Neural Medication Extraction: A Comparison of Recent Models in Supervised and Semi-supervised Learning Settings

Ali Can Kocabiyikoglu, Jean-Marc Babouchkine, Raheel Qader, François Portet

Neural Medication Extraction: A Comparison of Recent Models in Supervised and Semi-supervised Learning Settings

Ali Can Kocabiyikoglu,
Jean-Marc Babouchkine
Calystene SA, 38320 Eybens, France
a.kocabiyikoglu@calystene.com,
jm.babouchkine@calystene.com

François Portet
Univ. Grenoble Alpes, CNRS, Grenoble INP
LIG F-38000 Grenoble France
francois.portet@imag.fr

Raheel Qader
Lingua Custodia, Paris, France
raheel.qader@gmail.com

***Abstract*—Drug prescriptions are essential information that must be encoded in electronic medical records. However, much of this information is hidden within free-text reports. This is why the medication extraction task has emerged. To date, most of the research effort has focused on small amount of data and has only recently considered deep learning methods. In this paper, we present an independent and comprehensive evaluation of state-of-the-art neural architectures on the I2B2 medical prescription extraction task both in the supervised and semi-supervised settings. The study shows the very competitive performance of simple DNN models on the task as well as the high interest of pre-trained models. Adapting the latter models on the I2B2 dataset enables to push medication extraction performances above the state-of-the-art. Finally, the study also confirms that semi-supervised techniques are promising to leverage large amounts of unlabeled data in particular in low resource setting when labeled data is too costly to acquire.**

***Index Terms*—Medication information extraction, Natural Language Processing, Natural Language Understanding**

I. Introduction

In electronic health records (EHR) and other medical documents, drug information is often recorded in clinical notes, making it difficult for computerized applications to access this information as part of daily health care. Automatically extracting structured information related to drug prescriptions from medical free-texts is known as the medication extraction task. This task has attracted attention from the NLP community since the emergence of the I2B2 2009 medication extraction challenge [1]. In this challenge, the goal was to automatically label chunks of medication information from a whole clinical document. Figure 1 (left) shows an excerpt of a discharge summary in I2B2 from which information such as prescription duration and medication name should be extracted. Such kind of medication extraction system could be very useful to medical prescription writing software that are used to reduce the number of errors during the prescription, the transcription and the administration process of drugs.

Since the I2B2 2009 challenge, most work has used rule-based systems with the exception of a few hybrid ones. Some recent approaches using deep neural networks have shown promising results [2]–[4]. However, the task lacks a comprehensive comparison of deep learning methods over more traditional methods. Furthermore, the data provided within the I2B2 challenge is still very small for the needs of deep model training. Hence, methods able to leverage large amount of unlabeled clinical data (e.g., MIMIC III [5]) should be evaluated on this task. Among these methods, the recent pre-trained models have not been systematically studied for the task. Furthermore, recent semi-supervised techniques have also been rarely applied [6], [7].

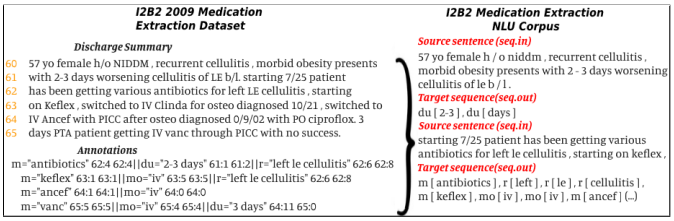

Fig. 1. Left: an I2B2 discharge summary excerpt and its medication annotations. Right: same excerpt segmented by sentence and formatted for sequence-to-sequence processing. (m = name; do = dose; mo = mode of administration; f = frequency; du = duration; r= reason)

In this paper, we explore the benefit of pre-trained models and semi-supervised learning to leverage non-annotated clinical documents for deep learning models. Indeed, for such a task, data annotation process requires medical experts and often patient data that has to be anonymised. This makes the process costly and limits the distribution of datasets. This is why methods resistant to OOV words such as BPE (Byte Pair Encoding) must be evaluated. Indeed, even in such a narrow drug prescription domain, vocabulary size and rare words for the input sequence are numerous [8]. This is why methods resistant to OOV words such as BPE (Byte Pair Encoding) must be evaluated.

Paper contributions. Our objective is to provide a comprehensive evaluation of the standard deep learning seq2seq methods (including Transformers) on the I2B2 2009 medication extraction task. We include a focused evaluation of pre-trained models and semi-supervised training to leverage unannotated medical corpora. The results show performances

above the state-of-the-art for pre-trained models and competitive performances of semi-supervised training.

Outline. The remaining of the paper presents a short review of the state-of-the-art before describing the methodology to pre-process the corpora and to perform supervised and semi-supervised learning. The experiment section then shows that the BlueBert pre-trained model can reach performance beyond the current state of the art. We then finish the paper by a short discussion and presentation of future work.

II. Related Work

The only known benchmarking efforts in medication extraction are the I2B2 2009 challenge [1] along with n2c2 shared task [9]. It is only recently that deep neural networks have superseded rule-based or hybrids models. Most significant progress has been made both thanks to deep models and also to the ability to leverage larger amount of data either through pre-trained embedding or through semi-supervised training. Regarding the pre-trained models, word embeddings trained on the MIMIC-III database have been used to improve slightly the state-of-the-art on the I2B2 medication extraction task [10]. The use of more performing pre-trained models, such as ELMo or BERT [11], [12] for word embeddings, have become prevalent in the biomedical NLP domain exceeding benchmarks on certain tasks [3], [13], [14]. For instance, Yang *et al.* [15], have compared 4 transformers models BERT, ALBERT, RoBERTa, and ELECTRA on a relation extraction task and showed the definite superiority of those. In a similar vein, Peng *et al.* [16] have proposed a benchmark setting (called BLUE) to evaluate pre-trained model in a clinical setting, showing that BERT model pre-trained on PubMed abstracts and MIMIC-III was superior (we refer to it as BlueBert in the rest of the paper). Recently, biomedical contextual embeddings have also been applied to improve the performance of adverse drug identification and the medication extraction task [9], [17].

From a semi-supervised point of view, Tao *et al.* [6] have proposed a semi-supervised system that achieved the current best overall performance on the I2B2 2009 medication extraction task by leveraging the non-annotated part of corpus (and also by using human annotations). More recently, Guzman *et al.* [18] proposed a system based on a LSTM model and transfer learning claiming state-of-the-art performance on extracting specific entities of the I2B2 dataset. However, their system is only partially sketched and thus difficult to reproduce. On other tasks, such as biomedical relation extraction, variational autoencoders [19] or event feature coupling generalization (EFCG) [20] have been proposed to benefit from unannotated datasets. In particular, [21] show that for the biomedical relation extraction task, a distantly supervised approach enables to produce large amounts of labeled but noisy data can be leveraged efficiently for data-driven approach. Despite, these recent progress, semi-supervised learning for medication extraction has only been applied by Tao *et al.* [6], which is to-date the state-of-the-art.

From this brief related work section, it is clear that deep neural network methods have shown significant progress in several bioNLP tasks. However, the ability to leverage larger amount of data has not been fully explored on the I2B2-2009 medication extraction task using recent methods. This paper is an attempt to not only propose a model taking advantage of the most recent NLP advances but also to provide comprehensive evaluation of state-of-the-art models on this task.

III. Method

The medication extraction problem is often addressed as a sequence labeling task requiring aligned BIO format. This is extremely costly to annotate and it prevents models from abstracting since they must stick to the surface form of the textual input. In this work, we approach the problem of medication extraction through encoder-decoder seq2seq models where input text is first abstracted by the encoder and then the labels and values are generated by the decoder. Figure 1 shows on the left side an excerpt of an I2B2 discharge summary. In that case the annotations consist of slot values indexed by position within the document (line:character index). The right side of Figure 1 shows the result of the I2B2 sentences adaptation. The document was sentence segmented and for each sentence the annotation was unaligned. This format is more realistic with respect what can be found in real clinical data were free-text segments might be only loosely related to electronic records.

A. Corpora and data pre-processing

The medication extraction models were trained on two publicly available datasets: the I2B2 2009 medication extraction dataset and the MIMIC-III dataset.

1) I2B2 dataset: The I2B2 dataset is composed of 1243 clinical documents. The annotation guidelines established for the I2B2 challenge focused on identifying the name of medication, their dosage, their mode of administration, frequency, duration and reason for administration in discharge summaries [1]. In the official distribution, 10 documents were annotated by medical experts and 251 by the research community. These latter 251 documents became the official test corpus for the evaluation [1]. The remaining 982 documents were un-annotated.

2) MIMIC-III unannotated dataset: MIMIC-III [5] is a large clinical dataset that does not have annotation data fit for supervised machine learning. But it does contain medication information both in the textual part of the dataset and in the database part (medication records). For the need of the semi-supervised learning, this corpus required some pre-processing to extract the relevant pieces of information related to drug prescriptions both from the textual point of view and the database point of view. To do so, we set up a simple algorithm: For each prescription in the database, the lines of the discharge summary of the same a patient's id were scored using regular expressions. Take the example Figure 2. The right side shows an extract of the prescription table while the left side shows some sentences of the patient's discharge summary that were scored. The line with 5pts had matched the most features of the Tacrolimus line. Thus, both the sentence and the prescription

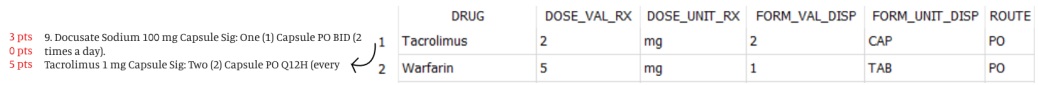

Fig. 2. Example matching of medication records database and sentences from MIMIC-III clinical texts

line are added to the annotated dataset. Prescriptions not matching any line of the discharge summary are not added to the unannotated dataset.

Data Split	I2B2-2019				MIMIC-III
	supervised			semi-supervised	semi-supervised
	community annotations	expert annotations	unannotated set	held out unannotated set	clinical notes
train	-	-	22,907 sents	4,655 sents	257,811 sents
validation		148 sents	2,158 sents	-	2,605 sents
test	4411 sents			-	-

TABLE I
FINAL DISTRIBUTION OF THE CORPORA FOR SUPERVISED AND UNSUPERVISED LEARNING.

At the end of the process, 962,252 lines of sentences related to the database records were extracted. It was then further filtered to remove similar examples as well as too long ones to obtain a final corpus of 260,416 loosely coupled sentences and database rows. It must be noted that the semantic information in the MIMIC III database is different from the I2B2 one. For instance, there is no reason or frequency in the MIMIC III records. However, such kind of information is often present in the textual part of MIMIC III. Thus, it can be concluded that this corpus is relevant for the I2B2 extraction task.

3) Preprocessing and Final Datasets Distribution: Since the seq2seq approach works at the sentence level, we extracted every medication sentence from the raw text of the MIMIC-III and i2b2 datasets using the *ClarityNLP* toolkit. The tokenization was based on the *Spacy* library. The language model was initialized with BERT pre-trained embeddings and then, we applied the sentence segmentation specialized for clinical documents to obtain a sentence-level segmentation and extracted the annotations in a seq2seq format as exemplified Figure 1.

To deal with out-of-vocabulary words (OOV), BPE (Byte Pair Encoding) codes [22] were learned from the MIMIC-III and I2B2 (test set excluded) text corpora.

The final distribution of the corpus is presented Table I. In our study, we used the official test set of 251 documents (4411 sentences extracted) annotated by the community to evaluate all the models. The train set was composed of 90% of the I2B2 unannotated documents. In the supervised learning approach, we used the freely-available MedExtractor system [23] which gave the second-best overall f-measure in the I2B2 challenge [1] to automatically annotate them. In the semi-supervised approach, we used a subset of the unannotated documents. The development set was composed of the 10 documents annotated by experts plus the remaining 10% of the unannotated documents that were automatically annotated. Regarding MIMIC, it was only used in a semi-supervised setting. Overall, the vocabulary size of MIMIC was 43k words while I2B2 was 18k words.

B. Supervised Methods

Following the recent deep learning methods applied on the I2B2 medication extraction task, we trained the initial encoder-decoder models using simple bi-directional LSTM models with attention [24]. This model is able to learn short and long dependencies in the input and can be trained on a reasonable amount of data. It is also surprisingly effective. We also included CNN models, since there are able to capture hierarchical relations between words and are quite efficient to train. We implemented the convolutional seq2seq model (conv-s2s) of Gehring *et al.* [25]. Finally, we included a transformer model [26] which are the current groundbreaking models.

For the pre-trained word embeddings models, we explored a BERT based model [12]. Furthermore, recently Zhu *et al.* [27] proposed a new algorithm for neural machine translation in which they exploit the BERT embeddings by extracting representations for an input sequence, and then fusing with each layer of the encoder and decoder through the attention mechanism. We called this model *bert-fused transformer*. Finally, since in many bioNLP tasks transformer-based embeddings models pretrained on clinical data have established new baselines [13], we included Biobert [13], clinical-bert [28] and BlueBert [16].

C. Semi-Supervised Approach

For the semi-supervised learning we used the approach of Qader *et al.* [7]. The approach considers two encoder-decoder models : one to extract semantics from text – called the Natural Language Understanding (NLU) model – and one to generate text from a semantic input – called the natural Language Generation (NLG) model. The approach considers three sets: a paired set of texts with their annotation, a unpaired set of texts (alone) and a unpaired set of semantics annotation (alone). The paired dataset is used to learn in a supervised manner both the NLU and NLG models. The unpaired sets are used by the two modules together. The text (resp. semantic) input is fed to the NLU (resp. NLG) models which outputs a semantic representation (res. a text) which is in turn send to the NLG (resp. NLU) which outputs a text (resp. a semantic representation). The difference between the input and output texts (resp. semantic) is used as a loss to optimize the two modules jointly. In this way, data that is not *paired* with annotation can be used for learning using this 'reconstruction' objective.

As NLU and NLG models are jointly learned, the losses of the NLG and NLU models for both paired and unpaired models could be denoted respectively as ${L^{\text{nlg}}}_{paired}$, ${L^{\text{nlu}}}_{paired}$, ${L^{\text{nlg}}}_{unpaired}$ and ${L^{\text{nlu}}}_{unpaired}$. These four losses are mixed together to perform the joint learning $L = \alpha\ {L^{\text{nlg}}}_{paired} +$

Model	F1	m	do	mo	f	du	r
LSTM*	0.78	**0.94**	**0.92**	0.93	0.89	**0.49**	0.50
LSTM(bpe)*	0.75	0.88	0.90	0.91	0.88	0.46	0.46
conv-s2s (bpe)†	0.68	0.87	0.83	0.84	0.76	0.38	0.41
transformer (bpe)†	0.75	0.92	0.88	0.89	0.84	0.47	0.50
Pre-trained Model	**F1**	**m**	**do**	**mo**	**f**	**du**	**r**
bert-base†	0.63	0.85	0.85	0.82	0.83	0.28	0.17
bert-fused -transformer†	0.74	0.90	0.87	0.89	0.83	0.47	0.50
clinical-bert†† base	0.75	0.82	0.76	0.75	0.76	0.33	0.45
biobert-base††	0.75	0.82	0.76	0.75	0.76	0.30	0.44
bluebert-†† base	**0.88**	0.92	0.88	**0.95**	**0.91**	0.46	**0.61**

TABLE II
F-MEASURE OF DIFFERENT MODELS ON THE I2B2 MEDICATION EXTRACTION TEST DATA. (*:SEQ2SEQ-PY LIBRARY , †:FAIRSEQ LIBRARY , ††: HTTPS://GITHUB.COM/THILINARAJAPAKSE/SIMPLETRANSFORMERS)

$\beta\ {L^{\text{nlu}}}_{paired} + \gamma\ {L^{\text{nlg}}}_{unpaired} + \delta\ {L^{\text{nlu}}}_{unpaired}$ where α , β , γ and $\delta \in [0, 1]$ are fine tuned empirically.

IV. EXPERIMENTS AND RESULTS

All the experiments were performed with two open-source seq2seq libraries for the experiments *seq2seq-py* from [29] and fairseq library from [30]. The vocabulary size of the training corpus was around 18k tokens without BPE and 10k with BPE. Seq2Seq-py configurations used negative log-likelihood (NLL) loss, with Adam optimizer. Learning rate was set to 0.001 with 2 bi-directional encoder-decoder layers, hidden size of 128 and 500 as embedding dimension. The dropout was set to 0.2 and gradients clipping was set to 2.0. For the fairseq experiments, LSTM architecture used cross entropy loss and nesterov accelerated gradient (NAG) as optimizer. Learning rate was fixed to 0.25 with 4 bi-directional encoder-decoder layers. Other hyper-parameters were kept the same according to the registered model configurations of fairseq library. We kept the name of the registered architectures of fairseq for reproducibility. The training continued up to 70 epochs for each model and the best model was chosen according to the validation loss.

Table II provides the overall macro average F-measure as well as those of slot labels for all models on the I2B2 test set (see Figure 1 for the meaning of each slot label). The upper part of the table contains the results of the standard supervised models: LSTM, CNN and Transformer. To deal with OOV and vocabulary size, most of these models has been tried with BPE leading to four models (more have been evaluated but were not reported due to lack of space).

It is clear that the simple LSTM with attention is difficult to beat since it got the highest F-measure for all slots. This might be due to the narrow domain and the lack of training data. For the same reason, using BPE does not bring any improvement. CNN and Transformer in NLP has been applied to domains with large training datasets. In this particular low-resource task, they failed to be efficient.

Use of pre-trained embeddings leads to diverse performances. `bert-base-uncased` did not succeed in specializing enough while `bert-fused` only reached comparable performances with standard supervised methods. Again, the lack of data might explain these low performance of the general purpose pre-trained models. However, Blue Bert (`bluebert-base-uncased`) which has been specifically pre-trained on medical documents (PubMed and MIMIC) reached the best F-measure (88%) and was particularly performing for the labels frequency (f) and reason (r) which are known to be particularly difficult.

corpus	α	β	γ	δ	F1	m	do	mo	f	du	r
mimic-bpe	1	0.1	1	0.1	0.55	0.83	0.77	0.79	0.75	0.37	0.39
mimic-no-bpe	1	0.1	1	0.1	0.64	**0.92**	0.88	0.90	0.85	**0.47**	**0.46**
i2b2-bpe	1	0.1	1	0.1	0.73	0.90	0.88	0.89	0.87	0.46	0.42
i2b2-no-bpe	1	0.1	1	0.1	**0.74**	0.91	**0.89**	**0.91**	**0.87**	0.43	0.44

TABLE III
F-MEASURE OF THE SEMI-SUPERVISED MODELS ON THE I2B2 TEST SET USING THE SIMPLE LSTM MODEL ARCHITECTURE.

Regarding the semi-supervised experiments, they have been based on the best simple LSTM from the supervised models. The results are presented table III. The values α, β, γ and δ have been fine tuned empirically. When MIMIC is used as unpaired data, the overall results are disappointing. This is due to the fact that MIMIC and I2B2 are still too divergent. When the unannotated I2B2 dataset is used as unpaired data instead of MIMIC, the performance increased. This is due to a good match between the training data and the test data. However, the performance does not reach the supervised one. Thus, it seems that for the task, using a pre-annotator like MedExtractor is more efficient than the semi-supervised strategy. However, for languages were such an extractor does not exist, the semi-supervised represents a good alternative.

Table IV summarizes the best results and compare them with the current state-of-the-art. We can see that a simple bi-LSTM model gives competitive results. Our BlueBert based pre-trained model beats the Tao *et al.* [6] model by a short margin. However, it is important to note that Tao *et al.* [6] used a larger set of human annotated training data whereas our approach only used automatically annotated ones. Furthermore, the BlueBert model shows great capability to extract the reason slot which has been reported as the most difficult to treat in the I2B2 challenge.

V. DISCUSSION AND FUTURE WORK

This paper presents a comprehensive evaluation of the state-of-the-art seq2seq models with and without pre-trained embeddings in a supervised and semi-supervised setting. The

System	F1	m	do	mo	f	du	r
Guzman et al. [18]	0.76	0.78	0.81	0.78	0.82	0.19	-
Tao et al. [6]	0.87	0.93	**0.94**	**0.95**	**0.94**	**0.68**	0.48
LSTM	0.78	**0.94**	0.92	0.93	0.89	0.49	0.50
bluebert-base	**0.88**	0.92	0.88	0.95	0.91	0.46	**0.61**

TABLE IV
F-MEASURE OF OUR TWO BEST MODELS AND THE TWO THE STATE-OF-THE-ART MODELS ON THE I2B2 TEST SET.

experiments show that even with limited training data, supervised seq2seq models seems to get high slot-label prediction performance for the medical extraction task. The impact of a pre-trained model on medical documents (here BlueBert) has proven to be particularly effective in handling the lexical rich slots such as the reason and frequency concepts which are among the hardest to extract in the I2B2 2009 dataset [1]. The interest of such pre-trained models are in-line with other recent research in BioNLP. For the supervised learning experiments our BlueBert model reach a f-measure of 88% in line with the current state-of-the-art method [6] (f-measure=87%) which used more annotated data and a complex semi-supervised pipeline. For the semi-supervised experiments, the results did not reach the state-of-the-art but the findings suggest that the unsupervised data were either too different from the test set or too small. Nevertheless, in case of low-resources settings, the approach could provide reasonable performances.

Future work includes how to better select and filter the unpaired datasets so that it is less noisy, closer to the target dataset and contains the most difficult cases (e.g., reason). Furthermore, we plan to extend this work to non-English data which are by far less resourced [31] and thus would benefit more from an unsupervised setting. Finally, an ongoing future work is using this approach to develop medical prescription recognition from spoken utterances [32]. This would have many applications, from harvesting large amount of spoken clinical data, to building medical assistants on smartphones. These applications would increase traceability in health care centers and could reduce the number of medication errors.

REFERENCES

[1] Ö. Uzuner, I. Solti, F. Xia, and E. Cadag, "Community annotation experiment for ground truth generation for the i2b2 medication challenge," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 519–523, 2010.

[2] X. Dai, S. Karimi, and C. Paris, "Medication and adverse event extraction from noisy text," in *Proceedings of the Australasian Language Technology Association Workshop 2017*, 2017, pp. 79–87.

[3] B. Fan, W. Fan, C. Smith *et al.*, "Adverse drug event detection and extraction from open data: A deep learning approach," *Information Processing & Management*, vol. 57, no. 1, p. 102131, 2020.

[4] Y. Si, J. Wang, H. Xu, and K. Roberts, "Enhancing clinical concept extraction with contextual embeddings," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1297–1304, 2019.

[5] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[6] C. Tao, M. Filannino, and Ö. Uzuner, "Fable: A semi-supervised prescription information extraction system," in *AMIA Annual Symposium proceedings*, vol. 2018. American Medical Informatics Association, 2018, p. 1534.

[7] R. Qader, F. Portet, and C. Labbé, "Semi-supervised neural text generation by joint learning of natural language generation and natural language understanding models," in *INLG 2019*, 2019, pp. 552–562.

[8] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, "Biowordvec, improving biomedical word embeddings with subword information and mesh," *Scientific data*, vol. 6, no. 1, pp. 1–9, 2019.

[9] S. Henry, K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner, "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 3–12, 2020.

[10] C. Tao, M. Filannino, and Ö. Uzuner, "Prescription extraction using crfs and word embeddings," *Journal of biomedical informatics*, vol. 72, pp. 60–66, 2017.

[11] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *NAACL2018*, New Orleans, Louisiana, 2018, pp. 2227–2237.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT 2019*, 2019, pp. 4171–4186.

[13] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[14] A. Mulyar and B. T. McInnes, "Mt-clinical bert: Scaling clinical information extraction with multitask learning," *arXiv preprint arXiv:2004.10220*, 2020.

[15] X. Yang, J. Bian, W. R. Hogan, and Y. Wu, "Clinical concept extraction using transformers," *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 1935–1942, 10 2020.

[16] Y. Peng, S. Yan, and 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*, 2019, pp. 58–65.

[17] S. Narayanan, K. Mannam, S. P. Rajan, and P. V. Rangan, "Evaluation of transfer learning for adverse drug event (ade) and medication entity extraction," in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 2020, pp. 55–64.

[18] B. Guzman, I. Metzger, Y. Aphinyanaphongs, H. Grover *et al.*, "Assessment of amazon comprehend medical: Medication information extraction," *arXiv preprint arXiv:2002.00481*, 2020.

[19] Y. Zhang and Z. Lu, "Exploring semi-supervised variational autoencoders for biomedical relation extraction," *Methods*, vol. 166, pp. 112–119, 2019.

[20] J. Wang, Q. Xu, H. Lin, Z. Yang, and Y. Li, "Semi-supervised method for biomedical event extraction," *Proteome Science*, vol. 11, no. 1, pp. 1–10, 2013.

[21] S. Amin, K. A. Dunfield, A. Vechkaeva, and G. Neumann, "A data-driven approach for noise reduction in distantly supervised biomedical relation extraction," *arXiv preprint arXiv:2005.12565*, 2020.

[22] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[23] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "Medex: a medication information extraction system for clinical narratives," *Journal of the American Medical Informatics Association*, vol. 17, no. 1, pp. 19–24, 2010.

[24] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *EMNLP 2015*, 2015, pp. 1412–1421.

[25] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *ICML 2017*, 2017, p. 1243–1252.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[27] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T.-Y. Liu, "Incorporating bert into neural machine translation," *arXiv preprint arXiv:2002.06823*, 2020.

[28] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[29] R. Qader, F. Portet, and C. Labbé, "Seq2seqpy: A lightweight and customizable toolkit for neural sequence-to-sequence modeling," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 7140–7144.

[30] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," *arXiv preprint arXiv:1904.01038*, 2019.

[31] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than english: opportunities and challenges," *Journal of Biomedical Semantics*, vol. 9, no. 12, 2018.

[32] A. C. Kocabiyikoglu, F. Portet, H. Blanchon, and J.-M. Babouchkine, "Towards spoken medical prescription understanding," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2019, pp. 1–8.