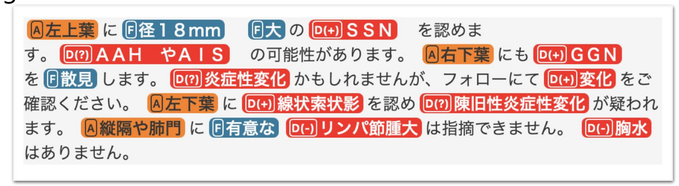
AMI Team at the NTCIR-16 Real-MedNLP Task

Satoshi Hiai, Shoji Nagayama, and Atsushi Kojima
Advanced Media, Inc.

Introduction

Medical Named Entity Recognition (NER) Task

- Recognition of named entities in medical text data



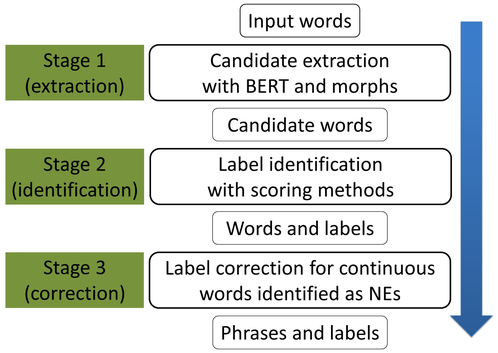
- The amount of Japanese medical text data is small
- Need for high-performance systems with few resources

Subtask1

- Participants construct NER methods with approximately 200 documents
- **BERT based approaches**
  - Pretrained language models have achieved excellent results for low resource situations
  - We use BERT pretrained on biomedical documents (**UTH-BERT**)

Subtask2

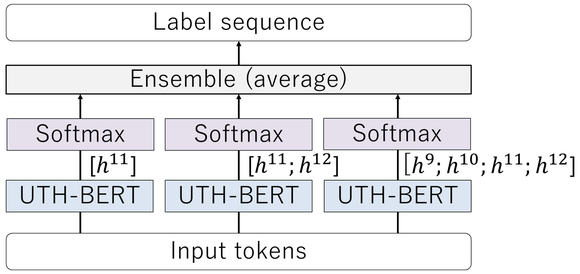
- Participants construct NER methods based on a guideline
  - Descriptions for the annotation of each tag
  - A handful of sample sentences
- Insufficient data for machine learning method, so we used a rule-based approach
- **Rule-based approach**
  - We used surfaces and syntactic patterns
  - We constructed a multistage method



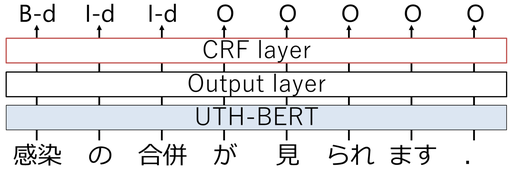
Methods

Subtask1: Two Methods Based on UTH-BERT

- **Ensemble of hidden layers from UTH-BERT**
  - UTH-BERT consists of 12 transformer layers
  - It has been reported that each layer of BERT captures different features [1]
  - We confirmed effective combinations of the layers for each named entity tag and ensembled them



- **UTH-BERT with a CRF layer**
  - We used CRF layer to consider tag sequences



[1] G. Jawahar et al. 2019. What does BERT learn about the structure of language? In Proc of ACL 2019.

Subtask2: Multistage Method

1. **Extraction**
   - We extracted nouns and words predicted as named entities by combination of MeCab and BERT
     - MeCab morph analyzer with biomedical dictionaries (Manbyo dictionary, Hyakuyaku, and comeJisyo)
     - UTH-BERT finetuned on sample sentences
2. **Identification**
   - We calculate entity scores for the extracted words using the formula below

$$score_{word} = \alpha * score_{dict} + \beta * score_{synt} + \gamma * score_{BERT}$$
$$(\alpha, \beta, and\ \gamma\ are\ weights\ for\ the\ scores)$$

     - $score_{dict}$
       - Whether biomedical dictionaries contain the word (0/1)
       - We augment the vocabulary in the dictionaries by a bootstrap method and word embeddings similarities
     - $score_{synt}$
       - Whether the word matches our regular expressions (0/1)
     - $score_{BERT}$
       - Softmax score of each label from UTH-BERT ([0-1])
3. **Correction**
   - We applied some rules to merge continuous words into one named entity
     - Normally, continuous words are merged into one phrase tagged as the last tag in the continuous tags

E.g. Brain + metastasis → Brain metastasis
<a> **<d>** **<d>**

Evaluation Results

Subtask1

		Development				FormalRun			
		CR-JA		RR-JA		CR-JA		RR-JA	
	Tag	Ensemble	Crf	Ensemble	Crf	Ensemble	Crf	Ensemble	Crf
F1-score	a	**69.61**	65.26	**100.0**	93.62	58.37	**58.43**	33.58	**89.16**
	d	**80.80**	70.88	**87.38**	82.26	**67.05**	**67.05**	7.88	**89.40**
	m-key	60.61	**62.86**	-	-	**70.63**	70.39	-	-
	m-val	**0.00**	**0.00**	-	-	**65.67**	**65.67**	-	-
	t-key	43.75	**46.15**	-	-	**35.76**	35.55	-	-
	t-test	**90.91**	80.85	**100.0**	**100.0**	**43.58**	43.38	0.00	**87.50**
	t-val	**50.00**	42.86	-	-	55.48	**55.68**	-	-
	timex3	85.00	**86.87**	**100.0**	**100.0**	**74.62**	74.39	24.49	**88.24**

Since the result on RR-JA of FormalRun is in stark contrast against Development, the Ensemble result may include formatting errors.

- Ensemble method outperformed CRF method overall
- As a future work, we will analyze the effectiveness of each layer against named entity classes more thoroughly

Subtask2

	Tag	CR-JA	RR-JA
F1-score	a	41.52	56.89
	d	41.68	68.45
	m-key	40.00	-
	m-val	22.38	-
	t-key	37.20	-
	t-test	28.17	81.25
	t-val	34.66	-
	timex3	35.02	74.42

- On CR-JA, the scores excluding <m-val> and <timex3> were the best scores among the participants
- On RR-JA, the scores excluding <a> were the best scores among the participants