encyclopedia. Based on this dataset, we formally define the task of sub-event relation inference from an encyclopedia, propose an experimental setting and evaluation metrics and evaluate some baseline approaches' performance on this dataset.

The main contributions of this paper are:

- We release a dataset[2] for studying sub-event relation inference from an encyclopedia.
- We formally define the task of sub-event relation inference from an encyclopedia and propose an experimental setting and evaluation metrics.
- We evaluate some baseline approaches' performance on this dataset.

2 SeRI Dataset

SeRI is a dataset we propose for studying sub-event relation inference. For constructing SeRI, we first need to find event articles on Wikipedia. In this paper, we use those 21,275 event articles identified by EventWiki [8] as our research targets. We keep the events with "*partof*" relation annotation in their infoboxes (shown in Fig. 1) and use the annotation as their sub-event relation annotations.

It should be noted that the *partof* annotations in infoboxes are mainly for direct sub-event relations. However, sub-event relations are transitional and there are many indirect sub-event relations that are not annotated in EventWiki. Therefore, we add the annotations of indirect sub-event relations through transitivity rules. The step is necessary to annotate some long-distance sub-event relations that are often not explicitly expressed in Wikipedia. For example, given the annotation from EventWiki that $sub(Battle\ of\ Netherlands, Battle\ of\ France)$,

[2] Please contact the first author to request the access to the dataset.

sub(*Battle of France*, *Western Front*) and *sub*(*Western Front*, *World War II*), according to the transitivity, long-distance sub-event relations such as *sub*(*Battle of Netherlands*, *World War II*) will be added to our annotation.

The resulting SeRI dataset has 3,917 Wikipedia articles (i.e., 3,917 events) and their sub-event annotations.

3 Task Overview

If an event e_i is a part of event e_j, we say e_i is e_j's *sub-event* and e_j is e_i's *sup-event*. There are many event articles (e.g., *World War II*) with reference links in an encyclopedia, as Fig. 2 depicts. The goal of this task is to harvest as much sub-event knowledge as possible from an encyclopedia. Strictly speaking, for an event e_i, we should identify all the other events in an encyclopedia to see if they are the sub-events or sup-events of e_i. In this sense, we must infer the sub-event relation over all the possible event pairs, which results in expensive cost of computation. Instead of the brutal-force solution, we propose a heuristic assumption that if an event e_i is a sub-event or sup-event of another event e_j, then e_i should be mentioned by the article of e_j or e_j should be mentioned by the article of e_i. This assumption is reasonable because if the sub-event relation holds between e_i and e_j then e_i and e_j should be strongly related and either of them should be mentioned by the other. Based on the assumption, we identify sub-event relations over the event pairs in which one event is mentioned (i.e., linked) by the article of the other event. For simplicity, we call such event pairs *candidate event pairs*.

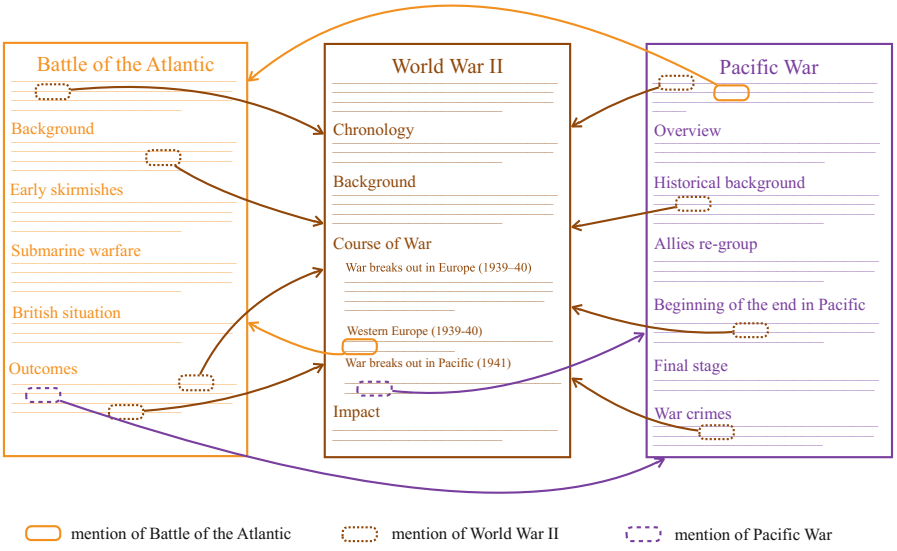

Fig. 2. An illustration of the basic structure of an encyclopedia in which articles are interconnected by the reference links.

We give the formal definition of the sub-event inference task: given a set of candidate event pairs $\mathcal{S} = \{\langle e_i, e_j \rangle\}$ in an encyclopedia, the goal is to identify all sub-event pairs $\langle e_i^*, e_j^* \rangle$ from $\mathcal{S}$ so that e_i^* is a sub-event or sup-event of e_j^*. In other words, for each pair $\langle e_i, e_j \rangle \in \mathcal{S}$, a model should be able to correctly predict its label $r \in \{none, sub, sup\}$ that indicates the sub-event relation between e_i and e_j. In our SeRI dataset, there are totally 7,373 candidate event pairs with relation label annotation $r \in \{none, sub, sup\}$.

4 Baseline Models

For the task of harvesting sub-event knowledge from an encyclopedia, how an event article links to (i.e., refers to) another is an important clue for inferring the sub-event relation between them. For example, it can be easily inferred that the *Battle of the Atlantic* is a sub-event of the *World War II* through analyzing the sentence corresponding to a reference link from *Battle of the Atlantic* to *World War II*, as Fig. 3 shows.

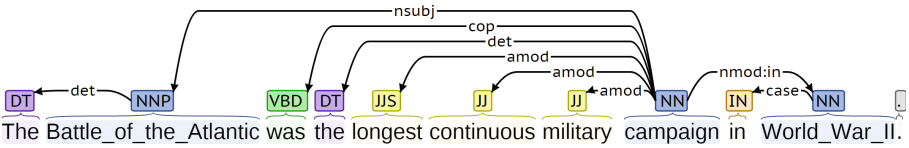

Fig. 3. The sentence that corresponds to the first link from *Battle of the Atlantic* to *World War II* in Fig. 2.

As Fig. 2 depicts, there may be multiple reference links from both directions between an event pair. Given the value of the reference links for sub-event relation inference, we propose a link-based classification model that predicts an event pair's sub-event relation through the reference links between them.

For an event pair $\langle e_i, e_j \rangle$ (i.e., an instance), we first find all the reference links between them. Then, we consider each link as a sub-instance and describe it using various features.

4.1 Features

As discussed previously, each link $e_i \xrightarrow{l} e_j$, which denotes the l^{th} link from the article of e_i to e_j, corresponds to a sentence that mentions e_j in the article of e_i. We extract the following features to describe the link:

- Context words of e_j: we extract the context words (i.e., last 5 words and next 5 words) surrounding the mention of e_j in the article of e_i as features. We denote this feature as $\boldsymbol{f_c}$.

- N-grams that start or end with the mention of e_j: we extract 2-, 3- and 4-grams that start or end with e_j as features. For generalization, e_j is replaced with *"TARGET"* in feature strings. We denote this feature as $\boldsymbol{f_g}$.
- Dependency parse: We extract all dependency arcs that contain e_j as features. In addition, we also use the part-of-speech tag of the word in a dependency arc as features. We denote this feature as $\boldsymbol{f_p}$.
- Section name: In addition to the sentence-level features that are commonly used by traditional relation extraction models, we propose to use the section name feature. For example, for the reference link from *World War II* to *Battle of the Atlantic* in Fig. 2, one can observe that the mention of *Battle of the Atlantic* is in the section called *Course of war* in the article of *World War II*. In this case, we extract "course of war" as the section name features. We denote this feature as $\boldsymbol{f_s}$.

After extracting the features of a reference link between an event pair $\langle e_i, e_j \rangle$, we can represent the features as a tuple to represent the link:

$$\boldsymbol{f} = (\boldsymbol{f_c}, \boldsymbol{f_g}, \boldsymbol{f_p}, \boldsymbol{f_s}, d) \tag{1}$$

where the last element $d \in \{\rightarrow, \leftarrow\}$ indicates the direction of the link.

We represent an event pair instance $\langle e_i, e_j \rangle$ by concatenating the features of all the links between the event articles in the pair to train the link-based classifier and infer the sub-event relation in the following way:

$$y^* = \arg\max_y P(y|\textsc{ConCat}_{l=1}^{L}(\boldsymbol{f}^{(l)})) \tag{2}$$

where l is the l-th link between e_i and e_j, $\boldsymbol{f}^{(l)}$ is the features of l-th link between e_i and e_j, and ConCat is the operation that concatenates the features of all the links between e_i and e_j.

4.2 Bi-directional Training Instance Generation

Although we do not consider the order of the items in a pair (i.e., we consider $\langle e_i, e_j \rangle$ and $\langle e_j, e_i \rangle$ are the identical pair), it is necessary to specify the order of an pair when generating training and test instances for a classifier because the label of the instance depends on the order[3]. For example, if the label of a pair instance (e_i,e_j) is *sub*, then the label of a pair instance (e_j,e_i) should be *sup*.

Therefore, for an event pair $\langle e_i, e_j \rangle$, we specify the order of the items to generate training and test instances. Specifically, for training instance generation, we generate bi-directional pair instances (i.e., 2 pair instances (e_i,e_j) and (e_j,e_i)). There are two reasons for the bi-directional training instance generation: First, the bi-directional training instances will provide different views for the model to learn to predict *sub* and *sup*, which is important because the test

[3] To distinguish from a *pair* which is denoted as $\langle e_i, e_j \rangle$ for which we do not consider the order of items, we say a *pair instance* denoted as (e_i,e_j)) when we consider the order of items in a pair.

instance could be in any order. Bi-directional training instances allow the model to handle the test instances with any order; Second, the bi-directional training instance generation can alleviate the imbalance of labels because the number of the training instances with *sub* and *sup* will be identical.

5 Experiments

5.1 Experimental Setting

As we mentioned before, SeRI has 3,917 Wikipedia articles (i.e., 3,917 events) and 7,373 candidate event pairs in total. Each candidate event pair has a label $r \in \{none, sup, sub\}$ indicating the sub-event relation. In our experiments, we use 80% of candidate event pairs as the training set and test on the remaining 20% of the pairs. Note that in the split, we split event pairs instead of event articles because it is very common that an event article's sub-event relation is not complete and it might have sub-event relation with new added event articles. Therefore, splitting event pairs is a more practical setting than splitting event articles. The label distribution is shown in Table 1.

Table 1. Label distribution in training and test set

	Train		Test	
Label	none	sub+sup	none	sub+sup
Number of instances	3,824	2,112	893	544
Ratio	64.4%	35.6%	62.1%	37.9%

In our experimental setting, we do not use any structured information (i.e., infoboxes) since we want a model to be general so that it can applied to any encyclopedia articles regardless of the existence of the infoboxes for discovering the sub-event knowledge that has not been explicitly expressed in an encyclopedia.

As traditional relation extraction tasks, we use Precision, Recall and F-score for evaluation:

$$Precision = \frac{\left|\{\langle e_i, e_j\rangle \mid \hat{r_{i,j}} = r^*_{i,j}\}\right|}{|\{\langle e_i, e_j\rangle | \hat{r_{i,j}} \in \{sup, sub\}\}|}$$

$$Recall = \frac{\left|\{\langle e_i, e_j\rangle \mid \hat{r_{i,j}} = r^*_{i,j}\}\right|}{\left|\{\langle e_i, e_j\rangle | r^*_{i,j} \in \{sup, sub\}\}\right|}$$

$$F\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where $\hat{r_{i,j}}$ and $r^*_{i,j}$ are the prediction and gold standard relation of $\langle e_i, e_j\rangle$ respectively.

5.2 Experimental Results

We conduct experiments test the baseline approaches mainly for answering the following three questions:

- Whether are the features used in the link-based classifier effective?
- Whether is generating bi-directional training instances necessary?
- How well a baseline model can perform?

We use Stanford CoreNLP [16] to do POS tagging and dependency parsing and maximum entropy classifier as the link-based classification model.

Table 2. Performance of baseline models. Uni- and Bi-directional indicate if the training instances are generated using bi-directional generation strategy.

Model		Uni-directional			Bi-directional		
		P	R	F	P	R	F
Link-based classifier	(1): Context	51.8	65.8	58.0	68.4	61.4	64.7
	(2): (1)+N-grams	52.6	72.4	61.0	70.6	63.1	66.6
	(3): (2)+Dependency parse	52.7	71.1	60.5	69.7	65.3	67.4
	(4): (3)+section name	63.0	76.8	69.3	77.8	73.7	75.7

Table 2 shows the performance of various approaches. For the link-based classifier with bi-directional training instance, the proposed features are all effective in inferring the sub-event relations, especially the section name features. The reason of the significant improvement (8.3% F-score gain) brought by the section name features is that they provide totally different views from the sentence-level features such as n-grams and can nicely address the cases where n-gram and parse features cannot help.

By comparing uni-directional and bi-directional training instance generation, we can easily observe the superiority of the bi-directional training generation strategy. For the models using the same feature set, the model adopting bi-directional training generation strategy outperforms the uni-directional counterpart by approximately 6.0% F-score gain. As we discussed before, the improvement is mainly owing to that bi-directional training instances provide more information for the model and alleviate the label imbalance problem.

According to Table 2, it is observed that the best baseline model can achieve a good performance of 75.7% F-score. Given this is a preliminary study on this task, we expect to see more results of various approaches by future work that studies on this dataset.

At last, we show the example of sub-event knowledge discovered by our model on the test set in Fig. 4. As observed, the results can be used for building an event hierarchy tree, which will be useful for event knowledge base construction and management as well as many other applications like knowledge inference and question answering.

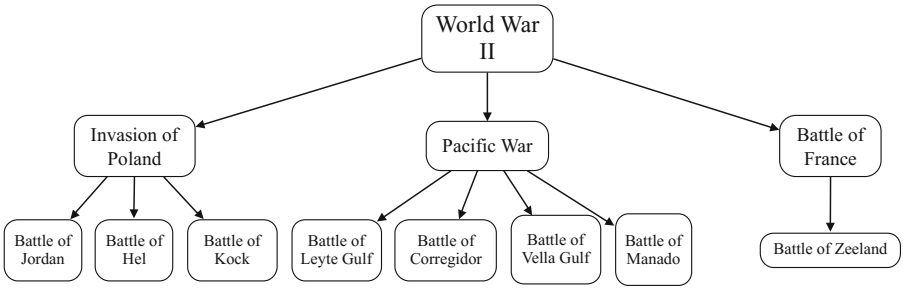

Fig. 4. An event tree derived from the predictions of our model on the test data.

6 Related Work

There is some previous work on extracting temporal and causal relation between events [1,3–5,7,9–11,17,18,24]. As for the studies regarding sub-events, most of them study either fine-grained events in a sentence or a document [2] or hierarchies of events (topics) in a text stream [12,13,23]. Even though they are very useful for understanding a document or the relation of events in a text stream, they do not focus on harvesting knowledge and the relations identified by their models are less likely to be used as general knowledge. In contrast, this paper mainly studies the acquisition of sub-event knowledge which can be used as world knowledge for a variety of applications.

Another research branch related to our task is hypernymy detection (e.g., [15]), whose goal is to discover the hypernymy between general words. In contrast to hypernymy detection, our targets are events which contain much richer information and sub-event relations that are more complicated because it is common that there is no co-occurrence of the mentions of two events in a sentence.

7 Conclusion and Future Work

In this paper, we study a novel dataset SeRI and define a new task – harvesting sub-event knowledge from an encyclopedia. We propose a link-based classification baseline model with features of encyclopedia-style documents, which achieves decent results on our dataset.

Following the preliminary study on event relations, we plan to study advanced approaches for this task and do more empirical studies on this dataset, and make an effort to keep enlarging the dataset. We expect more research could be conducted regarding this task, which would contribute to event knowledge discovery and event knowledge base construction.

Acknowledgments. The research work is supported by the National Science Foundation of China under Grant No. 61772040. The contact author is Zhifang Sui.

References

1. Abe, S., Inui, K., Matsumoto, Y.: Two-phased event relation acquisition: coupling the relation-oriented and argument-oriented approaches. In: Proceedings of the 22nd International Conference on Computational Linguistics, vol. 1, pp. 1–8. Association for Computational Linguistics (2008)
2. Araki, J., Liu, Z., Hovy, E.H., Mitamura, T.: Detecting subevent structure for event coreference resolution. In: LREC, pp. 4553–4558 (2014)
3. Chambers, N., Cassidy, T., McDowell, B., Bethard, S.: Dense event ordering with a multi-pass architecture. Trans. Assoc. Comput. Linguist. **2**, 273–284 (2014)
4. Chambers, N., Jurafsky, D.: Unsupervised learning of narrative schemas and their participants. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 2, pp. 602–610. Association for Computational Linguistics (2009)
5. Chambers, N., Jurafsky, D.: Unsupervised learning of narrative event chains. In: ACL, vol. 94305, pp. 789–797. Citeseer (2008)
6. Daniel, N., Radev, D., Allison, T.: Sub-event based multi-document summarization. In: Proceedings of the HLT-NAACL 03 on Text Summarization Workshop, vol. 5, pp. 9–16. Association for Computational Linguistics (2003)
7. Do, Q.X., Chan, Y.S., Roth, D.: Minimally supervised event causality identification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 294–303. Association for Computational Linguistics (2011)
8. Ge, T., Cui, L., Chang, B., Sui, Z., Wei, F., Zhou, M.: Eventwiki: a knowledge base of major events. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
9. Ge, T., Cui, L., Ji, H., Chang, B., Sui, Z.: Discovering concept-level event associations from a text stream. In: Lin, C.-Y., Xue, N., Zhao, D., Huang, X., Feng, Y. (eds.) ICCPOL/NLPCC -2016. LNCS (LNAI), vol. 10102, pp. 413–424. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50496-4_34
10. Hashimoto, C., Torisawa, K., Kloetzer, J., Oh, J.H.: Generating event causality hypotheses through semantic relations. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
11. Hashimoto, C., et al.: 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, Baltimore, USA. Association for Computational Linguistics, June 2014
12. Hu, L., Li, J., Zhang, J., Shao, C.: o-HETM: an online hierarchical entity topic model for news streams. In: Cao, T., Lim, E.-P., Zhou, Z.-H., Ho, T.-B., Cheung, D., Motoda, H. (eds.) PAKDD 2015. LNCS (LNAI), vol. 9077, pp. 696–707. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18038-0_54
13. Hu, L., et al.: Learning topic hierarchies for wikipedia categories. In: ACL (2015)
14. Im, S., Pustejovsky, J.: Annotating lexically entailed subevents for textual inference tasks. In: Twenty-Third International Flairs Conference (2010)
15. Levy, O., Remus, S., Biemann, C., Dagan, I., Ramat-Gan, I.: Do supervised distributional methods really learn lexical inference relations? In: HLT-NAACL, pp. 970–976 (2015)
16. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: ACL (System Demonstrations) (2014)

17. Mirza, P.: Extracting temporal and causal relations between events. In: ACL (Student Research Workshop), pp. 10–17 (2014)
18. Mirza, P., Tonelli, S.: An analysis of causality between events and its relation to temporal information. In: COLING, pp. 2097–2106 (2014)
19. Mulkar-Mehta, R., Welty, C., Hoobs, J.R., Hovy, E.: Using granularity concepts for discovering causal relations. In: Proceedings of the FLAIRS Conference (2011)
20. Pohl, D., Bouchachia, A., Hellwagner, H.: Automatic sub-event detection in emergency management using social media. In: Proceedings of the 21st International Conference Companion on World Wide Web, pp. 683–686. ACM (2012)
21. Shen, C., Liu, F., Weng, F., Li, T.: A participant-based approach for event summarization using twitter streams. In: HLT-NAACL, pp. 1152–1162 (2013)
22. Unankard, S., Li, X., Sharaf, M., Zhong, J., Li, X.: Predicting elections from social networks based on sub-event detection and sentiment analysis. In: Benatallah, B., Bestavros, A., Manolopoulos, Y., Vakali, A., Zhang, Y. (eds.) WISE 2014. LNCS, vol. 8787, pp. 1–16. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11746-1_1
23. Xing, C., Wang, Y., Liu, J., Huang, Y., Ma, W.Y.: Hashtag-based sub-event discovery using mutually generative lda in twitter. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
24. Yoshikawa, K., Riedel, S., Asahara, M., Matsumoto, Y.: Jointly identifying temporal relations with markov logic. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 1, pp. 405–413. Association for Computational Linguistics (2009)