Published in "CLEF 2020 conference", 22-25 September 2020, Thessaloniki, Greece, which should be cited to refer to this work.

Named entity recognition in chemical patents using ensemble of contextual language models

Jenny Copara[1,2,3], Nona Naderi[1,2], Julien Knafou[1,2,3], Patrick Ruch[1,2], and Douglas Teodoro[1,2]

[1] University of Applied Sciences and Arts of Western Switzerland, Geneva, Switzerland
[2] Swiss Institute of Bioinformatics, Geneva, Switzerland
[3] University of Geneva, Geneva, Switzerland
{firstname.lastname}@hesge.ch

Abstract. Chemical patent documents describe a broad range of applications holding key reaction and compound information, such as chemical structure, reaction formulas, and molecular properties. These informational entities should be first identified in text passages to be utilized in downstream tasks. Text mining provides means to extract relevant information from chemical patents through information extraction techniques. As part of the Information Extraction task of the Cheminformatics Elsevier Melbourne University challenge, in this work we study the effectiveness of contextualized language models to extract reaction information in chemical patents. We assess transformer architectures trained on a generic and specialised corpora to propose a new ensemble model. Our best model, based on a majority ensemble approach, achieves an exact F_1-score of 92.30% and a relaxed F_1-score of 96.24%. The results show that ensemble of contextualized language models can provide an effective method to extract information from chemical patents.

Keywords: Named-entity recognition, chemical patents, contextual language models, patent text mining, information extraction.

1 Introduction

Chemical patents represent a valuable information resource in downstream innovation applications, such as drug discovery and novelty checking. However, the discovery of chemical compounds described in patents is delayed by a few years [12]. Among the reasons, it could be considered the complexity of the chemical patent information sources [11], the recent increase in the number of chemical patents without manual curation, and the particular wording used in the domain. Narratives in chemical patents contain often concepts expressed in a way to protect or hide information, as opposed to scientific literature, for example,

where the text tends to be as clear as possible [34]. In this landscape, information extraction methods, such as Named Entity Recognition (NER), provide a suited solution to identify key information in patents.

NER aims to identify information of interest and their respective instances in a document [8, 24]. It has been often addressed as a sequence classification task, where a sequence of features, usually tokens, is used to predict the class of a text passage. One of the most successful approaches in sequence classification is Conditional Random Fields (CRF) [18, 32]. CRF was proposed to solve sequence classification problems by estimating the conditional probability of a label sequence given a word sequence, considering a set of observed features in the latter. It was established as the state-of-the-art in different NER domains for many years [19, 29, 20, 28, 9, 11, 37]. In the chemical patent domain, CRF was explored by Zhang *et al.* [39] in the CHEMDNER patent corpus [17]. Using a set of hand-crafted and unsupervised features derived from word embeddings and Brown clustering, their model achieved 87.22% of F_1-score. With similar F_1-score performance, Akhondi *et al.* [2] explored CRF combined with dictionaries in the biomedical domain in the tmChem tool [20] in order to select the best vocabulary for the CHEMDNER patent corpus. It has been shown [11] that recognizing chemical entities in the full patent text is a harder task than in titles and abstracts, due the peculiarities of the chemical patent text. Evaluation in full patents was performed using BioSemantics patent corpus [1] through neural approaches based on the Bidirectional Long-Short Term Memory (BiLSTM) CRF [10] and the BiLSTM Convolutional Neural Network (CNN) CRF [38] architectures, with performance of 82.01% and 85.68% of F_1-score, respectively. It is worth noting that for the first architecture [10], the authors used word2vec embeddings [23] to represent features, while in the latter [38], the authors used ELMo contextualized embeddings [26].

Over the years, neural language models have improved their ability to encode the semantics of words using large amounts of unlabeled text for self-supervised training. They have initially evolved from a straightforward model [3] of one hidden layer that predicts the next word in a sequence, aiming to learn the distributed representation of words (i.e., the word embedding vector), to an improved objective function that allows learning from larger amounts of text [4], using higher computational resources and with longer training time. These developments have encouraged the seeking of language models able to bring high-quality word embeddings with lower computational cost (i.e., word2vec [23] and Global Vectors (GloVe) [25]). However, natural language still presented challenges for language models, in particular, concerning word contexts and homonyms. More recently, a second type of word embeddings have attracted attention in the literature, the so-called contextualized embeddings, such as ELMo, UMLFiT [14], GPT-2 [27], and BERT [7]. Particularly, the BERT architecture uses the attention mechanism to train deep bidirectional token representations, conditioning tokens on their left and right contexts.

In this work, we explore contextualized language models to extract information in chemical patents as part of the Named Entity Recognition task of

the Information extraction from Chemical Patents (ChEMU) lab [12, 13]. Pretrained contextualized languages models, based on the BERT-based architecture, are used as baseline model and fine-tuned on the examples of the ChEMU NER task to classify tokens according to the different entities. In the challenge, the corpus was annotated with the entities: *example_label*, *other_compound*, *reaction_product*, *reagent_catalyst*, *solvent*, *starting_material*, *temperature*, *time*, *yield_other*, and *yield_percent*. We investigate the combination of different architectures to improve NER performance. In the following sections, we describe the design and results of our experiments.

2 Methods and data

2.1 NER model

Transformers with a token classification on top. We assess five language models based on the transformers architecture to classify tokens according to the named-entities classes. The first four models are variations of the BERT model in terms of size and tokenization: *bert-base-cased*, *bert-base-uncased*, *bert-large-cased*, and *bert-large-uncased*. These models were originally pretrained on a large corpus of English text extracted from BookCorpus [40] and Wikipedia, with different number of attention heads for the base and large types (12 and 16 respectively). The fifth pretrained language model assessed is ChemBERTa[1], a RoBERTa-based transformer architecture [22], trained on a corpus of 100k Simplified Molecular Input Line Entry System (SMILES) [35] strings from the ZINC benchmark dataset [15].

Our models consist of BERT models specialised for NER, with a fully connected layer on top of the hidden states of each token. They are fine-tuned on the ChEMU Task 1 dataset, using the train and development sets provided. The fine-tuning is performed with a sequence length of 256 tokens, a warmup proportion of 0.1 (percentage of warmup steps with respect to the total amount of steps), and a batch size of 32. The tokenization process is driven by the original model's tokenizer, i.e., for the BERT-based models, WordPiece [36] is applied, while for the RoBERTa-based model, Byte-Pair-Encoding [30] is applied. The Adam optimizer is employed to optimize network weights [16]. The first four language models are fine-tuned for 10 epochs and a learning rate of $3e-5$. For ChemBERTa model, we conduct a grid search over the development set and found the best performance around 29 epochs of fine-tuning and a learning rate of $4e-5$. The implementations are based on the Huggingface framework. [2]

Ensemble model. Our ensemble method is based on a voting strategy, where each model votes with its predictions and a simple majority of votes is necessary to assign the predictions [5]. In other words, for a given document, our models

[1] `https://github.com/seyonechithrananda/bert-loves-chemistry`
[2] `https://huggingface.co/transformers/`

infer their predictions independently for each entity, then, a set of passages that received at least a vote is taken into consideration for casting votes. This means that, for a given document and a given entity, we end up with multiple passages associated with a number of votes, then, again for a given entity, the ensemble method will predict as positive all the passages that get the majority of votes. Note that each entity is predicted independently and that the voting strategy does allow the fact that a passage could have been labeled as positive for multiple entities at once.

Finally, in order to decide on the optimal composition of the ensemble model, we used the development set and compute all possible ensemble predictions using the above methodology. As we had 7 models in total, we tried every possible combination from 2 to 7 models. We retained the ensemble composition with the best overall F_1-score and used it for the test set. Originally, the ensemble model giving the best F_1-score was combining *bert-large-uncased*, *bert-base-cased*, CRF, *bert-base-uncased* and the CNN model (5 models). However, due to the size of the test set (approximately 10k patent snippets), we had to discard the large models of the ensemble strategy due to their much higher algorithmic complexity and the time constraints. The retained models in the ensemble were then *bert-base-cased*, *bert-base-uncased* and the CNN model.

Baseline. We consider two models for our baseline: CRF and CNN. For the CRF model, a set of standard features in a window of ± 2 tokens are created without taking into account part-of-speech tags, neither gazetteers. The features used are token itself, lower-cased word, capitalization pattern, type of token (i.e., digit, symbol, word), 1-4 character prefixes/suffixes, digit size (i.e., size 2 or 4), combination of values (digit with alphanumeric, hyphen, comma, period), binary features for upper/lower-cased letter, alphabet/digit char and symbol. Please refer to [6, 9] for further details on the features used. The CRF classifier implementation relies on the *CRFSuite*.[3]

The CNN model [21] for NER relies on incremental parsing with Bloom embeddings, a compression technique for neural network models dealing with sparse high-dimensional binary-coded instances [31]. The convolutional layers use residual connections, layer normalization and maxout non-linearity. The input sequence is embedded in a vector compounded by Bloom embeddings modeling the characters, prefix, suffix and part-of-speech of each word. Convolutional filters of 1D are used over the text to predict how the next words are going to change. Our implementation relies on the spaCy NER module, [4] using the pretrained transformer *bert-base-uncased* for 30 epochs and a batch size of 4. During the test phase, we fixed the max size of the text to 1.5M due to RAM memory limitations.

[3] `http://www.chokkan.org/software/crfsuite/`
[4] `https://spacy.io`

2.2 Data

The data in ChEMU Task 1 (NER) is provided as snippets sampled from 170 English patents from the European Patent Office and the United States Patent and Trademark Office [12, 13]. Gold annotations were provided for training (900 snippets) and development (250 snippets) sets for a total of 20,186 entities. The annotation was done in the BRAT standoff format. Fig. 1 shows an example of a snippet with annotations for several entities, including *reaction_product* (two annotations), *starting_material* and *temperature*.

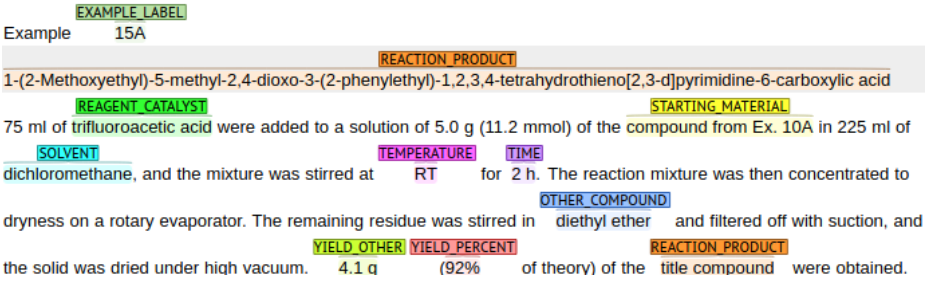

Fig. 1. Data example with annotations for the ChEMU NER task.

During the development phase, we used the official development set as our test set. The official training set was split into train and development sets in order to train the weights and tune hyper parameters of our models, respectively. As a result of this new setting, 800 snippets were available in train set, 100 in the development set and 225 in test set. Table 1 shows the entity distribution during the development phase. The majority of the annotations come from *other_compound*, *reaction_product* and *starting_material*, covering the 52% of entities in the development phase. In contrast, *example_label*, *time* and *yield_percent* entities represent 17% of entities in the development phase.

2.3 Evaluation metrics

The metrics used to evaluate the models are precision, recall, and F_1-score. As it can be seen in the example of Fig. 1, each entity has a span that is expected to be identified by the NER models as well as the correct entity type. The evaluation for the challenge is established under strict and relaxed span matching conditions [12, 13]. The exact matching condition takes into account the correct identification of both, span and entity type. On the other hand, the relaxed matching condition evaluates how accurate is the predicted span concerning the real. Our models are evaluated with the ChEMU web page system for the official results [5] and with the BRAT Eval tool for the offline analyses [6].

[5] `http://chemu.eng.unimelb.edu.au/`
[6] `https://bitbucket.org/nicta_biomed/brateval/src/master/`

Table 1. Entity distribution in the development phase based on the official training and development sets. Test set is the official development set. Dev set is random set extracted from the official training set.

Entity	**Train** (count/%)	**Dev** (count/%)	**Test** (count/%)	**All** (count/%)
example_label	784/5	102/5	218/6	1104/5
other_compound	4095/28	545/29	1080/28	5720/28
reaction_product	1816/13	236/12	506/13	2558/13
reagent_catalyst	1135/8	146/8	289/8	1570/8
solvent	1001/7	139/7	250/7	1390/7
starting_material	1543/11	211/11	413/11	2167/11
temperature	1345/9	170/9	346/9	1861/9
time	928/6	131/7	252/7	1311/6
yield_other	940/7	121/6	261/7	1322/7
yield_percent	848/6	107/6	228/6	1183/6
All	14435/100	1908/100	3843/100	20186/100

3 Results and discussion

In this section, we present the results of our models in the development and official test phases. Additionally, we perform error analyses on the results of the test set used in the development phase for some relevant models.

3.1 Model's performance in the development phase

Table 2 shows the exact and relaxed overall F_1-scores for all the models explored by our team in the development phase of the ChEMU NER task. As we can see, the ensemble model outperforms all the individual models for both exact and relaxed metrics. On the other hand, despite being trained on a specialised corpus, ChemBERTa achieves the lowest performance. The reported results come from the ChEMU official evaluation web page except for the CNN, *bert-large-uncased*, and the ensemble models, which are provided by the BRAT Eval tool.

Table 2. Performance of the different models in the development phase in terms of F_1-score. *models evaluated using the BRAT Eval tool.

Metric	**CRF**	**CNN***	**bert-base**		**bert-large**		**Chem**	**Ensemble***
			cased	**uncased**	**cased**	**uncased***	**BERTa**	
exact	0.8722	0.8182	0.9140	0.9113	0.9079	0.9052	0.6810	**0.9285**
relaxed	0.9450	0.8820	0.9732	0.9719	0.9706	0.9910	0.8500	**0.9876**

The results of all models with respect to the individual entities are presented in Table 3. As for the overall results, the ensemble model outperforms the individual models for all entities apart from *time*, for which the bert-base-cased

presents the best performance. The highest improvement for the ensemble model is seen for the *reaction_product* and *starting_material* entities with over 12-point increase in F_1-score. Considering only the individual models, the bert-base models outperform the other individual models, including the bert-large models, for all the entities, apart from *starting_material*, for which the CNN model has the best performance.

Table 3. Evaluation results on the development set for the exact F_1-score metric.

Entity	**CRF**	**CNN**	**bert-base**		**bert-large**		**Chem BERTa**	**Ensemble**
			cased	**uncased**	**cased**	**uncased**		
example_label	0.9630	0.9526	0.9862	0.9817	0.9793	0.9769	0.9631	**0.9885**
other_compound	0.8762	0.7409	0.8953	0.8938	0.8947	0.8925	0.7850	**0.9052**
reaction_product	0.7535	0.8425	0.8586	0.8515	0.8410	0.8427	0.5957	**0.8807**
reagent_catalyst	0.8330	0.8557	0.8595	0.8355	0.8498	0.8468	0.4673	**0.8946**
solvent	0.8949	0.7517	0.9447	0.9451	0.9407	0.9426	0.5945	**0.9545**
starting_material	0.7253	0.8229	0.8072	0.8153	0.7995	0.7813	0.4405	**0.8470**
temperature	0.9796	0.6397	0.9842	0.9842	0.9827	0.9841	0.8105	**0.9855**
time	0.9900	0.8533	**1.0000**	0.9941	0.9941	0.9941	0.8141	0.9980
yield_other	0.9046	0.9448	0.9905	0.9924	0.9811	0.9848	0.7135	**0.9943**
yield_percent	0.9913	0.9693	**0.9978**	**0.9978**	0.9913	0.9892	0.7131	**0.9978**

The ensemble model achieves the best performance for the *time*, *yield_other* and *yield_percent* entities. We believe this is due to the patterns observed for them in the training and test data. For example, for the *yield_percent* entity, the pattern is mostly a number followed by the percentage symbol ('%'). Similarly, for the *time* entity, the instances usually appear as a number followed for a time-indicator word. On the other hand, the *reaction_product*, *reagent_catalyst* and *starting_material* entities show the lowest performance, with 88.07%, 89.46% and 84.70% of F_1-score, respectively. These entities are of chemical types, often molecule strings (e.g., 4-(6-Bromo-3-methoxypyridin-2-yl)-6-chloropyrimidin-2-amine) [12, 13]. As our models did not include a post-processing step, as proposed in [33], these entities were sometimes recognized partially as a result of the language model sub-word tokenization process.

During the development phase, we also investigate the performance of ChemBERTa. As ChemBERTa is a language model trained on the chemical domain, it is expected to achieve competitive results. However, for the NER downstream task in chemical patents, the results go in a different direction. As shown in Table 3, ChemBERTa obtains the lowest results among all the explored models for both exact and relaxed metrics. We believe that the size of the corpus used to train the other explored language models has led to better chemical entity representations. Additionally, as the task aims to identify other entities than molecules, the ChemBERTa model naturally fails as its train set is only based on SMILES strings.

3.2 Model's performance in the test phase

In the official test phase, 9,999 files containing snippets from chemical patents were available for evaluating the models. We submitted 3 official runs: run 1, based on the baseline CRF model; run 2, based on the bert-base-cased model; and run 3, based on the ensemble model. Table 4 shows the official performance of our models for the exact and relaxed span matching metrics in terms of F_1-score. The ensemble model achieves 92.30% of exact F_1-score, yielding more than 11-point improvement over our baseline and at least 1-point improvement over the best individual contextualized language model (bert-base-cased). It outperforms run 1 and run 2 for all the entities in both exact and relaxed metrics. We believe that the performance difference between the CRF model and the ensemble model is due mostly to the fact that language models based on attention mechanisms are able to provide better contextual feature representations without the specific design of hand-crafted features as in the case of CRF.

Table 4. Official performance of our models in terms of F_1-score for the exact and relaxed metrics.

Entity	**CRF**		**bert-base-cased**		**Ensemble**	
	exact	**relaxed**	**exact**	**relaxed**	**exact**	**relaxed**
example_label	0.9190	0.9367	0.9617	0.9730	0.9669	0.9784
other_compound	0.8310	0.9029	0.8780	0.9608	0.8920	0.9653
reaction_product	0.6462	0.7689	0.8593	0.9378	0.8766	0.9322
reagent_catalyst	0.7598	0.8035	0.8791	0.9082	0.9022	0.9176
solvent	0.8299	0.8323	0.9444	0.9491	0.9541	0.9541
starting_material	0.4957	0.6752	0.8413	0.9343	0.8701	0.9394
temperature	0.9499	0.9688	0.9692	0.9902	0.9729	0.9877
time	0.9698	0.9843	0.9868	0.9967	0.9879	0.9978
yield_other	0.8984	0.8984	0.9799	0.9821	0.9842	0.9865
yield_percent	0.9705	0.9807	0.9936	0.9962	0.9974	0.9974
ALL	0.8056	0.8683	0.9098	0.9596	0.9230	0.9624

The 5-top best performing entities identified by our models are *example_label*, *temperature*, *time*, *yield_other*, *yield_percent*, which is similar to the results found in the development phase. For all of our submissions, the entity with lowest performance in the official test phase is *starting_material*, achieving 49.57%, 84.13% and 87.01% of exact F_1-score in the CRF, bert-base-cased and ensemble models, respectively. As we will see further in the error analyses section, this entity is often confused with the *reagent_catalyst* entity in the development phase. From the chemistry point of view, both starting material (reactants) and catalysts (reagents) entities are present at the start of the reaction, with the difference that the latter is not consumed by the reaction. These terms are often used interchangeably though, which could be the reason for the confusion. Despite the much larger size of the test set (approximately 10 times the size of the training

set), these results suggest that the test set has a similar entity distribution of the dataset provided in the development phase.

In Table 5 is shown a summary of the top ten official results, including our runs 2 and 3 (BiTeM team, ranked 6 and 7), the best model and the challenge baseline. If we consider the exact F_1-score metric, our ensemble model shows at least 3-point improvement from the ChEMU Task 1 NER baseline and more than 3-point behind the top 1. For the relaxed metric, our best model performs slightly better, showing more than 5-point improvement from the baseline and less than 1-point below the top system.

Table 5. Official BiTeM results compared to the best model and the BANNER baseline.

Rank	**Team**	**Precision**		**Recall**		**F_1-score**	
		exact	**relaxed**	**exact**	**relaxed**	**exact**	**relaxed**
1	Melaxtech	0.9571	0.9690	0.9570	0.9687	0.9570	0.9688
6	**BiTeM (run 3)**	0.9378	0.9692	0.9087	0.9558	0.9230	0.9624
7	**BiTeM (run 2)**	0.9083	0.9510	0.9114	0.9684	0.9098	0.9596
10	Baseline (BANNER)	0.9071	0.9219	0.8723	0.8893	0.8893	0.9053

The performance of the ensemble model for all entities on test set in terms of precision, recall and F_1-score for both exact match and relax is presented in Table 6. The best precision and recall for the exact match metric are achieved for the *yield_percent* entity, reaching 99.74% and 99.74%, respectively. Overall, precision is always above 93% for the relaxed metric and at least 88% for the exact metric.

Table 6. Performance of the ensemble model for all entities on the test set in terms of precision, recall and F_1-score

Entity	**Precision**		**Recall**		**F_1-score**	
	exact	**relaxed**	**exact**	**relaxed**	**exact**	**relaxed**
example_label	0.9711	0.9827	0.9628	0.9742	0.9669	0.9784
other_compound	0.9197	0.9730	0.8659	0.9578	0.8920	0.9653
reaction_product	0.8942	0.9367	0.8596	0.9277	0.8766	0.9322
reagent_catalyst	0.9268	0.9435	0.8790	0.8931	0.9023	0.9176
solvent	0.9620	0.9620	0.9463	0.9463	0.9541	0.9541
starting_material	0.8886	0.9545	0.8523	0.9247	0.8701	0.9394
temperature	0.9769	0.9901	0.9690	0.9852	0.9729	0.9876
time	0.9846	0.9956	0.9912	1.0000	0.9879	0.9978
yield_other	0.9776	0.9798	0.9909	0.9932	0.9842	0.9865
yield_percent	0.9974	0.9974	0.9974	0.9974	0.9974	0.9974

Lastly, our CRF baseline achieves 80.56% of exact F_1-score, while the competition baseline, which is based also on CRF, but customized for biomedical NER,

taking into account features, such as part-of-speech, lemma, Roman numerals, names of the Greek letters, achieves 88.93% [19]. Indeed, we believe those features give the advantage to the competition baseline as they could better characterize chemical entities.

3.3 Error analysis

As the gold annotations for the test set are not available, we perform the error analysis on the official development set (used as our test set in the development phase, see Table 1). Fig. 2 shows the confusion matrix for the ensemble predictions for the exact metric. As we can see, most confusion occurred for the *starting_material* entity, which is mostly confused with *reagent_catalyst*, and for the *reaction_product* entity, which is mistaken for *other_compound*. As mentioned previously, these entities - material/reactant and catalyst/reagent, and product/compound - are often used interchangeably in chemistry passages, which is likely the reason for the model's confusion.

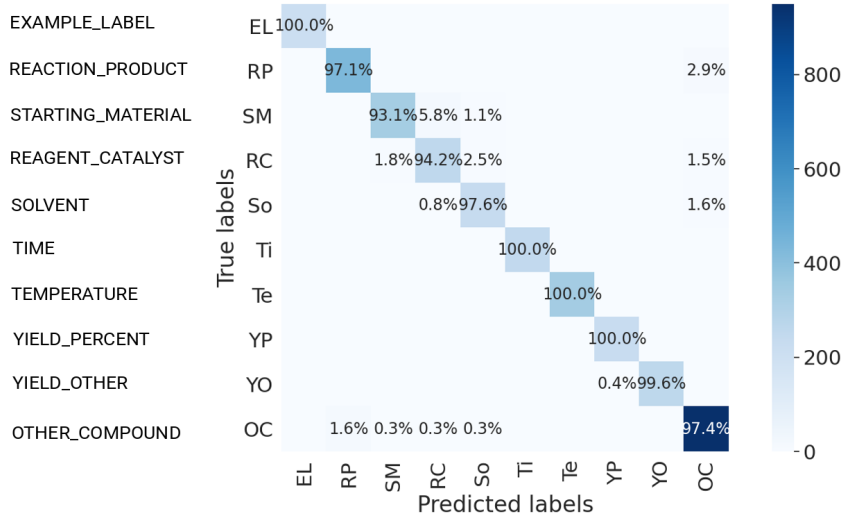

Fig. 2. Normalized confusion matrix for the ensemble model predictions on the official development set. Only exact matches are considered.

The error analysis of the incorrectly identified spans by the ensemble model shows that in almost 78.8% of the cases, the predicted entity was longer in length, for example, *sodium thiosulfate aqueous* instead of *aqueous* and *concentrated hydrochloric acid* instead of *hydrochloric acid*. The entities that are partially detected are mainly *starting_material*, which is inconsistently annotated in some

cases, as *Intermediate 13/6/21* (predicted as *13/6/21* by the ensemble model), and in some cases as only the number, such as *3* (predicted as *Intermediate 3* by the ensemble model). 42.3% of the span errors were multi-word entities.

Fig. 3 shows how different models detected a reagent catalyst entity described by a long text span. It seems that entities with longer text span, such as *reagent_catalyst*, *other_compound*, *reaction_product*, and *starting_material*, are less likely to be correctly detected by the contextualized language models. The bert-large-uncased and ChemBERTa models did not detect any token of the entity while both bert-large-cased and bert-base-cased models were able to only partially detect the entity. Particularly, the larger nature of the BERT large models was not translated into more effective representations for these entities.

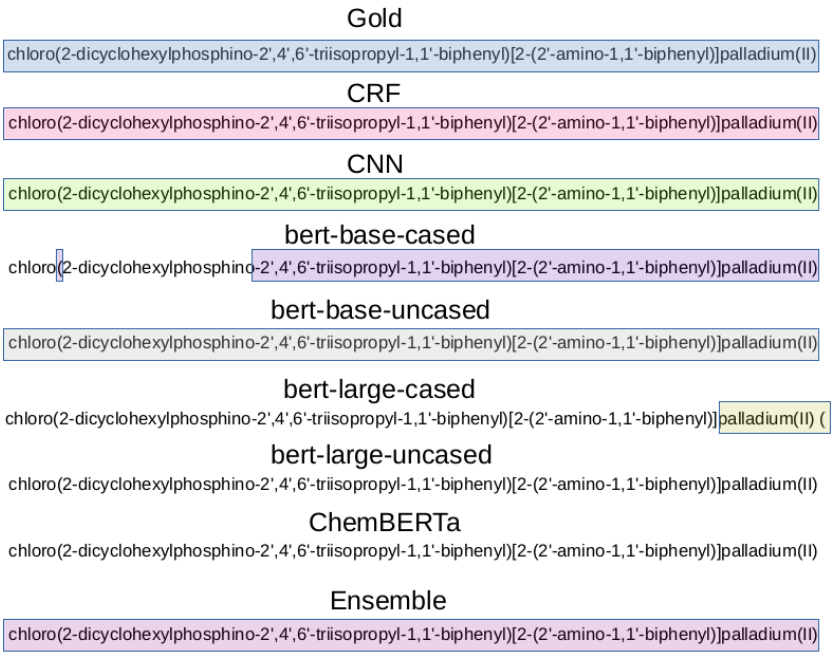

Fig. 3. An example of predictions by different models for (reagent_catalyst) annotation. The span detected by each model is color-coded.

Figure 4 shows the comparison of the span errors of the ensemble and BERT-base-cased models based on the length of entities (in character). While most errors of both models are focused on smaller entities, the BERT-base-cased model makes more mistakes than the ensemble model in detecting the spans of the longer entities. We believe this effect could be also related to the sub-word tokenization process of transformers. The combination of models smooths the effect in the ensemble model.

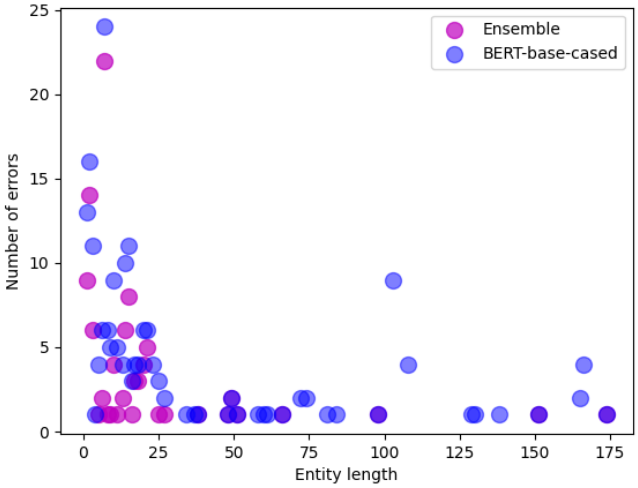

Fig. 4. Number of span errors by the ensemble and BERT-base-cased models based on the length of the entities (in character).

4 Conclusions

In this task, we explored the use of contextualized language models based on the transformer architecture to extract information from chemical patents. The combination of language models resulted in an effective approach, outperforming the baseline CRF model but also individual transformer models. Our experiments show that without extensive pre-training in the patent chemical domain, the majority vote approach is able to leverage distinctive features present in the English language, achieving 92.30% of exact F_1-score in the ChEMU NER task. It seems that the transformer models are able to take advantage of natural language contexts in order to capture the most relevant features without supervision in the chemical domain. Our next step will be to investigate pre-trained models on large chemical patent corpora to further improve the NER performance.

References

1. Akhondi, S.A., Klenner, A.G., Tyrchan, C., Manchala, A.K., Boppana, K., Lowe, D., Zimmermann, M., Jagarlapudi, S.A.R.P., Sayle, R., Kors, J.A., Muresan, S.: Annotated chemical patent corpus: A gold standard for text mining. PLoS ONE **9**(9), e107477 (Sep 2014)
2. Akhondi, S.A., Pons, E., Afzal, Z., van Haagen, H., Becker, B.F., Hettne, K.M., van Mulligen, E.M., Kors, J.A.: Chemical entity recognition in patents by combining dictionary-based and statistical approaches. Database **2016** (2016)

3. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. Journal of machine learning research **3**(null), 1137–1155 (Mar 2003)
4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research **12**, 2493–2537 (Nov 2011)
5. Copara, J., Knafou, J., Naderi, N., Moro, C., Ruch, P., Teodoro, D.: Contextualized French Language Models for Biomedical Named Entity Recognition. In: Cardon, R., Grabar, N., Grouin, C., Hamon, T. (eds.) 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes. pp. 36–48. ATALA, Nancy, France (2020)
6. Copara, J., Ochoa Luna, J.E., Thorne, C., Glavaš, G.: Spanish NER with word representations and conditional Random Fields. In: Proceedings of the Sixth Named Entity Workshop. pp. 34–40. Association for Computational Linguistics, Berlin, Germany (Aug 2016)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
8. Grishman, R.: Twenty-five years of information extraction. Natural Language Engineering **25**(06), 677–692 (Sep 2019)
9. Guo, J., Che, W., Wang, H., Liu, T.: Revisiting embedding features for simple semi-supervised learning. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 110–120. Association for Computational Linguistics, Doha, Qatar (Oct 2014)
10. Habibi, M., Weber, L., Neves, M., Wiegandt, D.L., Leser, U.: Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics **33**(14), i37–i48 (Jul 2017)
11. Habibi, M., Wiegandt, D.L., Schmedding, F., Leser, U.: Recognizing chemicals in patents: A comparative analysis. Journal of Cheminformatics **8**(1) (Oct 2016)
12. 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: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (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 (2020)
13. He, J., Nguyen, D.Q., Akhondi, S.A., Druckenbrodt, C., Thorne, C., Hoessel, R., Afzal, Z., Zhai, Z., Fang, B., Yoshikawa, H., Albahem, A., Wang, J., Ren, Y., Zhang, Z., Zhang, Y., Hoang Dao, M., Ruas, P., Lamurias, A., M. Couto, F., Copara, J., Naderi, N., Knafou, J., Ruch, P., Teodoro, D., Lowe, D., Mayfield, J., Köksal, A., Dönmez, H., Özkırımlı, E., Özgür, A., Mahendran, D., Gurdin, G., Lewinski, N., Tang, C., T.McInnes, Bridget C.S., M., RK Rao., P., Lalitha Devi, S., Cavedon, L., Cohn, T., Baldwin, T., Verspoor, K.: An extended overview of the clef 2020 chemu lab: Information extraction of chemical reactions from patents. In: Proceedings

of the Eleventh International Conference of the CLEF Association (CLEF 2020) (2020)
14. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
15. Irwin, J.J., Shoichet, B.K.: Zinc – a free database of commercially available compounds for virtual screening. Journal of Chemical Information and Modeling **45**(1), 177–182 (2005), pMID: 15667143
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
17. Krallinger, M., Rabal, O., Lourenco, A., Perez, M., Pérez-Rodríguez, G., Vazquez, M., Leitner, F., Oyarzabal, J., Valencia, A.: Overview of the CHEMDNER patents task. Proceedings of the Fifth BioCreative Challenge Evaluation Workshop pp. 63–75 (01 2015)
18. 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. p. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
19. Leaman, R., Gonzalez, G.: Banner: An executable survey of advances in biomedical named entity recognition. In: Altman, R.B., Dunker, A.K., Hunter, L., Murray, T., Klein, T.E. (eds.) Pacific Symposium on Biocomputing. pp. 652–663. World Scientific (2008)
20. Leaman, R., Wei, C.H., Lu, Z.: tmChem: a high performance approach for chemical named entity recognition and normalization. Journal of Cheminformatics **7**(S1) (Jan 2015)
21. Lecun, Y.: Generalization and network design strategies. Technical Report CRG-TR-89-4, University of Toronto (June 1989)
22. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019)
23. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. p. 3111–3119. NIPS'13, Curran Associates Inc., Red Hook, NY, USA (2013)
24. Okurowski, M.E.: Information extraction overview. In: TIPSTER TEXT PROGRAM: PHASE I: Proceedings of a Workshop held at Fredricksburg, Virginia, September 19-23, 1993. pp. 117–121. Association for Computational Linguistics, Fredericksburg, Virginia, USA (Sep 1993)
25. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (2014)
26. 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). Association for Computational Linguistics (2018)
27. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)

28. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009). pp. 147–155. Association for Computational Linguistics, Boulder, Colorado (Jun 2009)
29. Rocktäschel, T., Weidlich, M., Leser, U.: ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics **28**(12), 1633–1640 (Apr 2012)
30. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1715–1725. Association for Computational Linguistics, Berlin, Germany (Aug 2016)
31. Serrà, J., Karatzoglou, A.: Getting Deep Recommenders Fit: Bloom Embeddings for Sparse Binary Input/Output Networks. In: Proceedings of the Eleventh ACM Conference on Recommender Systems. p. 279–287. RecSys '17, Association for Computing Machinery, New York, NY, USA (2017)
32. Sutton, C.: An introduction to Conditional Random Fields. Foundations and Trends® in Machine Learning **4**(4), 267–373 (2012)
33. Teodoro, D., Gobeill, J., Pasche, E., Ruch, P., Vishnyakova, D., Lovis, C.: Automatic ipc encoding and novelty tracking for effective patent mining. In: The 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access (2010)
34. Valentinuzzi, M.E.: Patents and scientific papers: Quite different concepts: The reward is found in giving, not in keeping [retrospectroscope]. IEEE Pulse **8**(1), 49–53 (2017)
35. Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences **28**(1), 31–36 (Feb 1988)
36. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv (2016)
37. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 2145–2158. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018)
38. 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 and Shared Task. pp. 328–338. Association for Computational Linguistics, Florence, Italy (Aug 2019)
39. Zhang, Y., Xu, J., Chen, H., Wang, J., Wu, Y., Prakasam, M., Xu, H.: Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning. Database **2016** (2016)
40. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). p. 19–27. IEEE Computer Society, USA (2015)