Melaxtech: A report for CLEF 2020 – ChEMU Task of Chemical Reaction Extraction from Patent

Jingqi Wang[1] Yuankai Ren[2] Zhi Zhang[2] and Yaoyun Zhang[1]

[1] Melax Technologies, Inc, Houston, TX, USA
[2] Nantong University School of Medicine, Nantong, Jiangsu, China
Yaoyun.Zhang@melaxtech.com

Abstract. This work describes the participation of the Melaxtech team in the CLEF 2020 – ChEMU Task of Chemical Reaction Extraction from Patent. The task consisted of two subtasks: (1) Named entity recognition to identify compounds and different semantic roles in the chemical reaction. (2) Event extraction to identify event-triggers of chemical reaction and their relations with the semantic roles recognized in subtask 1. We developed hybrid approaches combining both deep learning models and pattern-based rules for this task. Our approaches achieved state-of-art results in both subtasks, with the best F1 of 0.957 for entity recognition and the best F1 of 0.9536 for event extraction, indicating the proposed approaches are promising.

Keywords: named entity recognition, event extraction, chemical reaction.

1 Introduction

New compound discovery plays a vital role in the chemical and pharmaceutical industry.[1] Characteristics of compounds, such as their reactions and experimental conditions are fundamental information for chemical research and applications.[2] The latest information of chemical reactions is usually present in patents, and is embedded in free text.[3] The rapidly accumulating chemical patents urge automatic tools based on natural language processing (NLP) techniques for efficient and accurate information extraction.[4]

Fortunately, the CLEF 2020 – ChEMU Task takes the initiative to promote the chemical reaction extraction from patent by providing benchmark annotation datasets. Two

important subtasks are setup in this challenge: chemical named entity recognition (NER) and chemical reaction event extraction. In particular, the annotation scheme of this benchmark data extends from previous challenges of chemical information extraction[5][6] to recognize multiple semantic roles of chemical substances in the reaction. Moreover, keywords event triggers and their relations with each semantic role are also annotated and provided for this task. The CLEF 2020 – ChEMU Task will greatly facilitate the development of automatic NLP tools for chemical reaction in patents with community efforts.[7][8]

This work describes our participation in both subtasks in CLEF 2020 – ChEMU. We developed hybrid approaches combining deep learning models and pattern-based rules for the information extraction systems. Our approaches achieved top rank in both subtasks, with the best F1 of 0.957 for entity recognition and the best F1 of 0.9536 for event extraction, indicating the proposed approaches are promising.

2 Methods

Dataset: The dataset provided by the CLEF 2020 – ChEMU Task was split into training data, development data and test data for open evaluation. For subtask 1, it was annotated with 10 entity type labels describing different semantic roles in chemical reaction, including EXAMPLE_LABEL, STARTING_MATERIAL, REAGENT_CATALYST, REACTION_PRODUCT, SOLVENT, TIME, TEMPERATURE, YIELD_PERCENT, YIELD_OTHER, and OTHER_COMPOUND. For subtask 2, the event trigger words (such as "addition" and "stirring") were annotated, which were further split into labels of "REACTION_SETUP" and “WORK_UP”. Their relations with different semantic roles were also annotated. Following the semantic proposition definition, the Arg1 type was used to mark the relation between event trigger words and compounds. ArgM represented the auxiliary role of the event and was used to mark the relation between the trigger word and the temperature, time or output entity.

Information extraction

Pre-processing In this step, patent text is segmented into sentences by sentence boundary detection. The tokens are also identified and separated by a tokenization tool based on lexicons and regular expressions. Modules of sentence segmentation and tokenization in the CLAMP software[9] were applied in this study.

Pre-training language model on patents Diverse expressions of chemical reaction information in the free text make them very sparse to be represented and modeled.[10] The semantic distributed representations (i.e. multi-dimension vectors of float values) of text generated by deep neural networks, or deep learning methods, alleviated the sparseness challenge by dramatically reducing the dimensions of language representation vectors using a non-linear space.[10] Specifically, language models pre-trained on large scale unlabeled datasets embed linguistic and domain knowledge that can be transferred to downstream tasks, such as NER and relation extraction.[11] In this study, Bi-

oBERT,[12] a pre-trained biomedical language model (a bidirectional encoder representation for biomedical text), was used as the basis for training a language model of patents. Based on Bert,[13] a language model pre-trained on large scale open text, BioBERT was further refined on using the biomedical literature in PubMed and PMC. Consequently, BioBERT outperforms Bert on a series of benchmark tasks for biomedical NER and relation extraction.[12] For this study, BioBERT was retrained using text files provided by CLEF 2020 – ChEMU to tailor the language model to patent data. For convenience, the pre-trained language model is referred as Patent_BioBERT.

Subtask 1 - Named entity recognition Semantic roles in chemical reactions are recognized using a hybrid method. First, Patent_BioBERT was fine-tuned using the Bi-LSTM-CRF (Bi-directional Long-Short-Term-Memory Conditional-Random-Field) algorithm. Next, several pattern-based rules were designed based on manual observation of the training and development datasets and used in a post-processing step. For example, rules were defined to differentiate STARTING_MATERIAL and OTHER_COMPOUND based on the relative positions and total number of EXAMPLE_LABEL labels. Specifically, the essential logic to determine whether the chemical mentions at the beginning of a text are STARTING_MATERIAL or OTHER_COMPOUND is actually to determine whether there is a hierarchical structure in the narrative to describe multiple steps of chemical reactions. If there are multiple example_labels present in the text, chemical mentions at the beginning of the entire text are usually the final chemical to be produced and should be labeled as OTHER_COMPOUND, while chemical mentions at the beginning of each later example label are STARTING_MATERIAL in each sub-step of chemical reactions.

Subtask 2 - Event extraction: This subtask contains two steps. For the step of event trigger detection, it was also a NER task, and was addressed with a similar approach as in subtask one. For the relation extraction task, given entities annotated in sentences, it can be transformed into a classification problem. A classifier can be built to determine categories of all possible candidate relation pairs (e_1, e_2), where entities e_1 and e_2 are from the same sentence. We generated candidate pairs by pairing each event trigger and semantic role. In order to represent a candidate event trigger and semantic role pair in an input sentence, we used the semantic type of an entity to replace the entity itself. The mentions of entities are directly generalized by their semantic types in the sentences. A linear classification layer was added on top of the Patent_BioBERT model to predict the label of a candidate pair in sentential context. As mentioned above, Patent_BioBERT was essentially built on the basis of BERT. In detail, BERT adds a classification token [CLS] at the beginning of a sentence input, whose output vector was used for classification. As typical with BERT, we used a [CLS] vector as input to the linear layer for classification. Then a softmax layer was added to output labels for the sentence. Furthermore, some event triggers and their linked semantic roles were present in different sentences, or different clauses in long complex sentences. Their relations were not identified using the deep learning-based model. Therefore, post-processing rules were designed based on patterns observed in the training data, and applied to recover some of these false negative relations.

Subtask 2 – End-2-End: Overall, a typical cascade, or pipeline model was built for the end-2-end system, in which semantic roles and event triggers were first recognized together in a NER model, their relations were then classified in a relation extraction model.

Evaluation

Precision, recall and F1 were used for performance evaluation, as defined in Equations 1-3. Both exact and relax matching results are reported. The primary evaluation metric was the F1 score of exact matching. We used 10-fold cross-validation on the merged training and development datasets to optimize parameters for the models.

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \quad (1)$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \quad (2)$$

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

10-fold cross-validation was conducted using the training process. The final set of hyperparameters and values used in the study are dropout_rate 0.2, max_seq_length 310, hidden_dim 128, learning rate 5e-5, batch-size 24. Based on this, we implemented three approaches for comparison:

1. Fine-tuning Patent_BioBert: Among the 10 models generated in the 10-fold cross-validation, the model with the highest performance on 1-fold was selected and used for submission.

2. Ensemble of output results from 10 models generated from the cross-validation using majority voting. Due to time limitation, no complex methods or optimization were made for the ensemble model. A simple majority voting result based on outputs from the 10 models was used for the ensemble submission.

3. Merge-data: fine-tuning Patent_BioBert using the merged training and development datasets.

3 Results

Performances on the test dataset for each task, namely NER, event extraction and the end-to-end information extraction are illustrated in Table 1-3, respectively. Promising results were obtained for the current approaches, with the highest F1 of 0.957 for NER, 0.9536 for event extraction and 0.9174 for the end-to-end system, respectively.

Moreover, the detailed performances of the fine-tuning method for each entity types and relation types, and their overall performances on the development set are reported in Table 4-5. The overall F1 is 0.942 for NER and 0.953 for relation extraction.

However, the ensemble systems and systems built on the merged data of training and development sets yielded lower performances than the system fine-tuned on a 90%-10% split of the gold standard data. Especially, the Merge-data systems got the lowest performances among all three approaches. One potential reason was that the hyperparameters used in the fine-tuning model were not the optimal set for the merged data. Due to time limitation, only majority voting was applied in the ensemble systems, more investigations into this direction are needed in our future work.

Unfortunately, sharp drops are obtained in the end-2-end performances of two proposed methods - Ensemble and Merge-data (Table 3). After checking the workflow, unexpected errors happened in these two runs in the post-processing stage. The lexicon of WORK_UP was mistakenly used to boost the recall of WORK_UP by recovering missing mentions. Since WORK_UP and Reaction_Step share many words, a large part of Reaction_Step were also replaced by WORK_UP. Among the three runs, the pipelines of Ensemble and Merge-data had mislabeled WORK_UP, and got poor performances (F-measures drop from above 90% from basic Fine-tuning method to around 20%). Luckily, this mistake was detected and fixed in our later submissions for relation extraction.

Table 1. Performances of named entity extraction for chemical reaction. Both exact and relaxed matching results are reported.

Method	Exact			Relax		
	Precision	Recall	F1	Precision	Recall	F1
Fine-tuning	0.9571	0.957	0.957	0.969	0.9687	0.9688
Ensemble	0.9587	0.9529	0.9558	0.9697	0.9637	0.9667
Merge-data	0.9572	0.951	0.9541	0.9688	0.9624	0.9656

Interestingly, performances of the exact and relaxed matching criteria did not have sharp differences, which indicated that limited boundary errors occurred in the NER step. This validated that the preprocessing modules in CLAMP could efficiently segment sentences and split tokens.

4 Discussion

Novel compound discovery is vital in the chemical and pharmaceutical industry. Chemical reaction is essential to rigorous understanding of compound for further research and applications.

Table 2. Performances of event extraction for chemical reaction. Both exact and relaxed matching results are reported.

Method	Exact			Relax		
	Precision	Recall	F1	Precision	Recall	F1
Fine-tuning	0.9568	0.9504	0.9536	0.958	0.9516	0.9548
Ensemble	0.9619	0.9402	0.9509	0.9632	0.9414	0.9522
Merge-data	0.9522	0.9437	0.9479	0.9534	0.9449	0.9491

Table 3. Performances of end-to-end systems for chemical reaction extraction. Both exact and relaxed matching results are reported.

Method	Exact			Relax		
	Precision	Recall	F1	Precision	Recall	F1
Fine-tuning	0.9201	0.9147	0.9174	0.9319	0.9261	0.9290
Ensemble	0.2394	0.2647	0.2514	0.2429	0.2687	0.2552
Merge-data	0.2383	0.2642	0.2506	0.2421	0.2684	0.2545

Our participation in the CLEF 2020 – ChEMU task answers to the urgent call for high-quality information extraction tools for chemical reaction information in patents. Evaluation based on the open test dataset demonstrated that the proposed hybrid approaches are promising, with top ranks in the two subtasks. Valuable lessons are also learned in this process:

A detailed error analysis was conducted for future system improvement. One major type of errors was the confusion between REAGENT_CATALYST and STARTING_MATERIAL or between REAGENT_CATALYST and SOLVENT. Information structures in sentences and context were not sufficient to differentiate these semantic types. Another major error was related to the event trigger recognition. Many false positive event triggers were recognized, and REACTION_STEP and WORKUP were often confused with each other, especially for words frequently present in different contexts (e.g., added, stirring). Failing to recognize named entities correctly also affected the next relation extraction step. As for relation extraction, the majority of errors were caused by long distance relations intra or inter sentences. Although rules were applied to fix such errors, they also brought false positive instances. The precision and

recall were examined carefully and balanced for each rule, only rules that could improve the performance with high confidences were kept in the system.

Table 4. Performances of named entity extraction for chemical reaction on the development set. Both performances on each entity type and the overall performance are reported. Performances of the fine-tuning method is reported.

Entity type	Exact		
	Precision	Recall	F1
EXAMPLE_LABEL	0.979	0.986	0.982
REACTION_PRODUCT	0.899	0.904	0.902
REACTION_STEP	0.952	0.944	0.948
STARTING_MATERIAL	0.896	0.926	0.911
YIELD_OTHER	0.99	0.965	0.977
YIELD_PERCENT	0.972	1	0.986
REAGENT_CATALYST	0.938	0.905	0.921
SOLVENT	0.963	0.93	0.946
TEMPERATURE	0.935	0.96	0.947
WORKUP	0.931	0.93	0.931
OTHER_COMPOUND	0.947	0.939	0.943
TIME	0.983	0.991	0.987
Overall	0.943	0.941	0.942

The motivation behind the three implemented methods is that it is interesting to examine if there is a space of performance improvement if majority voting or a larger training dataset is used. The three methods actually shared with the same set of hyper-parameters. The same set of hyper-parameters, based on our current interpretation, is a curse to the final performances. The majority-voting and merge-data methods did not generate better performances as originally expected. More investigations need to be conducted for these two methods, with an additional validation set for fine-tuning. Yet, the sensitivity of the hyper-parameters in deep learning models is a long-standing problem that needs even more efforts to be alleviated.

Table 5. Performances of chemical reaction extraction on the development set. Both performances on each relation type and the overall performance are reported.Performances of the fine-tuning method is reported.

Relation type	Exact		
	Precision	Recall	F1
ARG1|REACTION_STEP|OTHER_COMPOUND	0.733	0.805	0.767
ARG1|REACTION_STEP|REACTION_PRODUCT	0.985	0.948	0.966
ARG1|REACTION_STEP|REAGENT_CATALYST	0.979	0.965	0.972
ARG1|REACTION_STEP|SOLVENT	0.975	0.9522	0.968
ARG1|REACTION_STEP|STARTING_MATERIAL	0.957	0.916	0.936
ARG1|WORKUP|OTHER_COMPOUND	0.965	0.961	0.963
ARG1|WORKUP|REACTION_PRODUCT	0	0	0
ARG1|WORKUP|SOLVENT	0.2	1	0.333
ARG1|WORKUP|STARTING_MATERIAL	0	0	0
ARGM|REACTION_STEP|TEMPERATURE	0.957	0.928	0.942
ARGM|REACTION_STEP|TIME	0.978	0.926	0.952
ARGM|REACTION_STEP|YIELD_OTHER	0.984	0.942	0.962
ARGM|REACTION_STEP|YIELD_PERCENT	0.982	0.943	0.962
ARGM|WORKUP|TEMPERATURE	0.893	0.909	0.901
ARGM|WORKUP|TIME	0.7	1	0.824
ARGM|WORKUP|YIELD_OTHER	0	0	0
ARGM|WORKUP|YIELD_PERCENT	0	0	0
Overall	0.963	0.944	0.953

Comparisons between the performances with and without post-processing rules showed that the applied rules only contribute to slight improvements to the overall performances on the development set (NER: 0.9389 vs. 0.9421; Relation: 0.9526 vs. 0.9534), despite careful data analysis was conducted to find potential improvements from heuristics. This demonstrated the generalizability power of the pre-trained language model,

and also indicated that more investigations are needed for heuristics and knowledge-based improvement.

Limitations and future work: Although the proposed approaches obtained promising performances of chemical reaction extraction, there are several limitations and further improvements in next steps. (1) Firstly, domain knowledge of different semantic roles and their relations was not leveraged in the current study, such as lexicons of REAGENT_CATALYST and SOLVENT. This may potentially resolve the confusion among different semantic labels. (2) Secondly, dependency syntactic information was not applied in the current approaches, such as conjunctive structures and header-dependent patterns. Such information was proved to be effective for relation extraction and would be integrated into the deep learning models to further improve the performance. (3) The currently rules to fix errors in event triggers were data driven, which appeared to be ad hoc given the limited gold standard dataset. Data argumentation approaches[14] will be applied in the next step to enrich the training data and the coverage of different context patterns, so as to make a clearer differentiation among event triggers.

5 Conclusions

This work describes the participation of the Melaxtech team on the CLEF 2020 – ChEMU Task of Chemical Reaction Extraction from Patent. We developed hybrid approaches combining both deep learning models and pattern-based rules for this task. Our approaches achieved state of the art results in both subtasks, indicating the proposed approaches are promising. Further improvement will also be conducted in the near future by integrating domain knowledge and syntactic features into the current framework. Data augmentation will also be investigated for annotation enrichment in a cost-saving way, to further improve the system generalizability.

References

1. Akhondi, S. A., Rey, H., Schwörer, M., Maier, M., Toomey, J., Nau, H., Bobach, C. (2019). Automatic identification of relevant chemical compounds from patents. Database, 2019.
2. Akhondi, S. A., Klenner, A. G., Tyrchan, C., Manchala, A. K., Boppana, K., Lowe, D., Kors, J. A. (2014). Annotated chemical patent corpus: a gold standard for text mining. PloS one, 9(9), e107477.
3. Senger, S., Bartek, L., Papadatos, G., & Gaulton, A. (2015). Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. Journal of cheminformatics, 7(1), 1–12.
4. Muresan, S., Petrov, P., Southan, C., Kjellberg, M. J., Kogej, T., Tyrchan, C., Xie, P. H. (2011). 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.

5. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., & Valencia, A. (2015). CHEMDNER: The drugs and chemical names extraction challenge. Journal of cheminformatics, 7(S1), S1.
6. Krallinger, M., Rabal, O., Lourenço, A., Perez, M. P., Rodriguez, G. P., Vazquez, M., Valencia, A. (2015). Overview of the CHEMDNER patents task. In Proceedings of the fifth BioCreative challenge evaluation workshop (pp. 63–75).
7. Nguyen, D. Q., Zhai, Z., Yoshikawa, H., Fang, B., Druckenbrodt, C., Thorne, C., Verspoor, K. (2020). ChEMU: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. In J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, & F. Martins (Eds.), Advances in Information Retrieval (pp. 572–579). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-45442-5_74
8. He, J., Nguyen, D. Q., Akhondi, S. A., Druckenbrodt, C., Thorne, C., Hoessel, R., Verspoor, K. (2020). Overview of ChEMU 2020: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. In A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020) (Vol. 12260). Lecture Notes in Computer Science.
9. Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. (2018). CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association, 25(3), 331–336. https://doi.org/10.1093/jamia/ocx132
10. Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. Journal of Artificial Intelligence Research, 63, 743–788.
11. Liu, F., Chen, J., Jagannatha, A., & Yu, H. (2016). Learning for biomedical information extraction: Methodological review of recent advances. arXiv preprint arXiv:1606.07993.
12. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.
13. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
14. Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., & Jin, Z. (2016). Improved relation classification by deep recurrent neural networks with data augmentation. arXiv preprint arXiv:1601.03651.