AIST3120 Introduction to Natural Language Processing
Final Project Report
24 April 2025

Enhancing DeBERTa for NER with CRF Layer and Data Augmentation

CHU, Yan Yee
1155192092

FUNG, Choi Wing
1155193277

MA, Hoi Tung
1155192994

Abstract

In this project, we explore the effects of incorporating a Conditional Random Field (CRF) layer and data augmentation into the DeBERTa model for the Named Entity Recognition (NER) task. Although DeBERTa achieves a strong baseline F1 score of 0.9158, it does not account for label transition dependencies, which can enhance sequence modeling. To address this, we integrate a CRF layer and apply data augmentation techniques—specifically synonym replacement and token masking—to increase training data diversity and improve robustness. Using the CoNLL-2003 dataset, we evaluate four configurations: the baseline model, DeBERTa with CRF, DeBERTa with data augmentation, and a combined model. While CRF and augmentation individually provide marginal or no improvements, the combined model achieves the highest F1 score of 0.9191. These results suggest that the two methods complement each other, enhancing both generalization and sequence consistency in NER.

I. Introduction

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) as it is applied in many applications namely information retrieval, question answering, and summarization. The NER systems detect entities such as persons, organizations, and locations within text. Recent advances in this field are achieved by transformer-based deep learning models (Li et al., 2022).

This project investigates how well these advanced models perform for NER by building on a strong baseline: DeBERTa, a state-of-the-art transformer known for its architectural improvements. To improve performance further, we experiment with two key enhancements—integrating a CRF layer for better sequence tagging and applying data augmentation to increase the model's adaptability.

The CoNLL-2003 English dataset, which is a widely used benchmark for NER tasks, is used in this project for model evaluation. We examined four approaches involving the baseline DeBERTa model, DeBERTa with a CRF layer and data augmentation respectively, and a combined model with both enhancements. It is observed that the combined approach outperformed the others by achieving an F1 score of 0.9191 on the CoNLL-2003 dataset. This result highlights the strengths of combing CRF and data augmentation.

In the following sections, we survey related work in traditional and deep learning-based NER, describe our proposed methodology in detail, and present our experimental findings and analysis.

II. Related work

The following part will mention the traditional approaches briefly and give a survey on the deep-learning based approaches for NER. Then, we will explain why we selected DeBERTa as our baseline model and how we considered CRF and data augmentation as our enhancement methods.

1. Traditional techniques for NER

Over several decades, there have been many different approaches to NER with continuous improvement to surpass the traditional benchmarks. According to Li et al. (2022), the conventional approaches can be classified into three main categories including rule-based, unsupervised learning, and feature-based supervised learning systems. The rule-based approach requires handcrafted semantic and syntactic rules, which need human resources in domain expertise, and the system will become difficult to generalize to other domains (Jehangir et al., 2023). The supervised learning approach requires labeled data samples and feature engineering for NER systems (Li et al., 2022). In the early stage of NER, CRF-based NER proposed by McCallum and Li in the supervised learning approach was widely used as it achieved 84.04% F-score for English datasets in CoNLL-2003 (Li et al., 2022). However, the state-of-the-art achievements in NER are now driven and dominated by deep-learning based approaches in recent years.

2. Deep Learning (DL) for NER

DL-based models benefit from continuous real-valued vector representations and semantic composition in non-linear transformation to automatically learn meaningful features from raw data (Li et al., 2022). Therefore, the deep learning approaches actively redefine the benchmarks for NER and give a brilliant performance.

From surveys by Li et al. (2022) and Jehangir et al. (2023), it is found that the taxonomy of DL-based NER is defined in a similar way to systematically identify the structure of NER systems. From Figure I, the model can be separated into three main parts: distributed representations for input, context encoder and tag decoder.

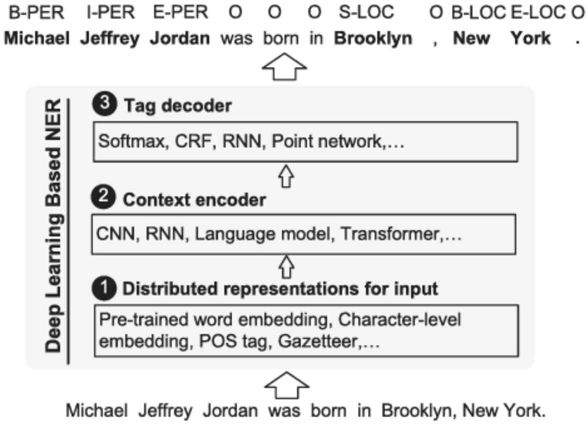
B-PER I-PER E-PER O O O S-LOC O B-LOC E-LOC O
Michael Jeffrey Jordan was born in Brooklyn , New York .
Deep Learning Based NER
3 Tag decoder
Softmax, CRF, RNN, Point network,...
2 Context encoder
CNN, RNN, Language model, Transformer,...
1 Distributed representations for input
Pre-trained word embedding, Character-level embedding, POS tag, Gazetteer,...
Michael Jeffrey Jordan was born in Brooklyn, New York.

Figure I: DL-based NER (Li et al., 2022)

Regarding the distributed representations for input, initial approaches used pre-trained word embeddings like Word2Vec and GloVe, while it is supplemented with character-level CNNs and RNNs such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) to increase the F-score of the NER systems (Li et al., 2022). Nevertheless, with recent work having hybrid representations of the input by pre-trained contextualized language-model embeddings such as BERT and ELMo, the performance of NER models is further improved (Li et al., 2022).

Following the representation of input, the context dependencies in input sentences are captured by the context encoder. Before the transformer is proposed by Vaswani et al. (2017), convolutional neural networks, recurrent neural networks, and recursive neural networks are used. Then, the deep transformer embeddings with attention mechanisms including BERT and Generative Pre-trained Transformer (GPT), become a new paradigm of NER to replace the traditional ones (Li et al., 2022).

By evaluating the performance, a Bidirectional Transformer model in a cloze-style objective, accomplished 93.5% F-score on CoNLL-2003 (Baevski et al., 2019), a model with BERT, softmax and dice loss achieved 93.33% F-score on CoNLL-2003 (Li et al., 2019a) while a neural NER with BERT and global embeddings also have 93.47% of the score (Liu et al., 2019a). According to the research, it is found that input representation heavily determines the success of a NER model (Li et al., 2022). Therefore, these pre-trained contextualized embeddings with transformers perform well in a NER system. Although LSTM and BiLSTM are the common architecture for DL-based NER, it is found that the model with a transformer pre-trained on huge corpora will be a more effective encoder (Li et al., 2022). Additionally, these language model embeddings with deep transformers can be effectively further fine-tuned for a wide range of tasks including NER by adding an output layer (Li et al., 2022). The neural NER by Li et al. (2019b) having a simple architecture with the BERT model and softmax layer attained an F-score of 93.04% on CoNLL-2003.

Therefore, as the performance of NER systems with BERT achieves top-tier scores in performance benchmarks and can be tuned to perform the NER task, BERT and its variants are being considered as the baseline model of this project.

3. BERT and its variants

By considering the project requirement and the model performance, DeBERTa is chosen as our baseline model. The original BERT model is proposed by Devlin et al. (2019) with its variants such as RoBERTa (Liu et al., 2019b), DistilBERT (Sanh et al., 2019), and DeBERTa (He et al., 2020), to name but a few. Considering DeBERTa is one of the BERT's variants published 5 years ago, it satisfies the project requirement. In addition, according to the F1 score on the CoNLL-2003 benchmark, BERT (large), RoBERTa (large) and DeBERTa (large) achieved 92.8%, 93.4% and 93.8% respectively (He et al., 2020). It is shown that DeBERTa performs better on NER tasks owing to its improved contextual understanding and pre-training efficiency (He et al., 2020). Hence, it will be encouraging if the performance of DeBERTa can be further enhanced.

4. Conditional Random Fields (CRF) for Sequence Labeling

According to the summary of recent work in neural NER, the CRF layer is the most common choice for tag decoder, which is added during fine-tuning for specific tasks like NER to improve sequence labeling performance (Li et al., 2022). A study by Rosvall (2019) experimented on the effect of a CRF layer on the BERT model. The model of a BERT model with Feed Forward layer and Softmax is compared with a model replacing the Feed Forward layer with a CRF layer (Rosvall, 2019). The result showed that BERT with CRF consistently outperforms the original model by 0.25% (Rosvall, 2019). Even though BERT with deep transformers can already effectively capture both forward and backward dependencies in input sentences (He et al., 2020), adding a CRF layer in the tag decoder may still improve performance by modeling the dependencies between output labels. For example, CRF can prevent a ”B-PER” label followed by an ”I-ORG” label which will be invalid. In light of this, we consider applying a CRF layer to the DeBERTa model as an approach to enhance its performance on CoNLL-2003.

5. Data Augmentation for NER

Data augmentation is a technique used to increase the diversity of training data without collecting new data. Research by Kyaw (2022) used text augmentation including Finite State Transducer and Abstractive Text Summarization techniques to fine-tune BERT variant models. The F1 score of the BERT and the DistilBERT NER model on the Groningen Meaning Bank corpus improved by 0.3% and 0.7% respectively, but the score of the RoBERTa-based model decreased by 0.2% (Kyaw, 2022). As the methods in this research are complex to implement and require additional tools, data augmentation is considered but with different methods for DeBERTa in our approach.

III. Methods

The baseline of our project is the DeBERTA (Decoding-enhanced BERT with Disentangled Attention) model, which is a decoding-enhanced version of the BERT (Bidirectional Encoder Representations from Transformers) model. We conducted data augmentation using synonyms and added a Conditional Random Field (CRF) layer on top of the baseline.

1. BERT Model

The original BERT model is a multi-layer bidirectional Transformer encoder (Devlin et al., 2019) which adopts the original implementation in Vaswani et al. (2017) It improves the fine-tuning approach using "masked language model" (MLM), which allows the model to conduct bidirectional conditioning without allowing each word to "see itself". (Devlin et al., 2019) MLM chooses 15% of the tokens at random and replaces each of them with a [MASK] token 80% of the time, a random token 10% of the time and an unchanged token 10% of the time. The model then predicts the likely original token with cross entropy loss. (Devlin et al., 2019) To comprehend the relationship between sentences, the BERT model performed Next Sentence Prediction where the model chooses between the actual next sentence and a random sentence based on the previous sentence. (Devlin et al., 2019)

2. DeBERTa Baseline

DeBERTa differs from the BERT model by incorporating disentangled attention, enhancing the mask decoder and fine-tuning with Scale Invariant Fine-Tuning. (He et al., 2020)

2.1 Disentangled Attention

The standard self-attention mechanism in Transformers encodes either relative or absolute position information by summing position bias and embeddings. (Vaswani et al., 2017) This method only considers the content-to-content and content-to-position information of the token, thus disentangled attention is used in DeBERTa to involve the position-to-content information. (He et al., 2020) In the attention score formula, two vectors $H_i$and $P_{i,j}$ are used to represent the content and relative position embeddings respectively. The former is used with projection matrices to generate Q_c, K_c, and V_c, and the latter is projected to generate Q_r and K_r. Finally, the attention score is calculated by summing the content-to-content term $Q_i^c {K_j^c}^\top$, content-to-position term $Q_i^c {K_{\delta(i,j)}^r}^\top$ and the position-to-content term $K_j^c {Q_{\delta(j,i)}^r}^\top$. (He et al., 2020)

$$
\begin{gathered}
Q_c = HW_{q,c}, \quad K_c = HW_{k,c}, \quad V_c = HW_{v,c}, \\
Q_r = PW_{q,r}, \quad K_r = PW_{k,r} \\
\tilde{A}_{i,j} = Q_i^c {K_j^c}^\top + Q_i^c {K_{\delta(i,j)}^r}^\top + K_j^c {Q_{\delta(j,i)}^r}^\top \qquad (1) \\
H_o = \text{softmax}\left(\frac{\tilde{A}}{\sqrt{3d}}\right) V_c
\end{gathered}
$$

2.2 Enhanced Mask Decoder

Unlike the BERT model, which includes absolute positions in the input layer, DeBERTa incorporates them after all Transformer layers and before the softmax layer. This allows DeBERTa to focus on the relative positions in the Transformer layers and only consider the absolute positions as complementary information at the decoder. (He et al., 2020) This method makes sure that the absolute position incorporation will not hinder the model from learning from the relative positions thoroughly. (He et al., 2020)

2.3 Scale Invariant Fine-Tuning (SiFT)

Perturbation is originally applied to the word embedding to improve models' generalization as a virtual adversarial training. Since this increases the variance for bigger models thus causes instability, the new SiFT algorithm applies the perturbations to normalized word embeddings instead. (He et al., 2020) This is done by first normalizing the word embedding vectors into stochastic vectors, then applying perturbations to normalized embedding vectors. (He et al., 2020)

3. Data Augmentation

To make the DeBERTa model more robust and prevent overfitting, we propose using data augmentation to expand the training dataset. Our methods include synonym replacement and token masking. Since the augmented examples are slightly different from the existing samples, the decision boundaries will be strengthened in classification problems such as NER. (Shorten et al.,

2021) Although the word replacement avoids entity tokens, the augmentation diversifies dataset and improves model generalization.

3.1 Synonym Replacement

We adopted the synonym replacement method proposed by Wei et al. (2019). The synonyms are imported from the WordNet interface by the Natural Language Toolkit (Dai et al., 2020) We assigned 15% of the training data to synonym replacement, where up to two random tokens in each sample will be replaced with synonyms randomly chosen from wordnet.synsets. The percentage of words replaced is small, since replacing too many words may change the identity of a sentence(Wei et al., 2019), but enough to make a difference in the model.

3.2 Token Masking

The token masking method is inspired by the Random Deletion method by Wei et al. (2019) where we replaced the selected word with a [MASK] token instead of directly removing it. The model will be able to identify the position of the masked token, while the sentence can retain its original structure. This method makes the model robust to noise without the risk of rendering the sentence unintelligible as mentioned in the Random Deletion method (Wei et al., 2019). 10% of the training dataset is selected and up to two tokens are randomly chosen to be masked in each sentence.

Table I: Examples of data augmentation using synonym replacement and random masking.

Original	The cat sat on the mat
Synonym	The feline sat on the mat
Masking	The [MASK] sat on the [MASK]

4. Conditional Random Fields (CRF)

While DeBERTa processes the entire sentence on a semantic level, the expected label transitions are not learnt. Replacing the softmax function with CRF allows the conditional probability of the labels to be considered on a sentence level. (Lafferty et al., 2001) Although the label sequence in NER may not be as significant as it is in the part-of-speech classification demonstrated in Lafferty's experiment (2001), recognizing the transition patterns of entity labels can effectively strengthen the model.

Similar probabilistic sequence models such as Maximum entropy Markov models (McCallum et al., 2000) considers the probabilities of possible next label given the current label and observation sequence, which leads to the label bias problem where the model will prefer the label sequence of common word pairs. This problem can be solved by CRF, which accounts for whole state sequences at once (Lafferty et al., 2001) and avoids bias on word pairs. Another advantage of CRF is that it guarantees global maximum likelihood convergence in a purely probabilistic setting (Lafferty et al., 2001) due to the convexity of the loss function.

Unlike the softmax layer, which predicts the label of each token independently, CRF is a random field globally conditioned on the entire sentence (Lafferty et al., 2001). The complete details follows the procedures proposed by Lafferty et al. (2001). The conditional probability is

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{\exp\left(\text{score}(\mathbf{x}, \mathbf{y})\right)}{Z(\mathbf{x})}$$

where $x =< x_1, x_2, ..., x_T >$ is the natural language input, $y =< y_1, y_2, ..., y_T >$ is the possible entity tags. The $score(x, y)$ sums how well each token x matches the label y and the transition possibility from $y_i - 1$ to y_i. As for the denominator, $Z(x)$ sums the exponential scores of all possible y given x. As for the loss function, it is given by

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = -\log P(\mathbf{y} \mid \mathbf{x}) = \log Z(\mathbf{x}) - \text{score}(\mathbf{x}, \mathbf{y})$$

which is used to find parameters that maximizes the log-likelihood of the training data.

IV. Experiments

We performed experiments on the CoNLL-2003 dataset to assess our Named Entity Recognition (NER) techniques under four configurations: a standard DeBERTa model, a DeBERTa model with a Conditional Random Field (CRF) layer, a DeBERTa model with data augmentation, and a combined model with both CRF and data augmentation. Details of the dataset, preprocessing, evaluation metrics, setup, results, and analysis follow.

1. Data

The CoNLL-2003 dataset (Tjong Kim Sang & De Meulder, 2003) is a benchmark for NER, consisting of news articles annotated with four entity types: Person, Organization, Location, and Miscellaneous. It includes training (14,041 sentences), validation (3,250 sentences), and test (3,453 sentences) sets, labeled using the BIO tagging scheme (e.g., `B-PER`, `I-PER`, `O`).

2. Experiment Details

Models were initialized from `microsoft/deberta-base` (He et al., 2021). Baseline and augmentation models used `AutoModelForTokenClassification`, while CRF models included a dropout layer (0.1), linear classifier, and CRF layer (PyTorch-CRF). Training used the Hugging Face `Trainer` API (Wolf et al., 2020) with:

- **Epochs**: 3 (Baseline, CRF, Data Augmentation); 5 (Data Augmentation + CRF).
- **Batch Size**: 8.
- **Learning Rate**: 2×10^{-5} (AdamW).
- **Weight Decay**: 0.01 (CRF models); 0 (others).

Implementation leveraged Hugging Face and PyTorch-CRF utilities.

We evaluated four configurations:

1. **Baseline**: DeBERTa fine-tuned on CoNLL-2003.
2. **CRF**: DeBERTa with a CRF layer to enforce coherent tag sequences.
3. **Data Augmentation**: Baseline DeBERTa with augmented data (15% synonym replacement, 10% masking).
4. **Data Augmentation + CRF**: CRF model with reduced augmentation (5% each).

In conclusion, the Data Augmentation + CRF model excels for NER but is limited by computational cost and dataset size. Future work could explore adaptive augmentation or larger datasets.

3. Results

Performance on the CoNLL-2003 test set is shown below.

Table II: Performance of Models on CoNLL-2003 Test Set

Model	Prec.	Recall	F1
Baseline	0.9079	0.9238	0.9158
CRF	0.9065	0.9236	0.9150
Data Aug.	0.9069	0.9220	0.9144
Data Aug. + CRF	**0.9118**	**0.9266**	**0.9191**

The Data Augmentation + CRF model achieved the highest F1 score (0.9191), outperforming Baseline (0.9158), CRF (0.9150), and Data Augmentation (0.9144). CRF models ensured sequence consistency but increased runtime, while augmentation improved robustness.

4. Analysis

The Data Augmentation + CRF model showed synergy between CRF's sequence modeling and augmentation's data diversity. Surprisingly, the CRF model alone slightly underperformed the Baseline (F1: 0.9150 vs. 0.9158), likely due to limited training data. High test loss in CRF models reflects their negative log-likelihood loss, not directly comparable to cross-entropy loss.

Ablation studies confirmed that moderate augmentation (5% each) outperformed aggressive augmentation (15% + 10%). The Data Augmentation + CRF model's success highlights the complementary nature of CRF and augmentation. An additional experiment with reduced augmentation (5% each) for the Data Augmentation model yielded an F1 score of 0.9162, slightly better than the Baseline.

V. Conclusion

The current examination shows that combining a Conditional Random Field (CRF) layer and with data augmentation significantly improves the ability of the DeBERTa model in Named Entity Recognition, achieving an F1 score of 0.9191 on the CoNLL-2003 dataset. This underscores the potential advantages of uniting traditional probabilistic approaches with modern deep learning techniques and highlights how the effectiveness of CRF in modeling sequential label dependencies and data augmentation works in tandem to improve model robustness. However, challenges such as high computational costs and restrictions driven by insufficient dataset sizes highlight the need for additional improvements. Future work focusing on adaptive augmentation methods and the use of larger datasets will ensure that we leverage the full potential of this method with a view to developing a more stable and scalable NER system for NLP applications.

VI. Contribution

CHU, Yan Yee	Experiments
FUNG, Choi Wing	Methods
MA, Hoi Tung	Related work

References

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., & Auli, M. (2019). Cloze-driven pretraining of self-attention networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* https://doi.org/10.18653/v1/d19-

Dai, H. J., Wang, F. H., Chen, C. W., Su, C. H., Wu, C. S., & Chiu, H. W. (2020). Deep learning–based natural language processing for screening psychiatric patients. *Journal of Biomedical Informatics, 108*, 103514. https://doi.org/10.1016/j.jbi.2020.103514

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805v2.* https://doi.org/10.48550/arXiv.1810.04805

He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654.* https://doi.org/10.48550/arXiv.2006.03654

Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). A survey on named entity recognition — datasets, tools, and methodologies. *Natural Language Processing Journal, 3,* 100017. https://doi.org/10.1016/j.nlp.2023.100017

Kyaw, Z. T. (2022). Data augmentation for name entity recognition (Master's thesis). Nanyang Technological University, Singapore. Retrieved from https://doi.org/10.32657/10356/161703

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML 2001 (pp. 282–289). https://repository.upenn.edu/cis_papers/159

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & Li, J. (2019b). A unified MRC framework for named entity recognition. *arXiv preprint arXiv:1910.11476.* http://arxiv.org/abs/1910.11476

Li, J., Sun, A., Han, J., & Li, C. (2022). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering, 34*(1), 50–70. https://doi.org/10.1109/tkde.2020.2981314

Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., & Li, J. (2019a). Dice loss for data-imbalanced NLP tasks. *arXiv preprint arXiv:1911.02855.* http://arxiv.org/abs/1911.02855

Liu, Y., Meng, F., Zhang, J., Xu, J., Chen, Y., & Zhou, J. (2019a). GCDT: A global context enhanced deep transition architecture for sequence labeling. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2431–2441.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019b). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692v1.* https://doi.org/10.48550/arXiv.1907.11692

McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of the 17th International Conference on Machine Learning (ICML)* (pp. 591–598). Morgan Kaufmann.

Rosvall, E. (2019). Comparison of sequence classification techniques with BERT for named entity recognition (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-261419

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.* https://doi.org/10.48550/arXiv.1910.01108

Shorten, C., & Khoshgoftaar, T. M. (2021). Text data augmentation for deep learning. *Journal of Big Data, 8*(1), 1–34. https://doi.org/10.1186/s40537-021-00451-5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*. https://doi.org/10.48550/arXiv.1706.03762

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Proceedings of EMNLP-IJCNLP 2019 Workshop on NLP4IF (pp. 638–644). https://arxiv.org/abs/1901.11196