Dynamic cross-sentential context representation for event detection

Dorian Kodelja, Romaric Besancon, Olivier Ferret

Dynamic Cross-Sentential Context Representation for Event Detection*

Dorian Kodelja, Romaric Besançon, and Olivier Ferret (✉)

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{dorian.kodelja,romaric.besancon,olivier.ferret}@cea.fr

Abstract. In this paper, which focuses on the supervised detection of event mentions in texts, we propose a method to exploit a large context through the representation of distant sentences selected based on coreference relations between entities. We show the benefits of extending a neural sentence-level model with this representation through evaluation carried out on the TAC Event 2015 reference corpus.

Keywords: Information extraction · event detection · global context.

1 Introduction

This study focuses on the supervised event extraction from text, which consists in identifying in texts the words or the sequences of words, called event mentions, that mark the presence of a predefined type of events. For instance, the word *pow-wow* for an event of type MEET in:

> Putin had invited Tony Blair to the **pow-wow** in Saint Petersburg's Grand Hotel Europe.

The best methods for achieving this task are generally based on neural models and operate at the sentence level, similarly to [13]. However, the sentence level is not always sufficient to get all the elements for detecting an event mention. Two main types of studies already explored the possibility to exploit information at a larger scale: on the one hand, methods that use document level information to perform event extraction at a local scale; on the other hand, methods that achieve event extraction globally at the document level through joint approaches [3,15,10,18]. Our work takes place among the first type of methods, which can be broken down into methods using specific information at the document scale between events [9,8] or event and entities [6] and methods exploiting a more global representation of documents, either through generic models such as *Doc2Vec* [4] or models specifically trained for the target task as in [19].

In this article, our contribution is a new method for taking into account the document context for event extraction. More precisely, we exploit the coreference links from the entities surrounding a candidate mention to dynamically build its context from selected event-related distant sentences. The representations of

* Work partly supported by ANR under project ASRAEL (ANR-15-CE23-0018) with the FactoryIA supercomputer supported by the Ile-de-France Regional Council.

those sentences are then integrated into a sentence-level model that, similarly to recent studies [13,11,17,1], is based on Graph Convolution Networks (GCN) [7].

2 Model

Classically, we frame event detection as a multi-class classification task for each word in a document. The label is either one of the 38 event types of the DEFT Rich ERE taxonomy [2] or the *NONE* label for the absence of event mention.

2.1 Intra-sentential GCN

Our intra-sentential model is a GCN relying on syntactic dependencies, similarly to [13]. In this model, we consider as candidate each word w_t in a sentence $S = (w_1, w_2, \ldots, w_n)$, where w_i is the i-th word in the sentence, associated with an entity type e_i (with $e_i = O$ if w_i is not an entity head). Each of these words is represented as a real-valued vector $X = x_1, x_2, \ldots, x_n$ built by concatenating three kinds of embeddings: a word embedding for representing the word itself, a position embedding for its relative distance to the candidate, and an entity embedding for its entity type e_i. A BiLSTM is applied to the target sentence S (focused on w_t through the position embedding) for producing a first contextual representation of each word. A GCN made of K convolution layers is then used for producing a contextual representation of each word taking into account the influence of distant words of S through up to K syntactic dependencies. It relies on a directed graph G where the nodes are the words of S and each edge (w_i, w_j) is associated with a label $L(w_i, w_j)$ corresponding to a syntactic dependency between w_i and w_j. The last step consists in aggregating the sequence $h^K_{w_1}, h^K_{w_2}, \ldots h^K_{w_n}$ at the last convolution layer into a final representation p_t of the target word w_t that can be fed to a dense layer with a softmax for the classification. [13] introduces a new pooling strategy that focuses on entities, with the assumption that entities carry a special interest for the task.

With a similar goal, we propose *syntactic pooling*, which also considers multiple specific words in the sentence while not requiring a prior annotation of named entities. In this case, the pooling is focused on the target word and all nouns (n), verbs (v), and adjectives (a) in the sentence:

$$p_t = maxpool(\{h^K_{w_t}\} \cup \{h^K_{w_i, 1 \leqslant i \leqslant n} : pos(w_i) \in \{n, v, a\}\}) \quad (1)$$

2.2 Cross-Sentential Context Representation

Contrary to work integrating a global representation of the document [4,19], we chose to take into account the context of a target sentence in a more selective way, both for improving the disambiguation of candidate event mentions and limiting the parameters of the model. For the task of event extraction, the presence of common named entities is a good indicator of the contextual association of the sentences since they are typically possible arguments of similar events (for

instance, different legal events concerning the same person), related events in a chronological succession (an injure event followed by a die event) or even two mentions of the same event. In the example given in the introduction, *pow-wow* is not a frequent word for a MEET event but the event is also mentioned with less ambiguous occurrences in the same document, in sentences sharing common entities, such as *Saint Petersburg*:

But the *Saint Petersburg* **summit** ended without any formal declaration on Iraq.

Context Representation Our context representation relies on the integration of a contextual representation of each entity mention e_i^j of the target sentence S^j. For selecting the context linked to e_i^j, we define the function $links(S^j, S^k, i)$, that gives the set of positions l in a context sentence S^k of its entity mentions that are in a coreference relation with the considered entity mention e_i^j:

$$links(S^j, S^k, i) = \{l : E(e_i^j) = E(e_l^k)\} \quad 1 \leqslant l \leqslant n \tag{2}$$

where $E(e)$ denotes the entity referred by the mention e. The context of e_i^j is then built from the set of pairs (context sentence, mention of $\mathrm{E}(e_i^j)$) defined as:

$$Links(S^j, i) = \left\{(S^k, l) : l \in links(S^j, S^k, i)\right\} \quad 1 \leqslant l \leqslant n,\ k \neq j \tag{3}$$

For each pair (S^k, l) of this context, we produce an input representation, noted $X^{k,l} = x_1^{k,l}, x_2^{k,l}, \ldots, x_n^{k,l}$, similar to the one in Section 2.1, except for the position embeddings: in this case, the position vector of each word of S^k represents the distance to the position l of the entity mention e_l^k. A BiLSTM is then applied to this input representation. Two extraction methods for the representation of each pair (S^k, l) are considered: the *Final* mode (eq. 4), which concatenates the final representations of the two LSTMs, and the *Mention* mode (eq. 5), which extracts the representations at the position of the entity mention e_l^k.

$$\textbf{Final} : h_{\mathbf{context}}(S^k, l) = [h_{\text{forward}}(x_n^{k,l}); h_{\text{backward}}(x_1^{k,l})] \tag{4}$$

$$\textbf{Mention} : h_{\mathbf{context}}(S^k, l) = [h_{\text{forward}}(x_l^{k,l}); h_{\text{backward}}(x_l^{k,l})] \tag{5}$$

Context Integration The context representation of the entity mention e_i^j is then integrated into the local context at two possible levels, as illustrated by Figure 1: either as an additional embedding in the local input representation of the entity mention, or as an additional node in the graph, associated with the node of the entity mention by a specific relation. For both integration modes, the expected representation is a vector that we obtain by aggregating the vectors of all contextual entity mentions through max-pooling:

$$context(e_i^j) = maxpool(\{h_{\mathbf{context}}(S^k, l) : (S^k, l) \in Links(S^j, i)\}) \tag{6}$$

For the integration as a node, we modify the dependency graph G by adding a node cn_i^j merging all the context representations of e_i^j and having $h^0_{cn_i^j} = context(e_i^j)$ as initial representation. We then define a new *Context* edge type between the

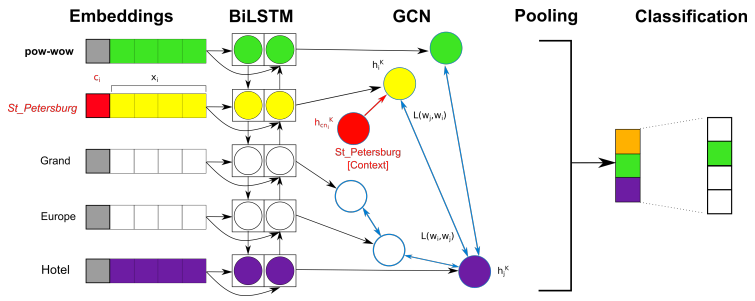

Fig. 1. Two solutions for integrating the context representation (in red) of an entity mention into the GCN model: at the input level or by adding a node to the graph.

local entity mentions and their context representation and add the corresponding edge (w_i^j, cn_i^j) in G. For the integration at the embedding level, the context representation is concatenated to the other embeddings. For the words having no context representation, a default representation $c_{default}$ is used, initialized randomly, and modified during training. The context vector defined in (6) is then generalized to all the words of the sentence with:

$$c_i^j = \begin{cases} context(e_i^j) & \text{if } |Links(S^j, i)| > 0 \\ c_{default} & \text{otherwise} \end{cases} \tag{7}$$

and the input sequence is redefined as $X^j = ([x_0^j, c_0^j], [x_1^j, c_1^j], \ldots, [x_n^j, c_n^j])$.

3 Experiments

3.1 Data and Preprocessing

Our training dataset is composed of 58 documents from the TAC 2015 training dataset and 288 documents from the DEFT Rich ERE (R2 V2 and V2) dataset. The validation set is composed of the remaining 100 documents from the TAC 2015 training dataset. We evaluate our proposed model on the test set of TAC 2015 to compare it with the graph model of [13].

We use the Stanford CoreNLP tool [12] for named entity recognition (NER), coreference resolution, and syntactic analysis to produce dependency graphs from its *Basic dependencies*. At the document level, an entity is defined as a group of mentions in coreference. For increasing the coverage of the automatic coreference system, we merge entities mentions of which are identical.

Example Generation To facilitate access to content-bearing words in the graph convolution, we filter some categories of words: punctuations, symbols, numbers, determiners, prepositions, conjunctions, or interjections. We also use a prediction mask: only nouns, verbs, and adjectives are associated with a predicted category; the other words are associated with the NONE class.

Hyperparameters The word embeddings are initialized with pretrained GloVe embeddings [14]. The position and the entity type embeddings are of size 50 while the dimensions of the local BiLSTM layer and the two graph convolution layers are 400 and 300 respectively. The embeddings for the words, entities, and distances are the same for the target sentences and the context sentences. The model is trained using SGD with momentum and batches of 10 examples. All average performances are computed on 10 runs with the same parameters.

3.2 Study of Model's Parameters

We first evaluate the influence of the different choices for the model's parameters:

- **Intra-sentential pooling**: *Syntactic*/*Entity*
- **Context representation extraction**: *Final*/*Mention*
- **Context representation integration**: *Embedding*/*Node*

We searched for the best values of these parameters on the validation set together with the values of less specific optimization parameters (learning rate, l2 regularization, dropout, momentum). Concerning the model's parameters, the best result is obtained using *Syntactic* pooling, *Final* extraction, and *Embedding* integration. These parameters are also the best, in general, in each tested configuration, but since we cannot show all results, we present in Table 1 the results for this best model, noted C-GCN, and the variations of this model when changing each of the other parameters.

Table 1. Performances on the validation set for the main model's parameters ($P_{avg.}$, $R_{avg.}$, $F_{avg.}$: average values from 10 runs of precision, recall, and F-score; F_{σ}: F-score standard deviation; $F_{max.}$: F-score maximal value).

	$P_{avg.}$	$R_{avg.}$	$F_{avg.}$	F_{σ}	$F_{max.}$
C-GCN	75.6	**50.4**	**60.5**	0.6	**60.4**
Pooling - *Entity*	74.8	49.2	59.3	0.9	60.2
Extraction - *Mention*	75.0	48.8	59.1	1.2	58.1
Integration - *Node*	**76.9**	48.1	59.1	1.2	59.3

We observe, with a weakly significant difference ($p = 0.058$), that the *entity* pooling is slightly worse than the *Syntactic* pooling, which indicates that the use of a larger set of context words benefits to an enriched representation of the target word. On the contrary, the *overall* pooling in [13] performs worse than the *entity* pooling while it also considers more words than that pooling. However, this difference may come from the use of different NER tools.

Concerning the context extraction, the poor results obtained with the *Mention* mode could also be related to the quality of the entities or to the fact that the final representations of the context sentences are more informative than the specific representations of the entity mentions. Finally, the integration of the context representation as a node does not degrade the results in a significant way but produces a less balanced performance between precision and recall.

3.3 Comparison with State-of-the-art

We compare in this section our proposed model to the original model from [13], noted GCN_{nguyen}, and to the best model of the TAC evaluation campaign, RPI_BLENDER, proposed by [5], based on a MaxEnt classifier using a large set of lexical, syntactic and entity features. To further prove the interest of having a specific context for each example, we train a model $C\text{-}GCN_{generic}$ that uses all the sentences of the document as context. In this case, there is no position embedding for the context sentences, and the same representation is used as an embedding for all the words in the considered sentence.

Table 2. Results on TAC 2015 test set ($F_{max./dev}$: F-score for the best parameters on the dev. set; for the two reference systems, P and R are *max./dev* values; average values for the others).

	P	R	$F_{avg.}$	F_{σ}	$F_{max./dev}$
RPI_BLENDER	75.2	47.7	–	–	58.4
GCN_{nguyen}	70.3	**50.6**	–	–	58.8
GCN_{repro}	**78.5**	47.0	58.7	0.8	59.1
$C\text{-}GCN_{generic}$	74.5	48.4	58.6	0.6	59.0
C-GCN	75.6	50.4	**60.5**	0.6	**60.4**

The results presented in Table 2 prove the interest of our proposition: our implementation of the GCN model, noted GCN_{repro}, achieve results similar to the ones reported by [13][1] and we obtain a gain of 1.8 F-score on this baseline when using the context representation ($p < 0.0001$). We also see that the integration of the context in $C\text{-}GCN_{generic}$ does not yield better results, which confirms our intuition on the interest of defining a context specific to each example.

4 Conclusion and Perspectives

We propose in this article a method allowing a neural model for event extraction to take into account a cross-sentential context. The approach consists in enriching the representation of entity mentions in a target sentence with a contextual embedding built using information from distant sentences where these entities also occur. The evaluation of the approach on the dataset TAC 2015 proves the interest of the method, with a significant gain over the initial model. One perspective would be to use an attention mechanism as an alternative to the max-pooling to aggregate the representations of all the mentions of an entity, which could lead to better discriminate and filter context sentences. Another one could be the use of a more elaborated GCN model able to take into account the type of the relations in the graph, such as Relational GCN [16].

[1] We note that we do not have exactly the same train/dev datasets because we also used the DEFT dataset as training, which can explain the slight gain in F-score.

References

1. Balali, A., Asadpour, M., Campos, R., Jatowt, A.: Joint event extraction along shortest dependency paths using graph convolutional networks. Knowledge-Based Systems **210**, 106492 (2020). https://doi.org/10.1016/j.knosys.2020.106492
2. Bies, A., Song, Z., Getman, J., Ellis, J., Mott, J., Strassel, S., Palmer, M., Mitamura, T., Freedman, M., Ji, H., et al.: A Comparison of Event Representations in DEFT. In: Fourth Workshop on Events. pp. 27–36 (2016)
3. Chen, Y., Yang, H., Liu, K., Zhao, J., Jia, Y.: Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms. In: 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018). pp. 1267–1276. ACL, Brussels, Belgium (2018). https://doi.org/10.18653/v1/D18-1158
4. Duan, S., He, R., Zhao, W.: Exploiting Document Level Information to Improve Event Detection via Recurrent Neural Networks. In: Eighth International Joint Conference on Natural Language Processing (IJCNLP 2017). pp. 352–361 (2017)
5. Hong, Y., Lu, D., Yu, D., Pan, X., Wang, X., Chen, Y., Huang, L., Ji, H.: RPI_BLENDER TAC-KBP2015 System Description. In: Proceedings of the 2015 Text Analysis Conference (2015)
6. Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., Zhu, Q.: Using Cross-Entity Inference to Improve Event Extraction. In: 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011). pp. 1127–1136. ACL (2011)
7. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations (ICLR 2017). Toulon, France (2017)
8. Kodelja, D., Besançon, R., Ferret, O.: Exploiting a More Global Context for Event Detection Through Bootstrapping. In: 41st European Conference on Information Retrieval (ECIR 2019). pp. 763–770 (2019)
9. Liao, S., Grishman, R.: Using Document Level Cross-Event Inference to Improve Event Extraction. In: 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010). pp. 789–797. ACL, Uppsala, Sweden (2010)
10. Liu, S., Liu, K., He, S., Zhao, J.: A Probabilistic Soft Logic Based Approach to Exploiting Latent and Global Information in Event Classification. In: Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16). AAAI Press, Phoenix, AZ, USA (2016)
11. Liu, X., Luo, Z., Huang, H.: Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation. In: 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018). pp. 1247–1256. Brussels, Belgium (2018)
12. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), System Demonstrations. pp. 55–60 (2014)
13. Nguyen, T.H., Grishman, R.: Graph Convolutional Networks with Argument-Aware Pooling for Event Detection. In: Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18). AAAI Press, New Orleans, LA, USA (2018)
14. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). pp. 1532–1543. ACL, Doha, Qatar (2014)

15. Reichart, R., Barzilay, R.: Multi-Event Extraction Guided by Global Constraints. In: 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2012). pp. 70–79. Montréal, Canada (2012)
16. Schlichtkrull, M., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I., Welling, M.: Modeling Relational Data with Graph Convolutional Network. In: 15th Extended Semantic Web Conference (ESWC 2018). pp. 593–607. Springer International Publishing, Heraklion, Crete, Greece (2018)
17. Yan, H., Jin, X., Meng, X., Guo, J., Cheng, X.: Event detection with multi-order graph convolution and aggregated attention. In: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). pp. 5766–5770. Hong Kong, China (2019)
18. Yang, B., Mitchell, T.M.: Joint Extraction of Events and Entities within a Document Context. In: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016). pp. 289–299. ACL, San Diego, California (2016)
19. Zhao, Y., Jin, X., Wang, Y., Cheng, X.: Document Embedding Enhanced Event Detection with Hierarchical and Supervised Attention. In: 56th Annual Meeting of the Association for Computational Linguistics (Short Papers) (ACL 2018). pp. 414–419. ACL, Melbourne, Australia (2018)