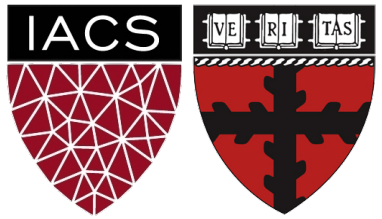
Lecture 15: Entity Linking

Discourse, Pragmatics, and Knowledge Graphs

Harvard

AC295/CS287r/CSCI E-115B

Chris Tanner



ONLY
BUILT
4
ENTITY
LINX...

ANNOUNCEMENTS

- HW4 has been released!
- HW2 and Phase 2 and Quiz 5 are being graded
- Research Project Phase 3 due Oct 28 (Thurs) @ 11:59pm
- I was able to get $100 AWS credit for each student

  (it must be used exclusively for research or HW in this course)

Some of today's slides are inspired from, based on, or directly from:
- **Laura Dietz (UNH)**
- **Alan Black & David Mortensen (CMU)**

Outline

What is entity linking?

Data and metrics

Workshop time

Modern approaches

Outline

What is entity linking?

Data and metrics

Workshop time

Modern approaches

Discourse & Pragmatics

Regarding language, we have mostly learned about:

- Individual words (representation and modelling)
- Large chunks of text (representation and modelling)

Yet, there's tons of nuanced, interconnected elements we wish to capture and model

Multiple levels* to language

Flashback to Lecture 2

Discourse — Understands structures and effects of interweaving dialog

Pragmatics — Understands how context affects meaning

Semantics

Syntax

Lexemes

Morphology

phonology

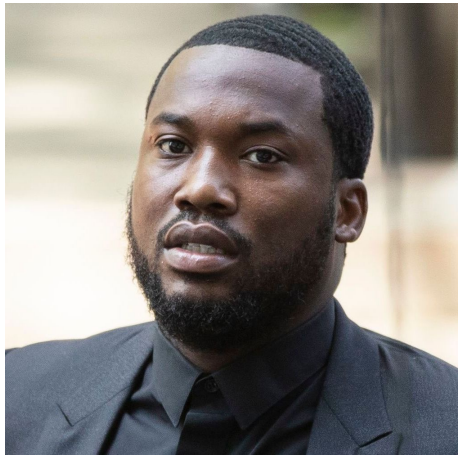
phonetics

speech

orthography

text

*



Discourse & Pragmatics

Pragmatics is a branch of linguistics dealing with language use in context (non-local meaning phenomena)

Ex 1: Can you pass the salt?

Ex 2: "Is he 21?" "Yea, he's 25"

Ex 3: "Are we still going out tonight?" "If I can get off by 5"

Ex 4: Key and Peele – Text Message Confusion (skit)

Too explicit for me to show in class

Discourse & Pragmatics

What does "context" mean?

Social context:

- Social identities, relationships, and setting

Physical context:

- Where? What objects are present? What actions?

Linguistic context:

- Conversation history, discourse context

Other:

- Shared knowledge, common knowledge

Discourse & Pragmatics

Discourse concerns the structures and effects in related sequences of sentences. Essential for dialogues, multi-party conversations.

Ex 1: "I said the blue shoes." "Oh, *blue*!"

Ex 2: "Can we discuss this later today?"

"When did you have in mind?"

"Does the afternoon work?"

"Well, I'm free after lunch"

"Shall we plan for 2pm?"

"That works"

Types of referring expressions

Indefinite noun phrases:

- I saw an incredible oak tree today

Definite noun phrases:

- I read about it in the New York Times

Pronominal mentions:

- Emily aced the quiz, as she expected

Nominal mentions and names:

- The amazing marathoner, Des Linden, is a true inspiration

Demonstratives:

- These pretzels are making me thirsty.

This, that, these, those

Discourse & Pragmatics

Types of referring expressions

Indefinite noun phrases:

Remember, in NLP, we devise feasible tasks to *approximate* the phenomenon (e.g., understanding discourse and pragmatics) we hope to eventually capture

Nominal mentions and names:

- The amazing marathoner, Des Linden, is a true inspiration

Demonstratives:

- These pretzels are making me thirsty.

Types of referring expressions

Indefinite noun phrases:

- I saw an incredible oak tree today

Definite noun phrases:

- I read about it in the New York Times

Pronominal mentions:

- Emily aced the quiz, as she expected

Nominal mentions and names:

- The amazing marathoner, Des Linden, is a true inspiration

Demonstratives:

- These pretzels are making me thirsty.

Entity Linking

The New York Times

Biden Backs Compromise to Win a Vast Social Agenda

The president's trip to Scranton, Pa., comes as Democrats close in on a deal to advance two bills carrying a scaled-back version of his domestic policy priorities.

Not logged in Talk Contributions Create account Log in

WIKIPEDIA
The Free Encyclopedia

Article Talk | Read View source View history | Search Wikipedia

Joe Biden

From Wikipedia, the free encyclopedia

"Joseph Biden" and "Biden" redirect here. For his tenure as president, see Presidency of Joe Biden. For his son Joseph Biden III, see Beau Biden. For other uses, see Biden (disambiguation).

Joseph Robinette Biden Jr.[a] (/ˈbaɪdən/ *BY-dən*; born November 20, 1942) is an American politician who is the 46th and current president of the United States. A member of the Democratic Party, he served as the 47th vice president from 2009 to 2017 under Barack Obama and represented Delaware in the United States Senate from 1973 to 2009.

Born and raised in Scranton, Pennsylvania, and later in New Castle County, Delaware, Biden studied at the University of Delaware before earning his law degree from Syracuse University in 1968. He was elected to the New Castle County Council in 1970 and became the sixth-youngest senator in U.S. history after he was elected to the United States Senate from Delaware in 1972, at age 29. Biden was the chair or ranking member of the Senate Foreign Relations Committee for 12 years and was influential in foreign affairs during Obama's presidency. He also chaired the Senate Judiciary Committee from 1987 to 1995, dealing with drug policy, crime prevention, and civil liberties issues; led the effort to pass the Violent Crime Control and Law Enforcement Act and the Violence Against Women Act; and oversaw six U.S. Supreme Court confirmation hearings, including the contentious hearings for Robert Bork and Clarence Thomas. He ran unsuccessfully for

Joe Biden

Official portrait, 2021

46th President of the United States

Incumbent

Assumed office
January 20, 2021

Vice President Kamala Harris

Preceded by Donald Trump

47th Vice President of the United States

In office
January 20, 2009 – January 20, 2017

President Barack Obama

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

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

Print/export
Download as PDF
Printable version

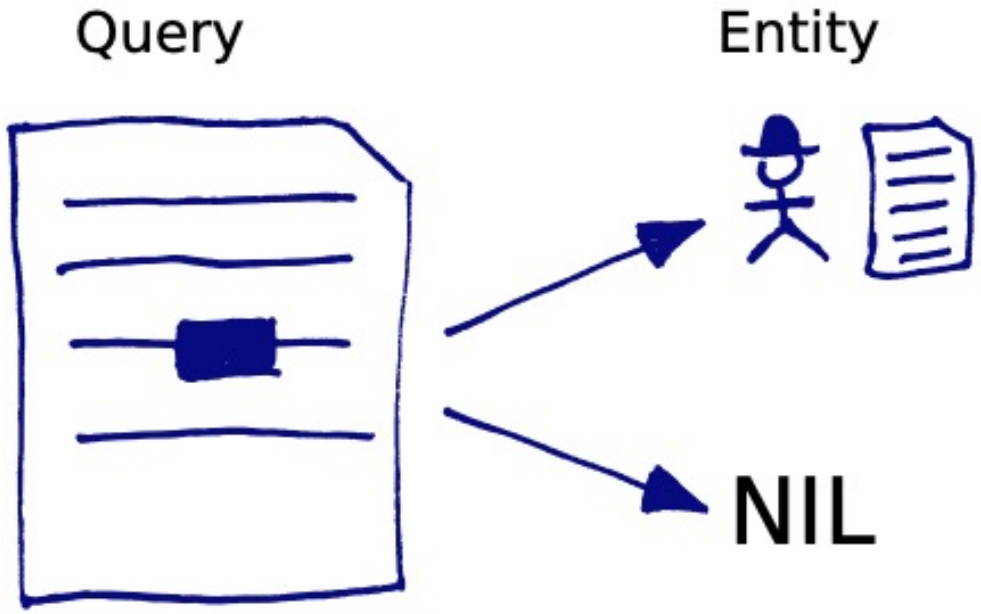
In other projects
Wikimedia Commons
Wikinews

Entity Linking

Rep. Pramila Jayapal of Washington, the Democratic leader of House progressives, described the discussions as “conversational,” but said the president was moving to bring the debate to an end.

“He is the closer, he is the convincer, the mediator in chief,” she said.

Entity Linking



Identify all named mentions (not nominal mentions) and disambiguate them by linking each of them to an external knowledge graph (KG) (e.g., Wikipedia's KG)

Entity Linking

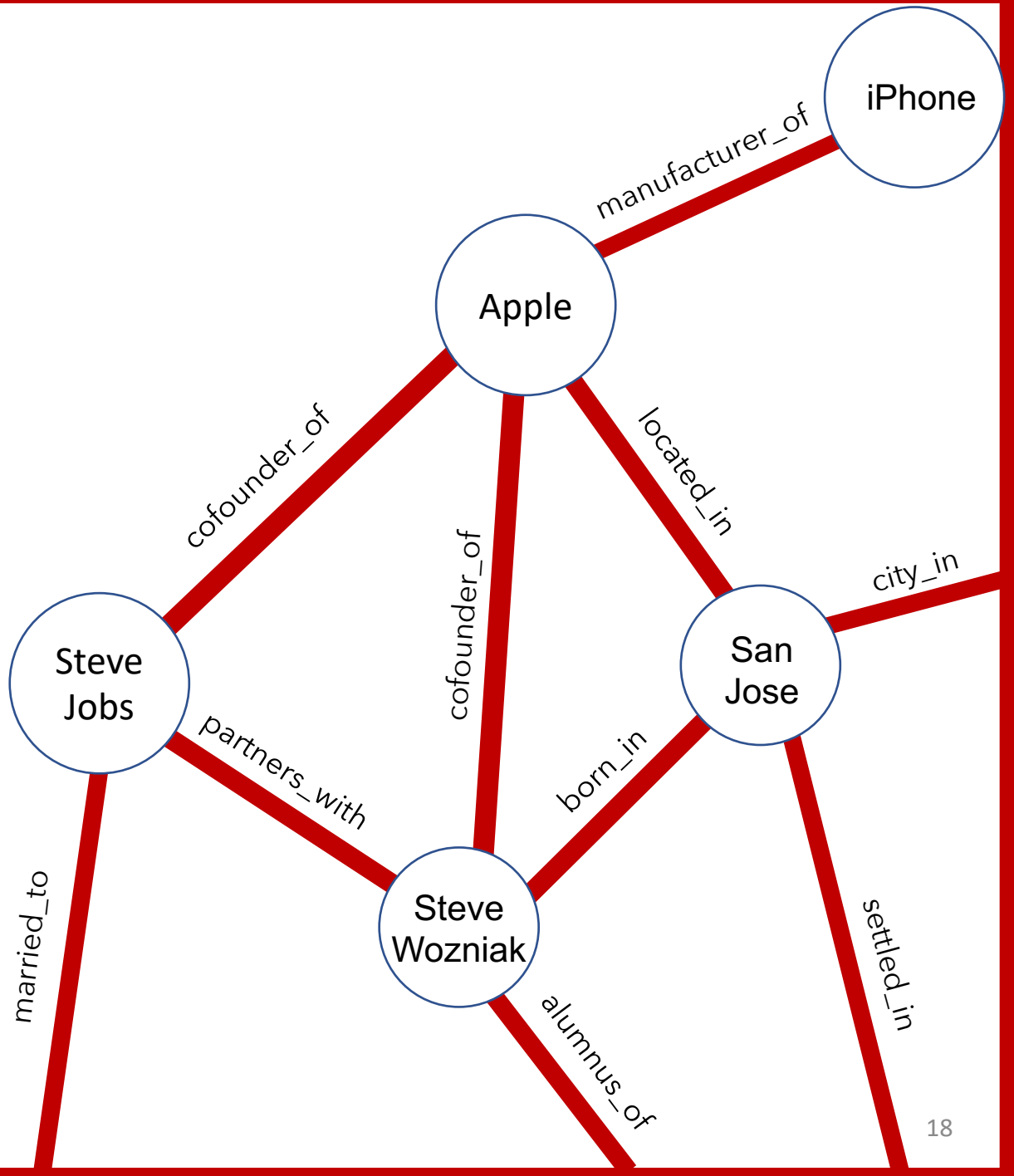
Entity Linking is also called Named Entity Disambiguation

Identify all named mentions (not nominal mentions) and disambiguate them by linking each of them to an external knowledge graph (KG) (e.g., Wikipedia's KG)

Entity Linking

Task Definition:

Determine which **named mentions** refer to which underlying real-world entities (within a knowledge graph).



Entity Linking

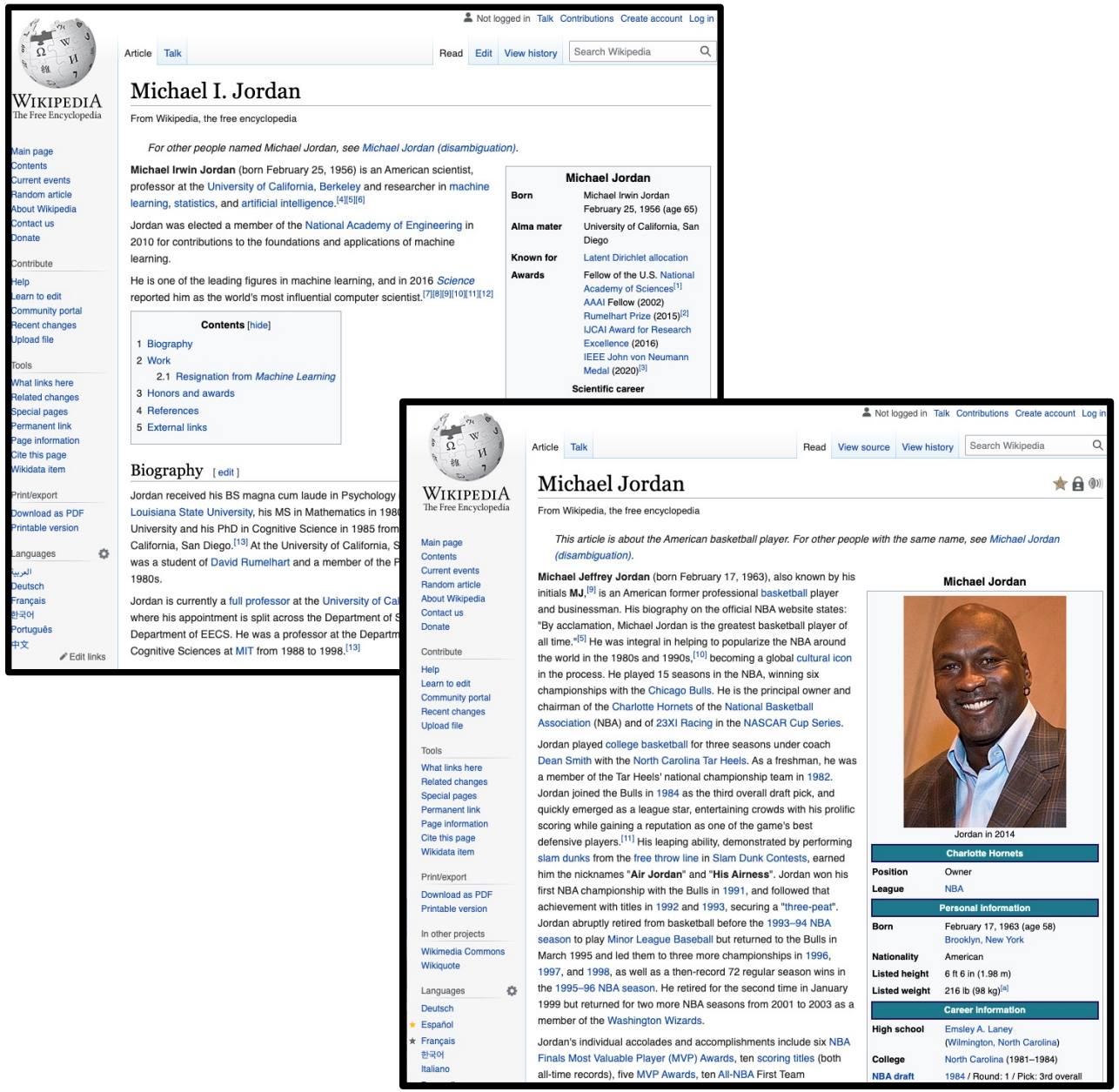
Michael Jordan, born in 1956, spoke with the Associated Press today after he announced a new $10M college scholarship fund

Document #271

Entity Linking

Michael Jordan, born in 1956, spoke with the Associated Press today after he announced a new $10M college scholarship fund

Document #271



Entity Linking

Michael Jordan (disambiguation)

From Wikipedia, the free encyclopedia

Michael Jordan (born 1963) is an American businessman and former professional basketball player.

Michael Jordan or **Mike Jordan** may also refer to:

Contents [hide]

1 People
  1.1 Sports
  1.2 Other people
2 Other uses
3 See also

People [edit]

Sports [edit]

- Michael Jordan (footballer) (born 1986), English goalkeeper
- Mike Jordan (racing driver) (born 1958), English racing driver
- Mike Jordan (baseball, born 1863) (1863–1940), American baseball player
- Mike Jordan (cornerback) (born 1992), American football cornerback
- Michael Jordan (offensive lineman) (born 1998), American football offensive lineman
- Michael-Hakim Jordan (born 1977), American professional basketball player
- Michal Jordán (born 1990), Czech ice hockey player

Other people [edit]

- Michael B. Jordan (born 1987), American actor
- Michael I. Jordan (born 1956), American researcher in machine learning and artificial intelligence
- Michael Jordan (insolvency baron) (born 1931), English businessman
- Michael Jordan (Irish politician), Irish Farmers' Party TD from Wexford, 1927–1932
- Michael H. Jordan (1936–2010), American executive for CBS, PepsiCo, Westinghouse
- Michael Jordan (mycologist), English mycologist

Entity Linking

"Entity Linking" is a two-stage process!

1. Mention detection (some people work on this exclusively)
2. Entity Linking

Outline

What is entity linking?

Data and metrics

Workshop time

Modern approaches

Outline

- What is entity linking?
- Data and metrics
- Workshop time
- Modern approaches

Annotations

Q1: What if our document only had “Michael”, not “Michael Jordan”?

Q2: What if our document mentions his name 10 times in the same doc?

Q3: What if our document contains an unpopular name, like Chris Tanner?

Annotations

Q1: What if our document only had “Michael”, not “Michael Jordan”?

A1: Only annotate the mentions that have sufficient context to make the task meaningful

Annotations

Q2: What if our document mentions his name 10 times in the same doc?

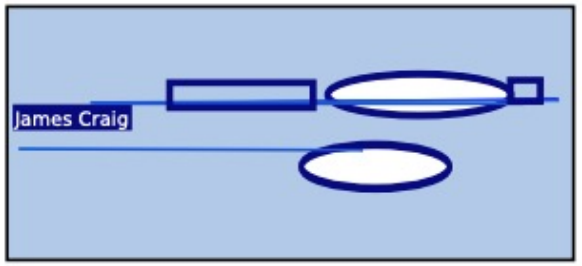
A2: Only annotate the first occurrence

Annotations

Q3: What if our document contains an unpopular name, like Chris Tanner?

A3: Can only annotate mentions that are present in the KG

Annotations

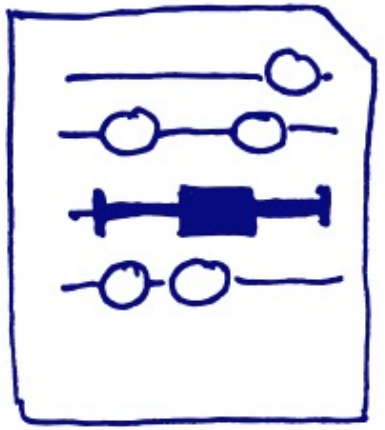


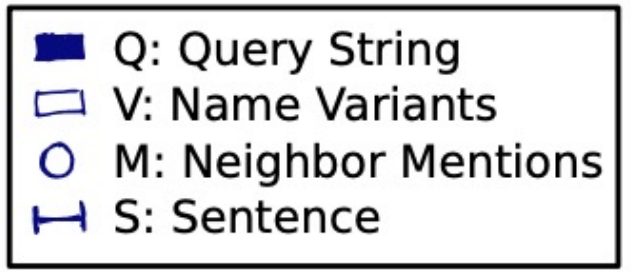
Name Variants:
- Within-doc Coreference

Neighbor Mentions:
- NER Tagger
- (Alternative Mention Detection)

Sentence:
- Term models

Symbol Notation:





Datasets

TAC KBP 2010

TACKBP-2010 is widely used for evaluating entity linking systems Ji et al. (2010).[4] Following prior work, we measure *in-KB* accuracy (P@1). There are 1,074 and 1,020 annotated mention/entity pairs derived from 1,453 and 2,231 original news and web documents on training and evaluation dataset, respectively. All the entities are from the TAC Reference Knowledgebase which contains 818,741 entities with titles, descriptions and other meta info.

AIDA/CoNLL

For training we used the biggest publicly available EL dataset, AIDA/CoNLL (Hoffart et al., 2011), consisting of a training set of 18,448 linked mentions in 946 documents, a validation set of 4,791 mentions in 216 documents, and a test set of 4,485 mentions in 231 documents.

Metrics

- Accuracy
- Recall @ N
- Micro F1
- Macro F1

Disambiguation-Only Approach

- Micro-Precision: Fraction of correctly disambiguated named entities in the full corpus.
- Macro-Precision: Fraction of correctly disambiguated named entities, averaged by document.

End-to-End Approach

- Gerbil Micro-F1 - strong matching: micro InKB F1 score for correctly linked and disambiguated mentions in the full corpus as computed using the Gerbil platform. InKB means only mentions with valid KB entities are used for evaluation.
- Gerbil Macro-F1 - strong matching: macro InKB F1 score for correctly linked and disambiguated mentions in the full corpus as computed using the Gerbil platform. InKB means only mentions with valid KB entities are used for evaluation.

Outline

- What is entity linking?
- Data and metrics
- Workshop time
- Modern approaches

Outline

What is entity linking?

Data and metrics

Workshop time

Modern approaches

Workshop time

How would you attempt entity linking?

Create a baseline system

Workshop time

Method 1: Popularity of Links

Step 1:
Build a dictionary of names for each entity.

Step 2:
Inspect all KB entities that have the query mention as a name variant.

Step 3:
Choose the entity with the most inlinks through this name.

Names and Links on Wikipedia

Northern Ireland

Northern Ireland (Irish: *Tuaisceart Éireann*[4] [ˈt̪ˠuəʃcəɾˠt̪ˠ ˈeːɾʲəɳˠ] (listen); Ulster Scots: *Norlin Airlann* or *Norlin Airlan*) is a part of the United Kingdom in the north-east of the island of Ireland. It is variously described as a country, province or region of the UK, amongst other terms.[5][6][7] Northern Ireland shares a border to the south and west with the Republic of Ireland. As of 2011, its

Unionism in Ireland

In Ireland, unionism, also called **Ulster unionism** in the context of Northern Ireland or – chiefly historically – Ulster specifically, refers to an ideology which favours the continuation of some form of political union between the islands of Ireland and Great

Unionists were in the majority in four counties of the Ulster (Antrim, Londonderry, Down and Armagh), and formed a large minority in the remaining counties of Fermanagh and Tyrone. Sir Edward Carson had expressly urged the new Prime Minister, Sir James Craig, to ensure absolute equality in the treatment of Catholics, so to guarantee the stability

James Craig, 1st Viscount Craigavon

From Wikipedia, the free encyclopedia

For other uses, see James Craig (disambiguation).

James Craig, 1st Viscount Craigavon, PC, PC (NI) (8 January 1871 – 24 November 1940), was a prominent Irish unionist politician, leader of the Ulster Unionist Party and the first Prime Minister of Northern Ireland. He was created a baronet in 1918 and raised to the Peerage in 1927.

The Right Honourable

The Viscount Craigavon
PC

Mining Name Variants and Neighbors

Northern Ireland

Northern Ireland

James Craig, 1st Viscount Craigavon

Sir James Craig

1st Viscount Craigavon

Ulster Unionists

Northern Ireland

Prime Minister of Northern Ireland

Unionism in Ireland

Irish Unionist

Unionism in Ireland

Ulster

Northern Ireland

James Craig, 1st Viscount Craigavon

Pros & Cons: Popularity of Links

Works for very popular entities
such as "Northern Ireland"

Fails for entities with confusable names
"James Craig", "Springfield", "Jaguar"

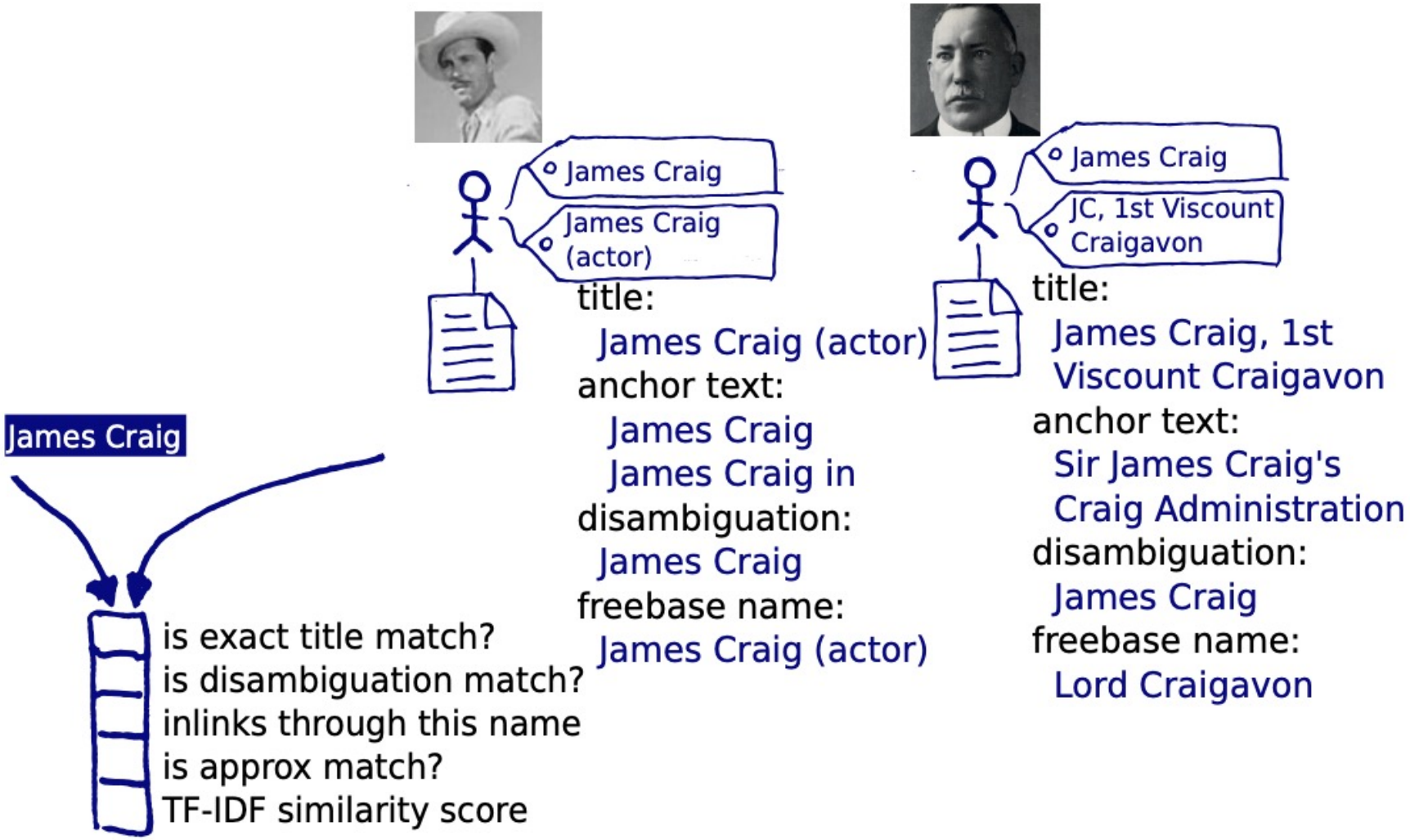
Method 2: Machine Learn Similarity

Step 1:
Collect different similarity features of query mention and entities

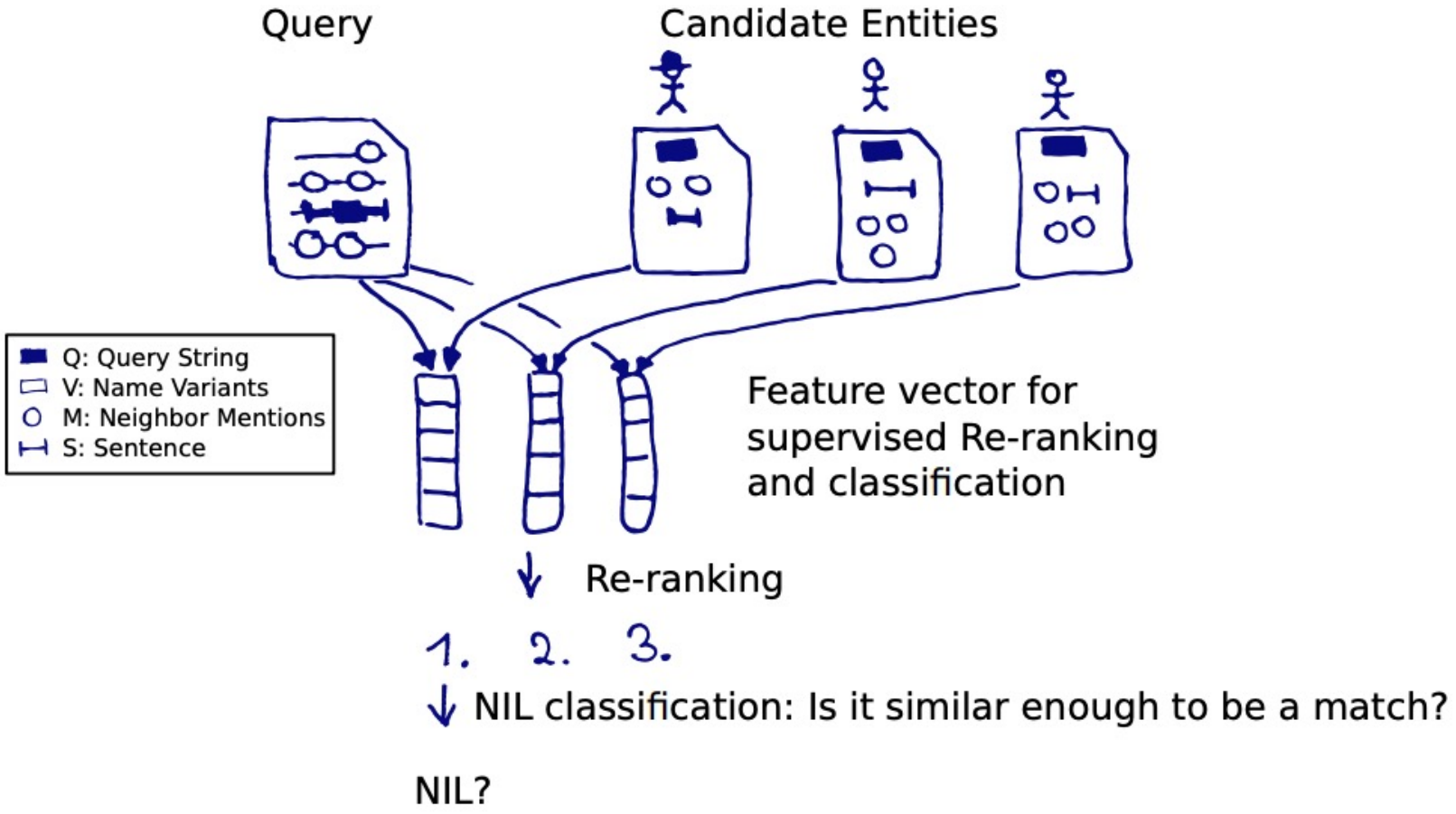
Step 2:
Machine learn the feature weights on training data (e.g. learning to rank)

Step 3:
Apply similarity to query and each entity, select the most similar entity.

Method 2: Similarity Features



Learn Similarity and NIL



Features: Name variants, Document Terms, Links, Popularity ...

Pros & Cons:Machine Learn Similarity

Pro: Combination of different indicators of similarity; option to predict "NILs".

Pro: Can incorporate name variants found in the text (coreference tools)

Con: Requires selection of a pool of candidate entities, which can be large ("John Smith").

Will still fail on "James Craig", because the wrong James has more anchor text matches.

Method 3: Context Disambiguation

Step 1:
Identify surrounding text, entities, etc.

Step 2:
Issue search query containing all of it.

Different Kinds of Context

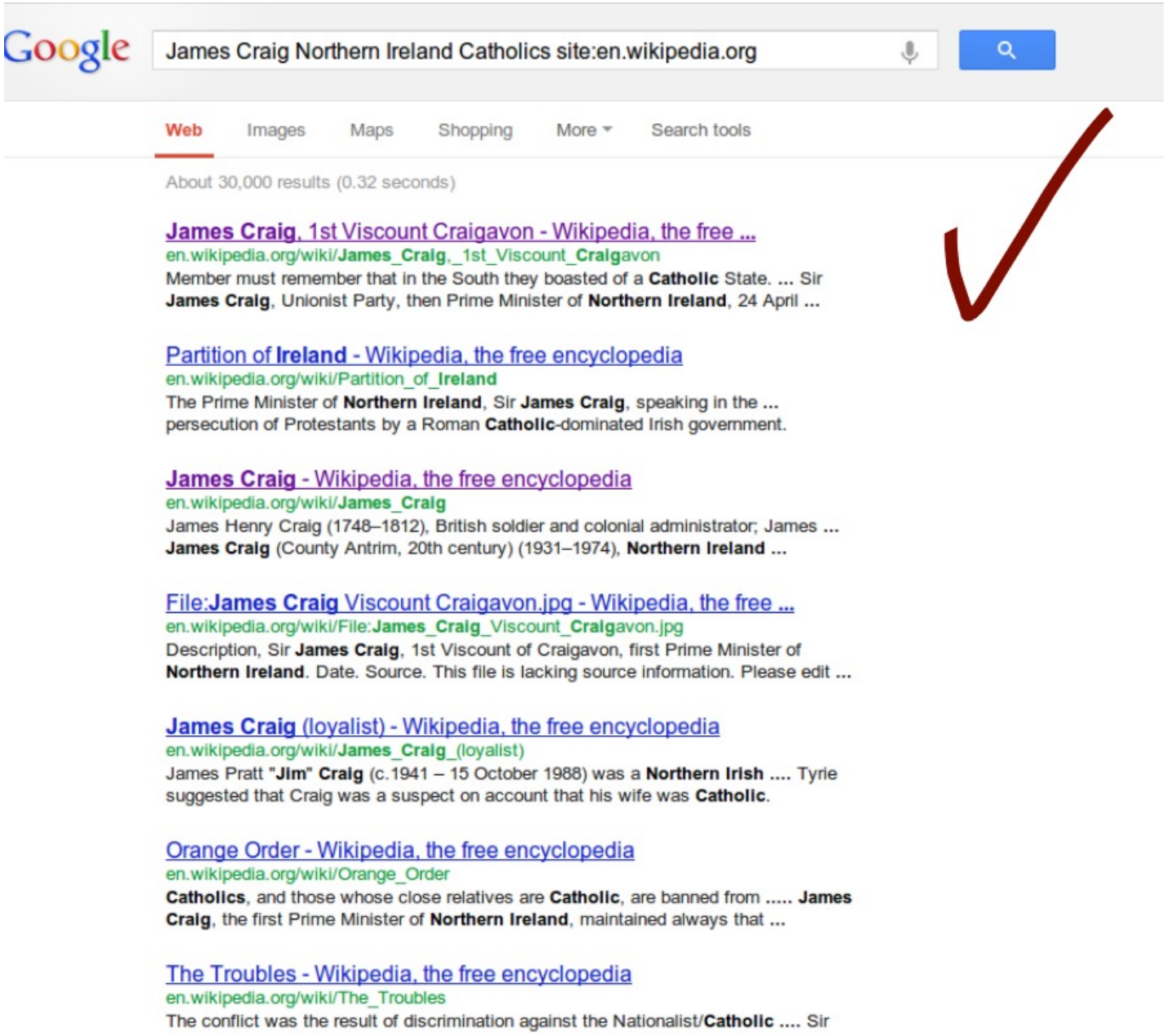
Example Query:

Northern Ireland has a population of about one and a half million people. At the time of partition in 1921 Protestants / unionists had a two-thirds majority in the region. The first Prime Minister of Northern Ireland, Sir James Craig, described the state as having 'a Protestant Parliament for a Protestant people.' The state effectively discriminated against Catholics in housing, jobs, and political representation.

http://cain.ulst.ac.uk/othelem/incorepaper09.htm

Search for:

James Craig + Name Variants + Neighbors + Sentence


Google
James Craig Northern Ireland Catholics site:en.wikipedia.org
Web
Images
Maps
Shopping
More
Search tools
About 30,000 results (0.32 seconds)
James Craig, 1st Viscount Craigavon - Wikipedia, the free ...
en.wikipedia.org/wiki/James_Craig,_1st_Viscount_Craigavon
Member must remember that in the South they boasted of a Catholic State. ... Sir James Craig, Unionist Party, then Prime Minister of Northern Ireland, 24 April ...
Partition of Ireland - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Partition_of_Ireland
The Prime Minister of Northern Ireland, Sir James Craig, speaking in the ... persecution of Protestants by a Roman Catholic-dominated Irish government.
James Craig - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/James_Craig
James Henry Craig (1748–1812), British soldier and colonial administrator; James ... James Craig (County Antrim, 20th century) (1931–1974), Northern Ireland ...
File:James Craig Viscount Craigavon.jpg - Wikipedia, the free ...
en.wikipedia.org/wiki/File:James_Craig_Viscount_Craigavon.jpg
Description, Sir James Craig, 1st Viscount of Craigavon, first Prime Minister of Northern Ireland. Date. Source. This file is lacking source information. Please edit ...
James Craig (loyalist) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/James_Craig_(loyalist)
James Pratt "Jim" Craig (c.1941 – 15 October 1988) was a Northern Irish .... Tyrie suggested that Craig was a suspect on account that his wife was Catholic.
Orange Order - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Orange_Order
Catholics, and those whose close relatives are Catholic, are banned from ..... James Craig, the first Prime Minister of Northern Ireland, maintained always that ...
The Troubles - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/The_Troubles
The conflict was the result of discrimination against the Nationalist/Catholic .... Sir

Method 3: Pros and Cons

Works for "James Craig"!

Problematic when neighbors are ambiguous:

"Lisa witnessed a shooting at Springfield high school".

(Unclear which "Lisa" and which "Springfield")

Method 3: Pros and Cons

Also problematic when neighbors don't provide enough disambiguation power

Example, all other James Craigs of Ireland which are less popular.

Workshop time

These approaches seem kind of manual and kludgy.

How can we use Deep Learning?

Outline

What is entity linking?

Data and metrics

Workshop time

Modern approaches

Outline

What is entity linking?

Data and metrics

Workshop time

Modern approaches

Outline

- What is entity linking?
- Data and metrics
- Workshop time
- Modern approaches

Modern approach

End-to-End Neural Entity Linking

Nikolaos Kolitsas *
ETH Zürich
nikos_kolitsas@hotmail.com

Octavian-Eugen Ganea *
ETH Zürich
octavian.ganea@inf.ethz.ch

Thomas Hofmann
ETH Zürich
thomas.hofmann@inf.ethz.ch

Motivation:

End-to-end EL is the realistic task and ultimate goal, but challenges in joint NER/MD and ED modeling arise from their different nature. Few previous methods tackle the joint task, where errors in one stage can be recovered by the next stage. One of

1) MD may split a larger span into two mentions of less informative entities:
B. Obama's wife gave a speech [...]
Federer's coach [...]
2) MD may split a larger span into two mentions of incorrect entities:
Obama Castle was built in 1601 in Japan.
The Kennel Club is UK's official kennel club.
A bird dog is a type of gun dog or hunting dog.
Romeo and Juliet by Shakespeare [...]
Natural killer cells are a type of lymphocyte
Mary and Max, the 2009 movie [...]
3) MD may choose a shorter span, referring to an incorrect entity:
The Apple is played again in cinemas.
The New York Times is a popular newspaper.
4) MD may choose a longer span, referring to an incorrect entity:
Babies Romeo and Juliet were born hours apart.

Table 1: Examples where MD may benefit from ED and viceversa. Each wrong MD decision (underlined) can be avoided by proper context understanding. The correct spans are shown in blue.

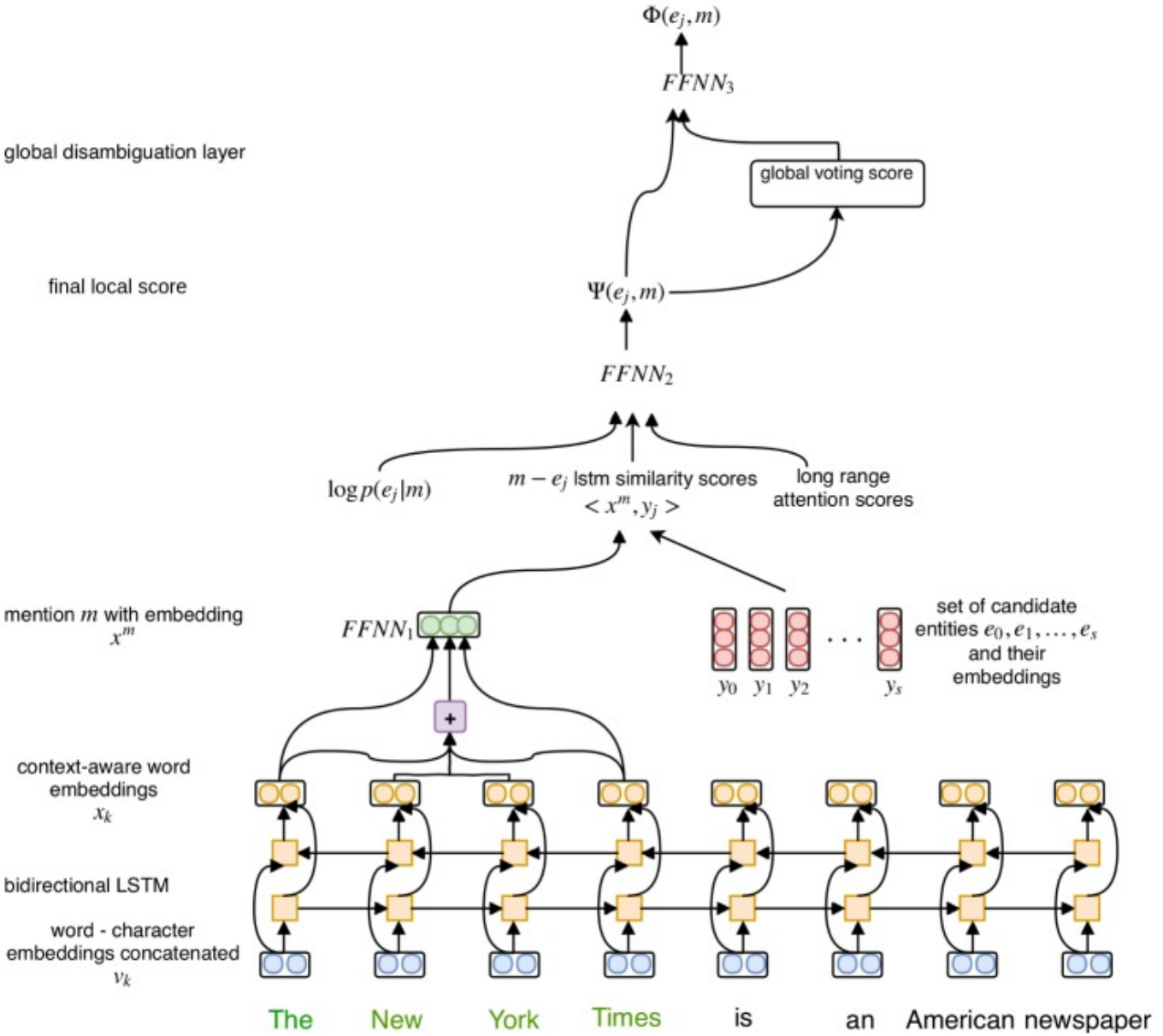


Figure 1: Our global model architecture shown for the mention *The New York Times*. The final score is used for both the mention linking and entity disambiguation decisions.

Modern approach

Scalable Zero-shot Entity Linking with Dense Entity Retrieval

Ledell Wu,[1] Fabio Petroni,[1] Martin Josifoski,[3*] Sebastian Riedel,[1,2] Luke Zettlemoyer[1]
[1]Facebook AI Research
{ledell, fabiopetroni, sriedel, lsz}@fb.com
[2]University College London
[3]Ecole Polytechnique Federale de Lausanne
martin.josifoski@epfl.ch

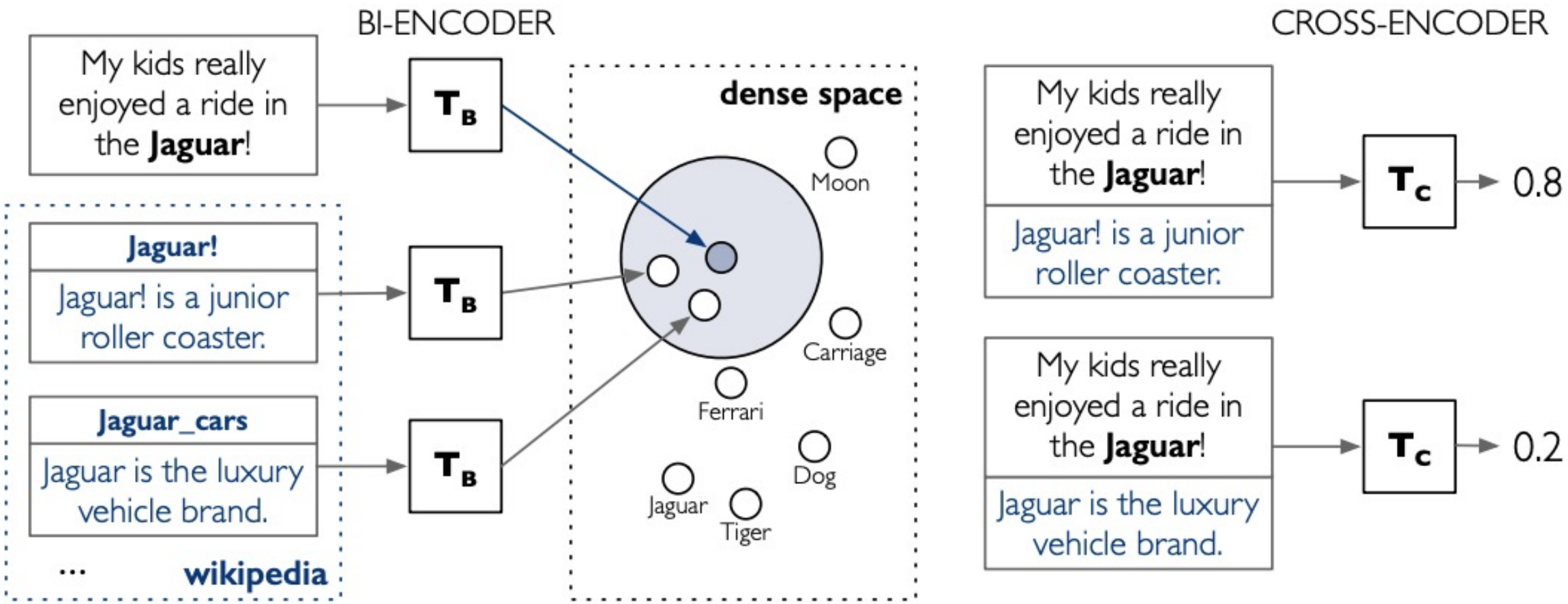


Figure 1: High level description of our zero-shot entity linking solution. From the top-left, the input gets encoded in the same dense space where all entities representations lie. A nearest neighbors search is then performed (depicted with a blue circle), k entities retrieved and supplied to the cross encoder. The latter attends over both input text and entities descriptions to produce a probability distribution over the candidates.