Using Semi-Supervised Learning and Wikipedia to Train an Event Argument Extraction System

Patrik Zajec and Dunja Mladenić
E-mail: patrik.zajec@ijs.si, dunja.mladenic@ijs.si
Jožef Stefan Institute and Jožef Stefan International Postgraduate School

Student paper

Keywords: event extraction, event argument extraction, semi-supervised learning, probabilistic soft logic

Received: June 2, 2021

The paper presents a methodology for training an event argument extraction system in a semi-supervised setting. We use Wikipedia and Wikidata to automatically obtain a small noisily labeled dataset and a large unlabeled dataset. The dataset consists of event clusters containing Wikipedia pages in multiple languages. The unlabeled data is iteratively labeled using semi-supervised learning combined with probabilistic soft logic to infer the pseudo-label of each example from the predictions of multiple base learners. The proposed methodology is applied to Wikipedia pages about earthquakes and terrorist attacks in a cross-lingual setting. Our experiments show improvement of the results when using the proposed methodology. The system achieves F1-score of 0.79 *when only the automatically labeled dataset is used, and F1-score of* 0.84 *when trained according to the methodology with semi-supervised learning combined with probabilistic soft logic.*

Povzetek: V prispevku predstavimo metodologijo za polnadzorovano učenje sistema, katerega naloga je ekstrakcija ključnih atributov dogodkov. Kot vir podatkov uporabimo prosto enciklopedijo Wikipedija in bazo znanja Wikidata. Iz obeh virov na avtomatski način pridobimo manjši del označenih podatkov ter večji del neoznačenih podatkov, sestavljenih iz gruč dokumentov, ki poročajo o posameznih dogodkih v več jezikih. Neoznačene podatki iterativno označimo s polnadzorovanim učenjem v kombinaciji z verjetnostno mehko logiko, ki napovedi za vsak primer iz večih predikcijskih modelov združi v eno samo psevdo-labelo. Predlagano metodologijo uporabimo na dogodkih iz Wikipedije na temo potresov in terorističnih napadov. Eksperimenti pokažejo izboljšanje rezultatov sistema, ki doseže F1 $0,79$*, ko je naučen samo na avtomatsko označenih podatkih, ter* $0,84$*, ko je naučen, z uporabo predlagane metodologije, s polnadzorovanim učenjem v kombinaciji z verjetnostno mehko logiko.*

1 Introduction

The event extraction task is usually divided into two subtasks, namely event type detection and event argument extraction. Event type detection aims to determine the type or topic of the event, while event argument extraction aims to extract arguments for some predefined argument roles associated with the topic of the event. In this paper, we focus on event argument extraction and assume that the topic of the event is known in advance. For example, consider an excerpt from a news article reporting on an earthquake:

*"On **April 22, 2019**, a **6.1** magnitude earthquake struck the island of Luzon in the Philippines, leaving at least **18** dead, 3 missing and injuring at least **256** others."* [1]

we are interested in extracting the arguments for the following roles: the magnitude of the earthquake, the time and date when the earthquake occurred, the number of victims, and the number of injured. The goal is to extract an argument for each role, which in our case is mentioned in the text. The output is a set of *(role, argument)* pairs for the given text, for example *(magnitude, **6.1**)*.

We formulate argument extraction as a supervised classification task, where the classifier is used to assign arguments into predefined roles. We assume that each document reports on a single event, where the topic of the event is either earthquake or terrorist attack and is known in advance. Since there is no dataset that contains the annotations for the roles we are interested in, one must be constructed to train the classification model. Manually constructing a labeled dataset is a resource-intensive process, so we instead develop a methodology to automate the labeling.

We start by automatically constructing a small, noisily labeled set of examples for each role by matching structured knowledge from Wikidata with text from Wikipedia pages. Then we use semi-supervised learning [1] with multiple base learners to increase the size of the labeled set. In

[1] Source: en.wikipedia.org/wiki/2019_Luzon_earthquake

each iteration, the most confident predictions for the examples from the unlabeled set are used to increase the training set by assigning pseudo-labels. We introduce an additional component that combines the confidence of multiple base learners for each example.

The contribution of this paper is a novel methodology for automatic construction of a multi-lingual labeled dataset and training of a classification system for event argument extraction in multiple languages.

2 Related work

Wikipedia is a commonly used source for dataset construction [2, 3, 4]. Most datasets are constructed, at least in part, under the supervision of a human annotator, since for most tasks the labels cannot be reliably generated automatically [5, 6]. The initial annotations are usually obtained using distant supervision [7], which is the automatic approach of matching knowledge from the knowledge base with free text, and later refined by the human annotator, while we use semi-supervised learning instead. As in our case, the common choice is to use Wikidata [8] as the knowledge source and Wikipedia [2] as the text source.

Wikidata often does not reflect all of the knowledge available in Wikipedia pages, and a considerable amount of work has been done on the automatic population of both Wikipedia infoboxes and Wikidata [9, 10, 11, 12]. Such approaches develop the models capable of extracting the structured knowledge from free text. However, since the task is limited to Wikipedia, they mostly rely on the structure of Wikipedia pages to achieve good performance. This is especially true for the language-independent models [11] and makes them not directly applicable to other domains, such as newswire data.

Event extraction, which includes both event detection and argument extraction, is an active research topic that is formulated at several levels (document and sentence level) and approached using different techniques [13, 14, 15]. The approaches are most commonly evaluated on the ACE 2005 corpus [16] and the MUC-4 corpus [17]. Recently, RAMS [18] and WikiEvents [13] corpora were introduced, both being manually annotated and containing only English documents. Further, there is no information about the documents that mention the same event which could be used to form the event clusters.

The use of event clusters as a source of redundancy has been explored previously [19, 20]. Approaches on cross-lingual argument classification [21] are mostly focused on transfer learning, where the model is first trained on the source language and later applied on the target language. None of the approaches build the event clusters using the documents from multiple languages and incorporates cross-linguality directly into the semi-supervised training.

[2] https://www.wikipedia.org/

The currently best-performing approaches for event argument extraction are based on machine reading comprehension [22] or conditional generation [13] and therefore do not train a specific classifier for each argument. We find them less suitable for the semi-supervised learning part of our approach as they require a kind of event templates that are usually manually constructed. We envision that such models can instead be trained or fine-tuned on the automatically labeled dataset that our methodology generates.

Pseudo-labeling can be considered one of the most intuitive and simple forms of semi-supervised learning, while still achieving competitive performance if approached correctly [23]. We have chosen a simple and extensible technique to combine the predictions of multiple base learners into a single pseudo-label based on [24].

3 Methodology

3.1 Problem definition

We are given a collection of events $\mathcal{E}$ and a collection of topics $\mathcal{T}$, where the topic of each event $e \in \mathcal{E}$ is exactly one topic $t \in \mathcal{T}$. An event e has occurred at a particular time and its event cluster $\mathcal{D}_e$ consists of documents in multiple languages, with a single document per language, reporting on it.

For each topic $t \in \mathcal{T}$ there is a set of argument roles $\mathcal{R}_t$ that are known in advance. For example, the members of $\mathcal{R}_{earthquake}$ are *magnitude*, *number of injured*, *number of deaths*. For each document d from the event cluster $\mathcal{D}_e$, there is a set of arguments that were already extracted from the text. The task is to assign at most one argument role from $\mathcal{R}_t$ to each extracted argument.

Note that there are multiple documents in the event cluster, each with its own set of arguments. For event e, a single argument role might have zero, one, or multiple different arguments assigned. It is possible, for example, that the earthquake caused no casualties, so there is no argument assigned to the role of *magnitude*. The documents from the same event cluster might have different arguments for the same role, since, for example, the reported number of injured could come from different sources. In addition, there may also be multiple arguments for the same role in the single document, since the assumption that each document specifically reports only on a single event is not always true in practice. Such example is when one magnitude refers to an actual earthquake that the event is about, while the other magnitude refers to a stronger earthquake that hit the same region years ago.

Choosing a single argument for each argument role is challenging and sometimes even impossible. The task is instead to assign all feasible arguments from the event cluster to a particular role, even if not all are directly related to the event.

3.2 Approach

There is no labeled dataset that would follow the required structure and contain the labels for argument roles. We develop a methodology to automatically build such dataset and train a classification model which is used to pseudo-label the dataset and once trained can be used to perform classification on new data.

First, we select the Wikidata entities that are instances of topics from $\mathcal{T}$ and class *occurence*. Each Wikidata entity links to Wikipedia pages in multiple languages that form an event cluster. We align the argument roles $\mathcal{R}_t$ with Wikidata properties and try to automatically match the value for each property with the text from the pages. The matching is performed automatically either by using anchor links or literal text matching for numerical values. This gives us the arguments from text assigned to particular argument roles.

As most of the Wikidata entities lack some property values and automatic matching is frequently ambiguous, we obtain a small, nosily labeled set which we refer to as a *seed set*. We further use the named entities from the pages as arguments and assign the ones that are not a part of the *seed set* to an *unlabeled set*. Each such argument is considered as an unlabeled example.

To obtain label more arguments we use semi-supervised learning with multiple base learners and pseudo-labeling. Each of the base learners is trained on the set of labeled arguments from the topic (or multiple topics) and the language assigned to it. The prediction probabilities for each of the unlabeled arguments are determined by combining the probabilities of all base learners. This is done either by averaging or by feeding the probabilities as approximations of the true labels into the component, which attempts to derive the true value for each argument and the error rates for each learner [24]. The arguments with probabilities above or below the specified thresholds are given a pseudo-label and added to the training set.

The entire workflow is repeated in each iteration until no new arguments are selected for pseudo-labeling. The result is an automatically labeled dataset that includes the given topics and labels for selected argument roles and a classification system that assigns the argument role to each argument.

3.2.1 Representing the arguments

Following the related work [25], we introduce two new special tokens, <*e*> and </*e*>. The context of each argument is converted to a sequence of tokens with the additional tokens that mark the beginning and end of the argument span in the context. For example, argument 6.1 from the sentence *"A 6.1 magnitude earthquake."* is represented with its context as *"A <e>6.1</e>magnitude earthquake."*. Such representation is feed through the pretrained version of the XLM Roberta model [26]. We use the implementation from the Transformers library [27] and use the last hidden state of <*e*> token as a representation of the argument. The XLM Roberta model remains fixed during the learning as we have observed that the representation from pretrained model is expressive enough for our purposes and it significantly speeds up the iterations.

3.2.2 Using multiple topics

Instead of grouping all arguments assigned to the same roles across topics we keep the information about the topic of the argument's event. Firstly because the way of expressing an argument role might be slightly different in different topics, secondly as shown by the experiments, such separation enables us additional supervision between the topics, and finally as we can use the all arguments from one topic as negative examples for some particular role in the other topic.

For two topics t and t' there is potentially a set of common roles and a set of distinct roles, appearing in only one of the topics. For role r, which appears in both topics, the base learner trained on t' can be used to make predictions for examples from t. By combining predictions from learners trained on t and t', we could get better estimates of the true labels of the examples. For the role r', which is specific to the topic t, all examples from the topic t' can be used as negative examples. Selecting reliable negative examples from the same topic is not easy, as we may inadvertently mislabel some of the positive examples.

3.2.3 Using multiple languages

In a sense, articles from different languages provide different views on the same event. The important arguments should appear in all the articles, as they are highly relevant to the event. The arguments for the roles such as location and time should be consistent across all articles, whereas this does not necessarily apply to other roles such as the number of injured or the number of casualties. Matching such arguments across the articles is therefore not a trivial task, and although a variant of soft matching can be performed, we leave it for future work and limit our focus only on the values that can be matched unambiguously. We can combine the predictions of several language-specific base learners into a single pseudo-label for examples where arguments can be matched across the articles.

3.2.4 Base learners

Each iteration starts with a set of labeled arguments X_l, a set of unlabeled arguments X_u and a set of base learners trained on X_l. Base learners are simple logistic regression classifiers that use the vector representations of arguments.

Each base learner $\bar{f}^r_{t,l}$ is a binary classifier trained on the labeled data for the role r from the topic t and the language l. Such base learners are *topic-specific* as they are trained on a single topic t. Base learners $\bar{f}^r_l$ are trained on the labeled data for the role r from the language l and all the topics with the role r. Such base learners are *shared* across topics, as they consider the arguments from all the topics as a single training set. We use the classification probability

of the positive class instead of hard labels, $\bar{f}^r_{t,l}(x), \bar{f}^r_l(x) \in [0, 1]$.

For each argument x from a news article with the language l reporting on the event e from the topic t we obtain the following predictions:

- $\bar{f}^r_{t',l}(x)$ for each $r \in \mathcal{R}_t$ and all such t' that $r \in \mathcal{R}_{t'}$, that is the probability that x is a argument for the role r, where r is a role from the topic t, using the *topic-specific* base learner trained on examples from the same language on the topic t' that also has the role r,
- $\bar{f}^r_{t,l'}(x)$ which equals $\bar{f}^r_{t,l'}(y)$ for each $r \in \mathcal{R}_t$ and for each language l' such that there is an article reporting about the same event e in that language and contains an argument y which is matched to x,
- $\bar{f}^r_l(x)$ for each $r \in \mathcal{R}_t$, using the *shared* base learner on arguments from all topics t' that have the role r.

3.2.5 Combining the predictions

Multiple base learners make predictions for each argument. We combine such predictions to a probability distribution over argument roles as a weighted average.

The weight of each base learner $\bar{f}$ is determined by its error rate $e(\bar{f})$ which is estimated using both unlabeled and labeled examples. We introduce the following logical rules (referred to as *ensemble rules* in [24]) for each of the base learners $\bar{f}^r$ predicting for x:

$$\begin{aligned} \bar{f}^r(x) \wedge \neg e(\bar{f}^r) &\rightarrow f^r(x) \\ \bar{f}^r(x) \wedge e(\bar{f}^r) &\rightarrow \neg f^r(x) \\ \neg\bar{f}^r(x) \wedge \neg e(\bar{f}^r) &\rightarrow \neg f^r(x) \\ \neg\bar{f}^r(x) \wedge e(\bar{f}^r) &\rightarrow f^r(x) \end{aligned}$$

Here, the truth values are not limited to Boolean values, but instead represent the probability that the corresponding ground predicate or rule is true. For a detailed explanation of the method, we refer the reader to [24].

We introduce a prior belief that the predictions of base learners are correct via the following two rules:

$$\bar{f}^r(x) \rightarrow f^r(x), \, and \, , \neg\bar{f}^r(x) \rightarrow \neg f^r(x).$$

Since each example can be a value for at most one role, we introduce a *mutual exclusion* rule:

$$\bar{f}^r(x) \wedge f^{r'}(x) \rightarrow e(\bar{f}^r).$$

The rules are written in the syntax of a Probabilistic soft logic [28] program, where each rule is assigned a weight. We assign a weight of 1 to all *ensemble rules*, a weight of 0.1 to all *prior belief* rules and a weight of 1 to all *mutual exclusion* rules. The inference is performed using the PSL framework [3]. As we obtain the approximations for all $x \in X_u$, we extend the set of positive examples for each role r with all x such that $f^r(x) >= T_p$ and the set of negative examples with all x such that $f^r(x) <= T_n$, for predefined thresholds T_p and T_n.

[3] https://psl.linqs.org/

4 Experiments

4.1 Dataset

To evaluate the proposed methodology, we have conducted experiments on two topics: *earthquakes* and *terrorist attacks*.

We have collected Wikipedia articles and Wikidata information of 913 earthquakes from 2000 to 2020 in 6 different languages, namely English, Spanish, German, French, Italian, and Dutch. We have manually annotated the arguments of 85 English articles using the argument roles *number of deaths*, *number of injured* and *magnitude*, which serve as a labeled test set and are not included in the training process. In addition, we have collected data from 315 terrorist attacks from 2000 to 2020 and articles from the same 6 languages.

4.2 Evaluation settings

The evaluation for each approach is performed on a labeled English dataset, where 76 arguments are labeled as number of deaths, 45 as number of injured and 125 as magnitude. The threshold values for the pseudo-labeling are set to $T_p = 0.6$ and $T_n = 0.05$. The approaches differ by the subset of base learners used to form the combined prediction and by the weighting of the predictions.

Single or multiple languages In a single language setting, only English articles are used to label the arguments and train the base learners. In a multi-language setting, all available articles are used and the arguments are matched across the articles from the same event.

Single or multiple topics In a single topic setting, only the arguments from the *earthquake* topic are used. In a multi-topic setting, the arguments from *terrorist attacks* are used as negative examples for *magnitude*, the base learners for the roles *number of deaths* and *number of injured* are combined as described in the section 3.2.4.

Uniform or estimated weights In the uniform setting all predictions of the base learners contribute equally, while in the estimated setting the weights of the base learners are estimated using the approach described in section 3.2.5.

4.3 Results

The results of all the experiments are summarized in Table 1. Since the test set is limited to the topic *earthquake* and English, only a subset of base learners was used to make the final predictions. We report the average value

Table 1: Results of all experiments. The column *Single iteration* reports the results of approaches where base learners were trained on the seed set only. Results, where base learners were trained in the semi-supervised setting with different weightings of the predictions, are reported in the columns *Uniform weights* and *Estimated weights*. The values of precision, recall, and F1 are averaged over all argument roles.

	Single iteration			Uniform weights			Estimated weights		
Model	P	R	F1	P	R	F1	P	R	F1
Single language, single topic	0.94	0.64	0.76	0.83	0.75	0.77	0.84	0.76	0.79
Multiple languages, single topic	0.94	0.64	0.76	0.82	0.74	0.76	0.83	0.75	0.77
Single language, multiple topics	0.91	0.76	0.83	0.83	0.83	0.83	0.86	0.83	0.84
Multiple languages, multiple topics	0.93	0.76	0.83	0.82	0.83	0.82	**0.84**	**0.84**	**0.84**

of precision, recall, and F1 across all argument roles. The probability threshold of 0.5 was used to determine the classification label.

Single iteration Approaches in which base learners are trained on the initial seed set for a single iteration achieve higher precision (0.94 compared to 0.84 achieved by *estimated weights* in *single language, single topic* setting) with the cost of a lower recall (0.64 compared to 0.76). We observe that they distinguish almost perfectly between the argument roles from the seed set and produce almost no false positives. Using one or more languages has almost no effect on the averaged scores when the number of topics is fixed. When using multiple topics, a higher recall is achieved without a significant decrease in precision. All incorrect classifications of the role *number on injured* are actually examples of the *number of missing* role that is not included in our set and likewise almost all incorrect classifications for the role *magnitude* are examples of the role *intensity on the Mercalli scale*. This could easily be solved by expanding the set of roles and shows how important it is to learn to classify multiple roles simultaneously.

Uniform and estimated weights Semi-supervised approaches in which base learners are trained iteratively trade precision in order to significantly improve recall (0.64 compared to 0.76 achieved by *estimated weights* in *single language, single topic* setting). Most of the loss of precision is due to misclassification between roles *number of deaths* and *number of injured*, similar to the example *"370 people were killed by the earthquake and related building collapses, including **228** in Mexico City, and more than 6,000 were injured."* where 228 was incorrectly classified as the number of injured and not the number of deaths. The use of multiple topics reduces misclassification between these roles and further improves recall as new contexts are discovered by the base learners trained on *terrorist attacks*.

Using the estimated error rates as weights for the predictions of base learners shows a slight improvement in performance. It may be beneficial to estimate multiple error rates for *topic-specific* base learners, as they tend to be more reliable in labeling arguments from the same topic. We believe more data and experiments are needed to properly evaluate this component. A major advantage is its flexibility, as we can easily incorporate prior knowledge about the roles or additional constraints on the predictions in the form of logical rules.

5 Conclusion

The proposed method avoids the need to manually annotate the data for event argument extraction and instead combines Wikipedia and Wikidata to obtain labeled data. Compared to the related work, the proposed methodology uses semi-supervised learning and integrates cross-lingual data into the learning process to enhance the pseudo-labeling supported by probabilistic soft logic. The resulting classification models, used for automatic labeling, can be readily used to extract the event arguments in new texts. It is also possible to train or fine-tune a stronger state-of-the-art model on the resulting dataset, extending it to new event arguments and languages beyond those included in the original training datasets. The experiments were performed on a relatively small dataset and show that the proposed direction seems promising. However, the more suitable test of our approach would be to apply it to a much larger number of topics and events, which we will do in the next step. Moreover, the current approach needs to be evaluated in more detail.

Acknowledgement

This work was supported by the Slovenian Research Agency under the project J2-1736 Causalify and cofinanced by the Republic of Slovenia and the European Union's H2020 research and innovation program under NAIADES EU project grant agreement H2020-SC5-820985.

References

[1] Jesper E. van Engelen and H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109:373–440, 2019. https://doi.org/10.1007/s10994-019-05855-6.

[2] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James

Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online, June 2021. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.200.

[3] Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. WikiReading: A novel large-scale language understanding task over Wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1545, Berlin, Germany, August 2016. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1145.

[4] Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 166–175, New York, NY, USA, 2019. Association for Computing Machinery. https://doi.org/10.1145/3292500.3330955.

[5] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-1034.

[6] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1514.

[7] Alessio Aprosio, Claudio Giuliano, and Alberto Lavelli. Extending the coverage of dbpedia properties using distant supervision over wikipedia. *CEUR Workshop Proceedings*, 1064, 01 2013. https://dl.acm.org/doi/10.5555/2874479.2874482.

[8] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September 2014. https://doi.org/10.1145/2629489.

[9] Dustin Lange, Christoph Böhm, and Felix Naumann. Extracting structured information from wikipedia articles to populate infoboxes. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, page 1661–1664, New York, NY, USA, 2010. Association for Computing Machinery. https://doi.org/10.1145/1871437.1871698.

[10] Florian Schrage, Nicolas Heist, and Heiko Paulheim. Extracting literal assertions for dbpedia from wikipedia abstracts. In *SEMANTiCS*, pages 288–294, 11 2019. https://doi.org/10.1007/978-3-030-33220-4_21.

[11] Nicolas Heist, Sven Hertling, and Heiko Paulheim. Language-agnostic relation extraction from abstracts in wikis. *Information*, 9(4):75, 2018. https://doi.org/10.3390/info9040075.

[12] Boya Peng, Yejin Huh, Xiao Ling, and Michele Banko. Improving knowledge base construction from robust infobox extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 138–148, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-2018.

[13] Sha Li, Heng Ji, and Jiawei Han. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online, June 2021. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.69.

[14] Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online, November 2020. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.73.

[15] Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski, and Marjorie Freedman. GAIA: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, Online, July 2020. Association

for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.11.

[16] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[17] Beth M. Sundheim. Overview of the fourth message understanding evaluation and conference. In *Proceedings of the 4th conference on Message understanding - MUC4 '92*. Association for Computational Linguistics, 1992. https://doi.org/10.3115/1072064.1072066.

[18] Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online, July 2020. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.718.

[19] James Ferguson, Colin Lockard, Daniel Weld, and Hannaneh Hajishirzi. Semi-supervised event extraction with paraphrase clusters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 359–364, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2058.

[20] Xiao Liu, Heyan Huang, and Yue Zhang. Open domain event extraction using neural latent variable models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2860–2871, Florence, Italy, July 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1276.

[21] Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. Cross-lingual structure transfer for relation and event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 313–325, Hong Kong, China, November 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1030.

[22] Jian Liu, Yufeng Chen, and Jinan Xu. Machine reading comprehension as data augmentation: A case study on implicit event argument extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2725, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.214.

[23] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning, 2021.

[24] Emmanouil A. Platanios, Hoifung Poon, Tom M. Mitchell, and Eric Horvitz. Estimating accuracy from unlabeled data: A probabilistic logic approach. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4364–4373, Red Hook, NY, USA, 2017. Curran Associates Inc. https://dl.acm.org/doi/10.5555/3294996.3295190.

[25] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1279.

[26] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747.

[27] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[28] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *J. Mach. Learn. Res.*,

18(1):3846–3912, January 2017. https://dl.acm.org/doi/10.5555/3122009.3176853.