Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3724–3731
Marseille, 20-25 June 2022

JaMIE: A Pipeline Japanese Medical Information Extraction System with Novel Relation Annotation

Fei Cheng[1*], Shuntaro Yada[2*], Ribeka Tanaka[3], Eiji Aramaki[2], Sadao Kurohashi[1]
[1]Kyoto University, Kyoto, Japan,
[2]Nara Institute of Science and Technology, Nara, Japan,
[3]Ochanomizu University, Tokyo, Japan
{feicheng, kuro}@i.kyoto-u.ac.jp,
{s-yada,aramaki}@is.naist.jp,
tanaka.ribeka@is.ocha.ac.jp

Abstract

In the field of Japanese medical information extraction, few analyzing tools are available and relation extraction is still an under-explored topic. In this paper, we first propose a novel relation annotation schema for investigating the medical and temporal relations between medical entities in Japanese medical reports. We experiment with the practical annotation scenarios by separately annotating two different types of reports. We design a pipeline system with three components for recognizing medical entities, classifying entity modalities, and extracting relations. The empirical results show accurate analyzing performance and suggest the satisfactory annotation quality, the superiority of the latest contextual embedding models. and the feasible annotation strategy for high-accuracy demand.

Keywords: Medical Information Extraction, Corpus Annotation, Relation Extraction, Open-access Toolkit

1. Introduction

Electronic medical record systems have been widely adopted in the hospitals. In the past decade, research efforts have been devoted to automated Information Extraction (IE) from raw medical reports. This approach should be able to liberate users from the burden of reading and understanding large volumes of records manually. While substantial progress has been made already in medical IE, it still suffers from the following limitations.

First, languages are the natural boundaries to hinder the existing research from being reused across languages. The development of the English corpora and approaches can less reflect the progress in other languages. Morita et al. (2013; Aramaki et al. (2014; Aramaki et al. (2016) present a series of Japanese clinical IE shared tasks. However, more semantic-aware tasks such as medical relation extraction (Uzuner et al., 2011) and temporal relation extraction (Bethard et al., 2017) are still undeveloped. **Second**, most existing medical IE datasets focus on general report content such as discharge summary, instead of more specific report types and diseases. Such settings potentially sacrifice the accuracy for analyzing specific report types, such as radiography interpretation reports.

In this work, we first propose a novel relation annotation scheme for investigating the medical and temporal relations in Japanese medical reports. Then, we intend to explore the correlation between the annotation efforts on specific report types and their analyzing accuracy, which is especially in demand for practical medical applications. Therefore, we target the comparison of analyzing two report types involved with the diseases of high death rates: (1) specific radiography interpretation reports of lung cancer (LC), (2) medical history reports (containing multiple types of reports relevant to a patient) of idiopathic pulmonary fibrosis (IPF). The relation annotation is based on the existing entities presented by Yada et al. (2020), which annotated the medical entities (e.g. *disease*, *anatomical*) and their modality information (e.g. *positive*, *suspicious*) in Japanese medical reports.

While rich English NLP tools for medical IE have been developed such as cTAKES (Savova et al., 2010) and MetaMap (Aronson and Lang, 2010), there are few Japanese tools available until MedEx/J (Aramaki et al., 2018). MedEx/J extracts only diseases and their negation information. In this paper, we present JaMIE: a pipeline **Ja**panese **M**edical **IE** system, which can extract a wider range of medical information including medical entities, entity modalities, and relations from raw medical reports.

In summary, we achieves three-fold contributions as following:

- We present a novel annotation schema for both medical and temporal relations in Japanese medical reports.
- We manually annotate the relations for two types of reports and empirically analyze their performance and desired annotation amount.
- We release an open-access toolkit JaMIE for automatically and accurately annotating medical entities ($F1$:95.65/85.49), entity modalities ($F1$:94.10/78.06), relations ($F1$:86.53/71.04) for two report types.

Although the annotated corpus is not possible to be opened due to the increase of anonymization level, the

[*]These authors contributed equally to this work

Category	Relation Type	Example
Medical	change(C, A)	The <A>intrahepatic bile ducts</A> are <C>dilated</C>.
	compare(C, *TIMEX3*)	<C>Not much has changed</C> since <TIMEX3>September 2003</TIMEX3>.
	feature(F, D)	No <F>pathologically significant</F> <D>lymph node enlargement</D>.
	region(A, D)	There are no <D>abnormalities</D> in the <A>liver</A>.
	value(T-*key*, T-*val*)	<T-key> Smoking</T-key>: <T-val>20 cigarettes</T-val>
Temporal	on(D, *TIMEX3*)	On <TIMEX3>Sep 20XX</TIMEX3>, diagnosed as <D>podagra</D>.
	before(CC, *TIMEX3*)	After <CC>visiting the cardiovascular department</CC>, she was hospitalized <TIMEX3>from April 11th to April 22nd, 2024</TIMEX3>.
	after(C, *TIMEX3*)	PSL 10mg/day had been kept since <TIMEX3>11 Aug</TIMEX3>, but it was <C>normalized</C>.
	start(M-*key*, *TIMEX3*)	<M-key>Equa</M-key> started at <TIMEX3>23 April</TIMEX3>.
	finish(R, *TIMEX3*)	On <TIMEX3>17 Nov</TIMEX3>, quitting <R>HOT</R>.

Table 1: The example of each relation type.

system code and trained models are to be released. [1]

2. Japanese Medical IE Annotation

2.1. Entity and Modality Annotation

We leverage an existing corpus (Yada et al., 2020) with entity and modality information annotated as the base for our relation annotation. The entity types are defined as following: Diseases and symptoms <D>, Anatomical entities <A>, Features and measurements <F>, Change <C>, Time <TIMEX3>, Test <T-test/key/val>, Medicine <M-key/val>, Remedy <R>, Clinical Context <CC>. The complete entity and modality definition refers to the original paper.

2.2. Relation Annotation

On the top of the entity and modality annotation above, we designed relation types between two entities. They can be categorized into *medical relations* and *temporal relations*. The example of each relation type is presented in Table 1.

2.2.1. Medical Relations

A relation(X, Y) denotes an entity of <X> type has a relation type toward another entity of the type <Y>, in which <X> and <Y> can be any entity type defined above (including the case that <X> is the same type as <Y>).

change: A <C> entity changes the status of another entity, the type of which can be <D>, <A>, <T/M-key>. A <C> is often presented as 'dilate', 'shrink', 'appear', etc.

compare: A <C> entity's change is compared to a certain point <Y>, typically <TIMEX3>.

feature: A <F> entity describes a certain entity <Y>. A <F> is often presented as 'significant', 'mild', the size (of a tumor), etc.

region: An entity of an object includes or contains another object entity (often <D> or <A>).

[1] https://github.com/racerandom/JaMIE/tree/demo

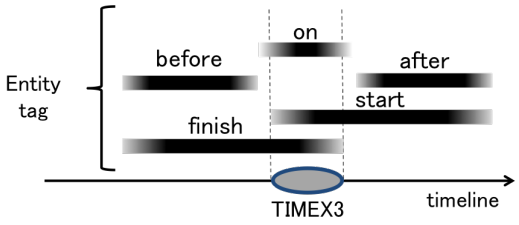

Figure 1: Visualization of temporal relations, i.e., *on*, *before*, *after*, *start*, and *finish*

value: The correspondence relation between <T/M-key> and <T/M-val>. In a rare case, however, other entities of the type <TIMEX3> and <D> may correspond to a value of a <X-key> entity.

2.2.2. Temporal Relations

Based on an existing medical temporal-relation annotation schema, THYME (Bethard et al., 2017), we propose a simplified temporal-relation set below. Note that any temporal relation is defined as a form relation$(X, TIMEX3)$, where the type of <X> can also be another <TIMEX3> entity. Figure 1 portrays a visualized comparison among the proposed temporal relations.

on: A <X> entity happens at the *meantime* of a time span described by a <TIMEX3> entity.

before: A <X> entity happens *before* a time span described by a <TIMEX3> entity.

after: A <X> entity happens *after* a time span described by a <TIMEX3> entity.

start: A <X> entity *starts* at a time span described by a <TIMEX3> entity.

finish: A <X> entity *finishes* at a time span described by a <TIMEX3> entity.

We show the XML-style radiography interpretation report example with the entity-level information and our relation annotation in Figure 2. The test '<T-test>CT scan<T-text>' is executed 'on' the day '<TIMEX3>July 26, 2016</TIMEX3>'. A disease '<D>right pleural effusion</D>' is observed in the 'region' of the anatomical entity '<A>the upper

Figure 2: An annotated radiography interpretation report example (translated into English). To be noticed, the translation may lead to unnatural annotation. For instance, 'after the surgery' in the second sentence is a specific temporal expression often used in Japanese clinical reports, while it look strange to be annotated with a time tag.

lobe of the lung</A>'. A '<F>new</F>' disease '<D>nodules</D>' is in the 'region' of '<A>the lung field</A>. The '<brel>' and '<trel>' tags distinguish the medical relations and temporal relations. JaMIE supports this XML-style format for training models or outputting system prediction. The complete annotation guideline is available. [2]

2.3. Annotation

In practice, we annotated two datasets: 1,000 radiography interpretation reports of LC and 156 medical history reports of IPF. We annotate all reports with two passes. One annotator conducted the first pass relation annotation for a report. In the second pass, the expert supervisor examined the annotation and led the final adjudication by discussing the inconsistency with the first pass annotator. This procedure is to balance the quality and cost, since it does not fully rely on the expert annotation. We separately calculate the Inter-Annotator Agreement (IAA) of the relation annotation (gold entity annotation based) between two independent annotators on the same five reports randomly selected from each report types. The radiography interpretation reports achieve the IAA with F1 95.19 and Accuracy 91.75%. The medical history reports achieve the IAA with F1 70.58 and Accuracy 70.35%. Considering the low agreement of temporal relation annotation reported by (Bethard et al., 2017), our annotation IAAs show that the over all annotation quality is guaranteed. The lower IAA in medical history reports also suggests that extracting relations from medical history reports is a more difficult task than radiography interpretation reports.

Table 2 shows the statistics of the relations annotation. Though the number of the medical history reports is relatively smaller, they usually contain more content per report and a wider coverage of entity types. Considering that the popular English 2010 i2b2/VA medical dataset contains 170 documents (3,106 relations) for training, our annotation scale are comparable with or even larger than it. The results show very different relation type distribution in the two types of reports. As the medical history reports of IPF can be viewed as the mixture of several types of reports such as radiography reports, examination reports, test results, etc., they show a more balanced coverage of relation types, while the radiography interpretation reports of LC are more narrowly distributed among the disease-relevant relation types such as 'region' and 'feature'.

Although our annotation experiment is conducted on Japanese medical reports, the annotation guideline is not limited to any specific languages.

3. System Architecture of JaMIE

Figure 3 shows the overview of our Japanese medical IE system with a pipeline process of three components: medical entity recognition, modality classification, and relation extraction. The over all implementation is based on the Pytorch Transformers [3].

3.1. Sentence Encoder

Recent medical IE research (Si et al., 2019; Alsentzer et al., 2019; Peng et al., 2019) suggests the contextual pre-trained models such as ELMO (Peters et al., 2018) and BERT (Devlin et al., 2019) markedly outperform traditional word embedding methods (e.g., word2vec, glove, and fastText). In our pipeline system, we adopt the Japanese pre-trained BERT as the sentence encoder for retrieving token embeddings.

Formally, a sentence $S = [x_0, x_1, x_2, ..., x_n]$ is encoded by a contextual BERT or word embedding with bidirectional Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) as:

$$X = Encoder([x_0, x_1, x_2, ..., x_n])$$

[2] Japanese:https://doi.org/10.6084/m9.figshare.16418787
English:https://doi.org/10.6084/m9.figshare.16418811 Some notations might be slightly different in the latest version.

[3] https://github.com/huggingface/transformers

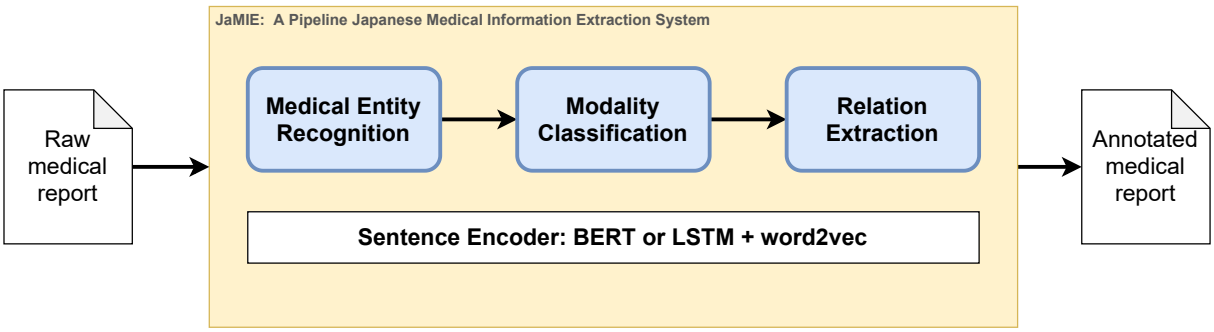

Figure 3: The overview of JaMIE.

3.2. Medical Entity Recognition

Medical entity recognition (MER) aims to predict the token spans of entities and their types from the text.
We formulate Medical Entity Recognition as sequential tagging with the BIO (begin, inside, outside) tags. The outputs are constrained with a conditional random field (CRF) (Lafferty et al., 2001) layer. For a tag sequence $y = [y_0, y_1, y_2, ..., y_n]$, the probability of a sequence y given X is the softmax over all possible tag sequences:

$$P(y|X) = \frac{e^{s(X,y)}}{\sum_{\hat{y} \in Y} e^{s(X,\hat{y})}}$$

where the score function $s(X, y)$ represents the sum of the transition scores and tag probabilities.
In practice, we adopt a CRF implementation PyTorch-crf[4] on the top of the sentence encoder.

3.3. Modality Classification

The modality classification (MC) component is to classify the modality types of the given entities. For a multi-token entity E_i predicted by the MER model, we represents the entity embedding as the element sum of embeddings in the entity span. To enrich the context for predicting assertion, we concatenate the entity embedding with the auxiliary entity type. The i-step modality prediction is:

$$y_i = softmax(fc([E_i; E_i^{type}]))$$

where E_i denotes the i-th entity embedding, E_i^{type} denote the entity type embedding predicted by the MER model, and $fc(.)$ denotes a single full-connected layer.

3.4. Relation Extraction

The relation extraction (RE) component is to predict the relations and their types between two named entities. Semantic or temporal relation extraction has been widely explored by recent work (Cheng and Miyao, 2017; Bekoulis et al., 2018; Cheng and Miyao, 2018; Zhang et al., 2020; Cheng et al., 2020; Zhong and Chen, 2021). For pursuing efficiency, we formulate the relation extraction problem as the multiple head selection (Zhang et al., 2017) of each entity in the sentence. Given each entity E_i in the sentence, the model predicts whether another entity E_j is the head of this token with a relation r_k. The probability of a relation between two entities is defined as:

$$P(E_j, r_k|E_i; \theta) = sigmoid(fc(E_j, r_k, E_i))$$

where $fc(.)$ denotes a single full-connected layer. An additional 'N' relation presents no relation between two tokens. The final representation of an entity E_i is the concatenated embeddings of the entity, entity type, and modality type.

4. Experiments

4.1. Settings

For each dataset, we conduct the 5-fold cross-validation to evaluate the performance of our system. 10% training data is split as the validation set for tuning best hyper-parameters and checkpoints. In each stage in the pipeline, the current component is trained with the gold inputs. The Japanese text is segmented into tokens by MeCab (Kudo et al., 2004).We adopt NICT Japanese BERT[5] as the sentence encoder. The following hyper-parameters are empirically chosen: training epoch as 10, batch size as 16, AdamW Optimizer with learning rate as 5e-5. The best checkpoints on the validation set are saved to produce test results. Our model is also compatible with other Japanese morphological analyzers and pre-trained models, such as: Juman++ (Tolmachev et al., 2018) and Ku-BERT [6].

4.2. Evaluation

Instead of applying the usual pipeline evaluation with the gold inputs at each stage, we are more interested in the practical performance of the system and adopt the joint evaluation (Zheng et al., 2017) as described in the following:

[4] https://pytorch-crf.readthedocs.io/en/stable/

[5] https://alaginrc.nict.go.jp/nict-bert/index.html

[6] https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese

1000 Radiography Interpretation Reports (LC)				156 Medical History Reports (IPF)			
Med REL	**#Num**	**Temp REL**	**#Num**	**Med REL**	**#Num**	**Temp REL**	**#Num**
region	6,794	on	696	region	631	on	1,583
change	689	start	5	change	465	start	219
feature	5,077	finish	2	feature	294	finish	43
value	2	after	3	value	1,932	after	22
compare	615	before	1	compare	229	before	14
Total	13,884			Total	5,432		

Table 2: The statistics of the relation annotation. '**Med**' and '**Temp**' denote the medical and temporal relations.

Report Type	**Encoder**	**MER F1**	**MC F1**	**RE F1**
Radiography Interpretation Reports (LC)	LSTM + word2vec	93.63	93.01	77.88
	BERT	**95.65**	**94.10**	**86.53**
	(Yada et al., 2020)	95.30	-	-
Medical History Reports (IPF)	LSTM + word2vec	82.73	75.26	60.42
	BERT	**85.49**	**78.06**	**71.04**

Table 3: The main results for automatically analyzing two types of reports.

- **Medical entity recognition** identifies medical entity from raw reports. We evaluate each {*entity, entity type*} to the reference.
- **Modality classification** classifies the modality types of the entities identified by the former stage. The evaluation is on each {*entity, entity type, modality type*}.
- **Relation extraction** extracts the relations between the entities identified by the former stages. The evaluation is on each triplet {*head entity, relation, tail entity*}.

We measure micro-F1 of the system prediction to the gold reference in each pipeline stage.

5. Experiment Results

5.1. Main Performance of JaMIE

Table 3 shows our system performance on two types of reports: radiography interpretation reports of LC and medical history reports of IPF. The radiography interpretation reports' performance suggests that by concentrating annotation efforts on a specific report type the system achieves high F1 with sufficient training data. Compared to 95.30 MER F1 reported by (Yada et al., 2020), our MER score outperforms their score by 0.35 with the additional CRF layer. The RE model obtains 86.53 F1 of the radiography interpretation reports and 71.04 F1 of the medical history reports.

We offer the baseline encoder with LSTM[7] upon word2vec embeddings (Mikolov et al., 2013) trained on Japanese Wikipedia. We observe significant drops in all three tasks, especially in the final relation extraction. In both radiography interpretation reports of LC and medical history reports of IPF, the BERT-based RE models leading 'LSTM + word2vec' by approximately 10 points F1. We suggest that solving relation extraction requires long-range information between entities. BERT naturally models such long-range dependency between any two tokens via the self-attention mechanism, while word2vec is trained with a fixed local window and LSTM could also accumulate fails over the long sequential actions.

Med REL	**RE F1**	**Temp REL**	**RE F1**
region	84.59	on	81.92
change	76.23	start	20.00
feature	90.16	finish	-
value	-	after	-
compare	80.86	before	-

Table 4: Each relation F1 (BERT-based) in the radiography interpretation reports of LC.

Med REL	**RE F1**	**Temp REL**	**RE F1**
region	71.73	on	70.48
change	58.66	start	49.33
feature	60.54	finish	12.02
value	83.12	after	-
compare	75.47	before	11.38

Table 5: Each relation F1 (BERT-based) in the medical history reports of IPF.

While the medical history reports contain broader relation types and the data size is relatively smaller, the system still obtains satisfactory performance. In addition, we present each relation F1 in Table 5. Except for three rarely appearing relations, i.e. 'finish', 'after' and 'before', the F1 scores on the other types are balanced and match the statistics in Table 2. As for the radiography interpretation report results in Table 4, the major re-

[7]LSTM and Word2vec hidden size equal to 256.

lations of 'region' and 'feature' relations achieve high performance with 84.59 and 90.16 F1. The moderate 'change', 'compare' and 'on' obtain satisfying 76.23 to 81.92 F1.

5.2. Correlation between Report Types and Demanding Annotation Efforts

Report Type	RE F1
Radiography Interpretation Reports	86.53
- with 39% training data	82.33
Medical history Reports	71.04

Table 6: The RE performance comparison between the radiography interpretation reports and medical history reports with comparable training size

One question is whether concentrating annotation efforts on a specific report type can quickly obtain high accuracy to meet the requirements of the practical applications. A valid approach is to compare the RE performance of two report types with the comparable annotation efforts i.e. training data size. The medical history report of IPF contains total 5,432 relations, which is approximately 39% of the radiography interpretation reports. We designed the experiment by reducing the train set of the radiography interpretation reports to the comparable 39% of the origin. The results in Table 6 show that even with comparable training size, the specific radiography interpretation reports lead the performance by 11.29 points F1.

To be clarified, the two results are still not exactly comparable due to the different relation distributions in two report types. However, the radiography interpretation reports more densely spread in the relation types such as 'region' and 'feature' (Table 2), which usually means less number of reports needed for achieving the similar over all accuracy compared to the medical history reports. In the scenario of demanding high accuracy for practical medical applications, the results suggest that the annotation strategy of starting from a specific type of report and gradually increasing the coverage of report types is more feasible.

6. System Application

6.1. User Interface

JaMIE provides an easy-to-use Command-Line Interface (CLI). We design our training/testing scripts similar to the official Transformers examples, in order to be friendly to the Transformers users. We demonstrate how to train/test a relation model with the following script:

```
$ # Training
$ python clinical_pipeline_rel.py \
$ --pretrained_model $JAPANESE_BERT \
$ --saved_model $MODEL_TO_SAVE \
$ --train_file $TRAIN_FILE \
$ --dev_file $DEV_FILE \
$ --batch_size 16 \
$ --do_train

$ # Testing
$ python clinical_pipeline_rel.py \
$ --saved_model $TRAINED_MODEL \
$ --test_file $TEST_FILE \
$ --test_out $TEST_OUTFILE \
```

6.2. Use Case

In the case of annotating raw medical reports with our trained model, users need to download our trained models from the JaMIE GitHub beforehand. Users then execute the pipeline 'test' scripts to annotate entities, modalities, and relations step by step. At each stage, the model will generate the prediction as the input of the next stage model. The prediction is presented in the same XML-style as shown in Figure 2.

Our medical IE annotation schema serves to encode a wide range of general medical information not limited to any specific disease, report types and languages. Users can manually annotate their medical reports by following our guideline. Users can apply the 'train' scripts to train the pipeline models on their newly annotated corpus for providing automatic annotation.

7. Conclusion

We propose a novel annotation schema for investigating medical and temporal relations between medical entities in Japanese medial reports. We empirically compare the annotation on two types of reports: specific radiography interpretation reports of LC and medical history reports of IPF. The system obtains overall satisfactory performance in three tasks, supporting the valuable findings of the good annotation quality, the feasible annotation strategies for targeting report types, and the superior performance of the contextual BERT encoder. The system code and trained models on our annotation are open-access.

In the future, we plan to stick to LC and IPF, cover more specific report types involved with LC, and increase the annotation amount of medical history reports of IPF.

Acknowledgements

We thank the reviewers for their helpful feedback. This work has been supported by the research grant: JPMW21AC500 from Ministry of Health, Labour and Welfare of Japan.

Bibliographical References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W., Jin, D., Naumann, T., and McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings. *CoRR*, abs/1904.03323.

Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T. (2014). Overview of the ntcir-11 mednlp-2 task. In

In Proceedings of the 11th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.

Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T. (2016). Overview of the ntcir-12 mednlpdoc task. In *In Proceedings of the 12th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*.

Aramaki, E., Yano, K., and Wakamiya, S. (2018). Medex/j: A one-scan simple and fast nlp tool for japanese clinical texts. In *MEDINFO 2017: Precision Healthcare Through Informatics: Proceedings of the 16th World Congress on Medical and Health Informatics*, volume 245, page 285. IOS Press.

Aronson, A. R. and Lang, F.-M. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Bekoulis, G., Deleu, J., Demeester, T., and Develder, C. (2018). Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

Bethard, S., Savova, G., Palmer, M., and Pustejovsky, J. (2017). SemEval-2017 Task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cheng, F. and Miyao, Y. (2017). Classifying temporal relations by bidirectional LSTM over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Vancouver, Canada, July. Association for Computational Linguistics.

Cheng, F. and Miyao, Y. (2018). Inducing temporal relations from time anchor annotation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1833–1843, New Orleans, Louisiana, June. Association for Computational Linguistics.

Cheng, F., Asahara, M., Kobayashi, I., and Kurohashi, S. (2020). Dynamically updating event representations for temporal relation classification with multi-category learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1352–1357, Online, November. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain, July. Association for Computational Linguistics.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Morita, M., Kano, Y., Ohkuma, T., Miyabe, M., and Aramaki, E. (2013). Overview of the ntcir-10 mednlp task. In *In Proceedings of NTCIR-10*.

Peng, Y., Yan, S., and Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. *CoRR*, abs/1906.05474.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). 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)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Si, Y., Wang, J., Xu, H., and Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

Tolmachev, A., Kawahara, D., and Kurohashi, S. (2018). Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium, November. Association for Computational Linguistics.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal*

of the American Medical Informatics Association, 18(5):552–556.

Yada, S., Joh, A., Tanaka, R., Cheng, F., Aramaki, E., and Kurohashi, S. (2020). Towards a versatile medical-annotation guideline feasible without heavy medical knowledge: Starting from critical lung diseases. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4565–4572, Marseille, France, May. European Language Resources Association.

Zhang, X., Cheng, J., and Lapata, M. (2017). Dependency parsing as head selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 665–676, Valencia, Spain, April. Association for Computational Linguistics.

Zhang, R. H., Liu, Q., Fan, A. X., Ji, H., Zeng, D., Cheng, F., Kawahara, D., and Kurohashi, S. (2020). Minimize exposure bias of Seq2Seq models in joint entity and relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 236–246, Online, November. Association for Computational Linguistics.

Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., and Xu, B. (2017). Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236, Vancouver, Canada, July. Association for Computational Linguistics.

Zhong, Z. and Chen, D. (2021). A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online, June. Association for Computational Linguistics.