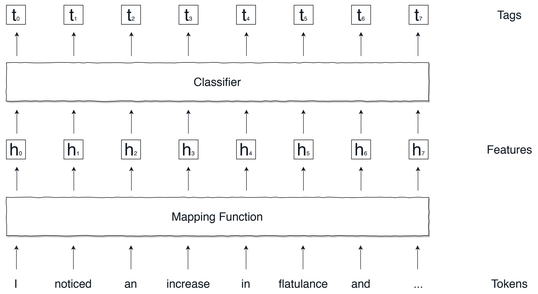
Recognizing Biomedical Names in Free Text

Xiang Dai

CSIRO Data61 and University of Sydney

Task and Prior Arts

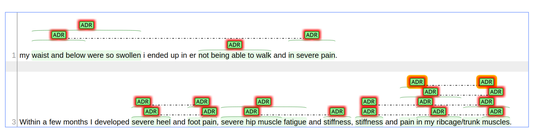
- The text, represented as a sequence of tokens, is taken as input, and entity mentions, each of which is represented as a set of token positions, are outputted. In addition, one entity category, such as disease, symptom, protein, DNA and so on, is assigned to each entity mention.
- The standard NER model – solving NER as a **sequence tagging** problem – assigns a tag to each token, indicating its role in the mention.



- Sequence taggers usually consist of two components: (1) mapping function converts the input sequence of tokens into a sequence of features vectors, each of which represents the corresponding token-in-the-context; (2) structural classifier predicts a sequence of tags given the input sequence of feature vectors.

Key Challenges

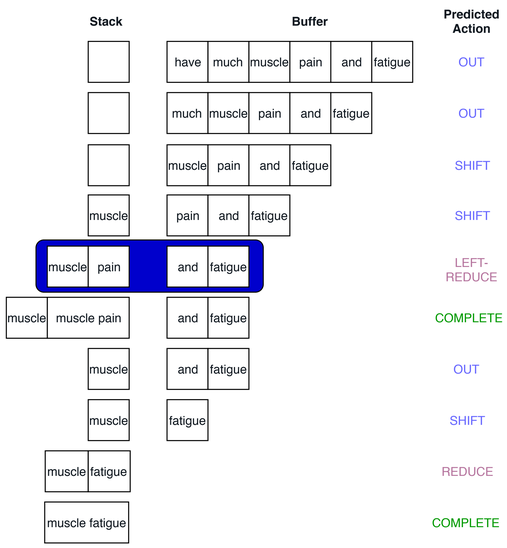
- Sequence tagger cannot recognize biomedical names that have **complex inner structures** – discontinuous and/or overlapping mentions.



- Training of deep neural models requires **large training set**, which is usually difficult to obtain in the biomedical domain.
- Discrepancy between **models pretrained on generic data** and biomedical data.

Transition-based model for discontinuous NER [1,2]

- The main motivation for recognizing discontinuous mentions is that they usually represent *compositional concepts* that differ from concepts represented by individual components.
- We propose a model based on the **shift-reduce parser** with six specialized actions and attention mechanism.
- The learning problem is then framed as: given the state of the parser, predict an action which is applied to change the state of the parser.



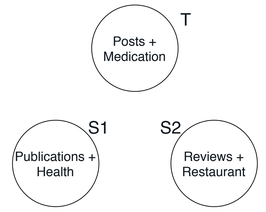
- Our model can effectively **recognize discontinuous mentions without sacrificing** the accuracy on continuous mentions.

References

1. Dai, Xiang: Recognizing complex entity mentions: A review and future directions. In SRW@ACL; 2018: 37-44.
2. Dai, Xiang, Karimi, Sarvnaz, Hachey, Ben, Paris, Cecile: An Effective Transition-based Model for Discontinuous NER. In ACL; 2020: 5860–5870.
3. Dai, Xiang, Karimi, Sarvnaz, Hachey, Ben, Paris, Cecile: Using similarity measures to select pretraining data for NER. In NAACL; 2019: 1460–1470.
4. Dai, Xiang, Karimi, Sarvnaz, Hachey, Ben, Paris, Cecile: Cost-effective Selection of Pretraining Data: A Case Study of Pretraining BERT on Social Media. In Findings of EMNLP; 2020: 1675–1681.
5. Dai, Xiang, Adel, Heike: An Analysis of Simple Data Augmentation for Named Entity Recognition. In COLING; 2020: 3861–3867.

Select in-domain Pretrain Data [3,4]

- Studies on domain-specific pretrained models show that, when in-domain data is used for pretraining, target task performance can be improved. However, the selection of in-domain data usually resorts to **human intuition**.
- We find that human intuition varies across NLP practitioners, especially regarding intersecting domains.



- We employ **cost-effective** measures to quantify similarity between source pretraining and target task data, and demonstrate that these measures are good predictors of the usefulness of pretrained models.

Data augmentation for NER [5]

- We design several **easy to adapt** data augmentation for NER, and show that simple data augmentation can improve performance even over strong baselines.

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