NTTD at the NTCIR-16 Real-MedNLP Task

Shuai Shao, Gongye Jin, Daisuke Satoh, Yuji Nomura

NTT DATA Corporation, Japan

NTT DATA Trusted Global Innovator

Overview

- We participated in the **Subtask1-CR-JA & Subtask1-RR-JA**, which were NER tasks with limited labeled data of Japanese medical documents.
- We trained 2 models respectively with **augmented training data** to extract named entities from the provided Japanese case reports and radiographic reports.
- From the aspect of **Entity-F1** of all entities, our models ranked **2nd in Subtask1-CR-JA** and **3rd in Subtask1-RR-JA**.

Challenges and Related Works

- Challenges:
  - Deep Learning applications need a huge amount of data, and Japanese medical documents are relatively difficult to acquire and annotate.
  - The inconsistency in a small dataset may affect models' output. [1]
- Approaches:
  - To assure the annotation quality
  - To reduce the necessary data volume
- Existing approaches for dealing with small dataset in NER:

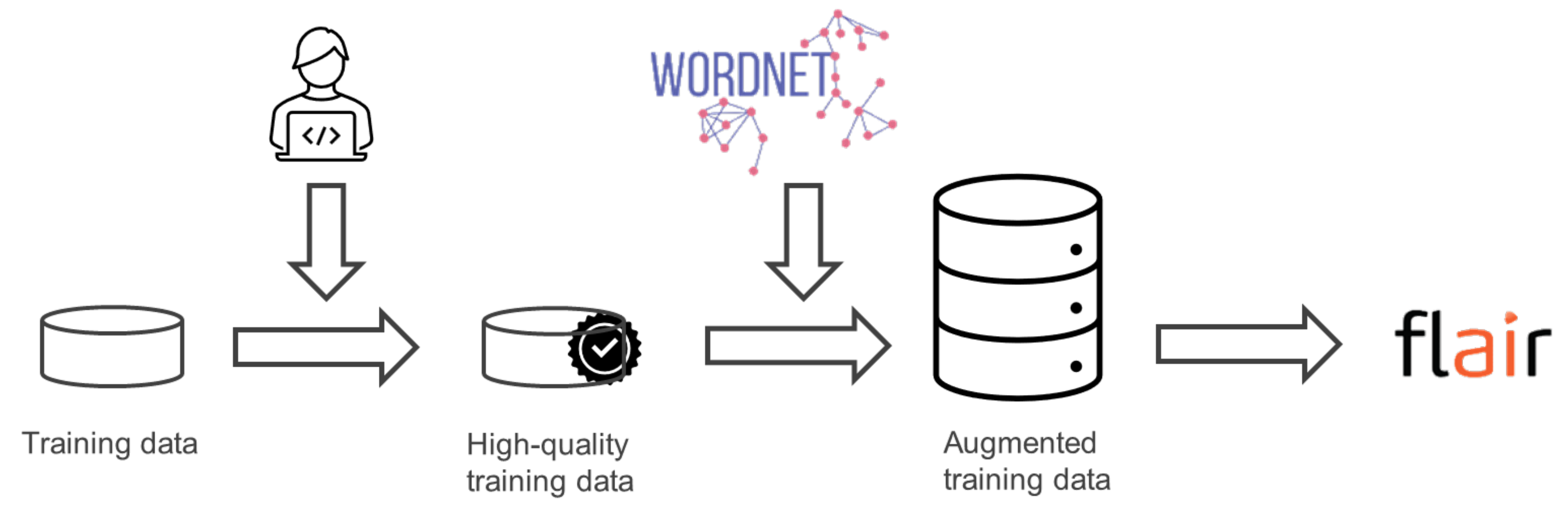
Semi-Supervised Learning

Transfer Learning

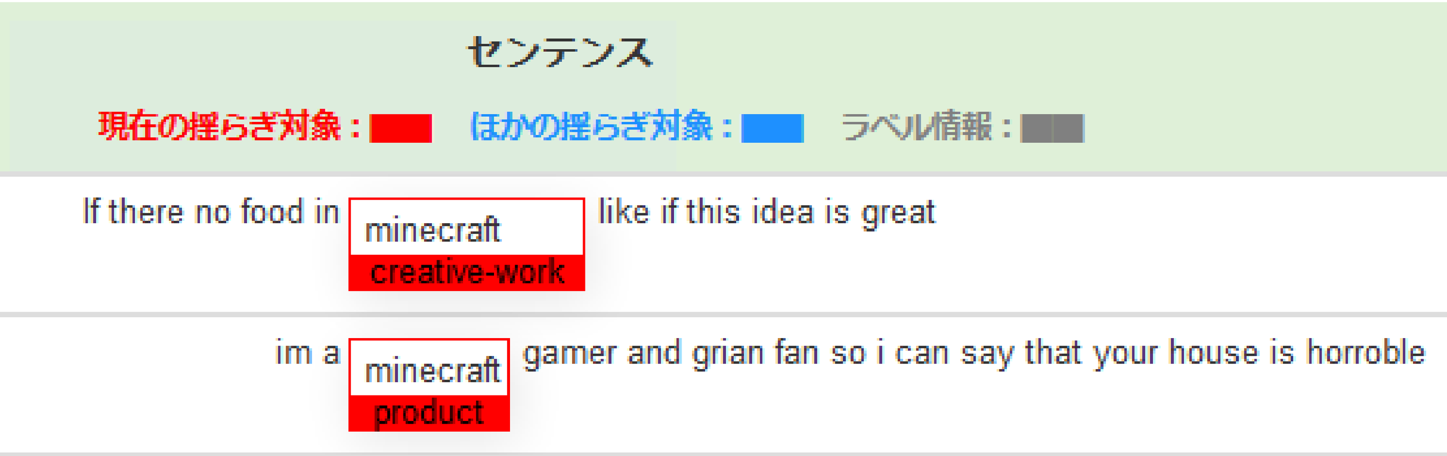
Active Learning

Highly depending on the volume of unlabeled data or the domain relevance between the source and target data.

Our Approach



- **Annotation Inconsistency Detection**
  - Automatically detecting entities labeled with different tags in KWIC format
  - Manually correcting inappropriate tags



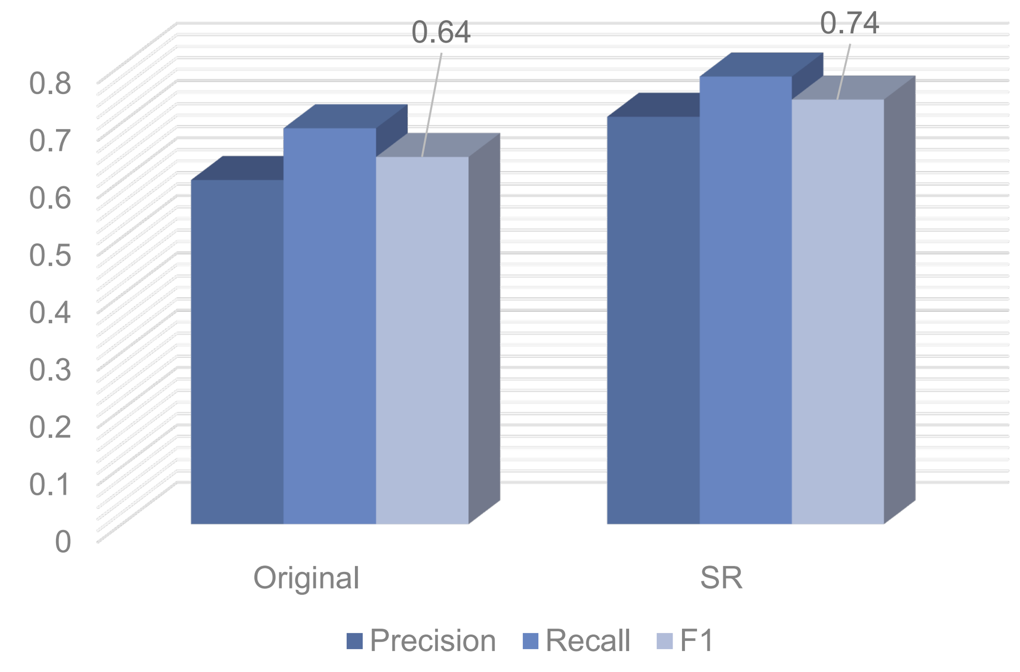
Labels	Sentences
a	<a>口腔粘膜および口唇に広範囲に糜爛 Widespread erosions of <a>oral mucosa and lips
a	<a>顔の一部と口腔粘膜にびらんを認めた Erosions on part of the <a>face and oral mucosa.
(missing)	豆腐と比べて口腔粘膜からの吸収性が良く Better absorption from the oral mucosa compared to tofu

- **Data Augmentation by Synonym Replacement** [2]
  - Using a binomial distribution to determine whether each token should be replaced.
  - Using synonyms from WordNet to replace a token.

	Sentences
Original	今回は眼瞼周囲の浮腫, 紫斑と呼吸困難のため緊急入院した。 This time, the patient was urgently hospitalized because of periocular edema, purpura and dyspnea.
Augmented by SR	今回は目縁周囲の水症,紫斑と呼吸作用波乱の惻巧事変入院した。 This time, the patient was hospitalized for the obedient incident of periorbital hydrops, purpura and respiratory disturbance.

Results

- Results of our 3-fold validation experiment (dividing CR training data into 3 pairs of training and test sets by a ratio of 8:2)



- Official Results of Entity level

Subtask	P	R	F1	Rank
CR-JA	**62.26**	61.53	61.89	2
Others' Best	61.96	68.91	65.25	
RR-JA	86.96	87.09	87.03	3
Others' Best	89.07	89.45	89.26	

- Annotation inconsistency detection & simple synonym replacement can boost the NER model's performance even in the field which needs high level of expertise.
- Specific augmentation strategy on different tag types & ensemble of multiple augmentation methods should be considered.

References

[1] Qingkai Zeng, Mengxia Yu, Wenhao Yu, Tianwen Jiang and Meng Jiang. 2021. Validating Label Consistency in NER Data Annotation. arXiv preprint arXiv:2101.08698.

[2] Xiang Dai and Heike Adel. 2020. An Analysis of Simple Data Augmentation for Named Entity Recognition. In Proceedings of the 28th International Conference on Computational Linguistics.

[3] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 54–59. aclweb.org.

NTT DATA Trusted Global Innovator