VinAI at ChEMU 2020: An accurate system for named entity recognition in chemical reactions from patents

Mai Hoang Dao[1,2] and Dat Quoc Nguyen[1]

[1]VinAI Research, Vietnam
{v.maidh3, v.datnq9}@vinai.io
[2]Posts and Telecommunications Institute of Technology, Vietnam

Abstract. This paper describes our VinAI system for the ChEMU task 1 of named entity recognition (NER) in chemical reactions. Our system employs a BiLSTM-CNN-CRF architecture [6] with additional contextualized word embeddings. It achieves very high performance, officially ranking second with regards to both exact- and relaxed-match F_1 scores at 94.33% and 96.84%, respectively. In a post-evaluation phase, fixing a mapping bug which converts the column-based format into the `brat` standoff format helps our system to obtain higher results. In particular, we obtain an exact-match F_1 score at 95.21% and especially a relaxed-match F_1 score at 97.26%, thus achieving the highest relaxed-match F_1 compared to all other participating systems. We believe our system can serve as a strong baseline for future research and downstream applications of chemical NER over chemical reactions from patents.

Keywords: Named entity recognition; Chemical reactions; Patents; Neural network.

1 Introduction

The discovery of new chemical compounds plays an essential key role in the chemical industry. To disclose newly discovered chemical compounds, patent documents are often selected as the initial venues; and only a small fraction of these chemical compounds are published in journals, but this usually takes up to 3 years after the patent disclosure [14]. Thus patents containing critical and timely information about the new chemical compounds serve as starting pointers for chemical research in both academia and industry [1]. Due to a huge volume of new chemical patent applications [9], it is becoming increasingly important to develop automatic information extraction approaches for large-scale mining of chemical information from these patent documents [1].

Chemical named-entity recognition (NER) is a fundamental step for information extraction from chemical patents, supporting many downstream tasks such as chemical reaction prediction [12,17], chemical syntheses [13] and the like. The ChEMU—Cheminformatics Elsevier Melbourne University—task 1 provides participants with opportunities to develop automatic chemical NER systems from chemical reactions in chemical patents. This task is to identify crucial elements of a chemical reaction, including compounds, conditions and yields as well as their specific roles in the reaction. Details of this task can be found in the overview paper of the ChEMU lab [3].

In this paper, we present our VinAI team's system for the ChEMU task 1. Our system is based on the well-known BiLSTM-CNN-CRF architecture [6] with additional contextualized word embeddings. Our system officially obtains the second best performance results in terms of both exact- and relaxed-match F_1 scores at 94.33% and 96.84%, respectively. In a post-evaluation phase, fixing a column-`brat` conversion bug then helps our system to obtain even better results at 95.21% for exact-match F_1 and especially 97.26% for relaxed-match F_1. We thus obtain the highest relaxed-match F_1 score in comparison to all other participating systems. We also provide an ablation study to investigate the contributions of different types of input word representations in the full system, reconfirming the effectiveness of the contextualized word embeddings for chemical NER [18].

2 Task description

The ChEMU task 1 of "Named entity recognition" involves identifying chemical compounds and their specific types. In particular, the task assigns the label of a chemical compound according to the role which it plays within a chemical reaction. In addition to identifying chemical compounds, the task also requires identification of the label of the chemical reaction, the temperatures and reaction times at which the reaction is carried out as well as yields obtained for the final chemical product. The task defines 10 different entity type labels as listed in Table 1, involving both entity boundary prediction and entity label classification. See [3,10] for more details.

3 Our system

In this section, we present our VinAI system for the ChEMU task 1. We formulate this task as a sequence labeling problem with BIO tagging scheme. Following [18], our system employs the well-known BiLSTM-CNN-CRF model [6] with additional contextualized word embeddings.

Figure 1 illustrates the architecture of our participating system. In particular, our system represents each word token w_i in an input sequence $w_1, w_2, ..., w_n$ by a vector $\boldsymbol{v}_i$ which is resulted by concatenating the pre-trained word embedding, the CNN-based character-level word embedding [6] and the contextualized word

Table 1. Definitions of entity types. "Abbr." denotes abbreviation.

Label	Definition	Abbr.
STARTING_MATERIAL	A substance that is consumed in the course of a chemical reaction providing atoms to products.	ST
REAGENT_CATALYST	A reagent is a compound added to a system to cause or help with a chemical reaction.	RC
SOLVENT	A solvent is a chemical entity that dissolves a solute resulting in a solution.	SO
REACTION_PRODUCT	A product is a substance that is formed during a chemical reaction.	RP
OTHER_COMPOUND	Other chemical compounds that are not the products, starting materials, reagents, catalysts and solvents.	OT
TIME	The reaction time of the reaction.	TI
TEMPERATURE	The temperature at which the reaction was carried out.	TE
YIELD_PERCENT	Yield given in percent values.	YP
YIELD_OTHER	Yields provided in other units than %.	YO
EXAMPLE_LABEL	A label associated with a reaction specification.	EX

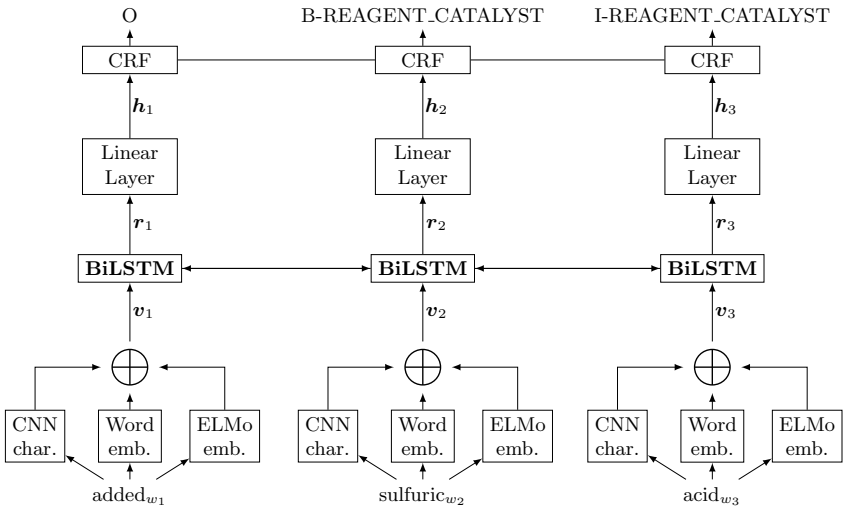

Fig. 1. Illustration of our participating system's architecture. This figure is drawn based on [18].

embedding of the word token w_i. Here, we utilize the pre-trained word embeddings released by [18], which are trained on a corpus of 84K chemical patents (1B word tokens) using the Word2Vec skip-gram model [7]. In addition, we also utilize the contextualized word embeddings generated by a pre-trained ELMo language model [11], which is trained using the same corpus of 84K chemical

patents [18].[1] Then vector representations $\boldsymbol{v}_i$ are fed into a BiLSTM encoder to extract latent feature vectors $\boldsymbol{r}_i$ for input words w_i. Each latent feature vector $\boldsymbol{r}_i$ is then linearly transformed into $\boldsymbol{h}_i$ before being fed into a linear-chain CRF layer for NER label prediction [5]. A cross-entropy loss is computed during training while the Viterbi algorithm is used for decoding.

4 Experiments

4.1 Experimental setup

Dataset: For system development, the ChEMU task 1 provides a corpus of 1125 chemical reaction snippets with gold standard NER annotations using the `brat` standoff format [15]. Although this corpus is pre-split into a training set of 900 snippets and a validation set of 225 snippets, participants are free to use this corpus in any manner they find useful when training and tuning their systems, e.g. using a different split or performing cross-validation. Thus we only employ the first 100 snippets in the provided validation set for validation,[2] and merge the remaining 125 snippets into the provided training set, resulting in a new training set of 1025 snippets in total. Following [18], we employ the OpenNLP toolkit [8] for sentence segmentation and the OSCAR4 tokenizer [4] to tokenize training and validation sentences, then convert these sentences into the CoNLL column-based format with the BIO tagging scheme.

Implementation: Our system is implemented based on the AllenNLP framework [2]. For training, we use exactly the same hyper-parameters used in [18] with the exception of using the batch size at 24. Pre-trained word embeddings and the pre-trained ELMo are fixed while other model parameters are updated during training. We train our system for 50 epochs and compute the standard exact-match F_1 score after each training epoch on the validation set. We select the model with the highest exact-match F_1 score on the validation set.

Evaluation phase: For the final evaluation phase, the ChEMU task 1 provides a raw test set consisting of 375 patent snippets. Each test snippet is sentence-segmented and tokenized using OpenNLP and OSCAR4, respectively. We then convert tokenized test sentences into the column-based format and apply our selected model to predict NER labels. We then use our own mapping script to convert the predicted BIO-based NER outputs into the `brat` standoff format, and submit the `brat`-formatted test outputs for evaluation.

Evaluation metrics: The ChEMU task 1 uses three metrics, namely precision, recall and F_1 scores for evaluation, under both "exact" and "relaxed" span matching conditions [16].

[1] `https://github.com/zenanz/ChemPatentEmbeddings`
[2] Sorted by file names: 0050–0690.

Table 2. Our official evaluation results (in %) on the test set, i.e. the predicted test outputs are submitted during the evaluation phase. The subscripts denote our ranking.

Entity label	Exact-match			Relaxed-match		
	P	R	F_1	P	R	F_1
STARTING_MATERIAL	93.24	91.14	92.18	96.40	94.10	95.24
REAGENT_CATALYST	88.54	90.48	89.50	91.84	94.04	92.93
SOLVENT	93.64	96.26	94.93	94.55	96.97	95.74
REACTION_PRODUCT	89.12	90.99	90.05	94.85	97.18	96.00
OTHER_COMPOUND	97.10	95.29	96.18	98.84	97.10	97.96
TIME	98.89	98.67	98.78	100.0	99.56	99.78
TEMPERATURE	95.54	94.44	94.99	99.01	98.68	98.84
YIELD_PERCENT	99.74	99.74	99.74	99.74	99.74	99.74
YIELD_OTHER	97.68	95.68	96.67	97.91	95.91	96.90
EXAMPLE_LABEL	91.10	87.97	89.50	94.07	90.83	92.42
Overall	94.62_2	94.05_3	94.33_2	97.07_1	96.61_3	96.84_2

Table 3. Our post-evaluation results (in %) on the test set, i.e. the predicted test outputs, which are resulted from fixing the column-`brat` conversion bug, are submitted right after the evaluation phase.

Entity label	Exact-match			Relaxed-match		
	P	R	F_1	P	R	F_1
STARTING_MATERIAL	93.56	91.98	92.77	96.71	94.94	95.82
REAGENT_CATALYST	90.47	92.26	91.36	92.22	94.05	93.12
SOLVENT	94.09	96.73	95.39	94.55	97.20	95.85
REACTION_PRODUCT	90.68	92.16	91.42	95.51	96.96	96.23
OTHER_COMPOUND	97.05	95.44	96.24	98.68	97.30	97.99
TIME	99.12	99.12	99.12	100.0	99.78	99.89
TEMPERATURE	95.70	94.44	95.00	99.01	99.01	99.01
YIELD_PERCENT	99.49	100.0	99.74	99.49	100.0	99.74
YIELD_OTHER	97.94	97.05	97.49	98.17	97.27	97.72
EXAMPLE_LABEL	97.38	95.70	96.53	97.67	95.99	96.82
Overall	95.38_2	95.04_2	95.21_2	97.37_1	97.16_1	97.26_1

4.2 Main results

Table 2 shows the official results of our system's outputs on the test set which is submitted during the evaluation phase. By employing a standard neural architecture, our system obtains a high performance which is officially ranked second among 11 participating systems, using both exact- and relaxed-match F_1 scores.

Table 4. Ablation "exact-match" results (in %) on the development set.

Model	P	R	F_1
Our system (full)	97.41	97.07	97.24
(a) w/o Word2Vec-based pre-trained word embeddings	96.32	96.66	96.49
(b) w/o CNN-based character-level word embeddings	97.24	97.01	97.13
(c) w/o ELMo-based contextualized word embeddings	96.25	96.19	96.22

Note that in the evaluation phase, we unfortunately were unaware of a bug in our mapping script which converts the predicted test outputs in the column-based format into the `brat` standoff format. Right after the evaluation phase, we fixed the bug, and reran our column-`brat` conversion script to produce a new submission, and then asked the ChEMU organizers to help evaluate the new submission. Table 3 details our post-evaluation results. Fixing the mapping bug helps improve our exact-match F_1 by 0.9% and our relaxed-match F_1 by 0.4%, absolutely; thus leading to the highest relaxed-match F_1 score compared to other participating systems.

4.3 Ablation study

Table 4 presents ablation tests over 3 factors of our system on the development set, including (a) removing the Word2Vec-based pre-trained word embeddings, (b) removing the CNN-based character-level word embeddings and (c) removing the ELMo-based contextualized word embeddings. Factor (a) degrades the exact-match F_1 score by 0.8%, while factor (b) and (c) degrade the exact-match F_1 score by 0.1% and 1.0%, respectively. The contribution of the CNN-based character-level word embeddings is not substantial because the pre-trained ELMo language model we employ also builds on character embeddings.

4.4 Error analysis

To understand the source of errors, we perform error analysis on the development set. Among 56 error cases in total, 34 cases are predicted with correct entity boundaries (i.e. exact span) but with incorrect labels (See the corresponding confusion matrix in Figure 2), while there are 17 cases corresponding with correct entity labels and overlapped inexact span. Figures 3 and 4 show examples of these two types of errors. In particular, Figure 3 shows an example of exact span and an incorrect label where a reagent catalyst entity of "HCL" is predicted as another compound type. The reason is probably because "HCL" and other popular chemical compounds such as "water", "citric acid" and the like play different/multiple roles in chemical reactions. Note that there is no error case corresponding with incorrect label and overlapped inexact span. The remaining 5 errors belong to the group of predicted entities in which their span is not overlapped with the span of any gold standard entity, i.e. non-chemical "O"-labeled words are predicted as REACTION_PRODUCT (RP) chemical entities as shown in the column O in Figure 2.

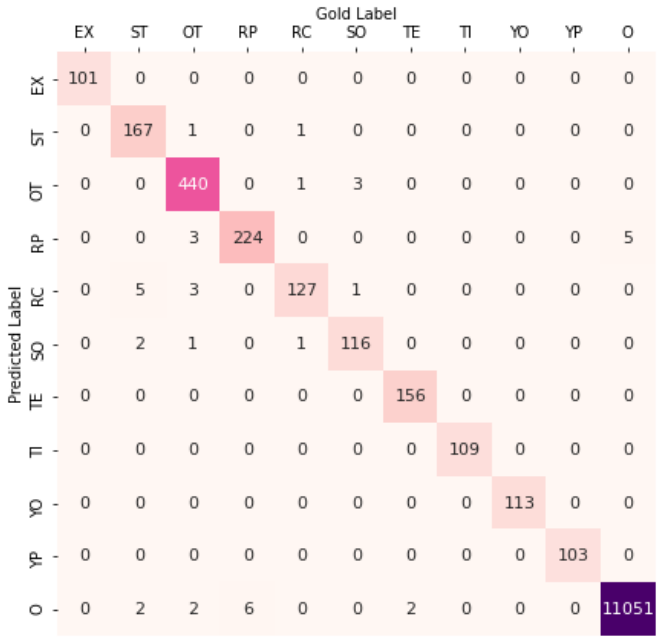

Fig. 2. Confusion matrix of our system on the development set w.r.t. the correct entity boundary prediction. Label abbreviations are presented in Table 1.

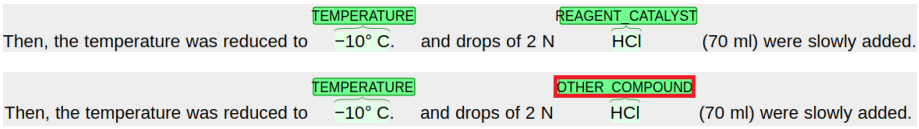

Fig. 3. An example of incorrect NER type prediction.

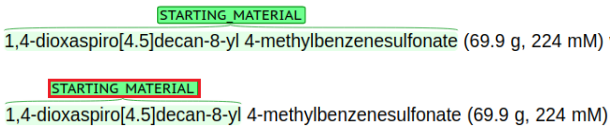

Fig. 4. An example of incorrect NER span prediction.

5 Conclusion

In this paper, we have presented our VinAI system for participating in the ChEMU task 1 of named entity recognition in chemical reactions from patents. We use a BiLSTM-CNN-CRF architecture with additional ELMo-based contextualized word embeddings to handle the task. Our system is officially ranked the second best performing one with regards to both the exact- and relaxed-match

F_1 scores. In addition, fixing the column-`brat` conversion bug then helps our system to obtain the highest relaxed-match F_1 score in a post-evaluation phase. We believe our system can serve as a strong baseline for future work on chemical NER in chemical reactions from patents.

References

1. Akhondi, S.A., Rey, H., Schwörer, M., Maier, M., Toomey, J.P., Nau, H., Ilchmann, G., Sheehan, M., Irmer, M., Bobach, C., Doornenbal, M.A., Gregory, M., Kors, J.A.: Automatic identification of relevant chemical compounds from patents. Database **2019**, baz001 (2019)
2. Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N.F., Peters, M., Schmitz, M., Zettlemoyer, L.S.: AllenNLP: A Deep Semantic Natural Language Processing Platform. In: arXiv:1803.07640 (2017)
3. He, J., Nguyen, D.Q., Akhondi, S.A., Druckenbrodt, C., Thorne, C., Hoessel, R., Afzal, Z., Zhai, Z., Fang, B., Yoshikawa, H., Albahem, A., Cavedon, L., Cohn, T., Baldwin, T., Verspoor, K.: Overview of ChEMU 2020: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. In: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020) (2020)
4. Jessop, D.M., Adams, S.E., Willighagen, E.L., Hawizy, L., Murray-Rust, P.: Oscar4: a flexible architecture for chemical text-mining. Journal of cheminformatics **3**(1), 1–12 (2011)
5. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289 (2001)
6. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1064–1074 (2016)
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
8. Morton, T., Kottmann, J., Baldridge, J., Bierner, G.: Opennlp: A java-based nlp toolkit. In: Proceeding of the 10th Conference of the European Chapter of the Association of Computational Linguistics (2005)
9. Muresan, S., Petrov, P., Southan, C., Kjellberg, M.J., Kogej, T., Tyrchan, C., Varkonyi, P., Xie, P.H.: Making every sar point count: the development of chemistry connect for the large-scale integration of structure and bioactivity data. Drug Discovery Today **16**(23-24), 1019–1030 (2011)
10. Nguyen, D.Q., Zhai, Z., Yoshikawa, H., Fang, B., Druckenbrodt, C., Thorne, C., Hoessel, R., Akhondi, S.A., Cohn, T., Baldwin, T., Verspoor, K.: ChEMU: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. In: Proceedings of the 42nd European Conference on Information Retrieval. pp. 572–579 (2020)
11. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237 (2018)

12. Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C., Laino, T.: "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chemical science **9**(28), 6091–6098 (2018)
13. Segler, M.H., Preuss, M., Waller, M.P.: Planning chemical syntheses with deep neural networks and symbolic ai. Nature **555**(7698), 604 (2018)
14. Senger, S., Bartek, L., Papadatos, G., Gaulton, A.: Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. Journal of cheminformatics **7**(1), 49 (2015)
15. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107 (2012)
16. Verspoor, K., Jimeno Yepes, A., Cavedon, L., McIntosh, T., Herten-Crabb, A., Thomas, Z., Plazzer, J.P.: Annotating the biomedical literature for the human variome. Database **2013**, bat019 (04 2013)
17. Yoshikawa, H., Nguyen, D.Q., Zhai, Z., Druckenbrodt, C., Thorne, C., Akhondi, S.A., Baldwin, T., Verspoor, K.: Detecting Chemical Reactions in Patents. In: Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association. pp. 100–110 (2019)
18. Zhai, Z., Nguyen, D.Q., Akhondi, S., Thorne, C., Druckenbrodt, C., Cohn, T., Gregory, M., Verspoor, K.: Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings. In: Proceedings of the 18th BioNLP Workshop. pp. 328–338 (2019)