Week 4: Named Entity Linking

Nhung Nguyen

slides courtesy of Phong Le

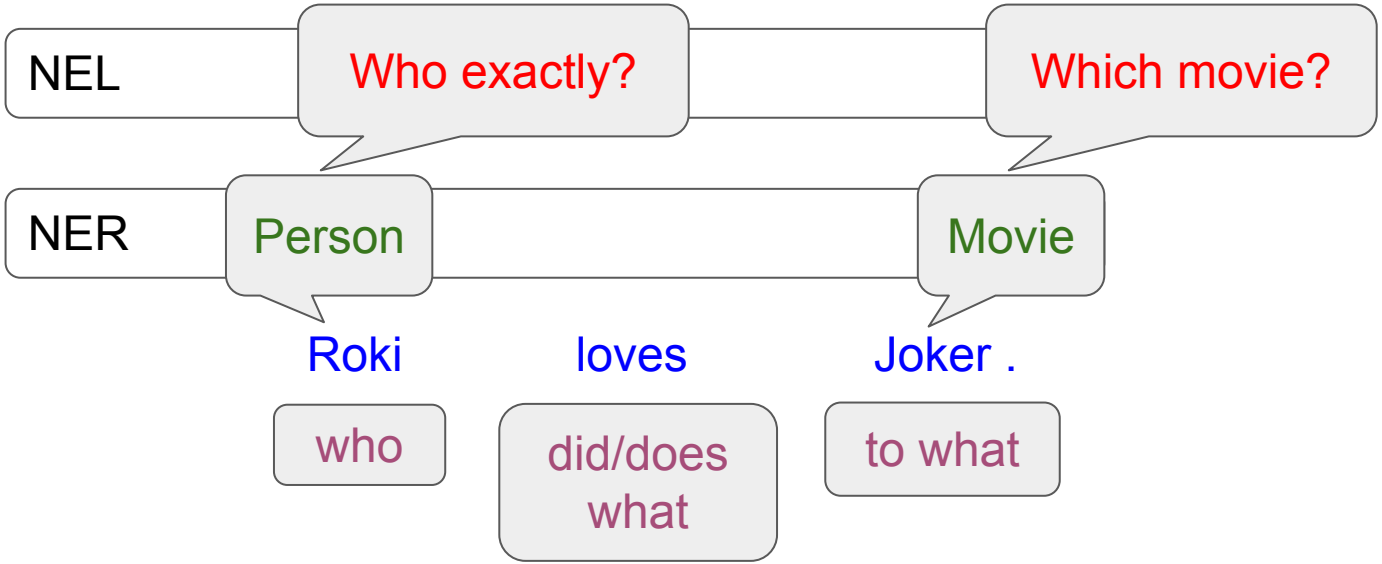
Intended learning outcomes

- Understand the importance of Named Entity Linking (NEL)
- Know several knowledge bases used in the task
- Know basic steps of NEL and their approaches

Materials

- Shen et al., *Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions.* TKDE14
- Ganea and Hofmann, *Deep Joint Entity Disambiguation with Local Neural Attention*, EMNLP 2017

Recap: Named entity recognition



- NER: identify *named entity mentions* and their *types*
- NER however doesn't tell us *who/what exactly the entities are*

Which entities?

- We need a knowledge base to store
  - entries for entities in interest (called entities in short)
  - relations between entities

→ depending on domains / tasks

- News/open domain:
  - Wikipedia (each Wikipedia article represents an entity)
  - Freebase, a set of triplets <subject, relation, object> (subjects and objects are entities)
  - Wikidata, DBpedia, …
- Biomedical
  - ICD-9, MedDRA
  - UMLS
  - SNOMED

Wikipedia

- about 5 million entities, each entities is described by an article

Entity

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

In other projects
Wikimedia Commons
Wikinews
Wikiquote
Wikisource

Print/export

Article Talk Read View source View history Search Wikipedia

Bill Gates

From Wikipedia, the free encyclopedia
(Redirected from Bill gates)

This article is about the co-founder of Microsoft. For other people of the same name, see Bill Gates (disambiguation).

William Henry Gates III (born October 28, 1955) is an American business magnate, software developer, investor, and philanthropist. He is best known as the co-founder of Microsoft Corporation.[2][3] During his career at Microsoft, Gates held the positions of chairman, chief executive officer (CEO), president and chief software architect, while also being the largest individual shareholder until May 2014. He is one of the best-known entrepreneurs and pioneers of the microcomputer revolution of the 1970s and 1980s.

Born and raised in Seattle, Washington, Gates co-founded Microsoft with childhood friend Paul Allen in 1975 in Albuquerque, New Mexico; it went on to become the world's largest personal computer software company.[4][a] Gates led the company as chairman and CEO until stepping down as CEO in January 2000, but he remained chairman and became chief software architect.[7] During the late 1990s, Gates had been criticized for his business tactics, which have been considered anti-competitive. This opinion has been upheld by numerous court rulings.[8] In June 2006, Gates announced that he would be transitioning to a part-time role at Microsoft and full-time work at the Bill & Melinda Gates Foundation, the private charitable foundation that he and his wife, Melinda Gates, established in 2000.[9] He gradually transferred his duties to Ray Ozzie and Craig Mundie.[10] He stepped down as chairman of Microsoft in February 2014 and assumed a new post as technology adviser to support the newly appointed CEO Satya Nadella.[11]

Since 1987, he has been included in the *Forbes* list of the world's wealthiest documented individuals.[12][13] From 1995 to 2017, he held the *Forbes* title of the richest person in the world all but four of those years.[1] In October 2017, he was surpassed by Amazon founder and CEO Jeff Bezos, who had an estimated net worth of US$90.6 billion compared to Gates' net worth of US$89.9 billion at the time.[14] As of November 9, 2019, Gates had an estimated net worth of US$107.1 billion, making him the second wealthiest person in the world, behind Bezos.

Later in his career and since leaving day-to-day operations at Microsoft in 2008, Gates pursued a number of philanthropic endeavors. He donated large amounts of money to various charitable organizations and scientific research programs through the Bill & Melinda Gates Foundation, reported to be the world's largest private charity.[15] In 2009, Gates and Warren Buffett founded The Giving Pledge, whereby they and other billionaires pledge to give at least half of their wealth to philanthropy.[16] The foundation works to save lives and improve global health, and is working with Rotary International to eliminate polio.[17]

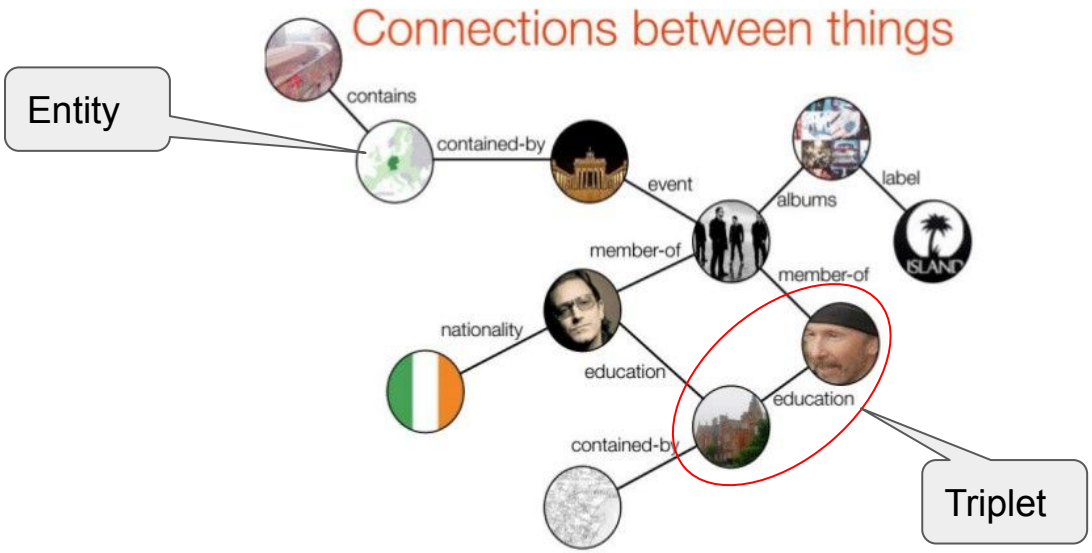
Bill Gates
KBE

Gates in 2014

Born	William Henry Gates III October 28, 1955 (age 64) Seattle, Washington, U.S.
Occupation	Software developer · investor · entrepreneur · philanthropist
Years active	1975–present
Known for	Co-founder of Microsoft
Net worth	US$108.8 billion (Jan 2020)[1]
Title	Co-chairman and co-founder of the Bill & Melinda Gates

Freebase

- Freebase is a knowledge graph (44 million nodes, 2.5 billion edges, Jan 2014)



UMLS - Unified Medical Language System

- Combine many biomedical vocabularies
- UMLS Metathesaurus:
  - Entities: more than 4 millions
  - Number of mentions: more than 15 millions
  - Number of distinct mentions: 13 millions

UMLS Terminology Services | About | Browse | Download | APIs | Tools | Help

UMLS
Metathesaurus Browser

Search

COVID19 (disease)

UMLS CUI: C5203670

Semantic Types: Disease or Syndrome

Definitions (1)

A viral disorder generally characterized by high FEVER; COUGH; DYSPNEA; CHILLS; PERSISTENT TREMOR; MUSCLE PAIN; HEADACHE; SORE THROAT; a new loss of taste and/or smell (see AGEUSIA and ANOSMIA) and other symptoms of a VIRAL PNEUMONIA. In severe cases ... **(MSH)**

Atoms (104)

Filter by Vocabulary
Reset Filters [x]

Search Atoms

Name	AUI	Vocabulary	Term Type	Code
COVID19 (disease)	A32341316	MTH	PN	NOCODE
COVID-19	A32282843	MSH	MH	D000086382
COVID19	A32282844	MSH	ET	D000086382
2019 Novel Coronavirus Disease	A32292069	MSH	ET	D000086382
2019 Novel Coronavirus Infection	A32280360	MSH	ET	D000086382
2019-nCoV Disease	A32286265	MSH	ET	D000086382
2019-nCoV Infection	A32290583	MSH	ET	D000086382
COVID-19 Virus Disease	A32281850	MSH	ET	D000086382
COVID-19 Virus Infection	A32285275	MSH	ET	D000086382
Coronavirus Disease 2019	A32292070	MSH	ET	D000086382
Coronavirus				

Broader Concepts (5)

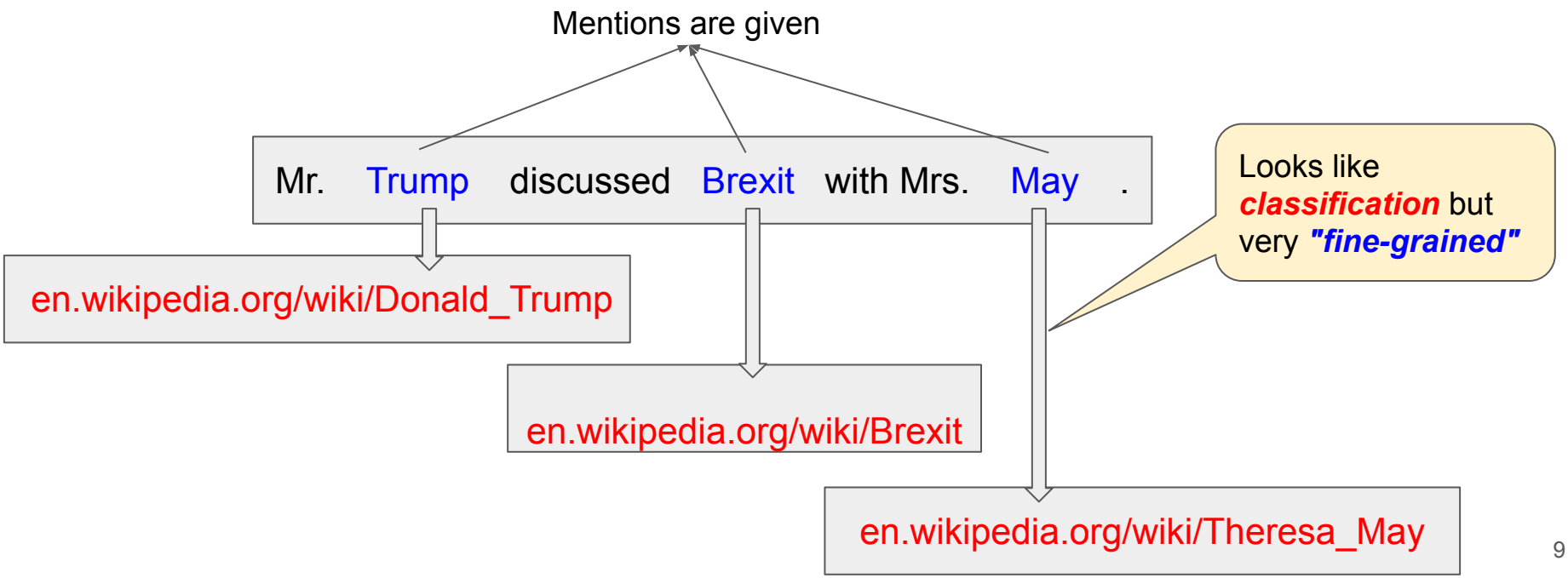
Coronavirus Infections
Emergency use of U07
Lower respiratory tract infection viral
Pneumonia, Viral
Respiratory Tract Infections

Narrower Concepts (6)

Dyspnea caused by Severe acute respiratory syndrome coronavirus 2
Fever caused by Severe acute respiratory syndrome coronavirus 2
Infection of upper respiratory tract caused by Severe acute respiratory syndrome coronavirus 2
Lower respiratory infection caused by SARS-CoV-2
pediatric inflammatory multisystem syndrome
Suspected COVID-19

Named entity linking (NEL)

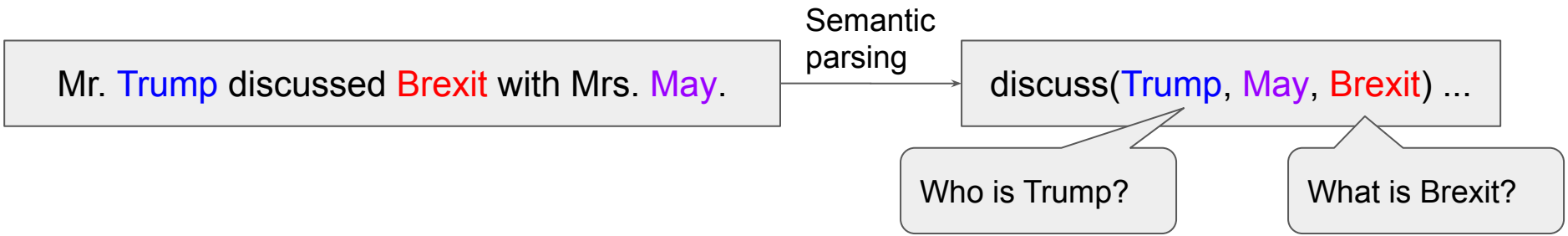
Named entity linking (NEL) is the task of linking a mention to the corresponding entity in a knowledge base (e.g., Wikipedia)



NEL is difficult

- A knowledge base can have a very large number of entities
  - Wikipedia: 5 millions
  - Freebase: 44 millions
  - UMLS: 13 millions
- An entity can have very different surface forms
  - Donald Trump: Trump, President, Snowflake-in-Chief
- A surface form can be linked to several entities
  - Trump: Donald Trump, Ivanka Trump
- Knowledge base can’t cover all entities or forms of entities in texts

NEL is important for natural language understanding



Question: Who discussed Brexit with Theresa May?

Answer:

- Trump ❌
- US President Donald Trump ✅

Entity embeddings

- We can learn entity embeddings in a similar way we learn word embeddings
- Using *Wikipedia* for entity annotations
  - We use pre-trained word embeddings (word2vec or glove)
  - Learn entity embeddings using (entity, context) pairs from Wikipedia
  - (We can jointly train entity embeddings and word embeddings)

...but partially blames the antitrust litigation during the time...

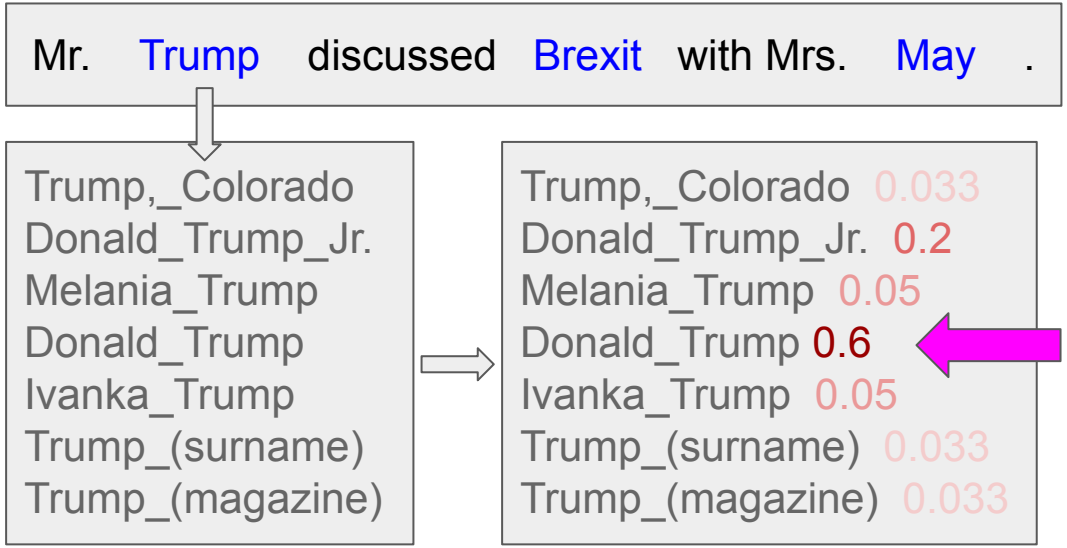
...but partially blames the wiki/United_States_antitrust_law litigation during the time...

Wikipedia2vec (demo)

General approach

- Generate candidates
  - Dictionary-based
  - Rule-based
  - Similarity
  - Information retrieval techniques, etc.
- Rank candidates
  - Classification
  - Neural-based

Example



Candidate generation
(A short list of candidates, < 100)

Candidate ranking

Step 1: Candidate generation

- Extracting an entity-alias dictionary from Wikipedia

Computer

From Wikipedia, the free encyclopedia

For other uses, see Computer (disambiguation).

A **computer** is a machine that can be instructed to carry out sequences of arithmetic or logical operations automatically via computer programming. Modern computers have the ability to follow generalized sets of operations, called programs. These programs enable computers to perform an extremely wide range of tasks. A "complete" computer including the hardware, the operating system (main software), and peripheral equipment required and used for "full" operation can be referred to as a **computer system**. This term may as

Boolean algebra

From Wikipedia, the free encyclopedia

For other uses, see Boolean algebra (disambiguation).

In mathematics and mathematical logic, **Boolean algebra** is the branch of algebra in which the values of the variables are the truth values *true* and *false*, usually denoted 1 and 0 respectively. Instead of elementary algebra where the values of the variables are numbers, and the prime operations are addition and multiplication, the main operations of Boolean

Algebra

From Wikipedia, the free encyclopedia

For the kind of algebraic structure, see Algebra over a field. For other uses, see Algebra (disambiguation).

Algebra (from Arabic: الجبر (*al-jabr*, meaning "reunion of broken parts"[1] and "bonesetting"[2]) is one of the broad parts of mathematics, together with number theory, geometry and analysis. In its most general form, algebra is the study of mathematical symbols and the rules for manipulating these symbols;[3] it is a

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

The quadratic formula expresses

...

Entity: https://en.wikipedia.org/wiki/Boolean_algebra
Alias: logical

Entity: https://en.wikipedia.org/wiki/Computer_program
Alias: programs

entity	alias	count
wiki/Computer_program	programs	120
wiki/boolean_algebra	logical	15
-	boolean	9
...	...	...

Step 1: Candidate generation (cont.)

- Extracting an entity-alias dictionary from Wikipedia

entity	**alias**	**count**
wiki/Computer_program	programs	120
wiki/boolean_algebra	logical	15
-	boolean	9
...	...	...

- An entity-alias dictionary gives us Pr(entity=e | alias/mention=m)

$$Pr(e|m) = \frac{count(e,m)}{\sum_{e'} count(e',m)}$$

smoothing?

Step 1: Candidate generation (cont.)

Selecting n (< 100) candidates by their Pr(entity|mention)

Mr. Trump discussed Brexit with Mrs. May .

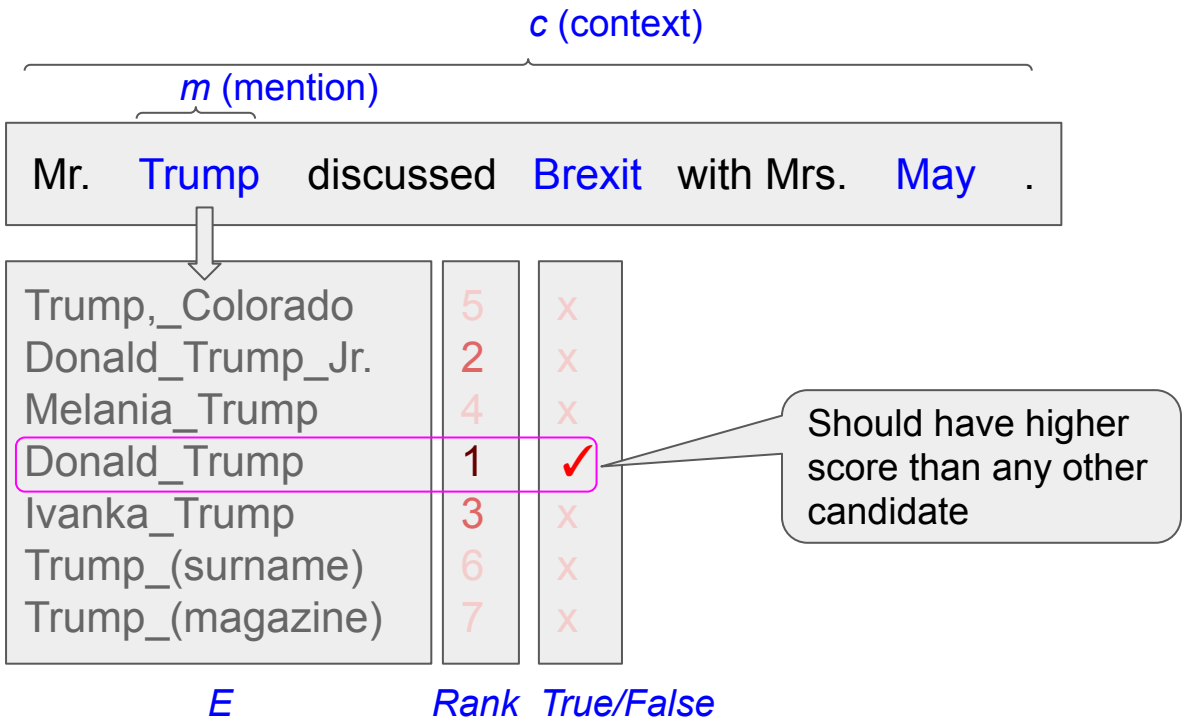
Trump,_Colorado	0.02
Donald_Trump_Jr.	0.2
Melania_Trump	0.02
Donald_Trump	0.6
Ivanka_Trump	0.05
Trump_(surname)	0.005
Trump_(magazine)	0.005
~~Trump_(video_game_player)~~	0.001
~~Trump_(card_games)~~	0.001
~~...~~	...

Pr(. | alias = "Trump")

Using word/entity embeddings?

Step 2: Candidate ranking

- Learning to rank



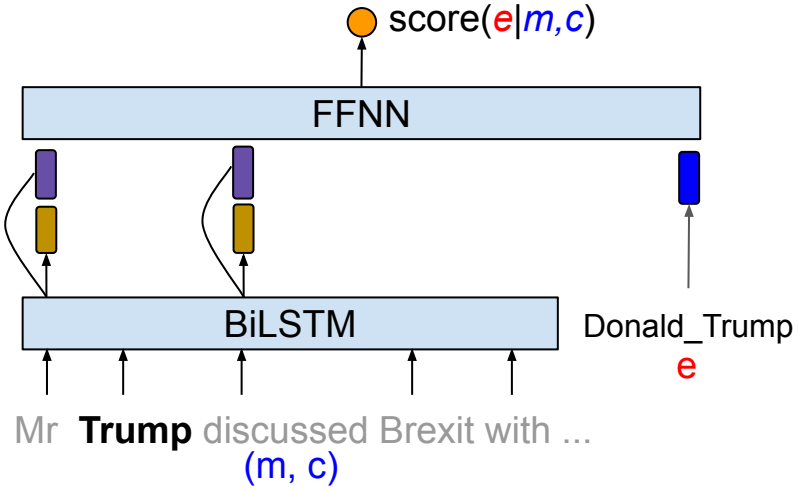
- Like NER, there are two approaches: *local* and *global*

Local approach

- For each mention m in context c, we have a set E of candidates $(e_1,\ldots,e_{|E|})$
- We want the correct entity e has the higher score among the candidates

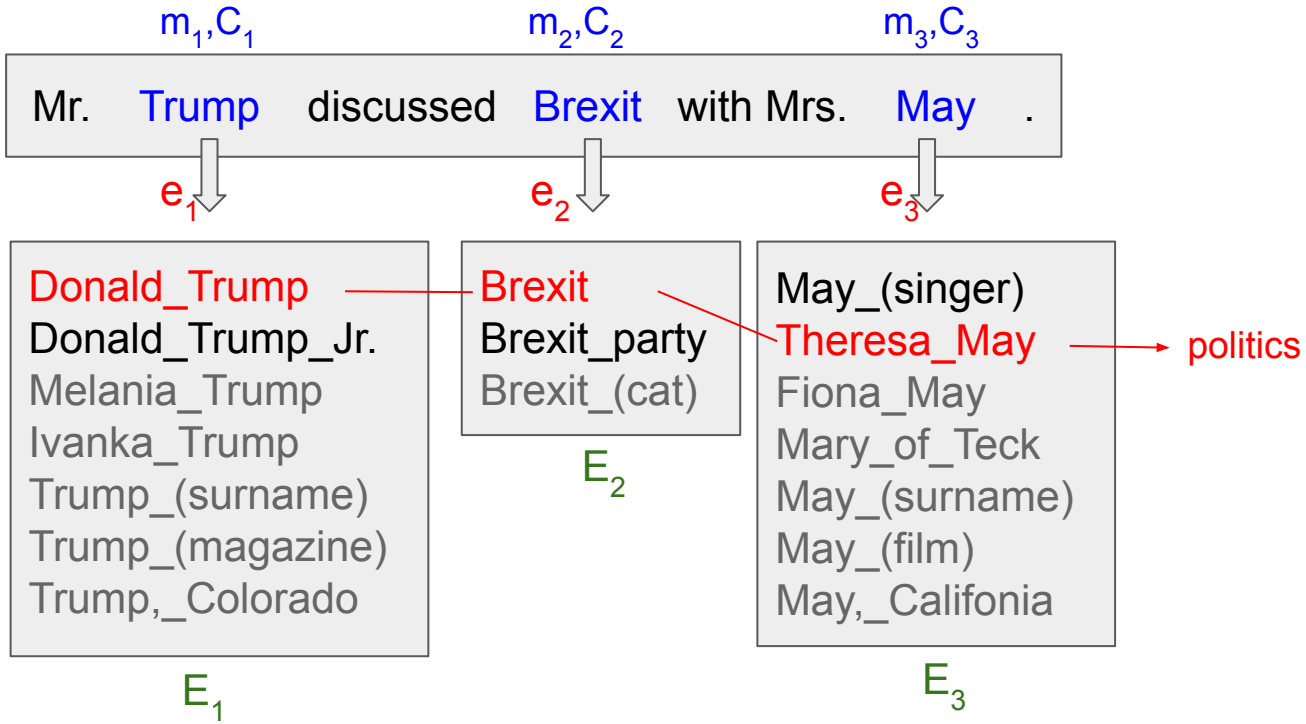
$$\text{score}(e|m,c) > \text{score}(e'|m,c) \text{ for all } e' \in E,\ e' \neq e$$

- Using BiLSTM



Global approach

- Coherency hypothesis: entities appearing together in a document should be *coherent*.



Global approach (cont.)

- $E = (e_1,...,e_n)$, $M = (m_1,...,m_n)$, $C = (c_1,...,c_n)$
- Fully connected graph CRF (Ganea & Hofmann, 2017): Score function of *all entities at onces*

Downstream applications

- Information retrieval
  - Linking ambiguous entity mentions in query to improve the search results
- Content analysis
  - Linking named entity mentions with a knowledge base across documents
- Question answering
  - exploit the entity linking technique to predict the types of questions and candidate answers, and obtain promising results
- Knowledge base (KB) construction/completion
  - Extract new facts/knowledge from texts, link them to existing KB

Summary

- NEL is an important task in language understanding and useful for many downstream applications
- NEL is challenging
- Two main steps of NEL: candidate generation and ranking
- For candidate ranking, we can implement local or global approach using CRF and/or neural networks.