The UWNLP system at SemEval-2018 Task 7: Neural Relation Extraction Model with Selectively Incorporated Concept Embeddings

Yi Luan Mari Ostendorf Hannaneh Hajishirzi
University of Washington
{luanyi, ostendor, hannaneh}@uw.edu

Abstract

This paper describes our submission for the SemEval 2018 Task 7 shared task on semantic relation extraction and classification in scientific papers. We extend the end-to-end relation extraction model of (Miwa and Bansal, 2016) with enhancements such as a character-level encoding attention mechanism on selecting pretrained concept candidate embeddings. Our official submission ranked the second in relation classification task (Subtask 1.1 and Subtask 2 Senerio 2), and the first in the relation extraction task (Subtask 2 Scenario 1).

1 Task Overview

The SemEval 2018 Task 7 Shared Task (Gábor et al., 2018) focuses on the task of recognizing the semantic relation that holds between scientific concepts. The task involves semantic relation extraction and classification into six categories specific to scientific literature: USAGE, RESULT, MODEL-FEATURE, PART_WHOLE, TOPIC, COMPARE. Two types of tasks are proposed: 1) identifying pairs of entities that are instances of any of the six semantic relations (extraction task), and 2) classifying instances into one of the specific relation types (classification task).

Consider the following input sentence: "[*Unsupervised training*] is first used to train a [*phone n-gram model*] for a particular domain." Given the concept pair [*Unsupervised training*] and [*phone n-gram model*], the relation extraction task is to identify whether there is a relation between the concepts, while the the relation classification task is to identity the relation as USAGE. Relation directionality is not taken into account for the evaluation of the extraction task. Directionality is taken into account when relevant for the classification task (5 out of the 6 semantic relations are asymmetrical). We will use this example throughout the paper to illustrate various parts of our system.

The SemEval 2018 Task 7 dataset contains 350 abstracts from the ACL Anthology for training and validation, and 150 abstracts for testing each subtask. Since the scale of the data is small for supervised training of neural systems, we introduce several strategies to leverage a large quantity of unlabeled scientific articles. In addition to initializing a neural system with pre-trained word embeddings, as in (Luan et al., 2017), we also try to incorporate embeddings of concepts that span multiple words. In neural models such as (Miwa and Bansal, 2016), phrases are often represented by an average (or weighted average) of the token's sequential LSTM representation. The intuition behind explicit modeling of multi-word concept embeddings is that the concept use may be different from that of its individual words. Due to the size of the dataset and the nature of scientific literature, a large number of the scientific terms in the test set have never appeared in the training set, so supervised learning of the phrase embeddings is not feasible. Therefore, we pre-trained scientific term embeddings on a large scientific corpus and provide a strategy to selectively incorporate the pre-trained embeddings into the relation extraction system.

2 System Description

2.1 Neural Architecture Model

Our system is an extension of (Luan et al., 2017) and (Miwa and Bansal, 2016) with LSTM RNNs that represent both word sequences and dependency tree structures, and perform relation extraction between concepts on top of these RNNs. As illustrated in Figure 1, it is composed of a 5 types of layers in a hierarchical neural model to encode context information. The first two layers (token, token LSTM) use the neural modeling framework in (Luan et al., 2017). The forward and backward dependency layers and the relation classification

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 788–792
New Orleans, Louisiana, June 5–6, 2018.

layer are based on (Miwa and Bansal, 2016). The concept selection layer is novel, to the best of our knowledge. The different layers are described in more detail below.

Token Layer. The token layer concatenates three types of vector space embeddings. *Word embeddings* are learned for words from a fixed vocabulary (plus the unknown word token), initialized using Word2vec pre-training with large scholarly corpora. The *character-based embedding* for a token is derived from its characters as the concatenation of forward and backward representations from a bidirectional LSTM. The character look-up table is initialized at random. The advantage of building a character-based embedding layer is that it can handle out-of-vocabulary words and equations, which are frequent in this data, all of which are mapped to "UNK" tokens in the Word Embedding Layer. *Word embeddings* are learned for words from a fixed vocabulary (plus the unknown word token), initialized using Word2vec pre-training with large scholarly corpora. A *feature embedding* is learned as a mapping from features associated with capitalization (all capital, first capital, all lower, any capital but first letter) and part-of-speech tags. The embeddings are randomly initialized and trained jointly with other parameters during supervised training.

Token LSTM Layer We apply a bidirectional LSTM at the token level taking the concatenated character-word-feature embedding as input. An LSTM hidden state generated in this layer is denoted as h^S.

Forward & Backward Dependency Layers Given the concept pair (C_l, C_r), the Forward Dependency Layer (generating h^F) traces from the closest common ancestor w_a (for example the word "*used*" in Fig. 1) to the headword w_j (word "*model*") of the right target concept C_r ("*phone n-gram model*"). The Backward Dependency Layer (generating h^B) traces from the ancestor to the headword w_i of the left concept C_l. We map the dependency relation into vector space and concatenate the resulting embedding to the embedding (h^S) of the headword of the concepts C_l or C_r for the backward and forward dependency layers, respectively. We concatenate the resulting bi-directional LSTM vector for the headwords together with the common ancestor in both Forward & Backward Dependency Layer as input to Relation Classification Layer

$$h^{DP} = [\overleftarrow{h^B_{w_i}}; \overrightarrow{h^B_{w_i}}, \overleftarrow{h^F_{w_j}}; \overrightarrow{h^F_{w_j}}; \overleftarrow{h^B_{w_a}}; \overrightarrow{h^B_{w_a}}; \overleftarrow{h^F_{w_a}}; \overrightarrow{h^F_{w_a}}] .$$

Concept Selection Layer The concepts in the task are mostly phrases rather than single words, in the SemEval Task 7. We therefore seek ways to obtain prior knowledge for those terms. We train a scientific concept extraction model using the state-of-the-art scientific neural tagging technique in (Luan et al., 2017), given the scientific concept annotation in the SemEval 2018 Task7 training data. We were able to achieve 79.8% F1 score (span level) to identify the scientific concepts. We then use the model to extract all scientific concepts in the ACL anthology and AI2 dataset (refer to Sec. 3). We keep all the concepts that occur more than 10 times in the whole corpus, which results in around 15k concepts. We treat each of the 15k concepts as an individual token and retrain word2vec embeddings v_k together with all other single words. At training time, given a scientific concept pair (C_l, C_r), we search through the 15k concepts to get all the concept candidates that have n-gram string match with C_l and C_r respectively (n is from 1 to the length of the target concept C). For example, for the concept *phone n-gram model*, the candidate concepts we get are {*phone n-gram*, *n-gram model*, *n-gram*, *model*, *phone*}. Since there may exist cases where no match could be found in the 15k concepts, we introduce a null vector $v_\varnothing$. $v_\varnothing$ is learned with other neural network parameters. Assume there are K concept candidates in the candidate list, we denote the embeddings for the concept candidates to be $V = \{v_1 \dots v_K, v_\varnothing\}$. The attention weights are calculated by $\alpha_{lk} \propto \exp(h^S_{C_l} W_{ATT} v_k)$, where $v_k \in V$. $h^S_{C_l}$ is the concatenation of bidirectional LSTM hidden states of the first and last word in C_l.[1] W_{ATT} is a parameter matrix for the bilinear score for $h^S_{C_l}$ and v_k. The final concept embedding v_{C_l} is $v_{C_l} = \sum_{v_k \in V} \alpha_{lk} v_k$. For a target concept C, if exact match exists in the 15K concepts, we set the pre-trained concept embedding to be v_{C_l}. We concatenate the resulting embedding for both concepts in the concept pair as input to the final classification layer ($v_C = [v_{C_l}; v_{C_r}]$).

Relation Classification Layer We concatenate the output of Forward & Backward Dependency Layer h^{DP} and Concept Embedding Selection Layer v_C as input to Relation Classification Layer.

[1] We also tried using the weighted average of all LSTM word embeddings in the span to calculate $h^S_{C_l}$; this yields a slightly worse result.

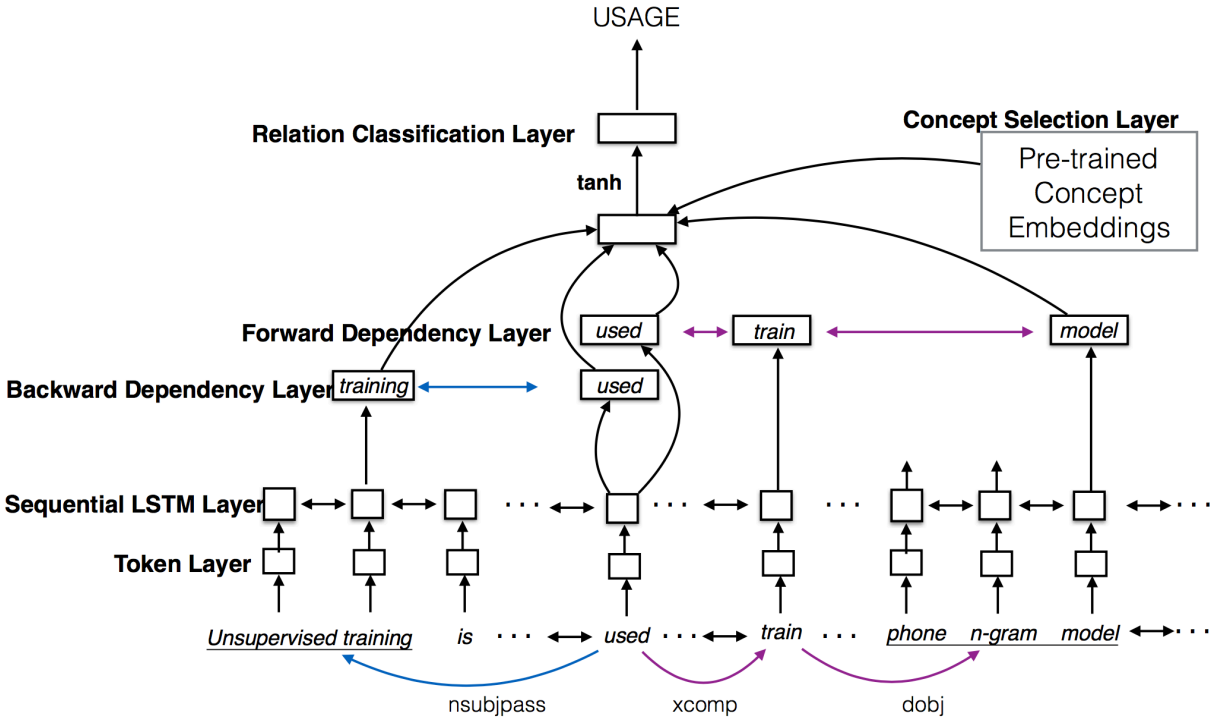

Figure 1: Neural relation extraction model with bidirectional sequential and dependency path LSTMs.

Besides, we also introduce a distance feature between the two concepts which indicates how many other concepts there are in between the target concept pairs. We concatenate the distance embedding with all the other features. The concatenated features are then projected down to a lower dimension through $tanh$ function and make the final prediction through a $softmax$ function.

3 Experimental Setup

External Data We use two external resources for pretraining word embeddings: i) the Semantic Scholar Corpus,[2] a collection of over 20 million research papers from which we extract a subset of 110k abstracts of publications in the artificial intelligence area; and ii) the ACL Anothology Reference Corpus, which contains 22k full papers published in the ACL Anothology (Bird et al., 2008).

Baseline We compare our model with a baseline that removes the Concept Selection Layer and replaces it with a weighted sum (using attention) of hidden states (from the Sequential LSTM Layer) for all words in a concept.

Implementation details All parameters are tuned based on dev set performance; the best parameters are selected and used for final evaluation.

[2]http://labs.semanticscholar.org/corpus/

For all experiments, we explore tuning with two different evaluation metrics: macro-F1 score and micro-F1 score.[3] We keep the pre-trained concept embedding fixed as additional input feature. The word embedding dimension is 250; the LSTM hidden dimension is 100 (for both sequential and dependency layer); the character-level hidden dimension is 25; and the optimization algorithm is SGD with a learning rate of 0.05. For Subtask 2, since 5 out of 6 relation types have directionality, we add relation label "_REVERSE" to all the 5 directional relations together with a "NONE" type, which result in 12 labels in total. For each epoch, we also randomly filter out some "NONE" samples with probability p during training, since the "NONE" type relation dominates the training set and would bias the model towards predicting "NONE" types. We tune p according to dev set, and use $p = 0.4$ for the final evaluation.

4 Experimental Results

Ablation Study Table 1 provides the results of an ablation study on the dev set showing the impact of removing different components of our system.

[3]The official evaluation is macro-F1, but since the number of instances in each class is highly unbalanced, the observed macro-F1 scores were unstable. We therefore introduce micro-F1 score for tuning and evaluation as well.

Model	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1
Our system	49.4	36.7	**42.1**	46.2	42.2	**44.1**
-DepFeat	38.2	39.6	39.0	45.2	41.9	43.0
-DistFeat	43.4	37.8	40.4	38.7	47.8	42.7
-DepLSTM	51.5	30.0	37.9	48.6	32.6	39.0
-Concept	36.2	41.8	38.8	37.6	46.5	41.6
Baseline	40.9	32.5	36.2	41.9	38.0	39.9

Table 1: Ablation study showing the impact of neural network configurations on system performance on the dev set for the relation classification task (Subtask 2, senerio 2). -DepFeat removes the input dependency relation embeddings from the Backward & Forward Dependency Layers. -DistFeat and -Concept omit the distance and concept selection features, respectively, from the final classification layer. -DepLSTM removes the Backward & Forward Dependency Layers entirely (using the LSTM embeddings in the weighted token average).

Looking at micro F1 scores, dependency path information is very important (performance dropped 11.5% without it), and the Concept Selection Layer is also important as it gives 2.5 absolute improvement. The Dependency relation feature and the distance feature also show 1-2 points gain. It is worth noticing that removing the Concept Layer (-Concept) does better than replacing it with the weighted sequential LSTM sum (Baseline). With the small amount of training data, it is difficult for the baseline system to learn a good transformation from word to phrase.

Competition Result The results of our system is in Table 2. We submit two sets of results, one tuned with micro F1 and the other with macro F1. It turns out that even though the official evaluation metric is macro F1 score, our model tuned by micro F1 gets better results in the final competition. In Subtask 1.1 and Subtask 2 scenario 2, we were the second place team with F1 score of 78.9% and 39.1% respectively. We were the first place in Subtask 2 scenario 1 with 50.0% F1.

5 Related Work

There has been growing interest in research on automatic methods to help researchers search and extract information from scientific literature. Past research has addressed citation sentiment (Athar and Teufel, 2012b,a), citation networks (Kas, 2011; Gabor et al., 2016; Sim et al., 2012; Do et al., 2013; Jaidka et al., 2014), summarization (Abu-Jbara and Radev, 2011) and some analysis of research community (Vogel and Jurafsky, 2012; Anderson et al., 2012). However, due to scarce hand-annotated data resources, previous work on information extraction (IE) for scientific literature is very limited. Most previous work focuses on unsupervised methods for extracting scientific terms such as bootstrapping Gupta and Manning (2011); Tsai et al. (2013), or extracting relations (Gábor et al., 2016). Luan et al. (2017); Augenstein and Søgaard (2017) applied semi-supervised learning and multi-task learning to neural based models to leverage large unannotated scholarly datasets for a scientific term extraction task (Augenstein and Søgaard, 2017).

Although not much supervised relation extraction work has been done on scientific literature, neural network techniqueshave obtained the state of the art for general domain relation extraction. Both convolutional (Santos et al., 2015) and RNN-based architectures (Xu et al., 2016; Miwa and Bansal, 2016; Peng et al., 2017; Quirk and Poon, 2017) have been successfully applied to the task and significantly improve performance.

Model	T1.1	T2-E	T2-C
Our system (Micro)	78.9	**50.0**	39.1
Our system (Macro)	78.4	49.3	37.0
Team-1	**81.7**	48.8	**49.3**
Team-2	76.7	37.4	33.6

Table 2: Competition result for the top 3 teams. The official evaluation metric is macro F1 score. T1.1 means Subtask 1.1, T2-E means Subtask 2 senerio 1 (extraction task), T2-C means Subtask 2 senerio 2 (classification task).

6 Conclusion

This paper describes the system of the UWNLP team submitted to SemEval 2018 Task 7. We extend state-of-the-art neural models for information extraction by proposing a Concept Selection module which can leverage the semantic information of concepts pre-trained from a large scholarly dataset. Our system ranked second in the relation classification task (subtask 1.1 and subtask 2 senerio 2), and first in the relation extraction task (subtask 2 scenario 1).

Acknowledgments

This research was supported by the NSF (IIS 1616112), Allen Distinguished Investigator Award, and gifts from Allen Institute of AI, Google, Ama-

zon, Samsung, and Bloomberg. We thank the anonymous reviewers for their helpful comments

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proc. Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. volume 1, pages 500–509.

Ashton Anderson, Dan McFarland, and Dan Jurafsky. 2012. Towards a computational history of the ACL: 1980-2008. In *Proc. ACL Special Workshop on Rediscovering 50 Years of Discoveries*. pages 13–21.

Awais Athar and Simone Teufel. 2012a. Context-enhanced citation sentiment detection. In *Proc. Conf. North American Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. pages 597–601.

Awais Athar and Simone Teufel. 2012b. Detection of implicit citations for sentiment detection. In *Proc. ACL Workshop on Detecting Structure in Scholarly Discourse*. pages 18–26.

Isabelle Augenstein and Anders Søgaard. 2017. Multi-task learning of keyphrase boundary classification. In *Proc. Annu. Meeting Assoc. for Computational Linguistics (ACL)*. pages 341–346.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan R Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, Yee Fan Tan, et al. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proc. Language Resources and Evaluation Conference (LREC)*.

Huy Hoang Nhat Do, Muthu Kumar Chandrasekaran, Philip S Cho, and Min Yen Kan. 2013. Extracting and matching authors and affiliations in scholarly documents. In *Proc. ACM/IEEE-CS Joint Conference on Digital libraries*. pages 219–228.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 Task 7: Semantic relation extraction and classification in scientific papers. In *Proc. Int. Workshop on Semantic Evaluation (SemEval)*.

Kata Gabor, Haifa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the ACL anthology corpus for the automatic analysis of scientific literature. In *Proc. Language Resources and Evaluation Conference (LREC)*.

Kata Gábor, Haïfa Zargayouna, Isabelle Tellier, Davide Buscaldi, and Thierry Charnois. 2016. Unsupervised relation extraction in specialized corpora using sequence mining. In *International Symposium on Intelligent Data Analysis*. Springer, pages 237–248.

Sonal Gupta and Christopher D Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proc. IJCNLP*. pages 1–9.

Kokil Jaidka, Muthu Kumar Chandrasekaran, Beatriz Fisas Elizalde, Rahul Jha, Christopher Jones, Min-Yen Kan, Ankur Khanna, Diego Molla-Aliod, Dragomir R Radev, Francesco Ronzano, et al. 2014. The computational linguistics summarization pilot task. In *Proc. Text Analysis Conference*.

Miray Kas. 2011. Structures and statistics of citation networks. Technical report, DTIC Document.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. In *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proc. Annu. Meeting Assoc. for Computational Linguistics (ACL)*. pages 1105–1116.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Trans. Assoc. for Computational Linguistics (TACL)* 5:101–115.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proc. Meeting of the European Association of Computational Linguistics*. pages 1171–1182.

Cicero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proc. Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. pages 626–634.

Yanchuan Sim, Noah A Smith, and David A Smith. 2012. Discovering factions in the computational linguistics community. In *Proc. ACL Special Workshop on Rediscovering 50 Years of Discoveries*. pages 22–32.

Chen-Tse Tsai, Gourab Kundu, and Dan Roth. 2013. Concept-based analysis of scientific literature. In *Proc. ACM Int. Conference on Information & Knowledge Management*. ACM, pages 1733–1738.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the ACL anthology. In *Proc. ACL Special Workshop on Rediscovering 50 Years of Discoveries*. pages 33–41.

Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. 2016. Improved relation classification by deep recurrent neural networks with data augmentation. In *Proc. Int. Conf. Computational Linguistics (COLING)*. pages 1461–1470.