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, 2. Nara Institute of Science and Technology, 3. Ochanomizu University

LREC 2022 Marseille

1. Motivation

Less semantic-aware tasks such as relation extraction has been developed in Japanese medical domain natural language processing. In this work, we present two contributions:

- novel annotation of **medical** and **temporal relations** in **Japanese.**
- an **open-access toolkit** for accurately recognizing entities, modalities and relations from medical texts.

2. Japanese medical relation annotation

We conduct **novel relation annotation** over an existing Japanese medical corpus (Yada+, 2020) with following entity and modality information annotated:
Disease<D>, Anatomical<A>, Feature<F>, Change<C>, Time <TIMEX3>, Test<T-test/key/val>, Medicine<M-key/val>, Remedy <R>, Clinical Context<CC>.

On the top of the entity and modality annotation, we designed medical relations and temporal relations between two entities.

Medical Relation

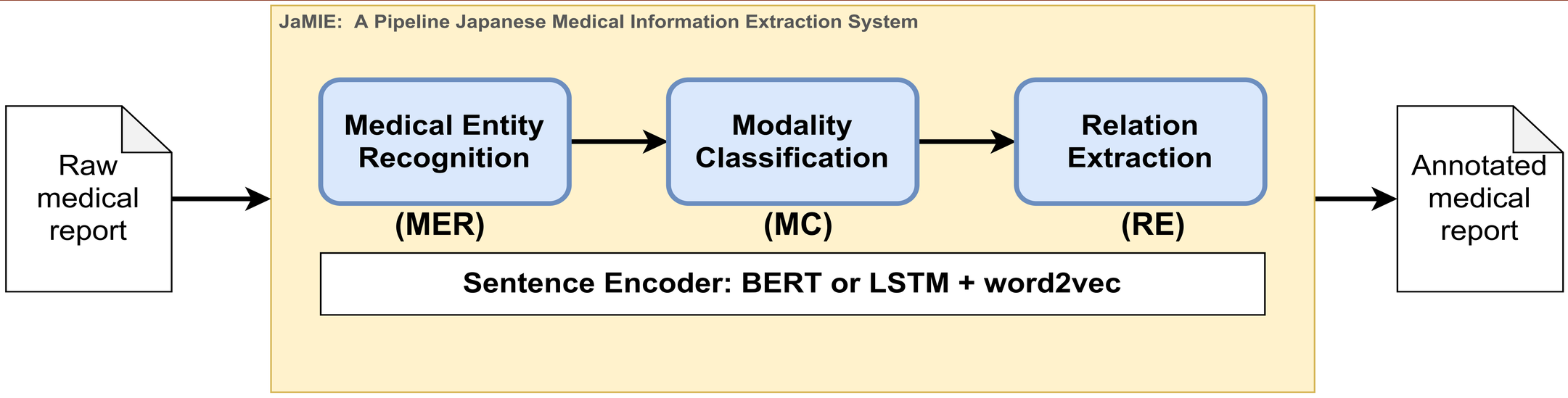
- **Change**
  "<A>intrahepatic bile ducts</A> are <C>dilated</C>"
- **Compare**
  "<C>not changed</C> since <TIMEX3> Sep. 2003</TIMEX3>"
- **Feature**
  "<F>pathologically significant</F> <D>lymph node enlargement</D>"
- **Region**
  "There are no <D>abnormalities</D> in the <A>liver</A>."
- **Value**
  "<T-key> Smoking</T-key>: <T-val>20 cigarettes</T-val>"

3. Japanese temporal relation annotation

Temporal Relation

- **On**
  "In <TIMEX3>Sep. 2003</TIMEX3>, diagnosed as <D>podagra</D>"
- **Before**
  "After <CC>visiting the cardiovascular department</CC>, she was hospitalized <TIMEX3>from April 11th</TIMEX3>."
- **After**
  "since <TIMEX3>11 Aug</TIMEX3>, PSL was <C>normalized</C>."
- **Start**
  "<M-key>Equa</M-key> started at <TIMEX3>23 April</TIMEX3>."
- **Finish**
  "On <TIMEX3>17 Nov</TIMEX3>, quitting <R>HOT</R>"

4. Pipeline System of JaMIE



5. The statistics of the relation annotation

In this work, we target two types of reports for annotating:
RIRLC denotes 1,000 Radiography Interpretation Reports of Lung Cancer.
MRIPF denotes 156 Medical Reports of Idiopathic Pulmonary Fibrosis.

Medical	# RIRLC	# MRIPF	Temporal	# RIRLC	# MRIPF
Change	689	465	**On**	696	1,583
Compare	615	229	**Before**	1	14
Feature	5,077	294	**After**	3	22
Region	6,794	631	**Start**	5	219
Value	2	1,932	**Finish**	2	43

6. Main results

Our BERT-based model achieves accurate analyzing performance, especially in the RIRLC data.

Report type	Encoder	MER	MC	RE
RIRLC	LSTM	93.63	93.01	66.88
	BERT	**95.65**	**94.10**	**86.53**
	Yada+ 2020	95.30	-	-
MRIPF	LSTM	82.73	75.26	60.42
	BERT	**85.49**	**78.06**	**71.04**

7. Additional comparison with comparable training data

Report type	Training data	RE
MRIPF	100%	71.04
RIRLC	Comparable size	**82.33**

8. User Interface

JaMIE provides an easy-to-use Command-Line Interface (CLI), which is similar to PyTorch Transformers. We demonstrate how to train/test a relation model with following scripts.

```
# Training
$ python clinical_pipeline_rel.py \
$ --pretrained_model $PRETRAINED_JAPANESE_BERT \
$ --saved_model $FINETUNED_MODEL \
$ --train_file $TRAIN_FILE \
$ --dev_file $DEV_FILE \
$ --batch_size 16 \
$ --do_train
# Testing
$ python clinical_pipeline_rel.py \
$ --saved_model $FINETUNED_MODEL \
$ --test_file $TEST_FILE \
$ --test_out $TEST_OUTPUT \
```