RGAT at SemEval-2023 Task 2: Named Entity Recognition Using Graph Attention Network

Abir Chakraborty
Microsoft
Abir.Chakraborty@microsoft.com

Abstract

In this paper, we (team RGAT) describe our approach for the SemEval 2023 Task 2: Multilingual Complex Named Entity Recognition (MultiCoNER II). The goal of this task is to locate and classify named entities in unstructured short complex texts in 12 different languages and one multilingual setup. We use the dependency tree of the input query as additional feature in a Graph Attention Network along with the token and part-of-speech features. We also experiment with additional layers like BiLSTM and Transformer in addition to the CRF layer. However, we have not included any external Knowledge base like Wikipedia to enrich our inputs. We evaluated our proposed approach on the English NER dataset that resulted in a clean-subset F1 of 61.29% and overall F1 of 56.91%. However, other approaches that used external knowledge base performed significantly better.

1 Introduction

Named Entity Recognition (NER) is an important natural language processing (NLP) task that deals with extracting relevant information from various types of queries by classifying each token/word of the query into one of the pre-defined classes called named entities. This has wide range of applications in industries like Retail, Healthcare and Manufacturing. For example, in retail domain, we would like to extract information from customer feedback that is instrumental in identifying current drawbacks of a product and direction for further improvement. Similarly, factory logs in a plant containing information about a specific machine and sensor with domain specific terminologies need to be understood to take necessary actions in time. However, the task becomes challenging as the set of entity classes can vary greatly from one domain to another domain.

Most of the NER datasets that are routinely used for benchmarking NER algorithms contain complete or almost complete sentences with relatively few entity classes (e.g., CoNLL-2003 has four classes and WNUT-2017 has six classes). However, in real world applications there could be many granular entities with a very short sentence that may not have any context information at all. As a result, performance of the models may significantly suffer and that is what we have also observed in our current study. The present task, MultiCoNER II (Fetahu et al., 2023b) is a continuation of the previous year's SemEval-2022 task (Malmasi et al., 2022b) that was created for detecting semantically ambiguous and complex entities in short and low-contextual settings for 11 languages (i.e., English, Spanish, Dutch, Russian, Turkish, Korean, Farsi, German, Chinese, Hindi, and Bengali). MultiCoNER II similarly has created tasks for 12 languages, namely, English, Spanish, Ukranian, Hindi, Bangla, Chinese, Swedish, Farsi, French, Italian, Portugese and German with 66 fine-grained entity classes spanning 6 broad groups of Product, Medical, Location, Creative Works, Groups and Person. The number of classes is significantly higher compared to last year (only 12 classes) that increases the difficulty of this task. Another challenging aspect of this task is added simulated errors to the test set during evaluation to make it more realistic and difficult.

Our approach lies in encoding the graph associated with the dependency parsing of the input sentence and create token (node) level representations that take care of both the neighborhood (connected edges) and dependency type (edge type) of each token. We add this feature with the other standard features like token embedding from a large language model like BERT or RoBERTA and part-of-speech embeddings. Based on our overall fine-grained macro-averaged F1-score of 56.91, we secured 23rd rank in the English track of the competition. It is to be noted that our approach heavily relies on the availability of dependency structure of the input query and thus cannot be extended easily

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 163–170
July 13-14, 2023 ©2023 Association for Computational Linguistics

to other languages. It is also another shortcoming of this approach especially where the queries are not well-formed English sentences.

The organization of the paper is as follows. In the next section we provide a detailed literature survey on the techniques employed for NER. Next, we present the details of the proposed approach. Subsequently, the model predictions and comparisons with other baseline methods are discussed. Finally, conclusions are drawn and scope for future works is outlined.

2 Related Work

Ever since it's introduction in 1996 (Grishman and Sundheim), NER has been an active field of research that has evolved considerably with the emergence of new techniques. Initial approaches to NER were limited to rule based methods followed by SVM, HMM and CRF (Li et al., 2020). With the advent of deep learning based techniques, there are RNN (Collobert et al., 2011; Huang et al., 2015; Lample et al., 2016; Shen et al., 2017; Zheng et al., 2017; Zhou et al., 2017) and CNN (Chiu and Nichols, 2016; Aguilar et al., 2017; Strubell et al., 2017) based models that tried to find the best representation of each token. Other than the token embeddings, additional features created are in terms of spelling, context and gazetteer (Huang et al., 2015) and it was found that incorporation of prior knowledge substantially boosted the performance. Similarly, the BiLSTM-CNN model by Chiu and Nichols (2016) uses additional word-level (capitalization, lexicons) and character-level features (4-dimensional vector representing the type of a character: upper case, lower case, punctuation, other) in addition to the token embeddings. Recently, Li et al. (2021) proposed Modularized Interaction Network (MIN) that generated the best result for CoNLL2003 (Sang and Meulder, 2003), WNUT2017 (Derczynski et al., 2017) and JNLPBA datasets. However, none of these approaches created features based on the dependency relations amongst different tokens.

As the Transformer architecture (Vaswani et al., 2017) started replacing many RNN/CNN based encoding, BERT (Devlin et al., 2019) based embeddings found applications for NER (Yamada et al., 2020; Schneider et al., 2020; Shaffer, 2021; Ushio and Camacho-Collados, 2021). All these approaches are based on large pre-trained language models that have seen large monlingual or multilingual corpus during training and later fine-tuned for NER task. Similar studies are carried out by Agarwal et al. (2021) where entity-switched datasets are created where different national origins play a role. Pfeiffer et al. (2020) studied transfer learning in a multilingual setup by creating invertible adapters.

While most of the NER tasks involve well-formed sentences as inputs the real-world datasets can be very different with short texts having low context, emerging domain-specific entities with long-tails and complex entities like noun phrases, infinitives or gerunds. These aspects are discussed in great details by Meng et al. (2021) where solutions are proposed to fuse gazetteer knowledge with learnable weights. Fetahu et al. (2021) extended this approach to multilingual and code-mixed settings along with a mixture-of-experts model on multilingual transformer. Continuing on this approach, Malmasi et al. (2022b) created the MultiCoNER task (Malmasi et al., 2022a) that further enhanced the task with expanded language base.

There are not many studies on the application of Graph Neural Network for NER. Cetoli et al. (2017) investigated the utility of dependency parser and applied stacked graph convolution network (GCN) on Ontonotes 5.0 dataset and obtained an improvement of 2% over a similar Bidirectional LSTM (Bi-LSTM) model. Yu et al. (2020) and Sui et al. (2022) further extended this approach for nested NER using a bi-affine model. Luo and Zhao (2020) introduced a bipartite graph to link the outer and inner named entities along with a Bi-LSTM and GCN model. Similarly, Hanh et al. (2021) proposed a combination of contextual features from XLNet and global features from GCN for improving model performance. GCN based features are also explored with Transformer and Bi-LSTM with CRF layer for biomedical domain (Celikmasat et al., 2022) and Chinese NER (Zhang and Peng, 2022).

Our approach is motivated by graph based approaches applied to Aspect Based Sentiment Analysis (ABSA), e.g., GCN based model of Zhang et al. (2019); Zhou et al. (2022) and Sun et al. (2019), graph attention network (GAT) based model of Huang and Carley (2019) and Wang et al. (2020) where the latter modified the original dependency tree to create an aspect-oriented dependency tree that was used further in a relational GAT (RGAT) where different relations contributed differently in the computation of nodal representations.

Entity	Train		Validation		Test	
	Count	Percentage	Count	Percentage	Count	Percentage
Person	19786	36.6 %	1007	36.6%	291458	36.2%
Creative Works	11163	20.7 %	584	21.2 %	169593	21.1 %
Group	9286	17.2 %	466	16.9 %	134168	16.7 %
Location	8371	15.5 %	392	14.2 %	130923	16.3 %
Product	3094	5.7 %	184	6.7 %	45086	5.6 %
Medical	2337	4.3 %	120	4.3 %	33105	4.1 %

Table 1: Distribution of named entities across the train, validation and test sets. The percentages are computed without considering the "O" tag.

3 Data

We conducted our experiment on the dataset provided by the SemEval 2023 Task 2 organizers (Fetahu et al., 2023b) covering 12 languages, namely, English, Spanish, Ukranian, Hindi, Bangla, Chinese, Swedish, Farsi, French, Italian, Portugese and German with 66 fine-grained entity classes spanning 6 broad groups of Product, Medical, Location, Creative Works, Groups and Person. In the English language track, there are 16778, 871 and 249980 examples in the training, validation and test set, respectively. The distributions of different entity classes across the train, validation and test split are shown in Table 1. More details on the dataset can be found in Fetahu et al. (2023a).

4 Methodology

We have modified the RGAT approach (Wang et al., 2020) which was originally targeted for aspect polarity prediction. The approach utilizes the dependency structure of the input sentence which captures the grammatical relations by connecting the words with the corresponding dependency type. To bring out the relations of different words with the aspect word(s) an Aspect Oriented Dependency Tree (AODT) was proposed by Wang et al. (2020) where the root of the original dependency tree was shifted to the target aspect word followed by pruning some of the unnecessary relations. However, this approach is not exactly applicable for the current study since we do not know the entity words, a priori. However, knowing that these words are generally nouns (ignoring gerunds) we have randomly chosen one of the nouns present in the sentence as a surrogate entity word and modified the dependency tree based on this word.

4.1 Relational Graph Attention Network

AODT can be represented by a graph structure where each node is a word and the edges between them are represented by the dependency relation, e.g., nominal subject, adverbial modifier, etc. Following Wang et al. (2020), given a neighborhood of a node $\mathcal{N}_i$, the node embeddings can be iteratively updated using multi-head attention (with K attentional heads) as

$$h_{att_i}^{l+1} = concat_{k=1}^{K} \sum_{j \in \mathcal{N}_i} \alpha_{ij}^{lk} W_k^l h_j^l, \tag{1}$$

$$\alpha_{ij}^{lk} = attention(i, j), \tag{2}$$

where $h_{att_i}^{l+1}$ is the attention head of node-i at layer $l+1$ and α_{ij}^{lk} is the normalized attention coefficient computed by the k-th attention at layer l and W_k^l is an input transformation matrix.

In addition to the attention head of word-i a relational head is also computed for this node as

$$h_{rel_i}^{l+1} = concat_{m=1}^{M} \sum_{j \in \mathcal{N}_i} \beta_{ij}^{lm} W_m^l h_j^l, \tag{3}$$

$$g_{ij}^{lm} = \sigma(relu\,(r_{ij} W_{m1} + b_{m1})\, W_{m2} + b_{m2}) \tag{4}$$

$$\beta_{ij}^{lm} = exp(g_{ij}^{lm}) / \sum_{j \in \mathcal{N}_i} exp(g_{ij}^{lm}) \tag{5}$$

where r_{ij} denotes the relation embedding between node-i and j and M is the number of relational heads. The final representation of each word (node) is a concatenation of the attention and relational embeddings:

$$x_i^{l+1} = concat(h_{att_i}^{l+1}, h_{rel_i}^{l+1}) \tag{6}$$

$$h_i^{l+1} = relu\left(W_{l+1} x_i^{l+1} + b_{l+1}\right) \tag{7}$$

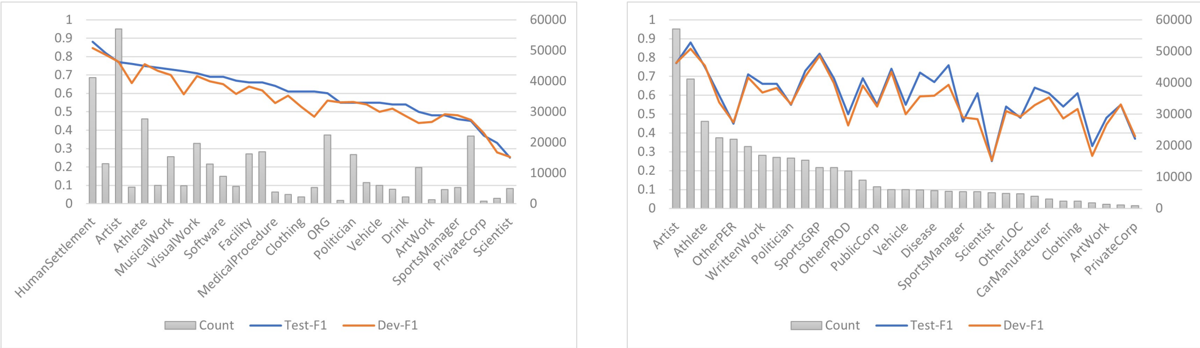

Figure 1: Fine-grained F1-score (primary axis) on the validation and test dataset along with the count of each tag from the test dataset (secondary axis). Left subfigure: sorted by F1-score and right: sorted by the count of each tag.

4.2 Named Entity Recognition

While the RGAT model utilizes only the root representation to predict the sentiment polarity, here, we use the node representation h_i^{l+1} for node-i to predict the corresponding NER tag. To further improve the model capability, we explore four different layers where h_i^{l+1} is provided as input, namely, (1) Linear layer, (2) RNN layer (we used Bi-LSTM), (3) Transformer layer and (4) CRF layer (Lafferty et al., 2001). For a linear layer, h_i^{l+1} is directly fed into a fully connected layer with softmax activation function that generates probabilities over all the NER classes, $\mathbf{C}$

$$p(y_i = \mathbf{C}) = softmax(W_p h_i^{l+1} + b_p). \tag{8}$$

where y_i is NER target label for the word-i and W_p, b_p are learnable parameters. In addition, we have explored having a Bi-LSTM and Transformer (Vaswani et al., 2017) layer along with a final CRF layer.

For all the three previous cases, we have a final layer that generates a score matrix $\mathbf{P}$ over all the tag classes, where P_{ij} is the score of the $j-$th tag for the $i-$th input token. The CRF layer learns a transition probability matrix $A \in \mathbb{R}^{K+2 \times K+2}$ where K is the number of tag classes (and +2 indicates one tag each for the start and the end marker).

For an input sequence $\mathbf{x}$ of length T and the corresponding tags $\mathbf{y} = \{y_1, y_2, \dots, y_T\}$ (with the start and end tag denoted by y_0 and y_{T+1}, the probability score is modified as

$$S(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{T} A_{y_i y_{i+1}} + \sum_{i=1}^{T} P_{i,y_i} \tag{9}$$

which is used to calculate the probability of the tag sequence as

$$p(\mathbf{y}|\mathbf{x}) = e^{S(\mathbf{x},\mathbf{y})} / \sum_{y' \in \mathbf{y}} e^{S(\mathbf{x},\mathbf{y}')} \tag{10}$$

which replaces Eq. 8. The loss for each word is calculated by the categorical cross-entropy loss

$$\mathcal{L}(\theta) = -\sum_{S \in \mathcal{D}} \sum_{w_i \in S} \log(p_{ij}) \tag{11}$$

where p_{ij} is the probability of the word-i with label-j where the word-i appeared in sentence-S and $\mathcal{D}$ represents the entire corpus. For CRF based models, the loss is calculated using the forward-backward algorithm.

4.3 Implementation Details

We use the bi-affine parser (Dozat and Manning, 2016) from AllenNLP for dependency parsing. While Wang et al. (2020) used the aspect words to orient the dependency tree, in our case, we cannot use that information. Instead, we randomly choose a noun word from the input sentence, if available, otherwise, we select the middle token about which the dependency tree is re-oriented. For all experiments, the embedding dimension for the dependency relation is set to 200 and the dropout is fixed at 0.3. The last hidden state of the pre-trained BERT[1] is used for the initial token representations which is subsequently fine-tuned. All models are trained using Adam optimizer (Kingma and Ba, 2014) with the default parameters. For all experiments we have used RGAT based feature extraction and BERT based token encoding. There are four variants of our model, namely, (1) RGAT-BERT, that does not use any other layer, (2) RGAT-BERT-CRF, that additionally uses a CRF final layer, (3)

[1] https://github.com/huggingface/transformers

Entity	Validation			Test		
	Precision	Recall	F1	Precision	Recall	F1
Person	0.8845	0.9331	0.9081	0.9163	0.9484	0.9321
Creative Works	0.7105	0.7308	0.7205	0.7598	0.7676	0.7637
Group	0.707	0.7409	0.7235	0.7872	0.7871	0.7872
Location	0.807	0.82	0.8134	0.8415	0.8665	0.8538
Product	0.5615	0.547	0.5541	0.5984	0.5816	0.5899
Medical	0.6292	0.6637	0.646	0.6585	0.6989	0.6781
Fine-grained (macro)	0.579	0.5728	**0.5691**	0.6250	0.6149	**0.6129**

Table 2: Model performance (coarse-grained) on the validation and test dataset across the entity classes. Also shown the fine-grained macro performance of the model on the validation and test dataset (used to compute the rank).

Model	CoNLL'03			WNUT'17		
	P	R	F1	P	R	F1
MIN + BERT	94.75	94.15	94.45	60.54	42.48	49.93
RGAT+Bi-LSTM+CRF (Ours)	92.68	90.62	91.64	63.6	34.5	44.7

Table 3: Model performance on the standard NER datasets

RGAT-BERT-BiLSTM-CRF, that uses a BiLSTM layer on the output of BERT before passing the output to a CRF layer and (4) RGAT-BERT-TRFMR-CRF that uses a Transformer layer instead of BiLSTM.

5 Results

In this section first we describe the performance on the test data of MultiCoNER-II followed by performance of our approach on other NER datasets. The best performance is obtained by a RGAT-BERT-BiLSTM-CRF model where the fine-grained macro average precision, recall and F1-score on the test data are 0.625, 0.615 and 0.613, respectively. The corresponding class-wise performance is shown in Fig. 1, where the top-3 classes are "HumanSettlement", "SportsGRP" and "Artist" and the bottom-3 classes are "PrivateCorp", "Symptom" and "Scientist". It is to be noted that "Artist" and "HumanSettlement" are amongst the top-3 classes in terms of the number of labels, whereas, "Symptom" and "PrivateCorp" are amongst the bottom-5 classes. Thus, it is tempting to conclude that the availability of data plays a significant role on the performance. However, as we can see on the right subfigure of Fig. 1, there is no definitive trend with respect to the number of available data points. We also show the coarse-grained performance on the validation and test dataset in Tab. 2. For the validation split, the macro average precision, recall and F1-score are 0.58, 0.57 and 0.57, respectively. We can see that overall the performance is better on the test dataset compared to the validation dataset.

We investigated if there is any correlation between the sentence length (ℓ) and model's performance and we can clearly see that the performance improves with increasing sentence length. In particular, we created 4 buckets with (1) $11 > \ell \geq 1$, (2) $14 > \ell \geq 11$, (3) $19 > \ell \geq 14$ and (4) $54 > \ell \geq 19$ where 1, 11, 14, 19 and 53 are the minimum, first quartile, second quartile, third quartile and the maximum sentence length, respectively. The corresponding macro average F1 scores for the four buckets are 0.57, 0.62, 0.66 and 0.68, respectively.

We have also compared the performance of our model on other standard NER datasets, namely, CoNLL 2003 (Sang and Meulder, 2003) and WNUT'17 (Derczynski et al., 2017). The results are shown in Tab. 3 where we have compared our model's performance with MIN (Li et al., 2021). It can be seen that our approach is quite competitive compared to MIN and MultiCoNER-II dataset is very different in terms of both complexity and context length that resulted in a dramatically inferior performance.

6 Conclusion

In this work we have explored the capability of Graph Attention Network based model for NER task on MultiCoNER-II dataset. In this approach we create features from the graph structure obtained by a dependency parser that are combined with other standard features like word embeddings and POS tags. However, we have not included any external knowledge base (typically WikiMedia data) and as a result the model's performance is clean-subset F1 of 61.29% and overall F1 of 56.91%. The exercise clearly brings out the importance of having external knowledge base and the next step would be to explore ways to optimally integrate external data with the rest of the architecture.

References

Oshin Agarwal, Yinfei Yang, Byron C Wallace, and Ani Nenkova. 2021. Entity-switched datasets: An approach to auditing the in-domain robustness of named entity recognition models.

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, Copenhagen, Denmark. Association for Computational Linguistics.

Gokberk Celikmasat, Muhammed Enes AKTURK, Yunus Emre ERTUNC, Abdul Majeed ISSIFU, and Murat Can GANIZ. 2022. Biomedical named entity recognition using transformers with bilstm + crf and graph convolutional neural networks.

Alberto Cetoli, Stefano Bragaglia, Andrew O'Harney, and Marc Sloan. 2017. Graph convolutional networks for named entity recognition. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 37–45, Prague, Czech Republic.

J. P. Chiu and E. Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. page 357–370.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch,. 12:2493–2537.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023a. MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer Enhanced Named Entity Recognition for Code-Mixed Web Queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023b. SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history.

Tran Thi Hong Hanh, Antoine Doucet, Nicolas Sidere, José G. Moreno, and Senja Pollak. 2021. Named entity recognition architecture combining contextual and global features. *CoRR*, abs/2112.08033.

Binxuan Huang and Kathleen Carley. 2019. Syntax-aware aspect level sentiment classification with graph attention networks. 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 5469–5477, Hong Kong, China. Association for Computational Linguistics.

Z. Huang, W. Xu, , and K. Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence.

G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. 2016. Neural architectures for named entity recognition.

Fei Li, Zheng Wang, Siu Cheung Hui, Lejian Liao, Dandan Song, Jing Xu, Guoxiu He, and Meihuizi Jia. 2021. Modularized interaction network for named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 200–209, Online. Association for Computational Linguistics.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A Survey on Deep Learning for Named Entity Recognition.

Ying Luo and Hai Zhao. 2020. Bipartite flat-graph network for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6408–6418, Online. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, volume 4, pages 142–147. Association for Computational Linguistics.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Kyle Shaffer. 2021. Language clustering for multilingual named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 40–45, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2680, Copenhagen, Denmark. Association for Computational Linguistics.

Yuan Sui, Fanyang Bu, Yingting Hu, Liang Zhang, and Wei Yan. 2022. Trigger-GNN: A trigger-based graph neural network for nested named entity recognition.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019. Aspect-level sentiment analysis via convolution over dependency tree. 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 5679–5688, Hong Kong, China. Association for Computational Linguistics.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. 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 4568–4578, Hong Kong, China. Association for Computational Linguistics.

XianKun Zhang and SiYuan Peng. 2022. Bg-efrl: Chinese named entity recognition method and application based on enhanced feature representation. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, page 234–245, Berlin, Heidelberg. Springer-Verlag.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236, Vancouver, Canada. Association for Computational Linguistics.

P. Zhou, S. Zheng, J. Xu, Z. Qi, H. Bao, and B. Xu. 2017. Joint extraction of multiple relations and entities by using a hybrid neural network.

Renjie Zhou, Zhongyi Xie, Jian Wan, Jilin Zhang, Yong Liao, and Qiang Liu. 2022. Attention and edge-label guided graph convolutional networks for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6499–6510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.