A pipelined approach to Anaphora Resolution in Chemical Patents

Ritam Dutt*[1], Sopan Khosla*[1] and Carolyn Rosé[1]

[1]*Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA*

Abstract

We present our pipelined approach for the sub-task of anaphora resolution in chemical patents as part of the ChEMU shared task at CLEF, 2021. Our approach consists of independently trained mention extraction and relation classification modules. For the former, we set up a BERT-CRF and leverage the BIO scheme to represent the mentions. We include a post-processing step after mention extraction to correct boundary errors and handle nested mentions. For relation classification, we develop a BERT-based model that captures the context between the two candidate mentions to predict the relation between the two. Our final submission ensembles BERT models pretrained on different types of clinical data and achieves a Strict F1 of 0.785 on the official test set.

Keywords

Information Extraction, Anaphora Resolution, Chemical Patents.

1. Introduction

Chemical patents play a crucial role in disseminating information about the synthesis, properties, and applications of new chemical compounds[1, 2]. The rapid publication pace over the past decade necessitates the need of automated techniques to extract semi-structured knowledge from the patent text [3, 4], such as components and process conditions corresponding to chemical reactions.

A key step in understanding how chemical reactions (in patent text) involves identifying anaphoric dependencies between entities mentioned in the reaction[3]. These dependencies involve co-reference relations where different surface mentions refer to the same chemical entity, or bridging relations where different entities interact amongst themselves in a particular manner. The first instance in Table 1 highlights a co-referent relation between N-methylpyrrolidone and NMP. Likewise, the second instance talks about how the stirring event transforms the mixture. We describe them in detail in §2.
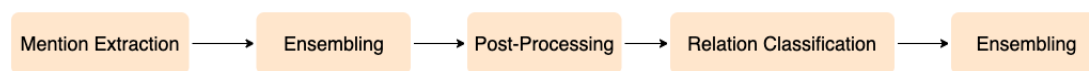

Figure 1: Flowchart outlining our pipelined methodology for anaphora resolution in chemical patents.

* Denotes equal contribution.

Table 1

Instances of text snippets and their corresponding anaphora relations.

Text snippet	Relation
[**N-methylpyrrolidone**]$_1$ [(**NMP**)]$_2$ was stirred for 1 day over CaH2 and finally distilled off.	CR(Ent_1, Ent_2)
[**The mixture**]$_1$ was stirred at room temperature for 1 day. A 2 molL aqueous solution of hydrochloric acid was added to [**the mixture**]$_2$.	TR(Ent_1, Ent_2)
[**Acetic acid (9.8 ml)**]$_1$ and [**water (4.9 ml)**]$_2$ were added to [**the solution of Compound (4) (0.815 g, 1.30 mmol)in THF (4.9 ml)**]$_3$. [**The mixture**]$_4$ was stirred for 3 hrs at 50^oC and then cooled to 0^oC.	RA(Ent_1,Ent_4), RA(Ent_2,Ent_4), RA(Ent_3, Ent_4)
[**The mixture**]$_1$ was extracted with [**ethyl acetate**]$_2$ for 3 times. [**The combined organic layer**]$_3$ was washed with water and saturated aqueous sodium chloride.	WU(Ent_1,Ent_3), WU(Ent_2,Ent_3)
[**Pyrazinecarboxylic acid (152.8 mg, 1.23 mmol, 1 eq)**]$_1$and [**H-Phe-OtBu-HCl(253.8 mg, 0.98 mmol, 0.8 eq)**]$_2$were charged into an [**eggplant flask**]$_3$	CN(Ent_1,Ent_3), CN(Ent_2,Ent_3)

We present a pipe-lined approach to solve anaphoric resolution in chemical patents, comprising of two key phases of Mention Extraction and Relation Classification. We perform ensembling after each of these two phase to reduce spurious corrleations and improve prediction. We also incorporate a post-processing module after extracting mentions to handle boundary issues, discontinuous, and nested spans. We describe our methodology in detail in §3 and Figure 1 illustrates a pictorial representation of our approach.

We provide the experimental details in §4 and present our results in §5. Our proposed approach achieves a performance of 0.804 F1 score and 0.785 F1 score on the validation and test set respectively for the strict matching paradigm, successfully beating the proposed baseline [3]. For relaxed or inexact match, our scores are substantially higher by almost a margin of 0.07 F1 scores. We conclude and present future ideas in §6.

2. Task Description

We focus on the sub-task of anaphora resolution in chemical patents as part of the ChEMU shared task at CLEF, 2021 [1]. The task of anaphora resolution seeks to identify the nature of anaphoric dependencies between mentions/expressions in chemical patents. Prior work [3] has investigated the following 5 anaphoric dependencies in chemical patents. We present instances of those in Table 1.

1. Coreference (CR): The relationship between expressions or mentions wherein they refer to the same chemical Mention.
2. Reaction Associated (RA): The relationship between a chemical compound and its immediate sources via a mixing/chemical process.

---

3. Transformed (TR) : The relationship between expressions or mentions, which have undergone physical changes (e.g., pH and temperature) but have the same chemical composition.

4. Work Up (WU): The relationship between chemical compounds that were used for isolating or purifying mentions, and their corresponding outputs.

5. Contained (CN): The association between chemical compounds and the equipment in which they are placed.

3. Methodology

We outline the details of a pipelined architecture to anaphora resolution in this section. Our approach consists of two major steps of Mention Extraction and Relation Classification. In the Mention Extraction phase, we identify all possible mentions from the patent text, whereas in the Relation Classification phase, we infer whether a given pair of mentions have any anaphoric dependency between them. We describe the neural architecture we have employed for these two phases.

3.1. Mention Extraction

Prior work has demonstrated the success of neural architectures in extracting chemical and bio-medical mentions [2], spans of chemical reaction [5] and the specific roles of mentions in a reaction [4].
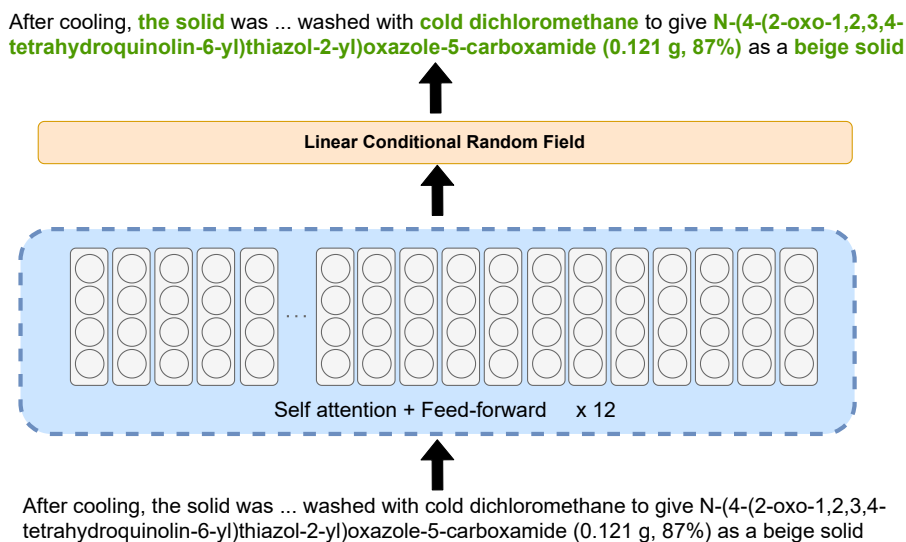

Figure 2: Extracting mentions using a transformer-Based encoder and CRF. The highlighted part in the text (in green) correspond to the extracted entities.

In this task, we consider any span of text snippet which was annotated either as an antecedent or an anaphora as a mention. Based on the annotation corpus of [3], mentions include quantified

chemical compounds (0.51 g of methanol, K2CO3 (300 mg, 2.2 mmol)), proper nouns (DMF, (2,6-dichloro-4-fluorophenyl)hydrazine hydrochloride), identifiers (5i, 4a), pronouns (it, they,) and noun phrases (the solvent, an autoclave). We note that approximately 3% of mentions in the dataset have dis-continuous spans, and we leverage post-processing techniques to deal with such spans.

We thus, model the task of mention extraction as a sequence labeling task. For this phase, we encode the longest contiguous span of text that includes these individual, discontinuous spans as the span of the given mention. Based on the recent success of transformer-based modules like [6] in information extraction [7, 8, 9], we employ a similar approach in our case. We use a transformer-based encoder to encode the mentions and then pass the encoding through a Linear Conditional Random Field (CRF) [10]. An overview of the Mention Extraction architecture is shown in Figure 2.

3.2. Relation Classification
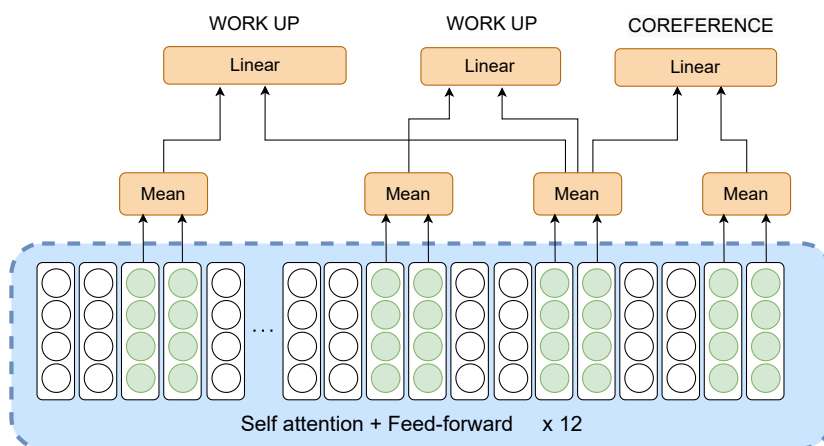
After cooling, **the solid** was ... washed with **cold dichloromethane** to give **N-(4-(2-oxo-1,2,3,4-tetrahydroquinolin-6-yl)thiazol-2-yl)oxazole-5-carboxamide (0.121 g, 87%)** as a **beige solid**

Figure 3: Relation classification between pairs of mentions. The parts of the text highlighted (in green) refer to to entity mentions and their corresponding embeddings are shaded. Mean pooling over the embeddings yield the final representation of the mention, and pairwise relation classification of the mentions are carried out.

We present an overview of the Relation Classification architecture in Figure 3. For a given pair of mentions, we define the context for a pair of mentions as the sequence of sentences that has the mention pair. We pass the context through a transformer Based encoder, and use mean-pooling over the individual mention tokens to obtain a representation corresponding to that mention. We concatenate the representations of the two mention spans and project it through a linear layer of 6 classes. These correspond to the 5 anaphoric dependencies and the No-Relation class if there is no dependency between the pair of mentions.

4. Experiments

4.1. Mention Extraction

For the task of mention extraction, we experiment with several transformer-Based encoder modules, such as BERT [6], Clinical BERT (trained on clinical notes) [11] and Pubmed-BERT (trained on Pubmed abstracts) [12]. Moreover, since chemical compounds are often several characters long, a single compound could be decomposed into several tokens. To circumvent this tokenization issue, we include the "LONG TOKEN" similar to [2], as a special token, to subsume the remaining tokens of a compounds beyond a certain length. For our experiments, the length is kept to 25. We use the BIO (Beginning Inside Outside) scheme to represent the mentions. For example "the residue is heated" will be converted to "B-ENT I-ENT O O".

We evaluate mention extraction in terms of precision, recall, and F1 score, for both exact (strict) and inexact (relaxed) match similar to [4]. We use the BRAT evaluation script provided by the organizers to compute the scores. We ran our models using the huggingface transformers library in PyTorch, with a batch size of 8, a learning rate of 1e-5, a dropout of 0.1, Adam optimizer, and patience of 5.

4.2. Post-Processing

To correct boundary errors and extract nested spans, we further post-process the output from the neural mention extractor using several rule-based sieves. The sieves were chosen after close inspection on the validation data and are described in detail in the subsequent section.

4.3. Relation Classification

For the task of relation classification, we experiment with several transformer-Based encoder modules, namely BERT-Base and BERT-large [6], Clinical BERT [11], Pubmed-BERT [12] and BioBERT [13]. Moreover, since we have to check anaphoric dependency between all possible pairs of entities during validation and testing, it is imperative to incorporate negative instances during training. Thus, all pairs of entities which do not have an anaphoric dependency between them are taken as negative instances and assigned the "NO RELATION" label. We also experiment by varying the proportion of negative instances during training. Similar to mention extraction, we use the BRAT evaluation script provided by the organizes. We find the precision, recall, and F1 score for the relation classification task, similar to [3].

We ran our relation classification models using the huggingface transformers library in PyTorch, with a batch size of 16, a learning rate of 3e-5, a dropout of 0.1, Adam optimizer, and patience of 5. We curated negative samples by pairing mentions that were not more than 10 mentions apart from each other in the patent document.

4.4. Ensembling

We perform ensembling twice, once after the mention extraction phase and once after the relation classification phase. We carry out majority voting over the outputs of five models and consider only those outputs which have been predicted by at least three models. The outputs

Table 2

Strict and Relaxed match of the Mention Extraction module using different tokenizers in terms of Precision, Recall, and F1 score.

Method	Strict			Relaxed		
	Precision	Recall	F1	Precision	Recall	F1
BERT-Base-Long	0.877	0.853	0.865	0.977	0.949	0.963
Clinical-BERT	0.897	0.872	0.884	0.981	0.953	0.967
Pubmed-BERT	0.888	0.860	0.874	0.969	0.939	0.954
Pubmed-BERT-Long	0.909	0.871	0.890	0.989	0.947	0.967
Clinical-BERT-Long	0.881	0.863	0.872	0.974	**0.955**	0.964
Ensemble	0.915	0.876	0.895	0.988	0.947	0.967
+ Post-processing	**0.943**	**0.903**	**0.922**	**0.991**	0.949	**0.970**

correspond to the extracted mention-span for mention extraction, and over pairs of extracted span and their corresponding relation label for relation classification. Ensembling has been proven to reduce spurious correlations and improve performance [14], and has been employed for several tasks [15, 16, 9].

5. Results and Analysis

5.1. Mention Extraction

We note the results for mention extraction in Table 2. At the outset, we observe that the models achieve almost 0.98 F1 score for the inexact (relaxed) match. However, they suffer almost a 0.09 drop in F1 score under the exact (strict) match evaluation. A majority of the errors occurs due to nested spans or discontinuous spans (see Example 4 in the post-processing sub-section). Another common issue is the omission or inclusion of tokens at the beginning/end of the span like (see Examples 1,2, and 3 in the post-processing sub-section), which we refer to as "boundary issues". Since the entities are multi-faceted and can range from simple noun-phrases and identifiers to complex chemical quantifiers, it makes the task exceedingly challenging.

We also observe that incorporating tokenizers corresponding to bio-medical and clinical domains perform slightly better than the uncased BERT model. The best performance is observed for PubMed-BERT-Long with F1 score of 0.890, as opposed to 0.865 for the uncased BERT-Long model. Inclusion of the "LONG TOKEN" is shown to benefit PubMed-BERT and BERT-uncased but fares worse for the Clinical-BERT model.

Unsurprisingly, ensembling over these 5 models achieves the highest score in the exact match setting across all three metrics. We use the entities extracted from the ensemble setting as the final ones and post-process them before the relation classification phase.

Table 3
Strict match for the Relation prediction task using the gold mentions and the predicted mentions after our post-processing step.

Method	Gold entities			Predicted entities		
	Precision	Recall	F1	Precision	Recall	F1
BioBERT-Long	0.915	0.912	0.914	0.823	0.765	0.793
Pubmed-BERT-Long	0.919	**0.916**	0.918	0.821	**0.770**	0.795
Clinical-BERT-Long	0.916	0.902	0.909	0.826	0.757	0.790
BERT-Base-Long	0.926	0.894	0.910	0.830	0.748	0.787
BERT-Large-Long	0.917	0.887	0.902	0.821	0.746	0.782
Ensemble	**0.939**	0.915	**0.927**	**0.842**	0.769	**0.804**

5.2. Post-Processing

For post-processing, we pass the outputs of the mention extraction module through several sequential rule-based sieves:

1. If the extracted mention ends with strings like {' and', ' under', ' or', ' over'}, remove them from the mention. E.g. [**Alcohol and**]$_1$ → [**Alcohol**]$_1$ **and**
2. If a mention is preceded by an article like {'a', 'the'}, include that article in the mention.
3. If the extracted mention ends with {' with', ' of', ' in'} and there is an adjoining mention after it, combine the two. E.g.
   [**ethanol in**]$_1$ [**the reaction mixture**]$_2$ → [**ethanol in the reaction mixture**]$_1$
4. We observed that often the patent documents refer to compounds with an ID. This ID is annotated as a coreferent mention to the actual compound. E.g. [**7-fluorobenzofuran-3(2H)-one**]$_1$ [**84c**]$_2$ [**(340 mg, 2.2 mmol)**]$_1$. But since our neural model can only extract contiguous mentions, it outputs [**7-fluorobenzofuran-3(2H)-one**]$_1$ [**84c (340 mg, 2.2 mmol)**]$_2$. In order to extract these coreferents from outputs with such patterns we identify instances that follow the below template. If predicted m2 starts with a word (w1) which follows regex *([0-9]+[a-z]+)*, and contains a second word (w2) which starts with a {'('}, then combine m2 with m1 excluding w1 which is separated out as the coreferent.

We find that this post-processing substantially improves the performance (from 0.895 F1 to 0.922 F1) on the official mention extraction *Strict* metrics (Table 2). Each of our sieves work towards increasing the number of exact matches between gold and system mentions. Furthermore, Sieve 3 and 4 also uncover new spans simultaneously impacting the Relaxed Match scores (from 0.967 F1 to 0.970 F1).

5.3. Relation Classification

We present the performance of models for the relation classification task for both the gold and predicted mentions in Table 3. We observe that pre-trained Pubmed-BERT and BioBERT fares slightly better than the uncased BERT-Base and BERT-Large model, highlighting again the

Table 4

Strict and Relaxed match for the complete Anaphora resolution task on the test set.

Relation	Support	Strict			Relaxed		
		Precision	Recall	F1	Precision	Recall	F1
Contained (CN)	148	0.773	0.689	0.729	0.856	0.764	0.807
Coreference (CR)	1491	0.757	0.582	0.658	0.895	0.688	0.778
Reaction Associated (RA)	1245	0.804	0.763	0.783	0.902	0.856	0.878
Transformed (TR)	166	0.942	0.885	0.913	0.942	0.886	0.913
Work Up (WU)	2576	0.846	0.845	0.845	0.919	0.918	0.919
All	5626	0.818	0.754	0.785	0.909	0.838	0.872

benefits of pre-training on clinical data. Moreover, our pair-wise relation classification approach achieves nearly a 0.91 F1 score for anaphoric relations using the gold mentions. We also observe empirically that including 100% of all negative examples during training achieves the highest performance. Models that are trained on 5% and 10% of the total negative samples achieve an F1-Score (gold entities) of 0.762 and 0.794 respectively, around 0.15 F1 points below their 100% counterpart. A majority of the misclassification errors takes place when an anaphoric dependency between a pair of mentions is predicted as "NO RELATION" and vice versa, since the negative classes account for around 87% of all labels. The only other instance of misclassification takes place where the RA relation is predicted as WU, since both describe associations between chemical compounds.

Unlike [3], our pair-wise approach can circumvent the problems of discontinuous and nested spans and hence we include those mentions. Nevertheless, we note how the errors in the mention extraction phase propagate downstream and downgrade the relation classification performance on the predicted entities. This results in an average score of 0.79 F1 for predicted entities (a drop in approximately 0.12 F1 points). While it would be prudent ideally to carry out the two phases in a joint fashion like [3] (to prevent cascading errors), the crucial post-processing step of fixing boundary issues and extracting additional nested mentions, necessitates the pipelined approach.

In fact, our architecture beats the transformer baseline that performs joint co-reference and bridging on both the validation and test set by 0.03 F1 score on the exact match metric. The boost for relaxed match is substantially higher, with our model outperforming the baseline by approximately 0.07 F1 score on both validation and test. Moreover, ensembling over the different models boosts the performance further by 0.01 F1 for both the gold and predicted entities.

We report the performance for the 5 individual anaphoric relations in Table 4. Coreference (CR) relations with their nuanced rules and long-term dependencies have the poorest performance [3], whereas the bridging relations being more local and specific in nature, fared considerably better. We acknowledge there is immense scope for improvement and posit how incorporating additional information like events or entity types can help bolster performance. We defer this exploration for future work.

Our final performance on the validation set was 0.804 F1 and 0.887 F1 for the strict and

relaxed match respectively. Likewise, our performance on the test set was 0.785 and 0.872 F1 for strict and relaxed match. We are currently ranked the first in the shared task.

6. Conclusion

Resolving anaphora dependencies in chemical patents plays a key role in understanding the nuances of how chemical reactions are described, and the interactions between participating entities. We describe a pipelined approach to address this challenge using independently trained mention extraction and relation classification modules. Such a design choice facilitates the inclusion of a rule-based post-processing module to handle boundary errors, and discontinuous/nested spans. We achieve a Strict F1 score of 0.785 and a relaxed F1 score of 0.872 on the official test set, significantly outperforming the baseline.

Acknowledgement

References

[1] S. A. Akhondi, H. Rey, M. Schwörer, M. Maier, J. Toomey, H. Nau, G. Ilchmann, M. Sheehan, M. Irmer, C. Bobach, et al., Automatic identification of relevant chemical compounds from patents, Database 2019 (2019).

[2] Z. Zhai, D. Q. Nguyen, S. Akhondi, C. Thorne, C. Druckenbrodt, T. Cohn, M. Gregory, K. Verspoor, Improving chemical named entity recognition in patents with contextualized word embeddings, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 328–338. URL: https://www.aclweb.org/anthology/W19-5035. doi:10.18653/v1/W19-5035.

[3] B. Fang, C. Druckenbrodt, S. A. Akhondi, J. He, T. Baldwin, K. Verspoor, ChEMU-ref: A corpus for modeling anaphora resolution in the chemical domain, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1362–1375. URL: https://www.aclweb.org/anthology/2021.eacl-main.116.

[4] D. Q. Nguyen, Z. Zhai, H. Yoshikawa, B. Fang, C. Druckenbrodt, C. Thorne, R. Hoessel, S. A. Akhondi, T. Cohn, T. Baldwin, et al., Chemu: named entity recognition and event extraction of chemical reactions from patents, in: European Conference on Information Retrieval, Springer, 2020, pp. 572–579.

[5] H. Yoshikawa, D. Q. Nguyen, Z. Zhai, C. Druckenbrodt, C. Thorne, S. A. Akhondi, T. Baldwin, K. Verspoor, Detecting chemical reactions in patents, in: Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association, Australasian Language Technology Association, Sydney, Australia, 2019, pp. 100–110. URL: https://www.aclweb.org/anthology/U19-1014.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186.

[7] S. Vashishth, D. Newman-Griffis, R. Joshi, R. Dutt, C. Rose, Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets, arXiv preprint arXiv:2005.00460 (2020).

[8] A. Thillaisundaram, T. Togia, Biomedical relation extraction with pre-trained language representations and minimal task-specific architecture, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 84–89. URL: https://www.aclweb.org/anthology/D19-5713. doi:10.18653/v1/D19-5713.

[9] J. W. Y. R. Z. Zhang, Y. Zhang, Melaxtech: A report for clef 2020–chemu task of chemical reaction extraction from patent, in: Working Notes of CLEF 2020–Conference and Labs of the Evaluation Forum, Vol. 2696. CEUR Workshop Proceedings, 2020.

[10] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data: proceedings of the 18th international conf. on machine learning, 2001, San Francisco, CA, USA (2001).

[11] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. URL: https://www.aclweb.org/anthology/W19-1909. doi:10.18653/v1/W19-1909.

[12] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 58–65. URL: https://www.aclweb.org/anthology/W19-5006. doi:10.18653/v1/W19-5006.

[13] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: A pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (Oxford, England) 36 (2020) 1234–1240.

[14] Z. Allen-Zhu, Y. Li, Towards understanding ensemble, knowledge distillation and self-distillation in deep learning, arXiv preprint arXiv:2012.09816 (2020).

[15] S. Khosla, Emotionx-ar: Cnn-dcnn autoencoder based emotion classifier, in: Proceedings of the sixth international workshop on natural language processing for social media, 2018, pp. 37–44.

[16] S. Khosla, R. Joshi, R. Dutt, A. W. Black, Y. Tsvetkov, Ltiatcmu at semeval-2020 task 11: Incorporating multi-level features for multi-granular propaganda span identification, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1756–1763.