

Week 4: Named Entity Recognition

Nhung Nguyen
slides courtesy of NaCTeM

Recap

- Contextualised word embeddings
- Sequence processing with RNN, LSTM, and BiLSTM

Intended learning outcomes

- The students will be able to understand the importance of the named entity recognition (NER) task and to recognise the task in practice
- The students will be able to describe different approaches to NER and explain their advantages/disadvantages

Materials

- Dan Jurafsky and James H. Martin. *Speech and Language Processing (3rd ed. draft)*. <https://web.stanford.edu/~jurafsky/slp3/> (Chap 8 and 9)

What is this snippet talking about?

Speaking at a Downing Street briefing on Monday evening, Mr Johnson added that the majority of people most in need of a vaccination in the UK might be able to get one by Easter.

And Prof Andrew Pollard - director of the Oxford vaccine group - said it had been "a very exciting day" and paid tribute to the 20,000 volunteers in the trials around the world, including more than 10,000 in the UK.

What is this snippet talking about?

Speaking at a **Downing Street** briefing on Monday evening, Mr Johnson added that the majority of people most in need of a vaccination in the **UK** might be able to get one by Easter.

And Prof Andrew Pollard - director of the Oxford vaccine group - said it had been "a very exciting day" and paid tribute to the 20,000 volunteers in the trials around the world, including more than 10,000 in the **UK**.

 Location

What is this snippet talking about?

Speaking at a **Downing Street** briefing on **Monday evening**, Mr Johnson added that the majority of people most in need of a vaccination in the **UK** might be able to get one by **Easter**.

And Prof Andrew Pollard - director of the Oxford vaccine group - said it had been "a very exciting day" and paid tribute to the 20,000 volunteers in the trials around the world, including more than 10,000 in the **UK**.



Location



Time

What is this snippet talking about?

Speaking at a **Downing Street** briefing on **Monday evening**, **Mr Johnson** added that the majority of people most in need of a vaccination in the **UK** might be able to get one by **Easter**.

And **Prof Andrew Pollard** - director of the Oxford vaccine group - said it had been "a very exciting day" and paid tribute to the 20,000 volunteers in the trials around the world, including more than 10,000 in the **UK**.



Location



Time



Person

What is this snippet talking about?

Speaking at a **Downing Street** briefing on **Monday evening**, **Mr Johnson** added that the majority of people most in need of a vaccination in the **UK** might be able to get one by **Easter**.

And **Prof Andrew Pollard** - director of the **Oxford vaccine group** - said it had been "a very exciting day" and paid tribute to the 20,000 volunteers in the trials around the world, including more than 10,000 in the **UK**.



Location



Time



Person



Organisation

How about this?

Remdesivir (GS-5734), an inhibitor of the viral RNA-dependent, RNA polymerase with in vitro inhibitory activity against SARS-CoV-1 and the Middle East respiratory syndrome (MERS-CoV),⁵⁻⁸ was identified early as a promising therapeutic candidate for Covid-19 because of its ability to inhibit SARS-CoV-2 in vitro.⁹

How about this?

Remdesivir (GS-5734), an inhibitor of the viral RNA-dependent, RNA polymerase with in vitro inhibitory activity against SARS-CoV-1 and the Middle East respiratory syndrome (MERS-CoV),⁵⁻⁸ was identified early as a promising therapeutic candidate for Covid-19 because of its ability to inhibit SARS-CoV-2 in vitro.⁹

Drug/Chemical

Beigel, John H et al. "Remdesivir for the Treatment of Covid-19 - Final Report." The New England journal of medicine vol. 383,19 (2020): 1813-1826. doi:10.1056/NEJMoa2007764

How about this?

Remdesivir (GS-5734), an inhibitor of the viral RNA-dependent, RNA polymerase with in vitro inhibitory activity against SARS-CoV-1 and the Middle East respiratory syndrome (MERS-CoV),⁵⁻⁸ was identified early as a promising therapeutic candidate for Covid-19 because of its ability to inhibit SARS-CoV-2 in vitro.⁹

■ Drug/Chemical

■ Gene/Protein

How about this?

Remdesivir (GS-5734), an inhibitor of the viral RNA-dependent, RNA polymerase with in vitro inhibitory activity against SARS-CoV-1 and the Middle East respiratory syndrome (MERS-CoV),⁵⁻⁸ was identified early as a promising therapeutic candidate for Covid-19 because of its ability to inhibit SARS-CoV-2 in vitro.⁹



Drug/Chemical



Gene/Protein



Virus

How about this?

Remdesivir (GS-5734), an inhibitor of the viral RNA-dependent, RNA polymerase with in vitro inhibitory activity against SARS-CoV-1 and the Middle East respiratory syndrome (MERS-CoV),⁵⁻⁸ was identified early as a promising therapeutic candidate for Covid-19 because of its ability to inhibit SARS-CoV-2 in vitro.⁹



Drug/Chemical



Gene/Protein



Virus



Disease

Named entity recognition (NER)

- The central question to understand a sentence:

Who did what to whom? (also ***where*** and ***when***)

- NER: identify *named entity mentions* and their *types* of interest

What is a named entity?

- A real world object (can be physical or abstract) that can be named
 - Person: Bill Gates, [your name]
 - Location: London, Mexico, [your PoB]
 - Time expression: 31/01/2020, [your DoB]
 - Monetary value: \$15, [your tuition fee]
 - ...
- For a specific task/domain, we care only some entity types
 - News: person, location, organisation, monetary, ...
 - Biomedical: gene/protein, drug name, symptom, disease, ...

What is an entity mention?

- An entity mention is a span of text (can be a single word or multiple words)

Speaking at a **Downing Street** briefing on **Monday evening**

This is a *mention*
of some *entity*

- More complicated case? Nested entities

John killed [[Mary] 's husband]

Approaches to NER

- Dictionary lookup
- Rule-based methods
- Machine learning-based methods

Isn't NER trivial?

Are all these necessary?

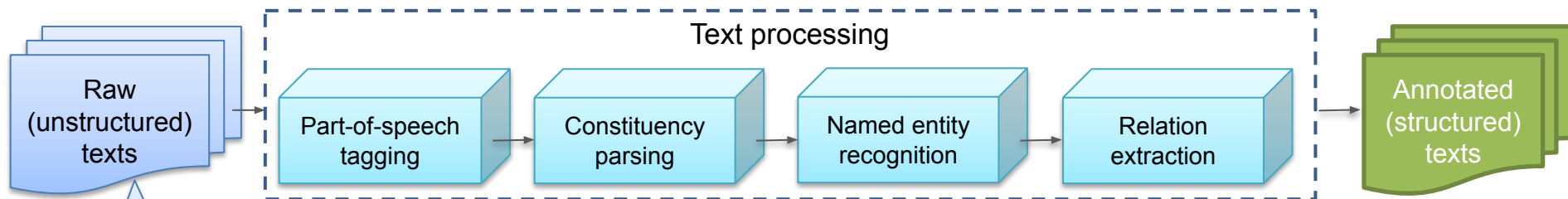
Why not just use dictionary lookup?

NER is not easy!

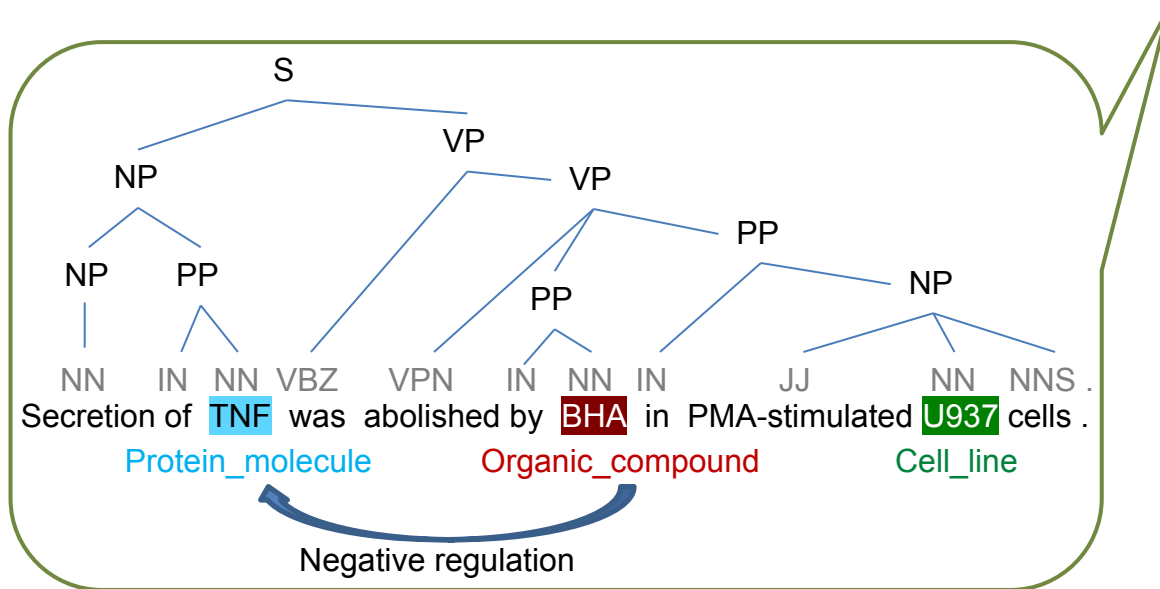
- Washington
 - Cities and Communities (many cities in USA or elsewhere)
 - Street name
 - Person name
 - School name?
- Cat
 - An animal
 - A gene/protein name: <https://www.uniprot.org/uniprot/P62577>
 - Abbreviation of terms, e.g., Common Admission Test

Context dependent!

Information extraction pipeline



. Secretion of TNF was abolished by BHA in PMA-stimulated U937 cells.



Summary

- NER is an important task in information extraction
- NER: need to detect **correct entity mention and category**
- NE categories are dependent on the downstream applications
- Three main approaches to NER