Leveraging Token-Based Concept Information and Data Augmentation in Few-Resource NER: ZuKyo-EN at the NTCIR-16 Real-MedNLP task

Joseph Cornelius, Oscar Lithgow-Serrano, Vani Kanjirangat, Fabio Rinaldi
Dalle Molle Institute for Artificial Intelligence Research (IDSIA)
Switzerland
{joseph.cornelius,fabio.rinaldi}@idsia.ch
{vanik,oscarwilliam.lithgow}@idsia.ch

Koji Fujimoto, Mizuho Nishio, Osamu Sugiyama, Kana Ichikawa
Department of Real World Data Research and Development Kyoto University
Japan
{kfb,sugiyama}@kuhp.kyoto-u.ac.jp
ichikawa@kuhp.kyoto-u.ac.jp
nishiomizuho@gmail.com

Farhad Nooralahzadeh, Aron Horvath, Michael Krauthammer
Department of Quantitative Biomedicine University of Zurich and University Hospital of Zurich
Switzerland
farhad.nooralahzadeh@uzh.ch
aronnorbert.horvath@uzh.ch
michael.krauthammer@uzh.ch

ABSTRACT

In this paper, we discuss our contribution to the NII Testbeds and Community for Information Access Research (NTCIR) - 16 Real-MedNLP shared task. Our team (ZuKyo) participated in the English subtask: Few-resource Named Entity Recognition. The main challenge in this low-resource task was a low number of training documents annotated with a high number of tags and attributes. For our submissions, we used different general and domain-specific transfer learning approaches in combination with multiple data augmentation methods. In addition, we experimented with models enriched with biomedical concepts encoded as token-based input features.

KEYWORDS

few-resource, transfer learning, named entity recognition

TEAM NAME

ZuKyo

SUBTASKS

Subtask1-CR-EN
Subtask1-RR-EN

1 INTRODUCTION

A significant part of physicians' working time is spent on documenting patient treatment, time that is lost for diagnostics and direct patient care. Therefore, the automatic processing of patient data has enormous potential to support physicians in their daily work. The automatic identification of diseases, anatomical characteristics, and medications is also required to conduct extensive statistical analysis of the course of a disease. The foundation of automatic document analysis is named entity recognition (NER). However, one of the biggest problems is obtaining large high quality labeled datasets, since expert annotations are very expensive. Therefore, a special interest exists in models that can produce a very accurate automatic recognition based on a small number of annotated samples.

The Real-MedNLP subtask 1 addresses the few resource problem for NER. The challenge is that the corpus consists of real clinical reports with a sample size of only up to 100 documents for training the model. This machine learning task is usually described as "few-resource machine learning".

Our team (ZuKyo) participated in the sub-task 1 " Few-resource Named Entity Recognition" of the NTCIR-16 Real-MedNLP task [11].

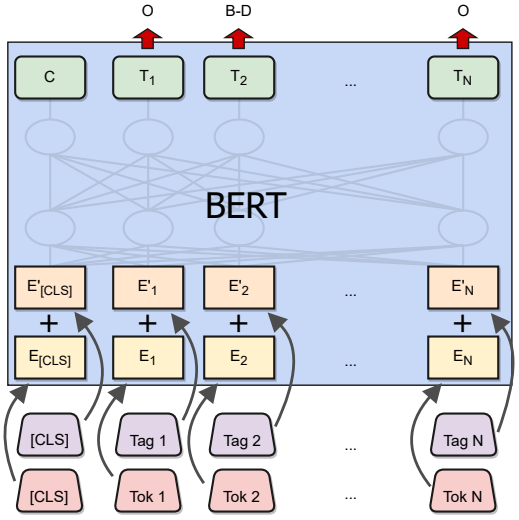

Figure 1: Extra tag variation of the BERT model. In addition to the tokens, an extra tag can be entered. The embedding vector representation of each input token is then obtained by concatenating the embeddings of the token and the extra tag.

2 RELATED WORK

In recent years, substantial research has been done on improving natural language tasks through data augmentation. There are several methods of data augmentation that are more sophisticated than simply adding noise to the training data. Ding et al. [4] showed a generative approach to generate novel NER data by first learning a model on sentences that were combined with their NER tags, and

Table 1: The number of medical records provided for subtask 1, divided into training, validation, and test datasets.

Dataset	MED-CR	MED-RR	Total
Train	88	62	150
Valid	22	10	32
Test	100	63	163

then using this model to create novel examples. In addition, it has been demonstrated that using training data from different domains can improve the robustness of the generated augmented data [2]. Kang et al. [8] were able to show that finding and replacing ULMS concepts with synonyms to create augmented data can improve performance on biomedical NER tasks.

One difficulty with tasks in the biomedical domain is that we have a technical language that uses a variety of different, strictly defined concepts that greatly expand the models' vocabulary. A common practice for dealing with this large terminological space is to use trained NER systems to add additional tags that map words to domain-specific concepts. These tags can be concatenated with the input strings or passed to the combined model's output layer with the vector representation of the additional information to an additional classification layer. Recently, it has been shown that we can improve the performance of transformer-based models by incorporating information from biomedical ontologies [5].

3 DATASET

The data provided by the organizers include radiological and case reports in English and Japanese. Since we used different approaches for English and Japanese, this paper only discusses methods based on English language data sets, and a separate paper describes the methods used for the Japanese data set.

The English radiology dataset (MedText-RR-EN) contains 15 different cases created by a total of 9 different radiologists. Each report contains the description of findings based on a single radiological image. The dataset is divided into 72 training documents and 63 test documents. The following entities were annotated in MedText-RR-EN:

- Diseases and symptoms (tag: d)
- Anatomical entities (tag: a)
- Time expressions (tag: timex3)
- Common names of clinical tests such as 'CT scan' (Tag: t-test)

The English case report dataset (MedText-CR-EN) contains a detailed description of a patient and their disease, as well as the temporal progression of the treatment. The reports come from different medical societies, influencing the type of description and the focus of the diseases treated. The dataset is divided into 100 training documents and 100 test documents. The MedText-CR-EN includes the same entities as the MedText-RR-EN, plus the following:

- Test entry (tag: t-key)
- Measured value of the test entry (tag: t-val)
- Medication names (tag: m-key)
- Medicine dosage (tag: m-val)

Table 2: A list of all extra tags, their meaning, source and the model that generates them.

Extra Tags	Meaning	Model	Source
A	Anatomy	OGER	UBERON
D	Diseases	OGER	CTD
M	Drugs	OGER	RxNorm
CARDINAL	Other numerals	SpaCY	*Learned*
DATE	Dates or periods.	SpaCY	*Learned*
ORDINAL	"first", "second", etc.	SpaCY	*Learned*
ORG	Companies, agencies, institutions	SpaCY	*Learned*
PERCENT	Percentage	SpaCY	*Learned*
QUANTITY	Measurements	SpaCY	*Learned*
TIME	Times smaller than a day.	SpaCY	*Learned*

3.1 Preprocessing

First, we decompose the XML articles into their individual sentences and convert the XML tag-based annotation into an IOB-2 format. This makes the data accessible to token classification methods. Only the tags specified in paragraph 3 are considered, all other tags contained in the original dataset are ignored.

This means that the example sentence given in XML:
`"No <d certainty="negative">pathological lymphadenopathy </d> is seen in the <a>mediastinum, hilar, or axilla</a>."`
is converted to a sentence in IOB-2 format as follows:
`"No[O] pathological[B-D] lymphadenopathy[I-D] is[O] seen[O] in[O] the[O] mediastinum[B-A] hilar[I-A] or[I-A] axilla[I-A] .[O]"`
here the IOB-2 tags are specified in the subscript. Thus, given n the number of official tags, the number of IOB-2 formatted tags is 2n+1.

From the training reports we separated reports for the validation, such that the validation reports contained ~15% of the total sentences from the original training set. The overall distribution of the training, validation and test data can be seen in Table 1.

3.2 Augmentation

To enlarge the training dataset we used augmentation based on the exchange of tags within the set of tags of the same type. First, for each tag type i, we created a set T^i containing all phrases t^i from the training dataset that are tagged with the respective tag type. Given a sentence s with $s = \{w_1, ..., w_{m-1}, t^i_m, t^i_{m+1}, ..., t^i_{m+k}, w_{m+k+1}, ..., w_n\}$, where w is a token and t^i a token tagged of type i. To obtain an augmented sentence s' of s we randomly draw a tagged phrase $t^{i\prime}$ of type i with $t^{i\prime} \neq t^i$ and replacing t^i with $t^{i\prime}$ such that $s' = \{w_1, ..., w_m - 1, t^{i\prime}_m, t^{i\prime}_{m+1}, ..., t^{i\prime}_{m+l}, w_{m+l+1}, ..., w_n\}$. Here, the length k of the original tagged phrase may differ from the length l of the randomly selected tagged phrase.

We experimented with different sized augmented datasets and used a training dataset with 10 augmented sentences for each sentence (x10) and one with 100 augmented sentences for each sentence (x100) in the final runs.

4 METHODS

In the following section, we present the methods used in NER and attribute identification task.

4.1 Bio and Clinical BERT

Both the MedText-RR-EN and the MedText-CR-EN datasets contain documents with many terms from the biomedical domain. Hence, domain-adapted models are likely to have an advantage. To this end, we use a Bio+ClinicalBERT model [1], initialized from BioBERT [9] and further trained on all notes from MIMIC III [7], a dataset with real-world electronic health records. BioBERT is a variation of the BERT model [3], whose training objective is to predict a randomly masked token based on its context and to predict the next sentence based on the previous sentence. The BioBERT model is initialized on the BERT model and trained on PubMed abstracts (PubMed) and PubMed central whole articles (PMC). Through the BioBERT and the Bio+ClinicalBERT, we obtain embeddings specific to the biological, medical, and clinical domains.

4.2 Extra Tags

For the BERT based models, we have a variant of the BERT model that allows us to add an extra tag as input in addition to the tokens. Here we allow the model to process additional information on the token level. This is achieved by having an extra embedding layer for the extra tags as shown in Figure 1. The full embedding vector representation of an input token is then obtained by concatenating the embeddings of the token and the extra tag.

The additional information provided is obtained, on the one hand, by a general domain learned NER model, namely the large EN-Core-Web model of SpaCy[1]. This model can determine entities such as date (DATE), measures (QUANTITY), or organizations (ORG). On the other hand, we used a model based on dictionary lookup and flexible matching, namely OntoGene's Biomedical Entity Recogniser (OGER) [6]. One advantage of this model is that we can use domain-specific dictionaries with several million terms. In this case, we used dictionaries containing anatomical (UBERON[2]), disease (CTD[3]), and drug (RxNorm[4]) terms. A detailed list of the extra tags can be found in the table 2.

4.3 RoBERTa

As an additional model, we used the pre-trained transformer language model RoBERTa [10]. Unlike the widely used transformer-based BERT model, RoBERTa is trained with a masked token that is not statically introduced into the data but dynamically injected in each mini-batch, thus increasing variability. In addition, the authors used larger mini-batches for pre-training, which improved perplexity. Furthermore, RoBERTa was not trained with the objective of predicting the next sentence, but with the objective of predicting the full sentences. Unlike the BERT model, we did not use a domain-specific version of the RoBERTa model.

4.4 Attributes identification

Besides identifying the span of text that refers to a Named Entity (NE), another important part of the annotation process is identifying and including, within the annotation, other aspects about the entity that are also conveyed in the text and related to the entity's type.

[1]https://spacy.io
[2]https://uberon.github.io
[3]http://ctdbase.org
[4]https://www.nlm.nih.gov/research/umls/rxnorm/index.html

Table 3: Statistics of Correlated Entity Tag and Attributes in Training Set

(Tag, Attribute)	Attribute Value	Number of Occurrences
(certainty, d)	general	1
	negative	148
	positive	462
	suspicious	191
(type, timex3)	date	26
	duration	2
	med	1
(state, t-test)	executed	19
	negated	2
	other	6
(state, cc)	other	15
(state, r)	other	1

Examples of these "aspects" are the degree of certainty about a reported disease, the nature of a time expression, the execution state of a test, and so on.

This information should be included as attributes of the corresponding NEs tags. The attribute names correspond to the NE's detected aspect and the attribute value to the aspect's value. For example: for the aspect *"degree of certainty of a reported disease"*, the attribute name is *"certainty"*, and the value can be *positive, negative, suspicious or general.*

First, we explored the training set and manually inspected randomly chosen samples. The objective was to try to detect some patterns that we could utilize to infer the attributes.

We found a lot of interesting correlations between the entity tags and the attributes. The statistics of such pairs are reported in Table 3.

It was observed that *certainty* attribute was mainly associated with entity tag *'d'*, the *type* attribute with *'timex3'* and *state* mainly with *'t-test'*. This in turn shows that detecting the attributes with their values requires the knowledge of both the text and the entity tags. Further, using the attribute *"certainty"* as a case study, we observed that just the presence of the word *"no/not"* was a good indicator of the attribute value. To advance this observation, we computed the ratio of cases where *"no/not"* was found in examples of the *"certainty"* attribute. The results showed that the difference of proportions correlates with the *"certainty"* value as follows: *certainty_negative*: 0.5946, *certainty_suspicious*: 0.1204, *certainty_positive*: 0.0649.

Based on this, we hypothesize that a basic classifier based on the presence of certain words could perform relatively well. Using individual sentences and considering already detected entities, we extracted examples for each entity type (e.g., D, CC, TIMEX3, T-TEST, etc.) and extracted the words (left and right window) next to the target term (named entity). Some examples are shown next:

- The patient was referred to our hospital with a chief complaint of hemoptysis , and was diagnosed with non - small cell lung cancer based on the findings of$_{\text{window}}$ the CT scan and bronchoscopy.
  - **Term:** non - small cell lung cancer

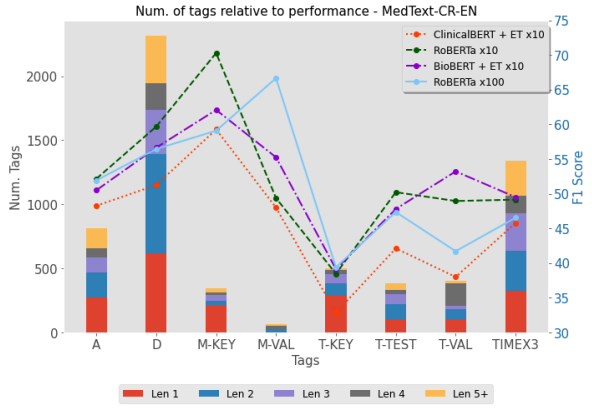

Figure 2: Number of tags per type in relation to the performance of the different Systems for the English MedText clinical reports.

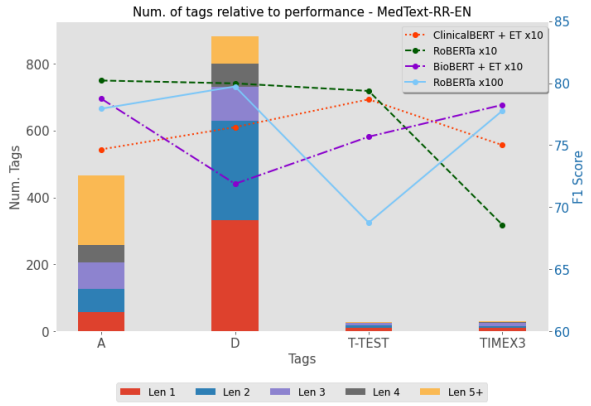

Figure 3: Number of tags per type in relation to the performance of the different Systems for the English MedText radiology reports.

  - **NE type:** D, **Attribute:** certainty_positive
- Histopathological findings showed scattered hepatocellular necrosis , which was thought to be coagulation necrosis , but there was no obvious sign of invasion of the lung cancer into the liver.window.
  - **Term:** invasion of the lung cancer into the liver
  - **NE type:** D **Attribute:** certainty_negative
- Case report of a 68 - year - old male patient In November 2011 , the patient underwent a right hemicolectomy and D2 dissection of the.window colon for ascending colon cancer with multiple liver and lung metastases .
  - **Term:** hemicolectomy
  - **NE type:** T-TEST **Attribute:** state_executed

We transformed the window's text and the identified named entity to features by applying TF-IDF. Those features and the encoded named-entity type were provided as inputs to a classifier that was trained to infer the expected attribute_value (e.g., certainty_positive).

5 EXPERIMENTS

The following section presents the official results of the English few-resource Named Entity Recognition subtask.

5.1 Named Entity Recognition

We have finetuned several pre-trained language models on a unified dataset of English MedText RR and CR. We used different domain-specific and general models as described in section 4, partly with an extended input through the extra tags. In addition, we trained all models on two augmented versions of the dataset as described in section 3.2.

The results of the four submitted models are significantly different with a maximum distance of 0.064 in F1 score, with the Bio+ClinicalBERT trained on the x10 augmented dataset on the CR and RR performing the worst with F1 scores of 0.457 and 0.758, respectively. We obtained the best results with the RoBERTa model trained on the x10 augmented data set with an F1 score of 0.521 for the CR data set and 0.800 for the RR obtained on the official test datasets. However, it is not clear whether the augmentation with the x10 or x100 versus the domain-specific models is preferable, since in the RR dataset the general RoBERTa model achieves the second-best place, whereas in the CR dataset the domain-specific BioBERT model performs second best. Both two ranked models were trained on the x100 dataset, while the best model was trained on the x10 dataset, but the difference in the F1 score is only 0.009 and 0.011 on the CR and RR datasets, respectively.

Both models trained on the x10 dataset were trained with a batch size of 8, a learning rate of 4e-05, and for one epoch. The models on the x100 dataset were trained with the same parameters but with a batch size of 32.

5.2 Attribute identification

We used the vectorized window text, term text and target attribute to train and evaluate a *Support Vector Machine* classifier (SVC). We experimented with different windows sizes, finding the best between 4 and 5 with a "weighted avg. F1" of 0.8070 (see fig. 4). For the TF-IDF vectorization we used a maximum of 300 features and 1 as a minimum threshold of the document frequency, no stop words filtering or any other pre-processing was applied. For the SVC we used a *radial basis function (rbf)* kernel.

The estimator was trained with 80% of the data and the remaining 20% was hold-out for evaluation. The results of evaluating the models in the test split during the training phase are shown in Table 6 for the RR model and in Table 5 for the CR model.

6 DISCUSSION

In the following, we will discuss our results and difficulties encountered in performing our studies.

6.1 Named Entity Recognition

In general, we found that the number of tags in the training data set correlated with the performance of the models as shown in Figures 2 and 3. That is, the more tags of a type were in the training dataset, the better the models generalized. Outliers are the drug names (M-KEY) and dosages (M-VAL). Presumably, dosage amounts

like 0.8 mg or 5 ml are easy to identify for the models due to the uniform nomenclature, this fact could also facilitate the recognition of the drug names.

Furthermore, contrary to our assumption, we could see that the size of the tagged phrases does not give any indication of the predictive performance of the models. Using the augmentation method, it is surprising that an extreme extension of the data set (x100) produces an almost unchanged performance of the models.

Finally, it is worth mentioning that while the results suggest that the more domain-specific the model, the worse the performance, this may be a fallacy, as the domain-specific models and the general models represent different architectures. In our training runs, we could see that with the same BERT architecture, the domain-specific models were significantly more accurate. Due to the limited time, we did not have the possibility to pre-train the RoBERTa models on domain-specific data.

6.2 Attribute identification

Attribute inference was an optional part of the task. Due to time constraints, we opted for a simple implementation using a classifier based on context words. Error analysis revealed that our approach most likely suffered from low-frequency context words in the training data: e.g., "subsided" or "(-)" as an indicator of certainty_negative. For example:

- Erythema around the nails **(-)** .
- Postoperatively, the pain of pancreatitis had **subsided** and glucose tolerance had improved.

There are other attribute types for which our approach did not achieve high accuracy. An example is "state_scheduled", where we observed that the words that are indicative of the scheduled state are far from the NE (i.e., out of the window scope). The following are two examples of such cases - the words that could indicate the "scheduled" state are marked in bold and the window used is underlined; as can be seen, the clues are out of the window scope:

- At a joint conference of the Department of Surgery , Internal Medicine , and Radiology , **it was decided** to perform a percutaneous transsplenic umbilical vein occlusion and splenic artery embolization.$_{\text{window}}$.
  - **Term:** splenic artery embolization
- Therefore , **we decided** to improve splenic function by first performing a PSE and then treating the gastric varices with the BRTO procedure.$_{\text{window}}$.
  - **Term:** BRTO

In post-submission experiments, we observed that a highly unbalanced attribute distribution (classes with thousands of examples and others with tens of examples) was causing high bias in our estimators, mainly in the model for the CR dataset (macro F1: 0.2978). After applying some techniques for addressing the imbalance, we obtained a better macro F1 score (0.346 in CR), and, more importantly, we got better scores for underrepresented classes.

In general, the major weakness of our approach was the low generalization shown when applied to the test data. This could be related to the high bias of our submitted models which were tuned to overrepresented attributes in the training set. Besides applying techniques to deal with unbalanced data, another possible approach would be to use embeddings. Instead of using textual representations that are prone to suffer more from term sparseness, word embeddings, even non-contextual embeddings, could help to reduce this effect and generalize better to text windows unseen during training. And finally, after our experiments, we think that a deep learning approach based on pretrained contextual embeddings could be a good option for future steps.

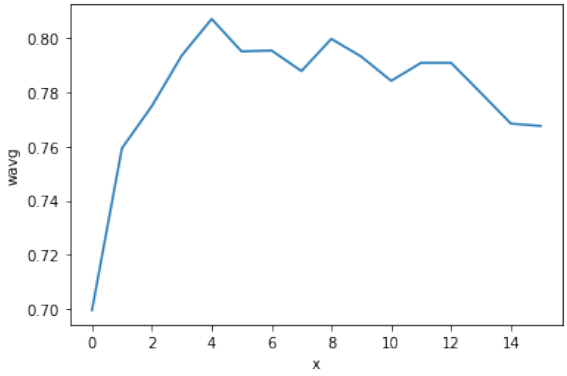

Figure 4: Performance of the attribute classifier for different window sizes: *y* axis corresponds to the F1-weighted average and *x* axis to the window size.

Table 4: Official and unofficial results of our systems compared.

System	Precision	Recall	F-Score
Validation Set (RR and CR joint)			
Bio+ClinicalBERT+ET_{x10}	0.687	0.739	0.712
BioBERT+ET_{x10}	0.691	0.742	0.716
BioBERT+ET_{x100}	0.686	**0.755**	**0.719**
RoBERTa$_{x10}$	**0.694**	0.740	0.717
Test Set (CR)			
Bio+ClinicalBERT+ET_{x10}	0.430	0.487	0.457
BioBERT+ET_{x100}	0.452	0.540	0.492
RoBERTa$_{x10}$	**0.494**	**0.550**	**0.521**
RoBERTa$_{x100}$	0.484	0.544	0.512
Test Set (RR)			
Bio+ClinicalBERT+ET_{x10}	0.729	0.788	0.758
BioBERT+ET_{x100}	0.770	0.808	0.789
RoBERTa$_{x10}$	**0.783**	**0.816**	**0.800**
RoBERTa$_{x100}$	0.773	0.801	0.787

7 CONCLUSIONS

In this paper, we presented several transformer-based models for the English NER subtask 1 of the Real-MedNLP shared task. We have described and evaluated the results of our four domain-specific and general models trained on an augmented version of the official dataset. Furthermore, we have described our approach to the optional attribute prediction task. Here we were able to show that attribute prediction is possible using a low-complexity SVC architecture. To conclude, using a transformer-based architecture and data

Table 5: Classifier evaluation in the CR test split. For brevity, classes with scores = 0.0 were omitted but considered in the summary rows.

Attribute	Precision	Recall	F-Score	sup.
no attribute	0.8762	0.9159	0.8956	618
certainty_positive	0.7169	1.0000	0.8351	347
state_executed	0.7183	0.7484	0.7330	310
type_age	1.0000	0.6667	0.8000	36
type_date	0.5723	0.9500	0.7143	100
type_med	0.9077	0.6941	0.7867	85
accuracy			0.7746	1708
macro avg	0.2995	0.3109	0.2978	1708
weighted avg	0.6928	0.7746	0.7246	1708

Table 6: Classifier evaluation in the RR test split

Attribute	Precision	Recall	F-Score	sup.
no attribute	1.000	0.901	0.948	172
certainty_negative	0.912	0.912	0.912	34
certainty_positive	0.804	0.827	0.815	104
certainty_suspicious	0.571	0.800	0.667	35
state_executed	1.000	1.000	1.000	2
state_other	1.000	1.000	1.000	5
type_date	1.000	1.000	1.000	3
accuracy			0.873	355
macro avg	0.898	0.920	0.906	355
weighted avg	0.892	0.873	0.879	355

augmentation, we achieved the second-best performance among the participating groups on both the English MedText-CR and -RR official test sets.

8 CONTRIBUTION

JC performed most of the experiments described in this paper, under the supervision of FR. OLS and VK designed and tested the module for the detection of attributes. All authors participated in several discussions and exchanged ideas. All authors read the paper and approved it.

REFERENCES

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 72–78. https://doi.org/10.18653/v1/W19-1909

[2] Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Data Augmentation for Cross-Domain Named Entity Recognition. *arXiv preprint arXiv:2109.01758* (2021).

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. *arXiv preprint arXiv:2011.01549* (2020).

[5] Lenz Furrer, Joseph Cornelius, and Fabio Rinaldi. 2021. Parallel sequence tagging for concept recognition. *BMC Bioinform.* 22-S, 1 (2021), 623. https://doi.org/10.1186/s12859-021-04511-y

[6] Lenz Furrer and Fabio Rinaldi. 2017. OGER: OntoGene's Entity Recogniser in the BeCalm TIPS Task. In *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop*. 175–182. http://www.biocreative.org/media/store/files/2017/BioCreative_V5_paper24.pdf

[7] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[8] Tian Kang, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. 2021. UMLS-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association* 28, 4 (2021), 812–823.

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[11] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. 2022. RealMedNLP: Overview of REAL document-based MEDical Natural Language Processing Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*. NII.