Samsung Research China - Beijing at SemEval-2023 Task 2: An AL-R Model for Multilingual Complex Named Entity Recognition

Haojie Zhang✉, **Xiao Li, Renhua Gu**
Xiaoyan Qu, Xiangfeng Meng, Shuo Hu, Song Liu
Samsung Research China-Beijing (SRC-B)
{tayee.chang,xiao013.li,renhua.gu,s0101.liu}@samsung.com

Abstract

This paper describes our system for SemEval-2023 Task 2 Multilingual Complex Named Entity Recognition (MultiCoNER II). Our team **Samsung Research China - Beijing** proposes an AL-R (Adjustable Loss RoBERTa) model to boost the performance of recognizing short and complex entities with the challenges of long-tail data distribution, out of knowledge base and noise scenarios. We first employ an adjustable dice loss optimization objective to overcome the issue of long-tail data distribution, which is also proved to be noise-robusted, especially in combatting the issue of fine-grained label confusing. Besides, we develop our own knowledge enhancement tool to provide related contexts for the short context setting and address the issue of out of knowledge base. Experiments have verified the validation of our approaches. In the official test result, our system ranked **2nd** on the English track in this task.

1 Introduction

Named Entity recognition is an important task in natural language processing. In low-context data(Malmasi et al., 2022a), semantic ambiguity, complex entities, multilingual and code-mixed settings make this task even more difficult. Recognizing complex entities in low-context situations was recently outlined by Meng et al. (2021). Multilingual and code-mixed settings were extended through Fetahu et al. (2021).

To address these issues, SemEval-2022 task 11 Multilingual Complex Named Entity Recognition(Malmasi et al., 2022b) received 34 system papers, the best of which were able to accurately recognise complex entities by incorporating external knowledge related to the data(Verlinden et al., 2021; Wang et al., 2022). However, when the entity recognition scenario is more complex, the top system (Wang et al., 2022) has drawbacks that make it difficult to cope.

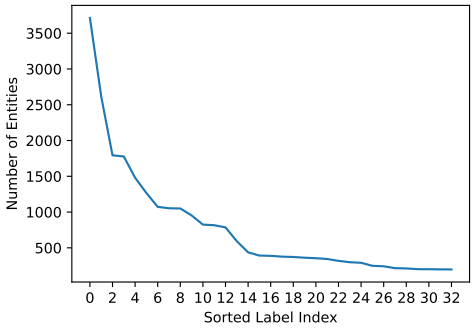

Figure 1: The long-tailed distribution of entity label for English dataset.

SemEval-2023 task 2 MultiCoNER II Multilingual Complex Named Entity Recognition (Fetahu et al., 2023b) aims to build a system to address the following problems: 1). The entities are complex and long-tailed illustrated by Figure 1. 2). The entities are out of knowledge-base entities. 3). The noisy scenarios like the presence of spelling mistakes and typos.

In this paper, we build an AL-R (Adjustable Loss RoBERTa) based system for MultiCoNER II task. Our system utilizes the RoBERTa-large (Liu et al., 2019) as the pre-trained model and fine tunes it on the English dataset (Fetahu et al., 2023a). For the long-tailed dataset and noise scene, we add the adjustable dice loss (Li et al., 2019b) to overcome the issue of long tail and help the system more robust. Because the dataset is low-context, we build a Knowledge Retrieval Module (Wang et al., 2022) via Wikipedia dump[1], from which we can perform sentence knowledge retrieval and entity knowledge retrieval for the dataset. In order to make our system more generalized, we integrate multiple AL-R models and let them jointly determine the entity label.

Extensive experiments and ablation studies on the English track show that our system is much

[1]https://dumps.wikimedia.org/

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 114–120

more superior to the baseline system, and the overall macro-F1 in the official test set is 0.8309[2].

2 Related Work

Named Entity Recognition (NER) is a basic natural language understanding task in information extraction (IE) (Grishman, 2015), which can be extensively used in many fields like information retrieval (Singhal et al., 2001), question answering (Antol et al., 2015), dialogues (Baxter, 2004) and so on. Pre-training language model (PLM) in recent years has achieved remarkable advance in NER tasks. However, it still faces some difficulty in the situation of short context and complex entities (Malmasi et al., 2022a,b), which has been formulated by SemEval-2022 Task 11 (Malmasi et al., 2022b). To address this issue, one classic approach is to use span-based or entity-aware PLM to gain enhanced token representations, like Ernie (Sun et al., 2019), SpanBERT (Joshi et al., 2020) and LUKE (Yamada et al., 2020).

Another classic approach is grouped by the different decoding methods, including pointer based (Bekoulis et al., 2018; Li et al., 2019a,c), token pairs based (Bekoulis et al., 2018; Yu et al., 2020) and span based (Eberts and Ulges, 2019). External knowledge (Verlinden et al., 2021; Wang et al., 2021, 2022) is also an effective approach to achieve top performance in the field of short text and complex entities. But when facing the situation of out of knowledge base, noisy scenarios, and long-tail data distribution, only by knowledge is vulnerable to address the new issue. In this paper, we still adopt external knowledge to build the backbone of our system, but to address the out of knowledge base problem, we build our own knowledge-enhancement tool. Li et al. (2019b) proposed self-adjusting dice loss for data-imbalanced NLP tasks and achieved significant performance boost. We inherit this idea in our system to overcome the issue of long-tail distribution in MultiCoNER II datasets.

3 Data

We train and test our models on the English track via `multiconer2-data` (Fetahu et al., 2023a). English dataset[3] of `multiconer2-data` contains a training set of size 16778, a dev set size of 800, and a test set of size 249980. It consists of 6 coarse-grained labels and 33 fine-grained labels.

[2]https://multiconer.github.io/results

[3]https://codalab.lisn.upsaclay.fr/competitions/10025

4 Methodology

We introduce a Knowledge-enhancement and Adjustable-loss Named Entity Recognition system to overcome the main challenges[4] paid much attention in this big event. First, to alleviate the insufficiency of the context of the input texts, which is critical for distinguishing and recognizing entities especially in the field of short texts, we feed the inputs into our Wikipedia-based Tool to get the knowledge-enhancement context. Second, to combat the performance degradation caused by the long-tail data distribution, we select and deliberately design the dice loss, here called adjustable-loss, to be the optimization object. Meanwhile, we introduce the label-smoothing technology to decrease the overfitting of some grained labels belonging to the same coarse labels. The architecture of our scheme is shown in Figure 2.

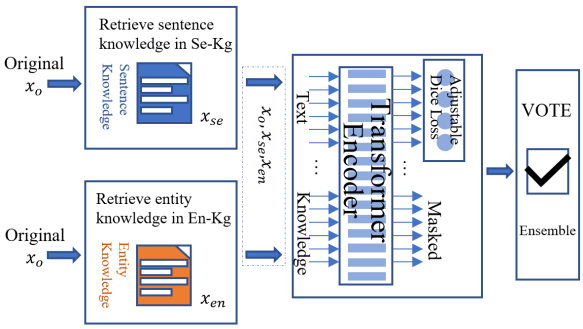

Figure 2: The illustration of our NER scheme.

4.1 Knowledge-enhancement Tool

Knowledge augmentation has demonstrated its effectiveness for the coarse-grained entity recognition task of short texts(Wang et al., 2022). For the fine-grained entity recognition task, we downloaded the latest version of Wikipedia dump, and built two separate knowledge retrieval modules using ElasticSearch[5]: a sentence retrieval module and an entity retrieval module. The illustration of our knowledge-enhancement tool is shown in Figure 3

Sentence Retrieval Knowledge Base (Se-Kg) The sentence retrieval knowledge base consists of two fields: sentence field, paragraph field. We create an index for each sentence in the Wikipedia dump as sentence field, the paragraph in which the sentence is stored as paragraph field. The wiki

[4]Please refer to https://multiconer.github.io/ for more details.

[5]https://www.elastic.co/cn/

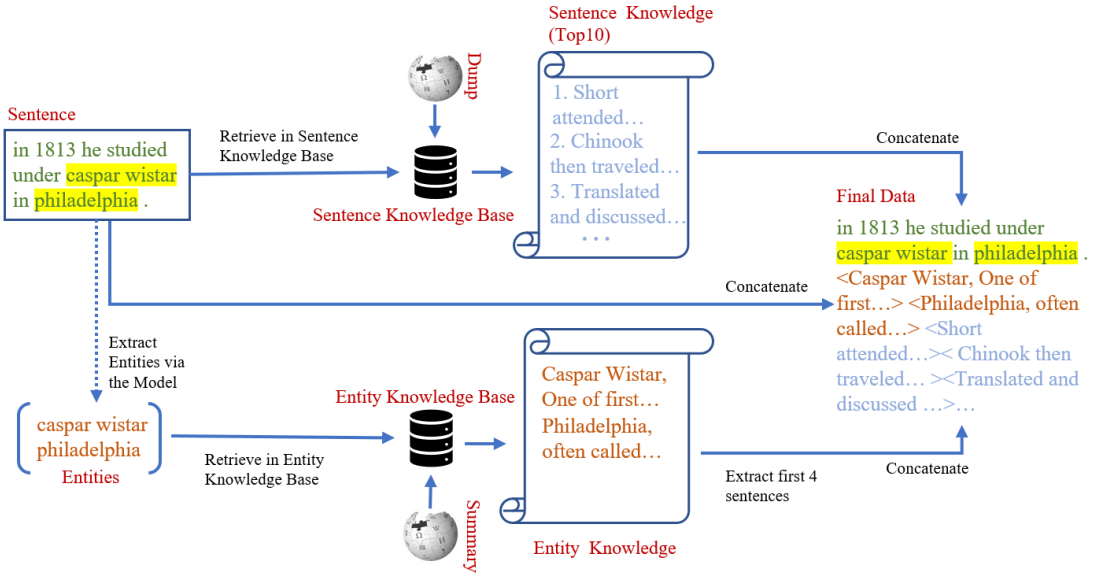

Figure 3: The illustration of our Knowledge-enhancement Tool.

anchors are marked to take advantage of the rich anchor information of Wikipedia. Indexes related to texts of the dataset can be retrieved in sentence field, and the content of the paragraph field will be enhanced as sentence knowledge.

Entity Retrieval Knowledge Base (En-Kg) The entity retrieval knowledge base has two fields: title field, paragraph field. We use the title of the page of the Wikipedia dump as title field, and the summary of the page is stored as paragraph field. The entities of the dataset are matched in title field, and the paragraph of the index which is matched will be the entity knowledge enhancement. If the matching fails, no entity knowledge augmentation is performed.

4.2 Basic NER Module

We mainly employ RoBERTa-large (Liu et al., 2019) as the encoder of the inputs. RoBERTa is an advanced variant of the bidirectional encoding representations transformer family, which has been usually employed as the feature extractor in many natural language understanding tasks in recent years (Liu et al., 2019; Vaswani et al., 2017).

Given an input sequence as $\{x_1, x_2, ..., x_n\}$, where x_i denotes i-th token of the input and n denotes the sequence length of the input, after feeding them into a RoBERTa-large model, we can obtain the representation vectors. Obviously, these representation vectors fuse the semantic information of the context of the input, according to the basic principle of the transformer-based encoder (Vaswani et al., 2017). On the top of our model, it's an adjustable loss layer which is deliberately designed to optimize the training objective. Instead of calculating the loss of all tokens of the input sequence, we use a masking to only pay attention on the loss of the real input text tokens not including those of the context.

4.3 Adjustable-loss

We propose an adjustable-loss RoBERTa model, which leverages RoBERTa as the backbone to encode the text input. In the field of sequence tagging, it is quite often to employ the classical conditional random field (CRF) layer to capture the dependency of output labels, which is proved to be effective to boost the performance (Lafferty et al., 2001; Wang et al., 2022). But in our system, instead of adopting a conditional random field loss as the optimization object, here we employ adjustable dice loss to combat the long-tail input data distribution. To fast the convergence of training process, we use a variant of dice loss.

Dice Loss It derivates from dice coefficient (DSC)[6] which generally is used to evaluate the similarity of two sets. As dice loss is a F1-score oriented loss, which is consistent with the final evaluation metric– overall macro F1 score[7]. So it is naturally to be a candidate of optimization objects.

[6] https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient

[7] https://en.wikipedia.org/wiki/F-score

It can be formulated as follows (Li et al., 2019b):

$$DL(x_i) = 1 - 2\frac{\mathbb{I}(p_{i1} > threshold) \cdot y_{i1}}{\mathbb{I}(p_{i1} > threshold) + y_{i1}} \quad (1)$$

When datasets consist of severely imbalanced categorical examples especially in the case of long-tail distribution, the optimization object is very easy to be dominated by the much more negative examples. As a result, the performance will be degraded severely. To overcome this problem, we use a variant of dice loss, which employs the soft probability instead of a trial indicator function. In addition, to address the distinguishing difficulty between the hard example and the easy example, we employ a focal-loss-like dice loss, which complies with the format of the famous focal loss (Lin et al., 2017). So a variant of dice loss is as follows (Li et al., 2019b):

$$DL(x_i) = 1 - \frac{2(1-p_{i1})^{\alpha} p_{i1} \cdot y_{i1} + \tau}{(1-p_{i1})^{\alpha} p_{i1} + y_{i1} + \tau} \quad (2)$$

where α is a hyper parameter, and τ is a smoothing factor, generally set as 1. The factor $(1-p_{i1})^{\alpha}$ is a modulation factor which will increase the weight coefficient of the hard example, which means the model will deliver more attention to them, and vice versa.

Considering that dice loss landscape is steep, it's very possible to step into the sharp minimum, as a result of being a little hard and slow to converge. We add the original cross-entropy loss to our dice loss, which will boost the convergence. So, the final loss will be as follows:

$$DL(x_i) = \beta \cdot L_{CE} + \gamma \cdot [1 - \frac{2(1-p_{i1})^{\alpha} p_{i1} \cdot y_{i1} + \tau}{(1-p_{i1})^{\alpha} p_{i1} + y_{i1} + \tau}] \quad (3)$$

where β and γ are hyper parameters and are both set as 1 in our system. As there are some hyper parameters in the loss, we also call it adjustable dice loss.

Label Smoothing In our experiments, we discover that some grained tail labels belonging to the same tail coarse labels such as PER are easily confusing by identification during inference process. An intuition is that the model has a strong confidence of the inference capability, hence causing overfitting. To overcome this problem, we introduce a label smoothing (LS) scheme (Müller et al., 2019). More specifically, we add some uniform distribution noise into the original one-hot distribution but keep it a real distribution.

Table 1: F1 score of different entity knowledge in the OriES-data-dev inferenced on the model AL-R-ES

Entity Knowledge	dev F1
3 sentence	0.8704
4 sentence	**0.8764**
5 sentence	0.8763
6 sentence	0.8737

4.4 Ensemble Method

Ensemble models (Dietterich, 2000) are used to boost the final prediction performance in our English track. We train a couple of models by setting different random seeds. Meanwhile, to keep a heterogeneity among different models, we introduce some Span-BERT-like[8] models such as ERNIE sourced from Baidu[9]. As ERNIE is pre-trained by some span masking tasks like phrase masking and entity masking (Sun et al., 2019), it will boost the entity extraction as one component of the ensemble models. In the end, we make a decision on the entities by majority voting.

5 Experiment

5.1 Data Prepare

Sentence Knowledge Enhancement The text contents of the original dataset (Ori-data) are sentence knowledge enhanced by Se-Kg to obtain the OriS-data.

Entity Knowledge Enhancement The already labeled entities of Ori-data are extracted, and matched with title field in En-Kg. The matched entity knowledge is supplemented into the OriS-data to produce the dataset OriES-data.

Entity Knowledge Truncation Experiment Considering that a sample may have multiple entities, the first 3 sentences, 4 sentences, 5 sentences and 6 sentences of entity knowledge in the OriES-data set are extracted as the knowledge supplement of the OriS-data respectively. We inference above different dataset on the model trained via OriES-data, and the results are shown in the table 1. Based on the results we select the first 4 sentences of entity knowledge (E4F) as the best entity knowledge supplement of OriS-data to acquire the dataset OriE4S-data.

[8] https://github.com/facebookresearch/SpanBERT
[9] https://github.com/PaddlePaddle/ERNIE

Table 2: Results of baseline model and ours models

	Model	Data set	Pre-Trained Model	dev F1	test F1
Baseline	Official Model	Ori-data	RoBERTa-base	0.6458	-
	Official Model-S	OriS-data	RoBERTa-base	0.7724	-
Ours	AL-R-S(wo/LS)	OriS-data	RoBERTa-base	0.7907	-
	AL-R-ES(wo/LS)	OriES-data	RoBERTa-base	0.8176	-
	AL-R-ES(wo/LS)	OriES-data	RoBERTa-large	0.8458	-
	AL-R-ES	OriES-data	RoBERTa-large	0.8650	0.8073
	AL-R-E4F	OriE4S-data	RoBERTa-large	0.8752	0.8101
	AL-R-E4F-Ensemble	OriE4S-data	RoBERTa-large	**0.8914**	**0.8309**

Test Set Augementation Since the entity span of the original test set (Ori-test) is not labeled, Ori-test is firstly subjected to sentence knowledge retrieval by Se-Kg to obtain the test set OriS-test, and the entity span is annotated through the model which is trained by OriS-data. Second, after matching the marked OriS-test entities in En-Kg, we extract the first 4 sentences of the matched entity knowledge to supplement OriS-data, and finally obtain the test set OriE4S-test.

5.2 Experiment Setting

We use RoBERTa-large model as the contextual embedder for our system. We use AdamW (Loshchilov and Hutter, 2017) optimizer to train our models with learning rate $\in$ {4e-5,2e-5,1e-5} and batch size $\in$ {64,32,16}. We use a linear warmup process of the first epoch before a linear learning rate scheduler of the rest epochs. The max sequence length is set as 512 and max epoch is set as 100.

5.3 Models Training

Baseline The official model[10] is selected as our baseline model but instead of using the available baseline model directly, we implement official baseline model by ourselves. The official model is composed of a pre-trained model and a CRF layer, in which we choose RoBERTa as the pre-trained model. We train the baseline model via Ori-data and OriS-data.

Single Model We train the single model AL-R via OriS-data, OriES-data and OriE4S-data, the pre-trained model of AL-R is RoBERTa. To demonstrate the effectiveness of LS, models AL-R without LS are also used for ablation study. Meanwhile, in order to compare fairly with the baseline we choose RoBERTa-base as the pre-trained model of AL-R without LS. Note that, in our models, the last three models all use LS.

[10] https://multiconer.github.io/baseline

6 Results

6.1 Main Result

In the table 2, AL-R-ES, AL-R-S, AL-R-ES(wo/LS) and AL-R-E4S are single models, and the AL-R-E4F-Ensemble is the multi-model ensemble model.

We observe that the dev F1 of Official Model trained without external knowledge is 12.6% lower than that of the Official Model-S trained with external knowledge. It demonstrates that external knowledge is particularly important for low-resource named entity recognition task.

The AL-R-S(wo/LS) and Official Model-S have the same pre-trained model and are both fine tuned via same dataset, but the dev F1 of AL-R-S(wo/LS) is $+1.83\%$ higher than that of Official Model-S, which indicates Dice-loss is more suitable for fine-grained entity recognition tasks than CRF loss, especially facing the long-tail data distribution.

AL-R-E4F-Ensemble achieves the best performance with macro-F1 0.8309 among all our models. It shows that model ensemble helps the system more robust, enabling it to handle complex entity recognition scenarios.

6.2 Ablation Study

Effect of Entity Knowledge After supplementing entity knowledge, the dev F1 value of AL-R-ES(wo/LS) is $+2.7\%$ higher than that of AL-R-S(wo/LS), which reveals that entity knowledge fits better with text than sentence knowledge and helps the model identify fine-grained entities.

Effect of Entity Knowledge Truncation The test F1 value of AL-R-E4S is the highest among all

single models, $+0.28\%$ higher than AL-R-ES. It means that the first 4 sentences of entity knowledge contain less noise and help the model to make better use of entity information.

Effect of Adjustable Dice Loss and LS Compared to Official Model-S, our primary AL-R-S(wo/LS) using only adjustable dice loss not including label smoothing, achieves a gain of $+1.8\%$ in F1. AL-R-ES which includes the label smoothing technology, achieves $+1.92\%$ in F1 in the development set compared to AL-R-ES(wo/LS). They prove the validation of our adjustable dice loss and label smoothing scheme.

7 Conclusion

In this paper, we build a system for complex fine-grained NER task. Our system assembles multiple AL-R models containing RoBERTa based pre-trained model and adjustable dice loss to overcome the issues of long-tailed dataset and noise entities. We construct the Knowledge Retrieval Module from which sentence knowledge augmentation and entity knowledge augmentation can be performed on low-context data. In MultiCoNER II task, our system ranks 2nd on English track.

Although our model is somewhat robust to noisy scenarios, in the future we will investigate better ways to deal with noisy entities.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Leslie A Baxter. 2004. Relationships as dialogues. *Personal relationships*, 11(1):1–22.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*, pages 1–15. Springer.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. *arXiv preprint arXiv:1909.07755*.

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. 2015. Information extraction. *IEEE Intelligent Systems*, 30(5):8–15.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019a. A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019b. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019c. Entity-relation extraction as multi-turn question answering. *arXiv preprint arXiv:1905.05529*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: a Large-scale Multilingual dataset for Complex Named Entity Recognition.

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)*. 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.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Amit Singhal et al. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Severine Verlinden, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2021. Injecting knowledge base information into end-to-end joint entity and relation extraction and coreference resolution. *arXiv preprint arXiv:2107.02286*.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. *arXiv preprint arXiv:2105.03654*.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, et al. 2022. Damo-nlp at semeval-2022 task 11: A knowledge-based system for multilingual named entity recognition. *arXiv preprint arXiv:2203.00545*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. *arXiv preprint arXiv:2005.07150*.