FRDC at NTCIR-16 Real-MedNLP Task

FRDC
Zhnogguang Zhong
Lu Fang
Yiling Cao
2022/6/17

FUJITSU

Tasks Participated

- Subtask1-CR-EN (NER)
- Subtask3-CR-EN (ADE)
- Subtask3-RR-EN (CI)

Few-resource NER

FUJITSU

- **Data augmentation**
  - **From: An Analysis of Simple Data Augmentation for Named Entity Recognition**
    - **Label-wise token replacement (LwTR):** we use a binomial distribution to randomly decide whether the token should be replaced. If yes, then randomly select another token with the same label
    - **Synonym replacement (SR):** replace the token with one of its synonyms retrieved from WordNet
    - **Mention replacement (MR):** randomly select another mention from the original training set which has the same entity type as the replacement
    - **Shuffle within segments (SiS):**first split the token sequence into segments of the same label. Thus, each segment corresponds to either a mention or a sequence of out-of-mention tokens. Then for each segment, we use a binomial distribution to randomly decide whether it should be shuffled.

	Instance												
None	She	did	not	complain	of	headache	or	any	other	neurological	symptoms	.	
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O	
LwTR	L.	One	not	complain	of	headache	he	any	interatrial	neurological	current	.	
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O	
SR	She	did	non	complain	of	headache	or	whatsoever	former	neurologic	symptom	.	
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O	
MR	She	did	not	complain	of	neuropathic	pain	syndrome	or	acute	pulmonary	disease	.
	O	O	O	O	O	B-problem	I-problem	I-problem	O	B-problem	I-problem	I-problem	O
SiS	not	complain	She	did	of	headache	or	neurological	any	symptoms	other	.	
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O	

Few-resource NER

FUJITSU

○ **Result of data augmentation (SR, MR, LwTR, SiS)**

model	Augmentation strategy	epoch	Dev set F1	Test set F1
BioBERT + CRF	None (baseline)	40	**0.626**	**0.573**
BioBERT + CRF	LwTR	40	0.633	0.565
BioBERT + CRF	SR	40	**0.635**	**0.591**
BioBERT + CRF	MR	40	0.625	0.573
BioBERT + CRF	SiS	40	0.631	0.559
BioBERT + CRF	SR+MR	40	**0.643**	**0.577**
BioBERT + CRF	LwTR+SR+SiS	40	0.612	0.5
BioBERT + CRF	All	40	0.637	0.578

Few-resource NER

FUJITSU

- Result of data augmentation (SR, MR, LwTR, SiS)

model	Augmentation	epoch	Test set F1
BioBERT + CRF	none	40	0.573
BioBERT + CRF	SR	40	0.591
BioBERT + CRF	all	40	0.578

BioBERT model is better

model	Augmentation	P	R	F1
BioBERT	no	0.6028	0.5974	0.6001
BioBERT	SR	0.596	0.6015	0.5988
BioBERT	all	0.5985	0.6072	0.6029

Few-resource NER

FUJITSU

- The 2 models submitted

model	Training set size	P	R	F1
BioBERT	2/3 data of all the dataset	0.6419	0.6539	0.6478
BioBERT	2/3 data of all the dataset + sr+mr+lwtr+sis	0.6463	0.6445	0.6454

Few-resource NER

FUJITSU

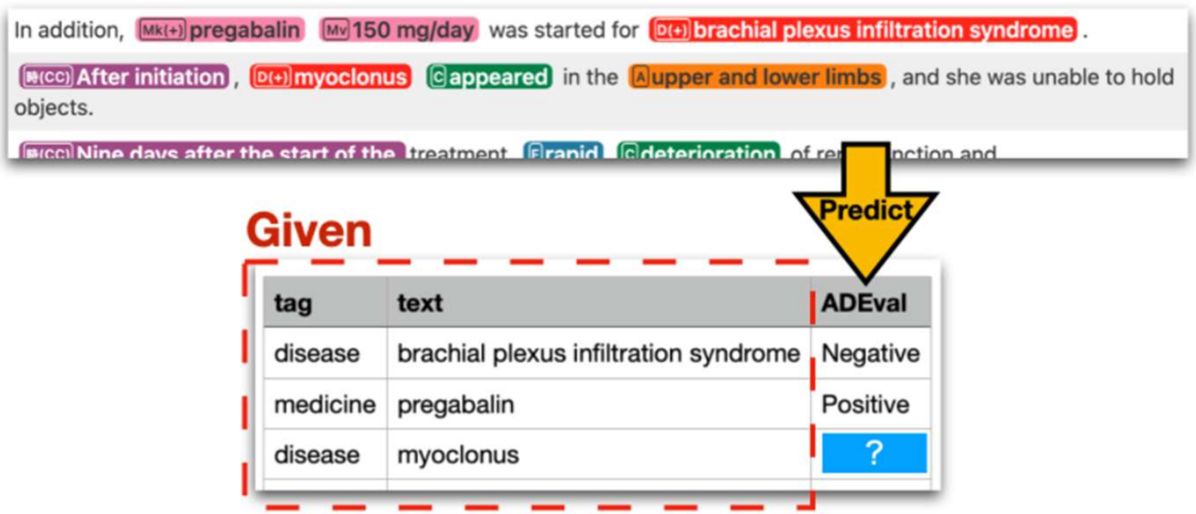
○ The result

	num of each tag	propotion	Entity-F1		rank
a	819	16.67%	51.56	49.72	3
d	2346	47.76%	55.28	58.74	3
m-key	344	7.00%	64.90	65.33	3
m-val	64	1.30%	66.67	71.19	1
t-key	524	10.67%	37.12	38.58	2
t-test	388	7.90%	48.26	47.06	2
t-val	427	8.69%	45.37	45.10	3
num of all tags	4912				
(All target entities)			38.55	40.58	4

Adverse Drug Event detection (ADE)

FUJITSU

- Extract adverse drug event (ADE) information from case reports and fill out a table
  - This subtask is especially designed for MedTxt-CR
  - Given an input report, the system extracts the ADE information from the report.
  - Four levels of ADE certainty (ADEval) are to be given:
    - 3 Definitely
    - 2 Probably
    - 1 Unlikely
    - 0 Unrelated
  - Two types of entities:
    - Disease and symptom names
    - Medication(drug) names



Adverse Drug Event detection (ADE)

- Method
  - Our method is based on fine-tuning the vocabulary adapted BERT model (VART) with multi-learning mechanism, where we transform the ADE task as a classification task
    - Input Sequence: *[CLS] context [SEP] entity_name*
    - Main task: Classification of four levels (ADEval = 0, 1, 2, 3)
    - Auxiliary task: Learning to classify the type of an entity (two types)
- Dataset
  - Training data (small and imbalance)
    - 148 reports

ADEval	0	1	2	3
Number of entities	1376	63	85	174

  - Test data:75 reports, 671 entities
  - Data Augmentation for training data
    - Punctuation insertion and synonym replacement methods

Adverse Drug Event detection (ADE)

FUJITSU

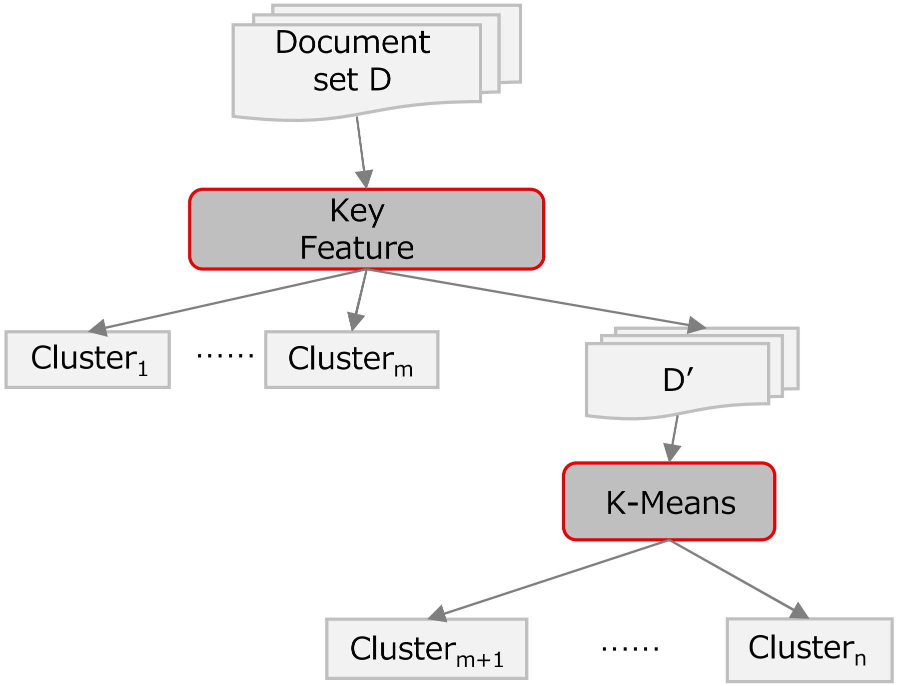
- Results of official run on test data
  - We achieved the second best F1-score on entity level evaluation of ADEval=1,3
  - Also, the second best F1-score on report level evaluation

MaskedGroupID		C8	I1	F2	H1
ADEval=0	P	96.42	97.02	95.39	96.57
	R	97.79	97.63	98.10	97.95
	F	97.10	97.32	96.73	97.25
ADEval=1	P	20.00	30.00	0.00	14.29
	R	5.26	31.58	0.00	5.26
	F	8.33	30.77	0.00	7.69
ADEval=3	P	47.62	100.00	40.00	60.00
	R	52.63	26.32	42.11	63.16
	F	50.00	41.67	41.03	61.54
Report-level	P	50.00	50.00	40.00	50.00
	R	77.78	88.89	44.44	66.67
	F	60.87	64.00	42.11	57.14

Case Identification (CI)

FUJITSU

- Two-step Document Clustering Method
  - Key Feature based clustering
    - Cluster document based on trusted features
  - K-means clustering
    - Document embedding
    - Data Augmentation
    - Data De-nosing



Key Feature Based Clustering

FUJITSU

Key Feature Extraction

Information: 9 documents in each cluster

Words	DF	TF	CF
in	71	273	8
the	70	627	8
lung	63	152	8
……			
12	9	10	1
atelectasis	8	11	1

Words	7<=DF<=9	TF	CF
image	9	9	5
impression	9	9	8
12	9	10	1
……			

Words= Number	7<=DF<=9	TF	CF
12	9	10	1
28	8	8	1
43	8	8	1
78	8	8	1
18	8	8	1

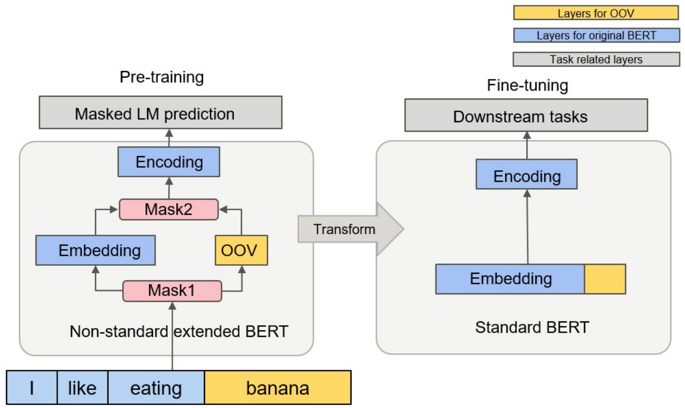
```
<article id="0" title="03">
There is <d certainty="positive">consolidation</d> <a>along the br
="positive">bronchiectasis</d> and <d certainty="positive">pleural
contradiction with the <d certainty="positive">lung cancer</d> you
The size of the <d certainty="positive">lesion</d> is <f>28 mm</f>
```

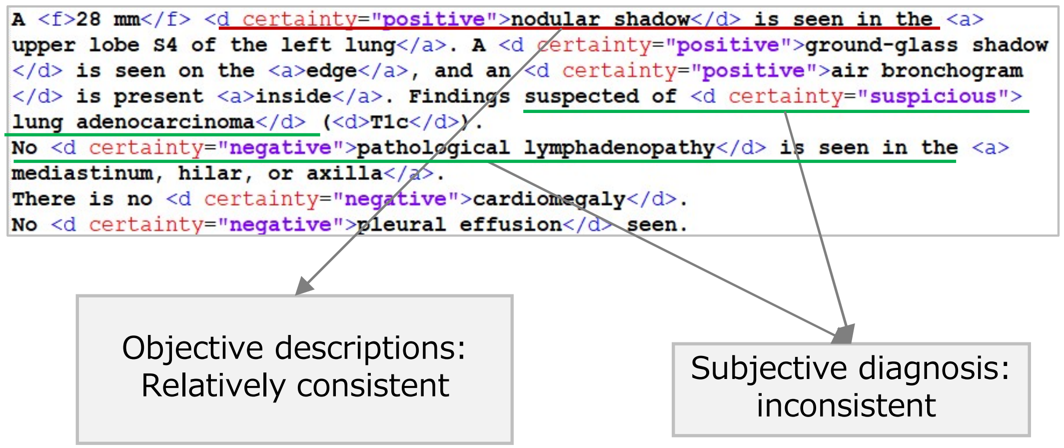
```
<article id="3" title="03">
An <d certainty="positive">irregular mass lesion</d> of <f>approximately 28 mm</f>
lingular segment S4</a>. The <a>interior</a> is <f>occupied</f> by <d certainty="p
"positive">expanded image</d> <f>partly</f> remains in the <a>bronchioles</a>. Sus
predominant lepidic growth pattern</f>. There is also a <d certainty="positive">li
suspect the presence of <d certainty="suspicious">inflammatory scars and fibrous c
infiltration</d> is not suspected. <r state="other">Surgical treatment</r> seems p
```

K-means Document Clustering

FUJITSU

- Document embedding
  - Vocabulary Adapted BERT Model
    - Adapt BERT model on the target training set by extending the BERT vocabulary
  - Use sentence-transformer to encode the document
- Data Augmentation for Pre-training BERT
  - Original documents + documents with stop-words removal
- Document De-noising
  - Input: original documents + de-noised document
    - Remove the subjective contents





Results on Training Set

FUJITSU

- Experiment results
  - Each number is the average normalized mutual info score (NMI) of 50 runs.

Pre-trained models
TAPT: pre-train BERT/BioBERT on the training set at first.

VART: vocabulary adapted BERT model

Pre-train / Input	txt	txt+rm-stop		
	txt		txt+de-noise	
	K-means			Key Feature +K-means
BERT	0.3634	0.3903	0.4427	0.7727
BioBERT	0.3396	0.4644	0.5421	0.8168
$TAPT_{BERT}$	0.4252	0.4252	0.5089	0.7995
$TAPT_{BioBERT}$	0.3924	0.4630	0.5322	0.8094
$VART_{BERT}$	**0.4943**	**0.5689**	**0.5909**	**0.8394**
$VART_{BioBERT}$	0.3443	0.4584	0.4952	**0.8335**

Dataset format
txt: original document
rm_stop: remove the stop words
de-noise: remove the subjective contents (de-noise)

Score (Normalized Mutual Info)	0.8724	0.8463	0.8468	0.8468	0.8468	0.8468	0.8581	0.8468	0.8468	0.8576	0.2172	0.7879

Thank you

FUJITSU