Recognizing Complex Entity Mentions A Review and Future Directions

Xiang Dai dai.dai@csiro.au
CSIRO Data61 and University of Sydney

Motivation

- Sequence tagging techniques can effectively recognize entity mentions that consist of contiguous tokens and do not overlap with each other.
- However, in practice, there are many domains, such as the biomedical domain, in which there are nested, overlapping, and discontinuous entity mentions that cannot be directly recognized by conventional sequence tagging models.
- We review the existing methods which are revised to tackle complex entity mentions, and discuss some directions that we are exploring.

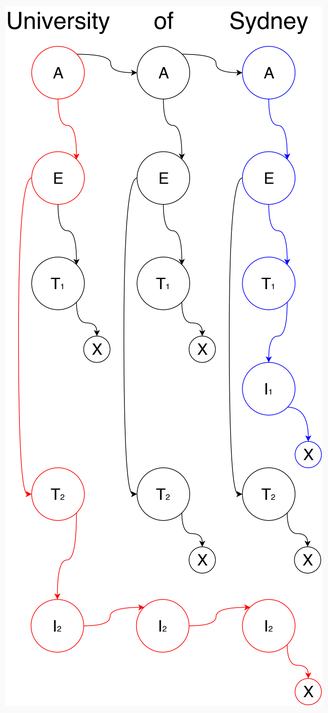
Token-level Approach

Based on conventional sequence tagging, with expanded BIO tag set or specialized strategies.
[1] used additional position indicators to represent discontinuous and overlapping entity mentions, which is shown in the following example. (Weaknesses: high ambiguity level. For example, this example can be decoded as having three mentions: '*intense pelvic pain*', '*back pain*' and '*pain*'.)

had	intense	pelvic	and	back	pain	.
O	BD	ID	O	B	BH	O

Sentence-level Approach

Instead of predicting the tag for each token, sentence-level approach predicts directly a combination of entity mentions within a sentence.
Lu and Roth [2] proposed a novel hypergraph to compactly represent possible nested mentions in one sentence, and one sub-hypergraph of the complete hypergraph can therefore be used to represent a combination of mentions in the sentence.



This hypergraph representation encodes two mentions: 'University of Sydney' and 'Sydney'. Each path starts with node of type A and ends with node of type X. It can represent one mention if it includes node of type I.

Complex Entity Mentions

Examples involving overlapping, discontinuous and nested entity mentions:

a) ... activation of the HIV-1 enhancer following ...

DNA: HIV-1 enhancer
Virus: HIV-1

In (a), '*HIV-1 enhancer*' and '*HIV-1*' are nested entity mentions.

b) ... had intense pelvic and back pain ...

ADE: intense pelvic pain
ADE: back pain

In (b), two Adverse Drug Events (ADEs): '*intense pelvic pain*' and '*back pain*' overlap, meanwhile, '*intense pelvic pain*' is a discontinuous mention.

Proposed Directions

- Combine neural network models with previous mentioned hypergraph representation.
- Sequence-to-sequence: Single encoder + multiple decoders.
- Active learning: Annotate dataset with complex mentions, solve the lack of training data issue.

References

[1] Alejandro Metke-Jimenez and Sarvnaz Karimi. Concept extraction to identify adverse drug reactions in medical forums: A comparison of algorithms. CoRR abs/1504.06936, 2015.
[2] Wei Lu and Dan Roth. Joint mention extraction and classification with mention hypergraphs. In Conference on Empirical Methods in Natural Language Processing, pages 857-867, Lisbon, Portugal, 2015.

Acknowledgements

The author would like to thank Sarvnaz Karimi, Ben Hachey and Cecile Paris for helpful advice on this work. The author also thanks Google for providing travel grant.