Information Extraction

Announcements

- Course evaluation: please fill out
- HW4 extended to noon, 12/4
- Thursday: bring your laptop!
  - Poetry generation
  - Final review
- Final exam: 12/21, final is cumulative
- What topics would you like to hear about again?

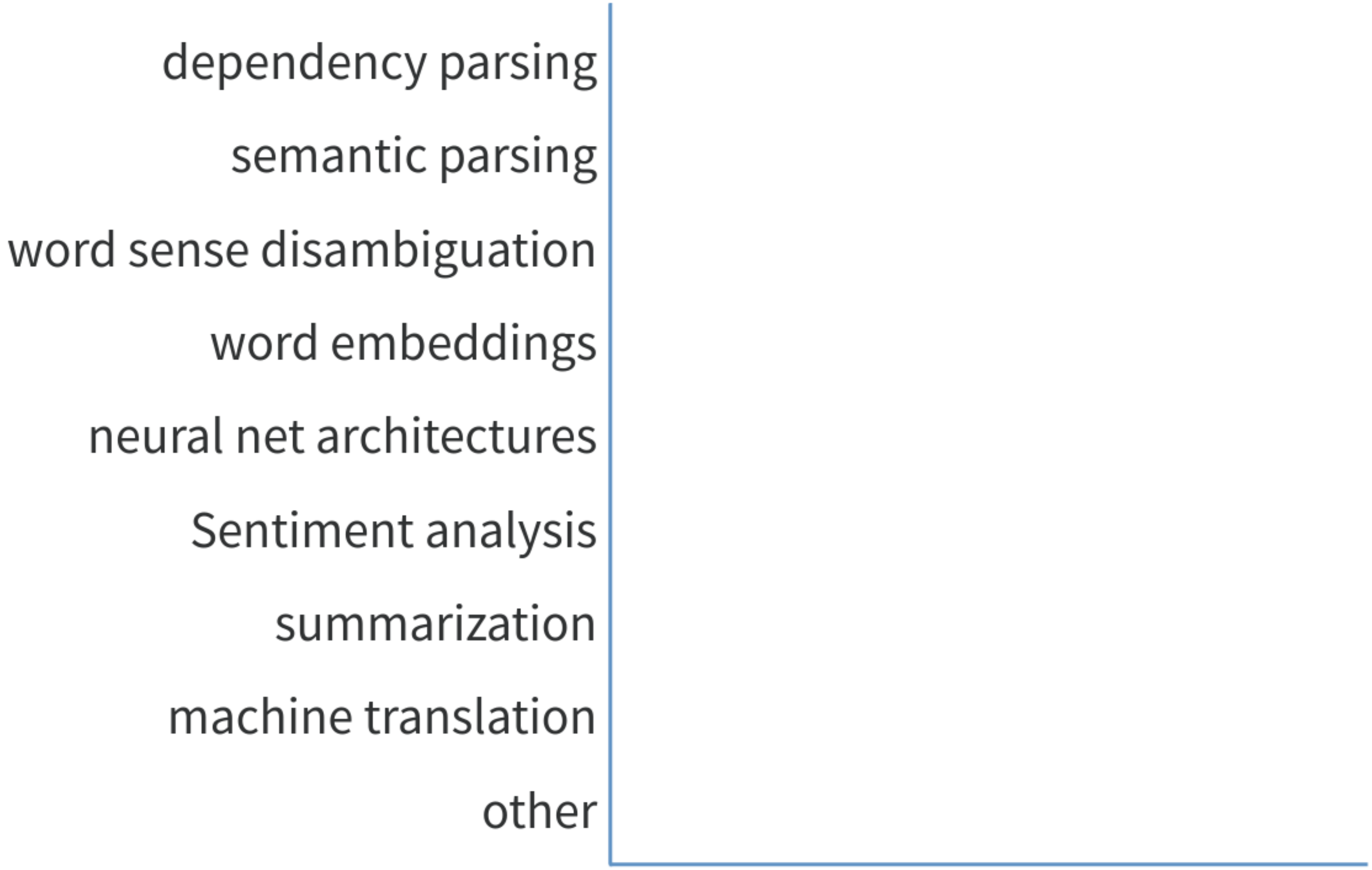
- Dan Jurafsky's talk: 5pm today, CEPSR auditorium.
  Hope to see you all there!
  "Does this Vehicle Belong to You?"

Enabling Connections in a Hyperconnected World through Emotion AI
Taniya Mishra, Affectiva

- Date: Wednesday, December 6th
- Time: 7-8pm
- Location: Room 504 of the Diana Center

- Abstract:
- We live in a hyperconnected world powered by smart devices. A big part of building connections is recognizing and responding to emotions. But our smart devices still lack this fundamental aspect of social communication, rendering our interactions with or through them superficial and limited. Now imagine if we could empower our devices with intelligence to recognize human emotions. The results would be transformative, ranging from empathetic chatbots to personalized digital signage to smart cars that ensure the comfort and safety of passengers by recognizing their emotions. Emotion AI — emotion estimation via artificial intelligence — can make this possible.

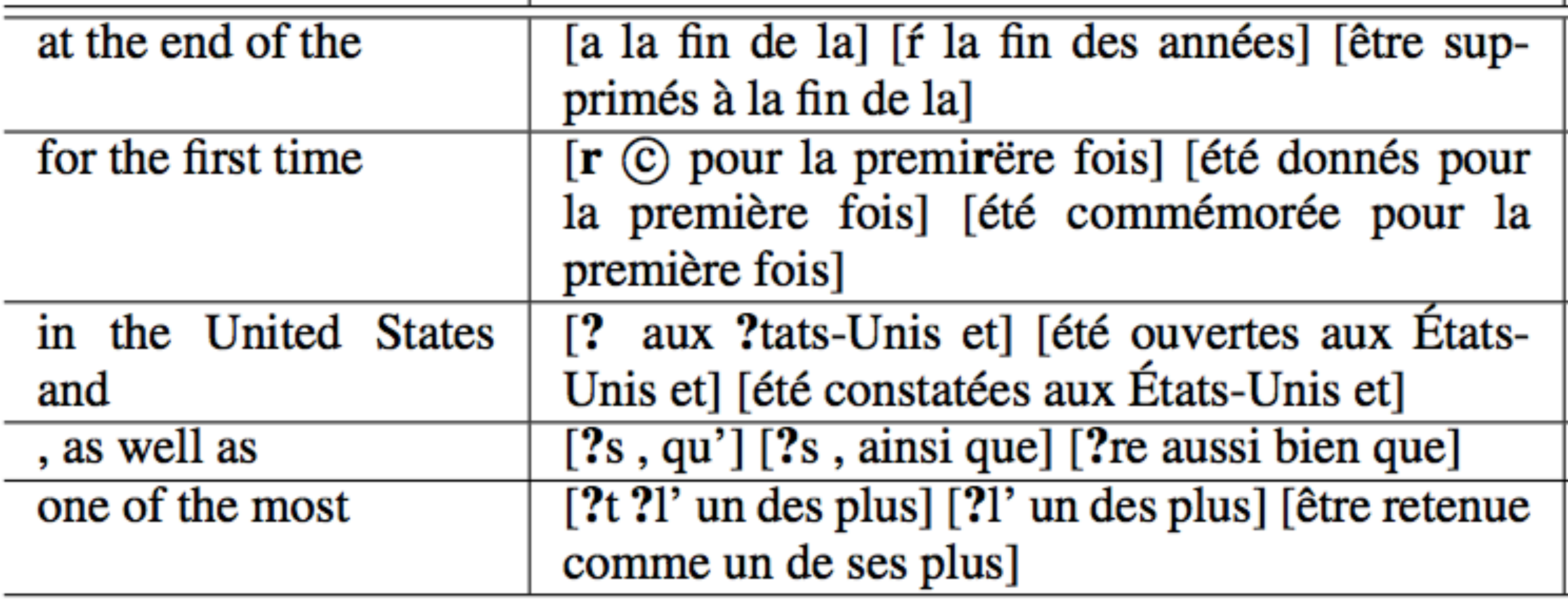
What topics would you like reviewd on Thursday?



Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

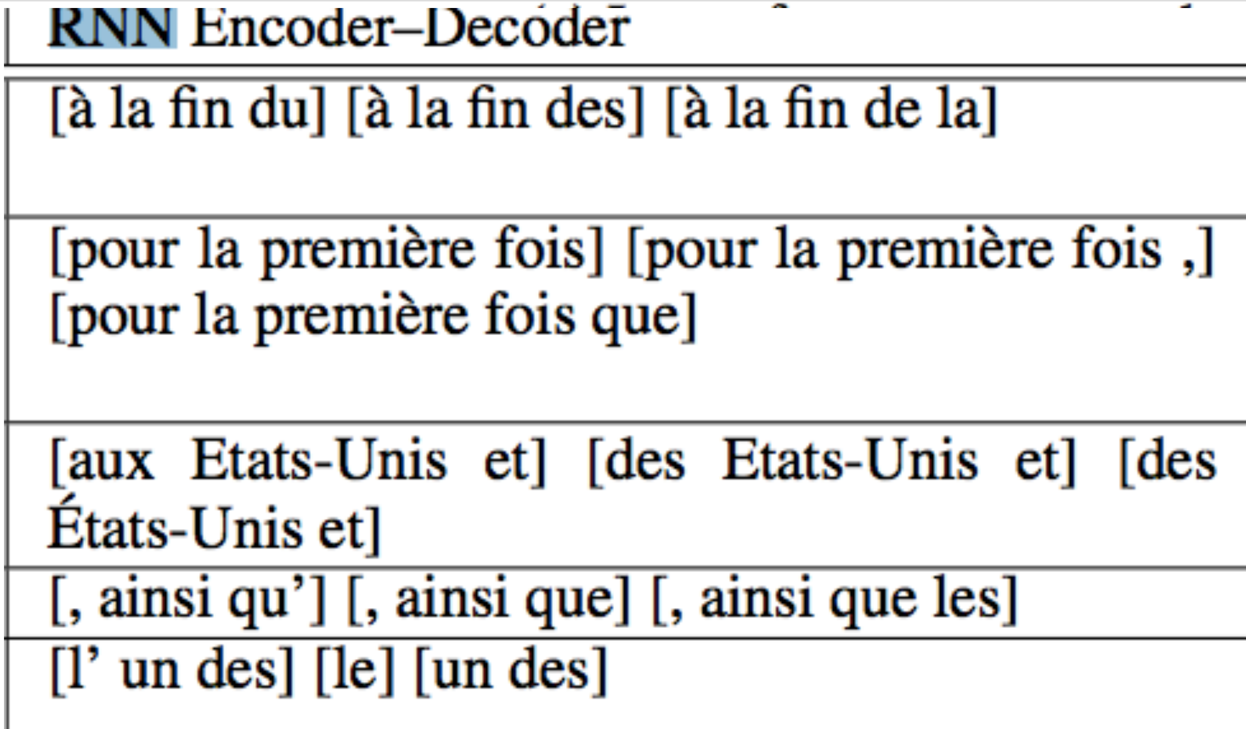
Other topics?

Clarifications from last time.



at the end of the	[a la fin de la] [ŕ la fin des années] [être supprimés à la fin de la]
for the first time	[**r** © pour la premirëre fois] [été donnés pour la première fois] [été commémorée pour la première fois]
in the United States and	[? aux ?tats-Unis et] [été ouvertes aux États-Unis et] [été constatées aux États-Unis et]
, as well as	[?s , qu’] [?s , ainsi que] [?re aussi bien que]
one of the most	[?t ?l’ un des plus] [?l’ un des plus] [être retenue comme un de ses plus]

Clarification: This is Phrase based MT (top) versus Neural MT (for identification of phrases) (bottom) as I presented last time.



RNN Encoder–Decoder
[à la fin du] [à la fin des] [à la fin de la]
[pour la première fois] [pour la première fois ,] [pour la première fois que]
[aux Etats-Unis et] [des Etats-Unis et] [des États-Unis et]
[, ainsi qu’] [, ainsi que] [, ainsi que les]
[l’ un des] [le] [un des]

Cho et al 2014

Performance

- Without attention, LSTM works quite well until a sentence gets longer than 30 words

- Attention does better, however, even with shorter sentences

- Other tricks in WMT 2017:
  - Improvements of 1.5 – 3 blue points (Edin)
  - Layer normalization, deeper networks (encoder depth of 5, decoder depth of 8)
  - Base Phrase Encodings (BPE)
    - THESE ARE ACTUALLY “BYTE PAIR ENCODING”. THEY ARE TO IDENTIFY SUB-WORDS. PRODUCE ONLY SUBWORDS SEEN IN THE TRAINING CORPUS. This restriction reduced vocabulary.
    - Reduced vocabulary improves memory efficiency
  - Data: parallel, back-translated, duplicated monolingual

Information Extraction

- Extraction of concrete facts from text
- Named entities, relations, events
- Often used to create a structured knowledge base of facts

- Kathy McKeown, a professor from Columbia University in New York City, took a train yesterday to Washington DC.

Named Entities

- Kathy McKeown$_{\text{per}}$, a professor from Columbia University$_{\text{org}}$ in New York City$_{\text{loc}}$, took a train yesterday to Washington DC$_{\text{loc}}$.

Named Entities, Relations

- Kathy McKeown$_{\text{per}}$, a professor from Columbia University$_{\text{org}}$ in New York City$_{\text{loc}}$, took a train yesterday to Washington DC$_{\text{loc}}$.

- Kathy McKeown from Columbia
- Columbia in New York City

Named Entities, Relations, Events

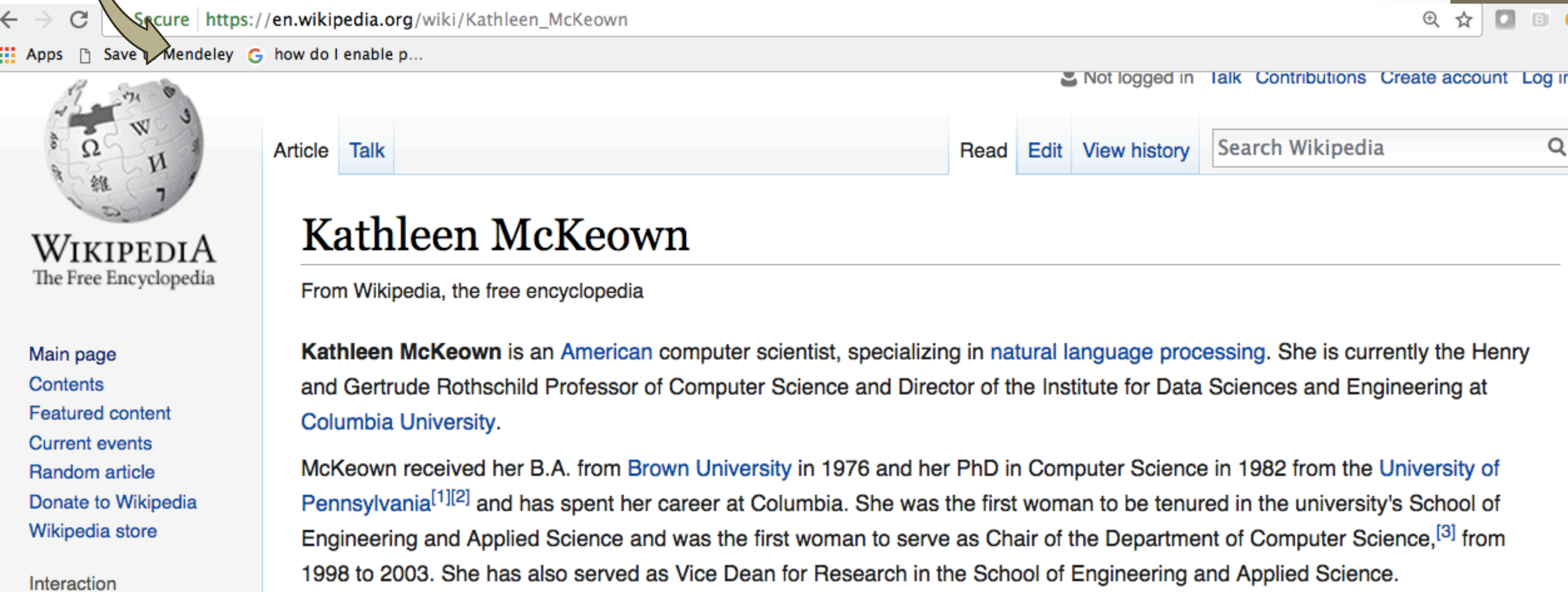
WIKIPEDIA
The Free Encyclopedia

- Kathy McKeown$_{\text{per}}$, a professor from Columbia University$_{\text{org}}$ in New York City$_{\text{loc}}$, took a train yesterday to Washington DC$_{\text{loc}}$.

- Kathy McKeown took a train (yesterday)

Entity Discovery and Linking

Kathy McKeown, a professor from Columbia University in New York City, took a train yesterday to Washington DC.



Kathleen McKeown

From Wikipedia, the free encyclopedia

Kathleen McKeown is an American computer scientist, specializing in natural language processing. She is currently the Henry and Gertrude Rothschild Professor of Computer Science and Director of the Institute for Data Sciences and Engineering at Columbia University.

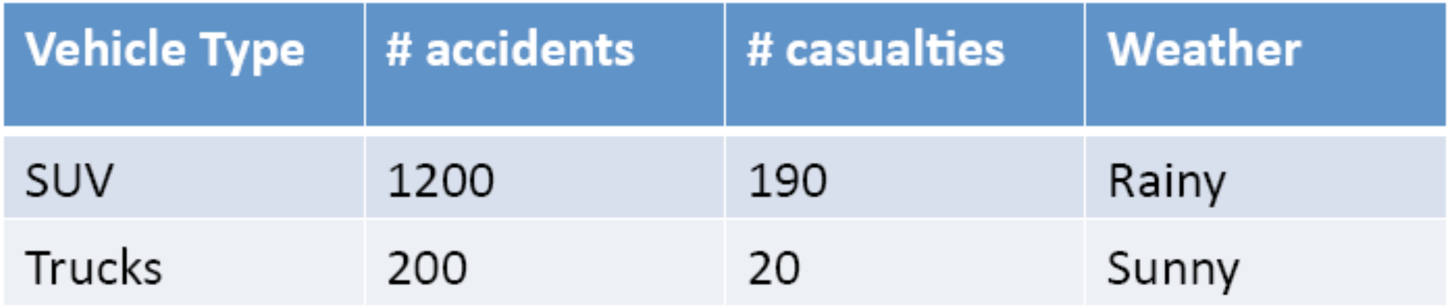
McKeown received her B.A. from Brown University in 1976 and her PhD in Computer Science in 1982 from the University of Pennsylvania[1][2] and has spent her career at Columbia. She was the first woman to be tenured in the university's School of Engineering and Applied Science and was the first woman to serve as Chair of the Department of Computer Science,[3] from 1998 to 2003. She has also served as Vice Dean for Research in the School of Engineering and Applied Science.

IE for Template Filling
Relation Detection

Given a set of documents and a domain of interest, fill a table of required fields.

- For example:

  Number of car accidents per vehicle type and number of casualties in the accidents.



Vehicle Type	# accidents	# casualties	Weather
SUV	1200	190	Rainy
Trucks	200	20	Sunny

Never-Ending Language Learner

Tom Mitchell

CMU

- Can computers learn to read?
- Browses the web and attempts to extract facts from hundreds of millions of web pages
- Attempts to improve its methods and accuracy
- To date, 50 million candidate facts at different levels of confidence
- http://rtw.ml.cmu.edu/rtw/

IE for Question Answering

Q: When was Gandhi born?
A: October 2, 1869

Q: Where was Bill Clinton educated?
A: Georgetown University in Washington, D.C.

Q: What was the education of Yassir Arafat?
A: Civil Engineering

Q: What is the religion of Noam Chomsky?
A: Jewish

State of the Art (English)

	F-measure
Named Entities (news)	89%
Relations (slot filling)	59%
Events (nuggets)	63%

Methods: Sequence labeling (MEMM, CRF), neural nets, distant learning

Features: linguistic features, similarity, popularity, gazeteers, ontologies, verb triggers

Where Have You Been Entity Discovery and Linking?

Grow with DEFT	2006-2011	2012-2017 *HENG JI, RPI*
Mention Extraction	Human (most)	Automatic
NIL Clustering	None	64 methods
Foreign Languages	Chinese (5%-10% lower than English)	**System for 282 languages (Chinese/Spanish comparable to/Outperform English); research toward 3,000 languages**
Document Size	-	500 →90,000 documents
Genre	News, web blog	**News, Discussion Forum, Web blog, Tweets**
Entity Types	PER, GPE, ORG	**PER, GPE, ORG, LOC, FAC, hundreds of fine-grained types for typing**
Mention Types	Name or all concepts (most)	Name, Nominal, Pronoun (for BeST)
KB	Wikipedia	Freebase → List only
Training Data	20,000 queries (entity mentions)	**500 → 0 documents; unsupervised linking comparable to supervised linking**
#(Good) Papers	62	110 (new KBP track at ACL); 6 tutorials at top conferences

Slide from Heng Ji

Approach for NER

- <PERSON>**Alexander Mackenzie</PERSON> , (<TIMEX >January 28, 1822 <TIMEX>** - <TIMEX>**April 17, 1892</TIMEX>), a building contractor and writer, was the second** Prime Minister of <GPE>**Canada</GPE> from ….**

- **Statistical sequence labeling techniques can be used – similar to POS tagging**
  - **Word-by-word sequence labeling**
  - **Example of features**
    - **POS tags**
    - **Syntactic constituents**
    - **Shape features**
    - **Presence in a named entity list**

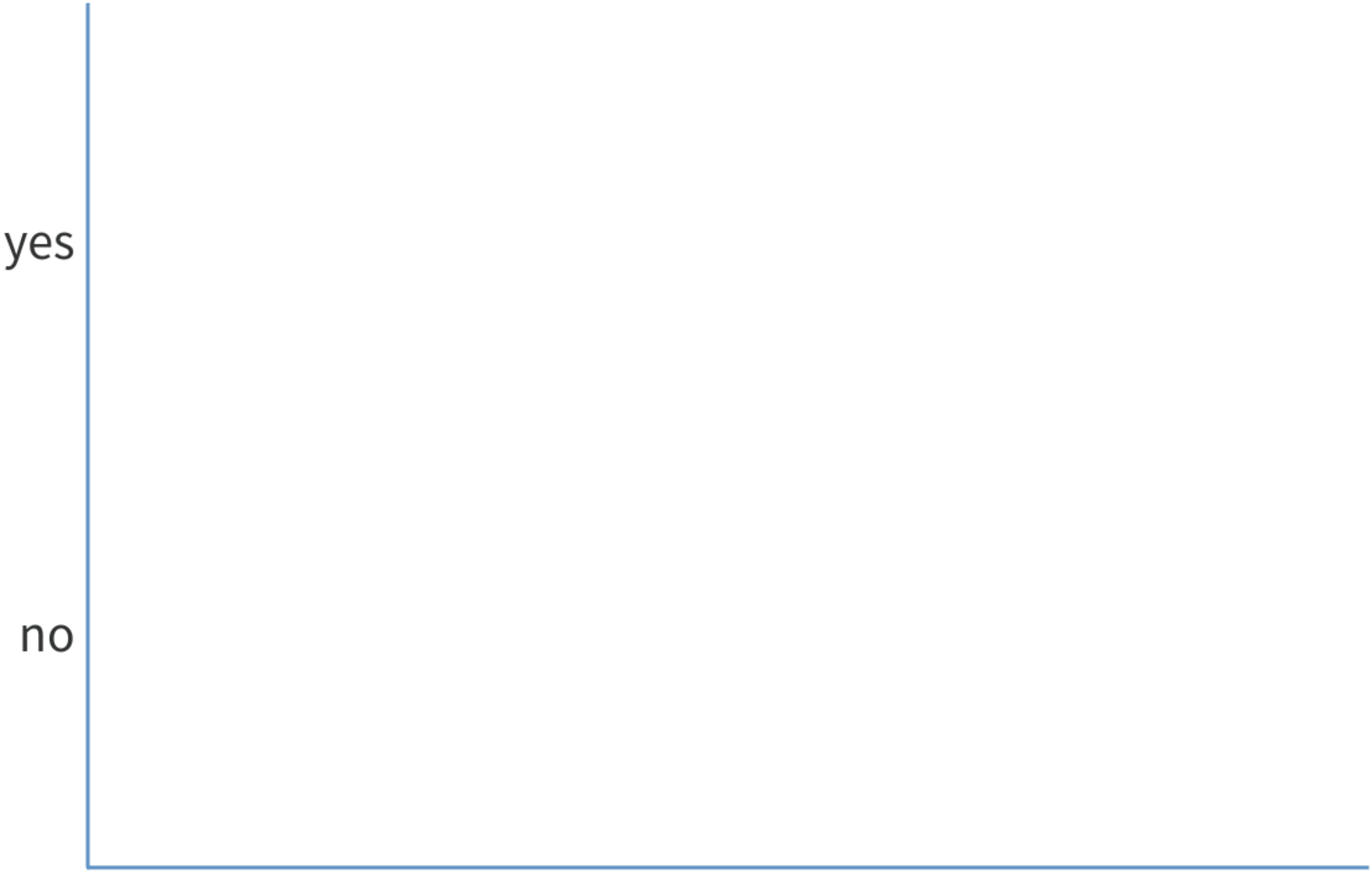
Supervised Approach for relation detection

- Given a corpus of annotated relations between entities, train a classifier:
  - A binary classifier
    - Given a span of text and two entities -> decide if there is a relationship between these two entities
- Features
  - Types of two named entities
  - Bag of words
  - POS of words in between
- Example:
  - A rented SUV went out of control on Sunday, causing the death of seven people in Brooklyn
  - Relation: Type = Accident, Vehicle Type = SUV, casualty = 7, weather = ?

Pattern Matching for Relation Detection

- Patterns:
  - "[CAR_TYPE] went out of control on [TIMEX], causing the death of [NUM] people"
  - "[PERSON] was born in [GPE]"
  - "[PERSON] was graduated from [FAC]"
  - "[PERSON] was killed by <X>"
- **Matching Techniques**
  - **Exact matching**
    - Pros and Cons?
  - **Flexible matching (e.g., [X] was .* killed .* by [Y])**
    - Pros and Cons?

Is rule-based exact matching still used? (take a guess)



Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

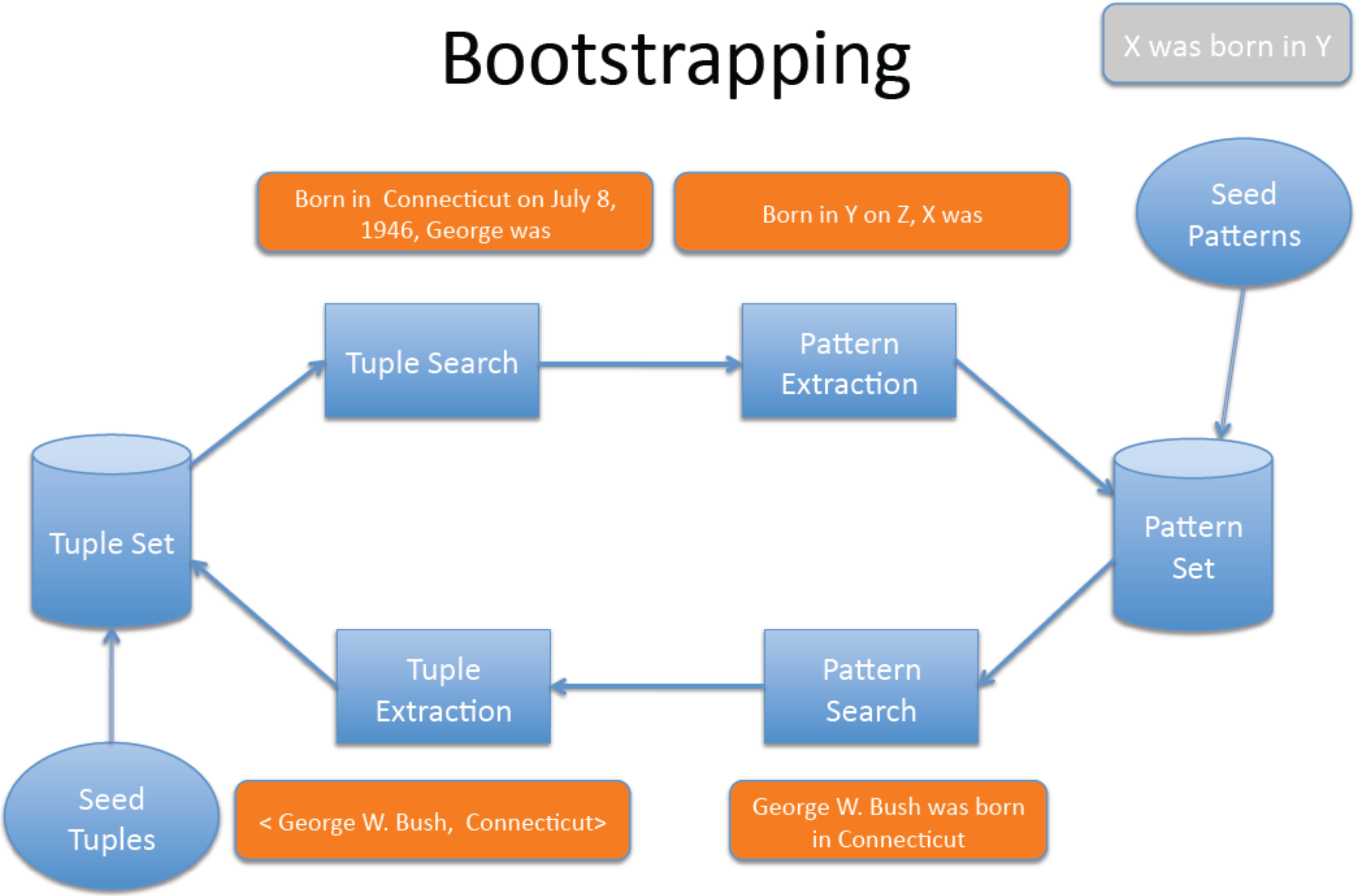
What problems would arise with flexible matching?

Pattern Matching

- How can we come up with these patterns?
- Manually?
  - Task and domain-specific
  - Tedious, time consuming, not scalable

- Machine learning, semi-supervised approaches

Bootstrapping



X was born in Y

Born in Connecticut on July 8, 1946, George was

Born in Y on Z, X was

< George W. Bush, Connecticut>

George W. Bush was born in Connecticut

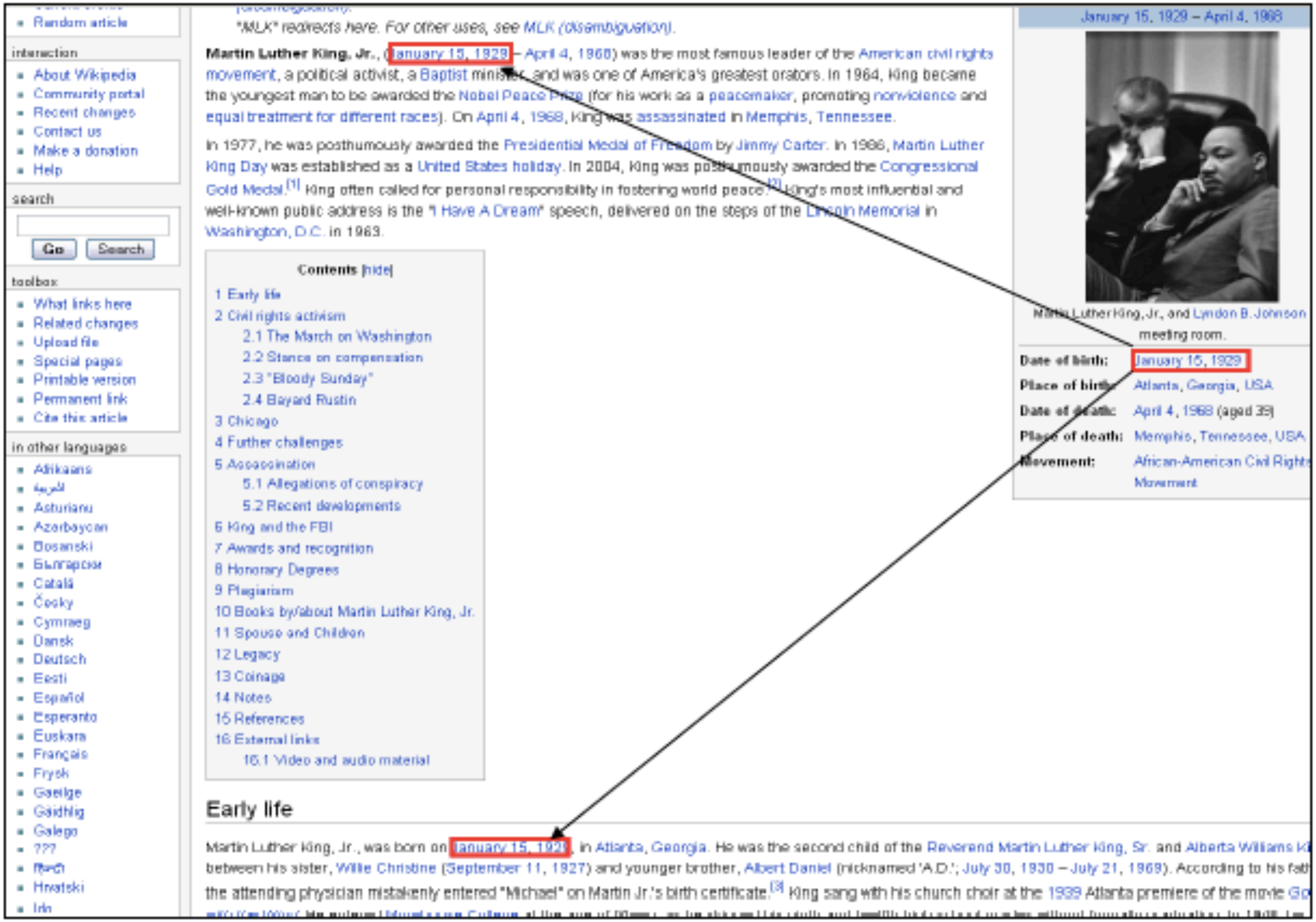
Task:
Produce a biography of [person]

1. Name(s), aliases:
2. *Date of Birth or Current Age:
3. *Date of Death:
4. *Place of Birth:
5. *Place of Death:
6. Cause of Death:
7. Religion (Affiliations):
8. Known locations and dates:
9. Last known address:
10. Previous domiciles:
11. Ethnic or tribal affiliations:
12. Immediate family members
13. Native Language spoken:
14. Secondary Languages spoken:
15. Physical Characteristics
16. Passport number and country of issue:
17. Professional positions:
18. Education
19. Party or other organization affiliations:
20. Publications (titles and dates):

Biography – two approaches

- To obtain high precision, handle each slot independently using bootstrapping to learn IE patterns.
- To improve the recall, utilize a biographical sentence classifier

Biography patterns from Wikipedia



"MLK" redirects here. For other uses, see MLK (disambiguation).

Martin Luther King, Jr., (January 15, 1929 – April 4, 1968) was the most famous leader of the American civil rights movement, a political activist, a Baptist minister, and was one of America's greatest orators. In 1964, King became the youngest man to be awarded the Nobel Peace Prize (for his work as a peacemaker, promoting nonviolence and equal treatment for different races). On April 4, 1968, King was assassinated in Memphis, Tennessee.

In 1977, he was posthumously awarded the Presidential Medal of Freedom by Jimmy Carter. In 1986, Martin Luther King Day was established as a United States holiday. In 2004, King was posthumously awarded the Congressional Gold Medal.[1] King often called for personal responsibility in fostering world peace.[2] King's most influential and well-known public address is the "I Have A Dream" speech, delivered on the steps of the Lincoln Memorial in Washington, D.C. in 1963.

Contents [hide]

January 15, 1929 – April 4, 1968

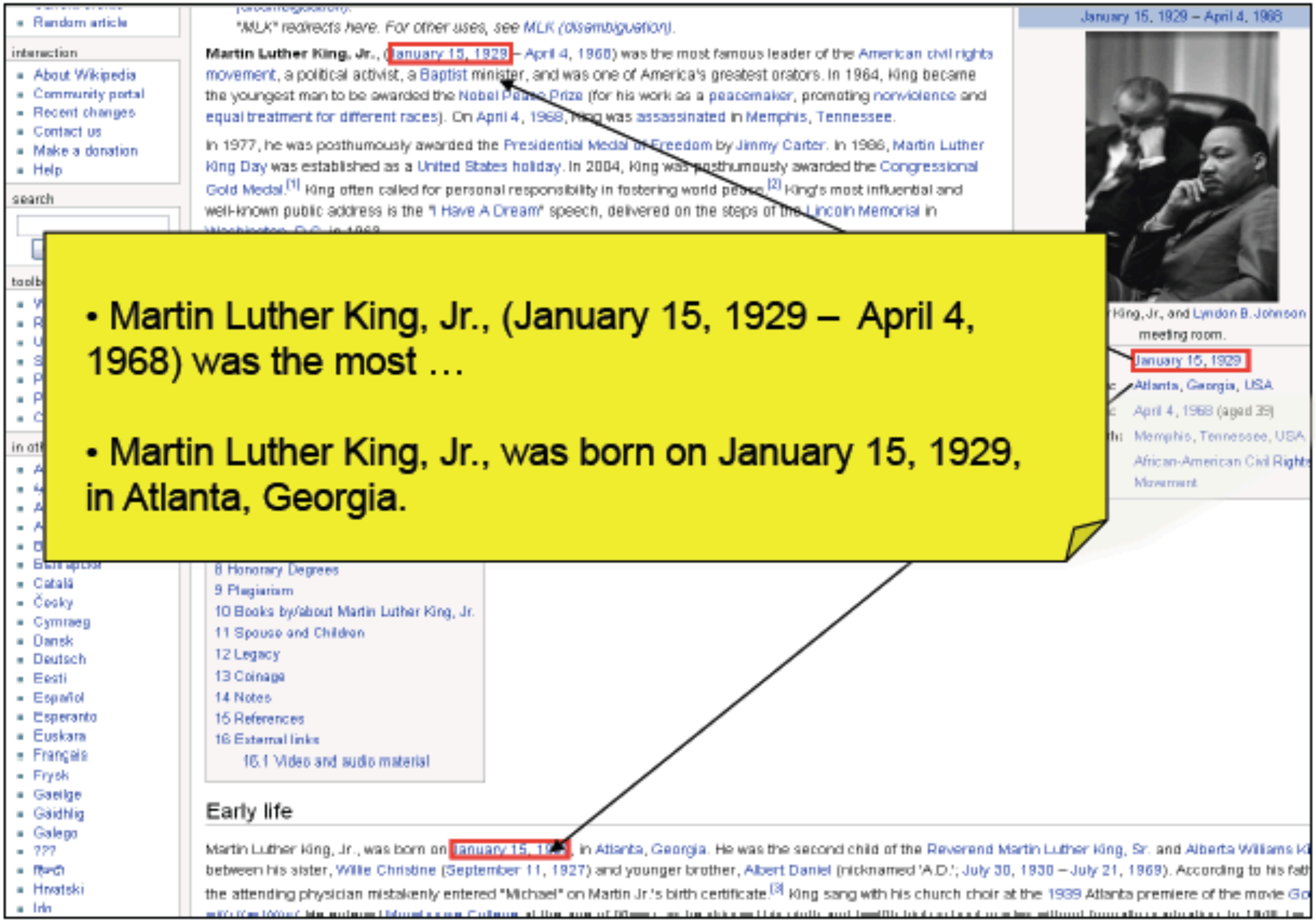
Martin Luther King, Jr., and Lyndon B. Johnson meeting room.

Date of birth:	January 15, 1929
Place of birth:	Atlanta, Georgia, USA
Date of death:	April 4, 1968 (aged 39)
Place of death:	Memphis, Tennessee, USA
Movement:	African-American Civil Rights Movement

Early life

Martin Luther King, Jr., was born on January 15, 1929, in Atlanta, Georgia. He was the second child of the Reverend Martin Luther King, Sr. and Alberta Williams Ki between his sister, Willie Christine (September 11, 1927) and younger brother, Albert Daniel (nicknamed 'A.D.'; July 30, 1930 – July 21, 1969). According to his fath the attending physician mistakenly entered "Michael" on Martin Jr.'s birth certificate.[3] King sang with his church choir at the 1939 Atlanta premiere of the movie Go

Biography patterns from Wikipedia



- Random article

interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact us
- Make a donation
- Help

search

- Català
- Česky
- Cymraeg
- Dansk
- Deutsch
- Eesti
- Español
- Esperanto
- Euskara
- Français
- Frysk
- Gaeilge
- Gàidhlig
- Galego
- ???
- हिन्दी
- Hrvatski
- Ido

"MLK" redirects here. For other uses, see MLK (disambiguation).

Martin Luther King, Jr., (January 15, 1929 – April 4, 1968) was the most famous leader of the American civil rights movement, a political activist, a Baptist minister, and was one of America's greatest orators. In 1964, King became the youngest man to be awarded the Nobel Peace Prize (for his work as a peacemaker, promoting nonviolence and equal treatment for different races). On April 4, 1968, King was assassinated in Memphis, Tennessee.

In 1977, he was posthumously awarded the Presidential Medal of Freedom by Jimmy Carter. In 1986, Martin Luther King Day was established as a United States holiday. In 2004, King was posthumously awarded the Congressional Gold Medal.[1] King often called for personal responsibility in fostering world peace.[2] King's most influential and well-known public address is the "I Have A Dream" speech, delivered on the steps of the Lincoln Memorial in Washington, D.C. in 1963.

January 15, 1929 – April 4, 1968

King, Jr. and Lyndon B. Johnson

meeting room.

January 15, 1929

Atlanta, Georgia, USA

April 4, 1968 (aged 39)

Memphis, Tennessee, USA

African-American Civil Rights Movement

Early life

Martin Luther King, Jr., was born on January 15, 1929, in Atlanta, Georgia. He was the second child of the Reverend Martin Luther King, Sr. and Alberta Williams Ki between his sister, Willie Christine (September 11, 1927) and younger brother, Albert Daniel (nicknamed 'A.D.'; July 30, 1930 – July 21, 1969). According to his fath the attending physician mistakenly entered "Michael" on Martin Jr.'s birth certificate.[3] King sang with his church choir at the 1939 Atlanta premiere of the movie Go

- Martin Luther King, Jr., (January 15, 1929 – April 4, 1968) was the most …
- Martin Luther King, Jr., was born on January 15, 1929, in Atlanta, Georgia.

Run NER on these sentences

- <Person> Martin Luther King, Jr. </Person>, (<Date>January 15, 1929</Date> – <Date> April 4, 1968</Date>) was the most…
- <Person> Martin Luther King, Jr. </Person>, was born on <Date> January 15, 1929 </Date>, in <GPE> Atlanta, Georgia </GPE>.
- Take the token sequence that includes the tags of interest + some context (2 tokens before and 2 tokens after)

Convert to Patterns:

- <Target_Person> (<Target_Date> – <Date>) was the
- <Target_Person> , was born on <Target_Date>, in
- Remove more specific patterns – if there is a pattern that contains other, take the smallest > *k* tokens.
- ➔ <Target_Person> , was born on <Target_Date>
- ➔ <Target_Person> (<Target_Date> – <Date>)
- Finally, verify the patterns manually to remove irrelevant patterns.

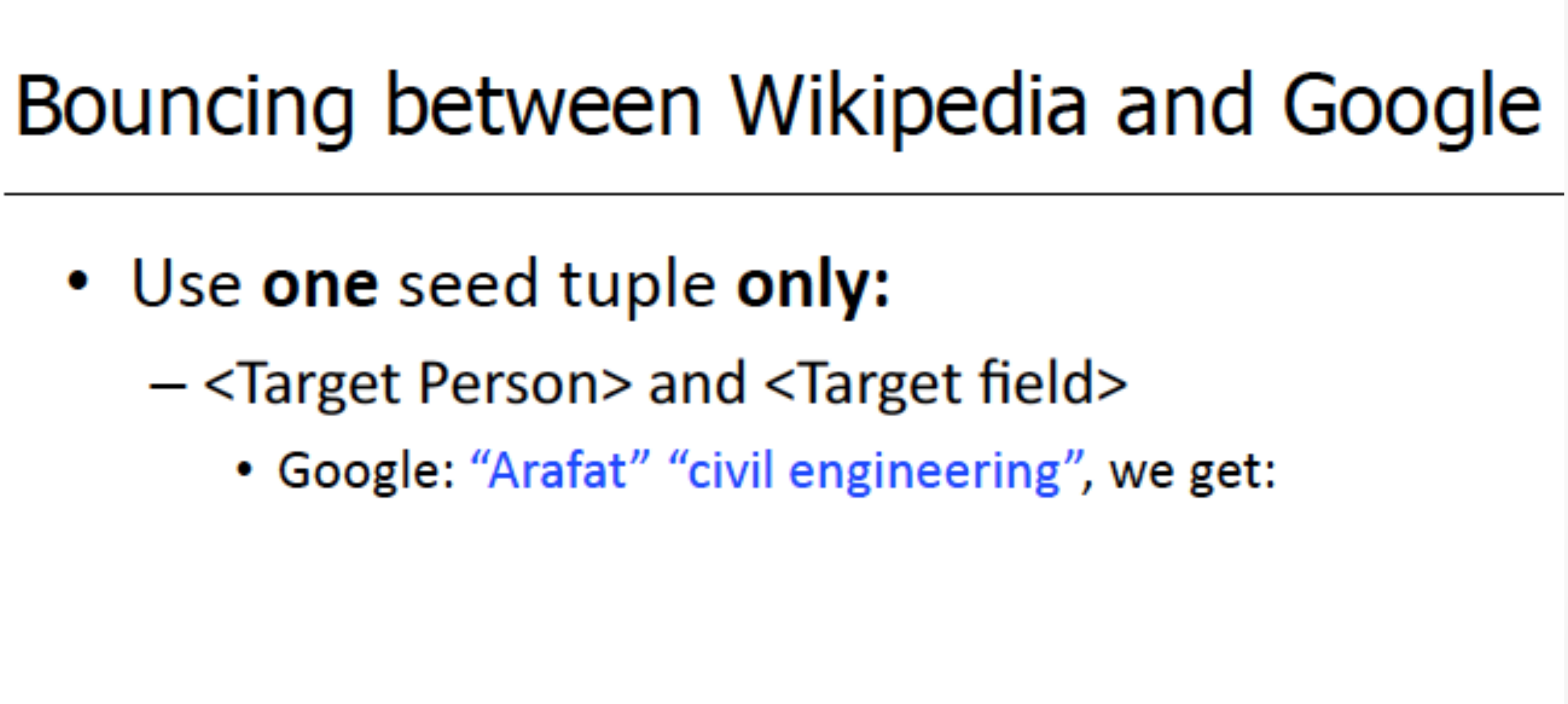
Examples of Patterns:

- 502 distinct place-of-birth patterns:
  - 600 <Target_Person> was born in <Target_GPE>
  - 169 <Target_Person> (born <Date> in <Target_GPE>)
  - 44 Born in <Target_GPE> , <Target_Person>
  - 10 <Target_Person> was a native <Target_GPE>
  - 10 <Target_Person> 's hometown of <Target_GPE>
  - 1 <Target_Person> was baptized in <Target_GPE>
  - ...
- 291 distinct date-of-death patterns:
  - 770 <Target_Person> (<Date> - <Target_Date>)
  - 92 <Target_Person> died on <Target_Date>
  - 19 <Target_Person> <Date> - <Target_Date>
  - 16 <Target_Person> died in <GPE> on <Target_Date>
  - 3 < Target_Person> passed away on < Target_Date >
  - 1 < Target_Person> committed suicide on <Target_Date>
  - ...

Biography as an IE task

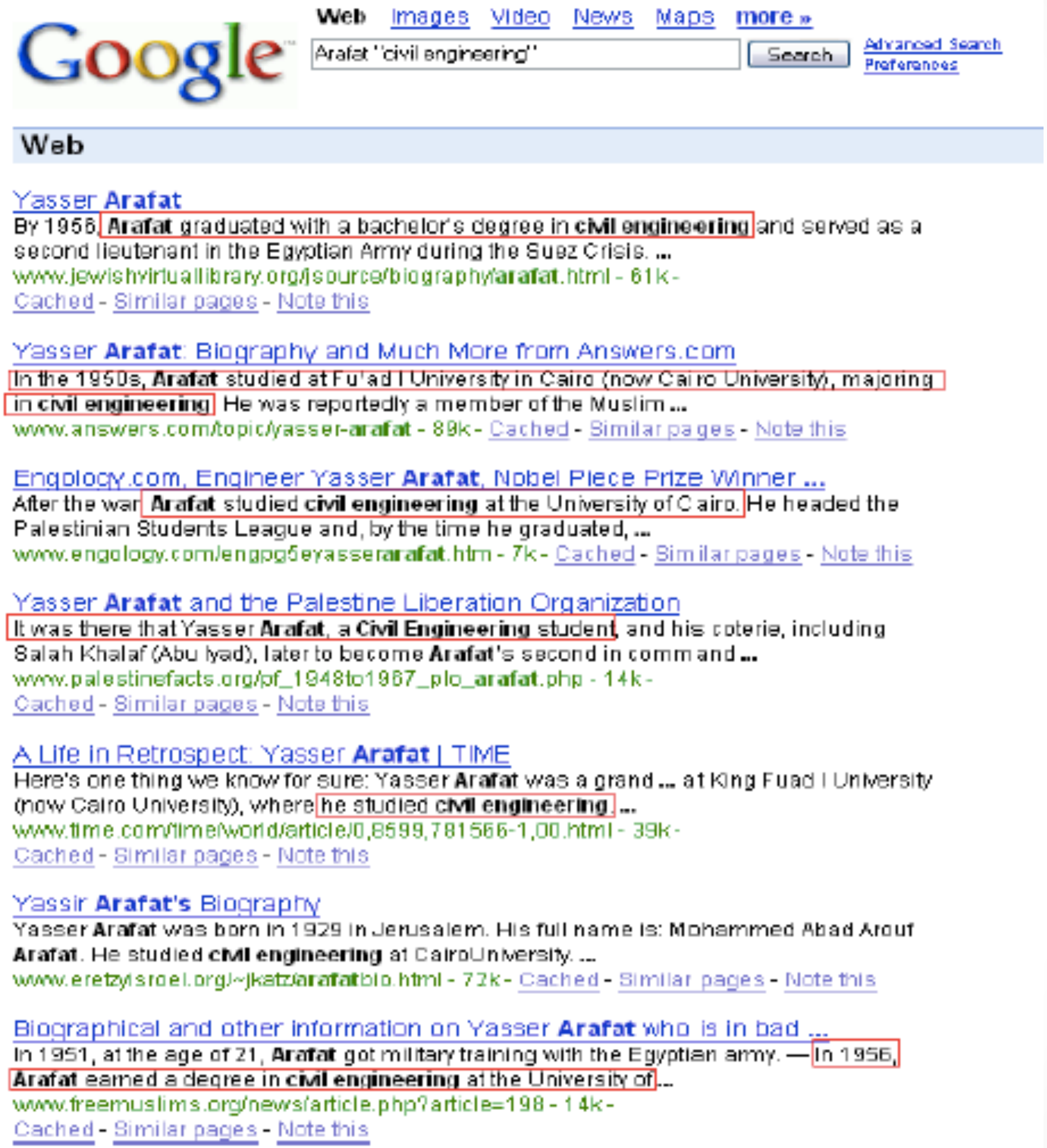
- This approach is good for the consistently annotated fields in Wikipedia: *place of birth, date of birth, place of death, date of death*

- Not all fields of interests are annotated, a different approach is needed to cover the rest of the slots



Bouncing between Wikipedia and Google

- Use **one** seed tuple **only:**
  - – <Target Person> and <Target field>
    - Google: “Arafat” “civil engineering”, we get:



Arafat "civil engineering"

Web

Yasser **Arafat**
By 1956, **Arafat** graduated with a bachelor's degree in **civil engineering** and served as a second lieutenant in the Egyptian Army during the Suez Crisis. **...**
www.jewishvirtuallibrary.org/jsource/biography/**arafat**.html - 61k - Cached - Similar pages - Note this

Yasser **Arafat**: Biography and Much More from Answers.com
In the 1950s, **Arafat** studied at Fu'ad I University in Cairo (now Cairo University), majoring in **civil engineering**. He was reportedly a member of the Muslim **...**
www.answers.com/topic/yasser-**arafat** - 89k - Cached - Similar pages - Note this

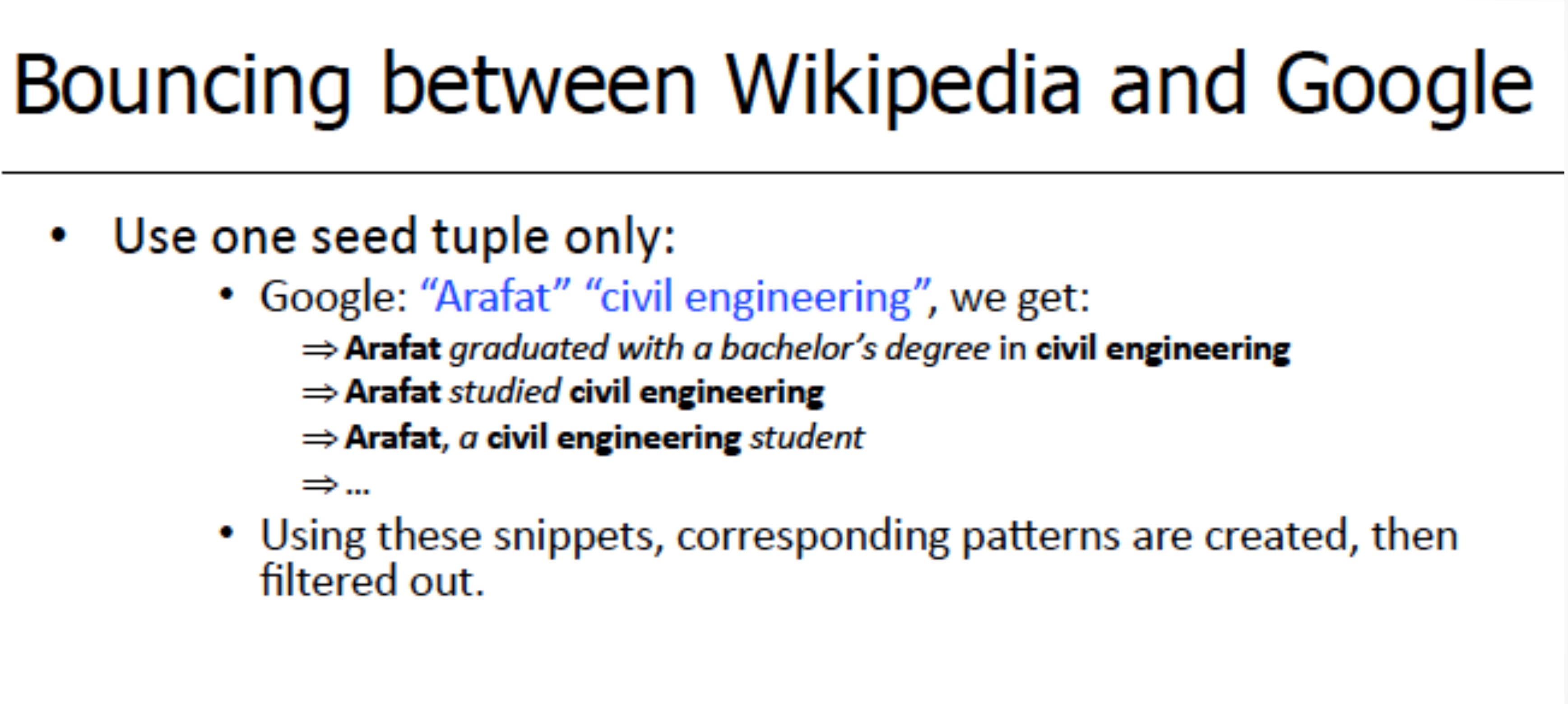
Engology.com, Engineer Yasser **Arafat**, Nobel Piece Prize Winner **...**
After the war, **Arafat** studied **civil engineering** at the University of Cairo. He headed the Palestinian Students League and, by the time he graduated, **...**
www.engology.com/engpg5eyasser**arafat**.htm - 7k - Cached - Similar pages - Note this

Yasser **Arafat** and the Palestine Liberation Organization
It was there that Yasser **Arafat**, a **Civil Engineering** student, and his coterie, including Salah Khalaf (Abu Iyad), later to become **Arafat**'s second in command and **...**
www.palestinefacts.org/pf_1948to1967_plo_**arafat**.php - 14k - Cached - Similar pages - Note this

A Life in Retrospect: Yasser **Arafat** | TIME
Here's one thing we know for sure: Yasser **Arafat** was a grand **...** at King Fuad I University (now Cairo University), where he studied **civil engineering**. **...**
www.time.com/time/world/article/0,8599,781566-1,00.html - 39k - Cached - Similar pages - Note this

Yassir **Arafat's** Biography
Yasser **Arafat** was born in 1929 in Jerusalem. His full name is: Mohammed Abad Arouf **Arafat**. He studied **civil engineering** at CairoUniversity. **...**
www.eretzyisroel.org/~jkatz/**arafat**bio.html - 72k - Cached - Similar pages - Note this

Biographical and other information on Yasser **Arafat** who is in bad **...**
In 1951, at the age of 21, **Arafat** got military training with the Egyptian army. — In 1956, **Arafat** earned a degree in **civil engineering** at the University of **...**
www.freemuslims.org/news/article.php?article=198 - 14k - Cached - Similar pages - Note this



Bouncing between Wikipedia and Google

- Use one seed tuple only:
  - Google: “Arafat” “civil engineering”, we get:
    - ⇒ **Arafat** *graduated with a bachelor’s degree* in **civil engineering**
    - ⇒ **Arafat** *studied* **civil engineering**
    - ⇒ **Arafat**, *a* **civil engineering** *student*
    - ⇒ …
  - Using these snippets, corresponding patterns are created, then filtered out.

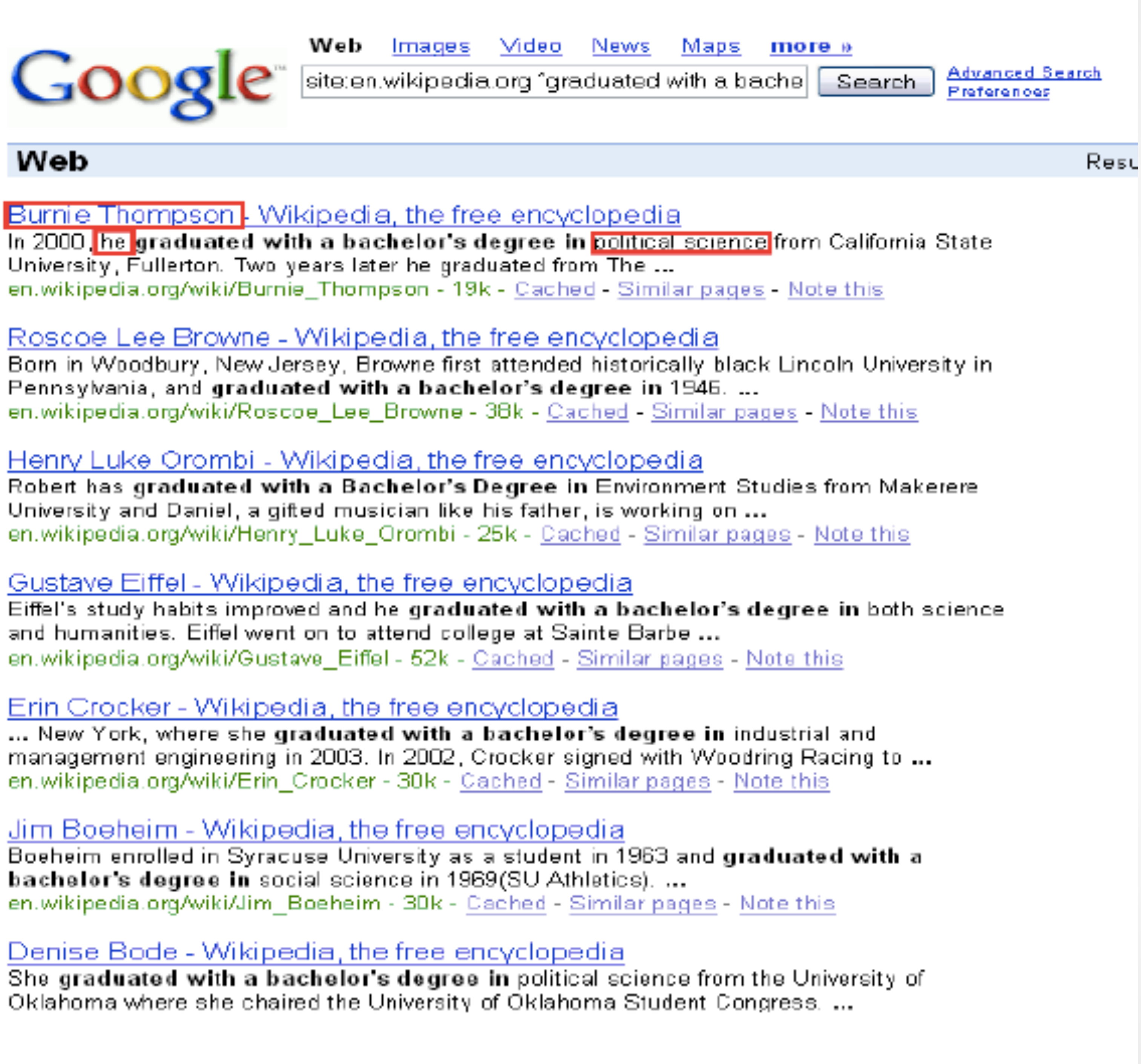
Bouncing between Wikipedia and Google

- Use one seed tuple only:
    - Google: “Arafat” “civil engineering”, we get:
        - ⇒ **Arafat** ***graduated with a bachelor’s degree*** in **civil engineering**
        - ⇒ **Arafat** *studied* **civil engineering**
        - ⇒ **Arafat**, *a* **civil engineering** *student*
        - ⇒ **…**
    - Using these snippets, corresponding patterns are created, then filtered out manually
    - Due to time limitation the automatic filter was not completed.

- To get more seed tuples, go to Wikipedia biography pages only and search for:
    - ***“graduated with a bachelor’s degree in”***
    - We get:

Bouncing between Wikipedia and Google

- **New seed tuples:**
  - “Burnie Thompson” “political science“
  - “Henrey Luke” “Environment Studies”
  - “Erin Crocker” “industrial and management engineering”
  - “Denise Bode” “political science”
  - …
- Go back to Google and repeat the process to get more seed patterns!



Burnie Thompson - Wikipedia, the free encyclopedia
In 2000, he **graduated with a bachelor's degree in** political science from California State University, Fullerton. Two years later he graduated from The ...
en.wikipedia.org/wiki/Burnie_Thompson - 19k - Cached - Similar pages - Note this

Roscoe Lee Browne - Wikipedia, the free encyclopedia
Born in Woodbury, New Jersey, Browne first attended historically black Lincoln University in Pennsylvania, and **graduated with a bachelor's degree in** 1946. ...
en.wikipedia.org/wiki/Roscoe_Lee_Browne - 38k - Cached - Similar pages - Note this

Henry Luke Orombi - Wikipedia, the free encyclopedia
Robert has **graduated with a Bachelor's Degree in** Environment Studies from Makerere University and Daniel, a gifted musician like his father, is working on ...
en.wikipedia.org/wiki/Henry_Luke_Orombi - 25k - Cached - Similar pages - Note this

Gustave Eiffel - Wikipedia, the free encyclopedia
Eiffel's study habits improved and he **graduated with a bachelor's degree in** both science and humanities. Eiffel went on to attend college at Sainte Barbe ...
en.wikipedia.org/wiki/Gustave_Eiffel - 52k - Cached - Similar pages - Note this

Erin Crocker - Wikipedia, the free encyclopedia
... New York, where she **graduated with a bachelor's degree in** industrial and management engineering in 2003. In 2002, Crocker signed with Woodring Racing to ...
en.wikipedia.org/wiki/Erin_Crocker - 30k - Cached - Similar pages - Note this

Jim Boeheim - Wikipedia, the free encyclopedia
Boeheim enrolled in Syracuse University as a student in 1963 and **graduated with a bachelor's degree in** social science in 1969(SU Athletics). ...
en.wikipedia.org/wiki/Jim_Boeheim - 30k - Cached - Similar pages - Note this

Denise Bode - Wikipedia, the free encyclopedia
She **graduated with a bachelor's degree in** political science from the University of Oklahoma where she chaired the University of Oklahoma Student Congress. ...

Bouncing back and forth

- Worked well for a fields such as education, publications, immediate family members, party, other organization activities

- Did not work well for other fields including religion, ethnic or tribal affiliations, previous domiciles -> too much noise

- Why is the bouncing idea better than using only one corpus?

Why is bouncing between two data sources better than just using one?

How are neural nets used for IE?

Organizing knowledge

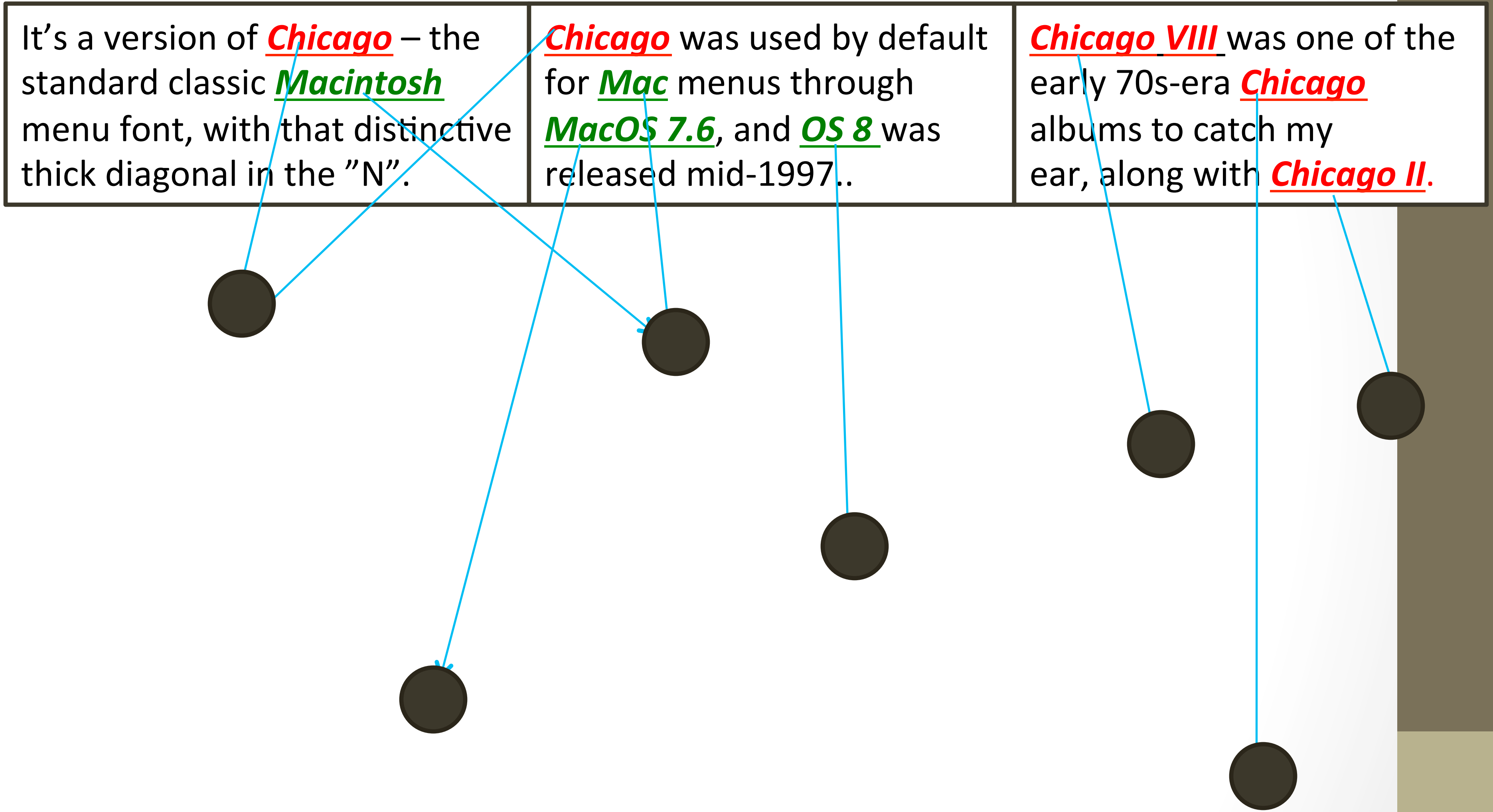
It’s a version of ***Chicago*** – the standard classic Macintosh menu font, with that distinctive thick diagonal in the ”N”.

Chicago was used by default for Mac menus through MacOS 7.6, and OS 8 was released mid-1997..

Chicago VIII was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.

Slide from Heng Ji

Cross-document co-reference resolution



It's a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the ”N”.

Chicago was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..

Chicago VIII was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.

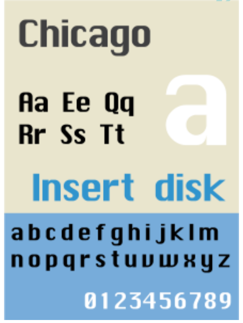
Slide from Heng Ji

Reference resolution: (disambiguation to Wikipedia)

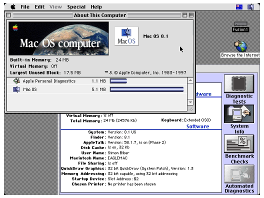
It’s a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the ”N”.

Chicago was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..

Chicago VIII was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.



Workbench2.0 94% full, 48K free, 789K in



Mac OS 8.1

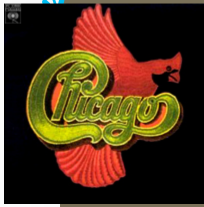
Built-in Memory: 24 MB
Virtual Memory: Off
Largest Unused Block: 17.5 MB

™ & © Apple Computer, Inc. 1983-1997

Apple Personal Diagnostics 1.1 MB
Mac OS 5.1 MB

Virtual Memory : is off
Total Memory : 24 Mb (24576 Kb)
Keyboard : Extended (ISO)

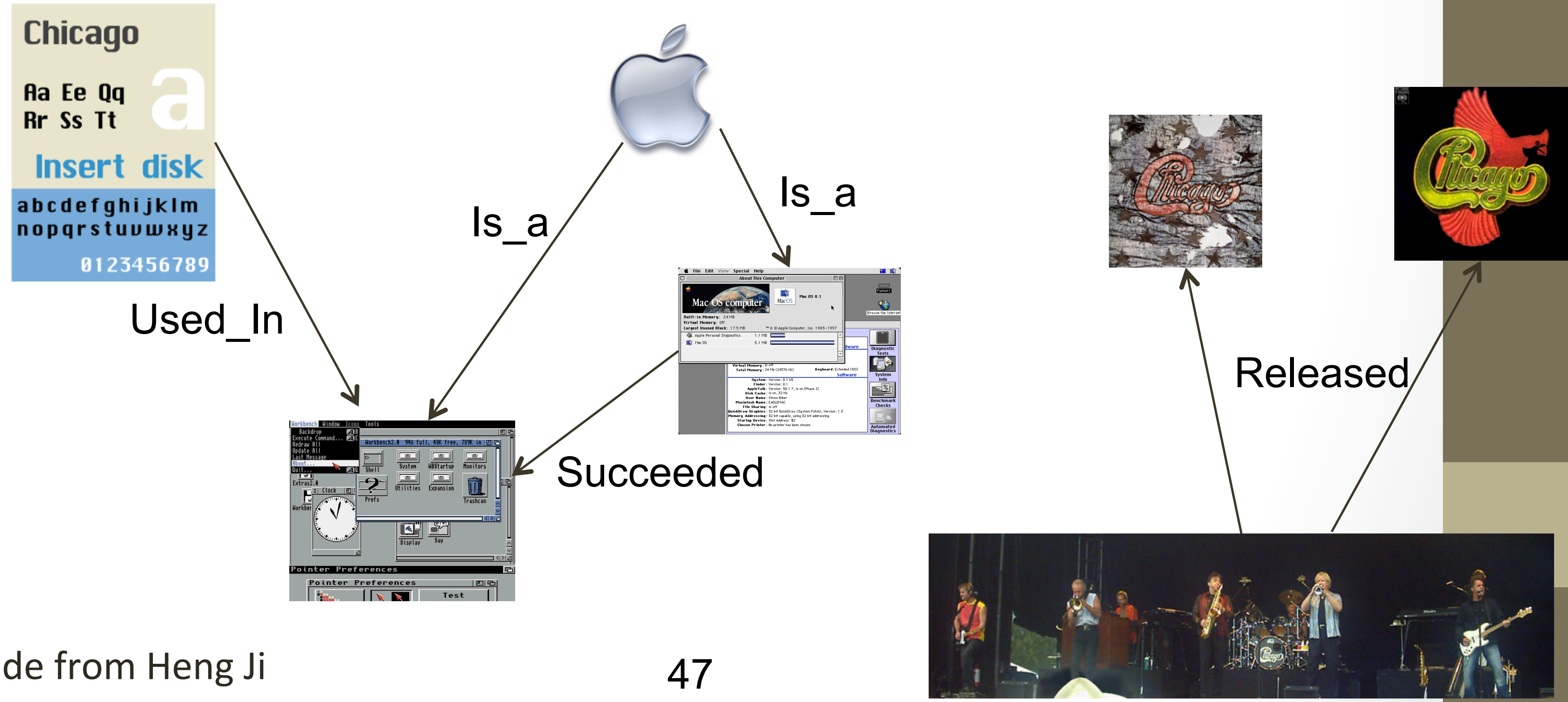
System : Version: 8.1 US
Finder : Version: 8.1
AppleTalk : Version: 58.1.7, is on (Phase 2)
Disk Cache : is on, 32 Kb
User Name : Simon Biber
Macintosh Name : EAGLEMAC
File Sharing : is off
QuickDraw Graphics : 32 bit QuickDraw (System Patch), Version: 1.3
Memory Addressing : 32 bit capable, using 32 bit addressing
Startup Device : Slot Address: $2
Chosen Printer : No printer has been chosen



Slide from Heng Ji

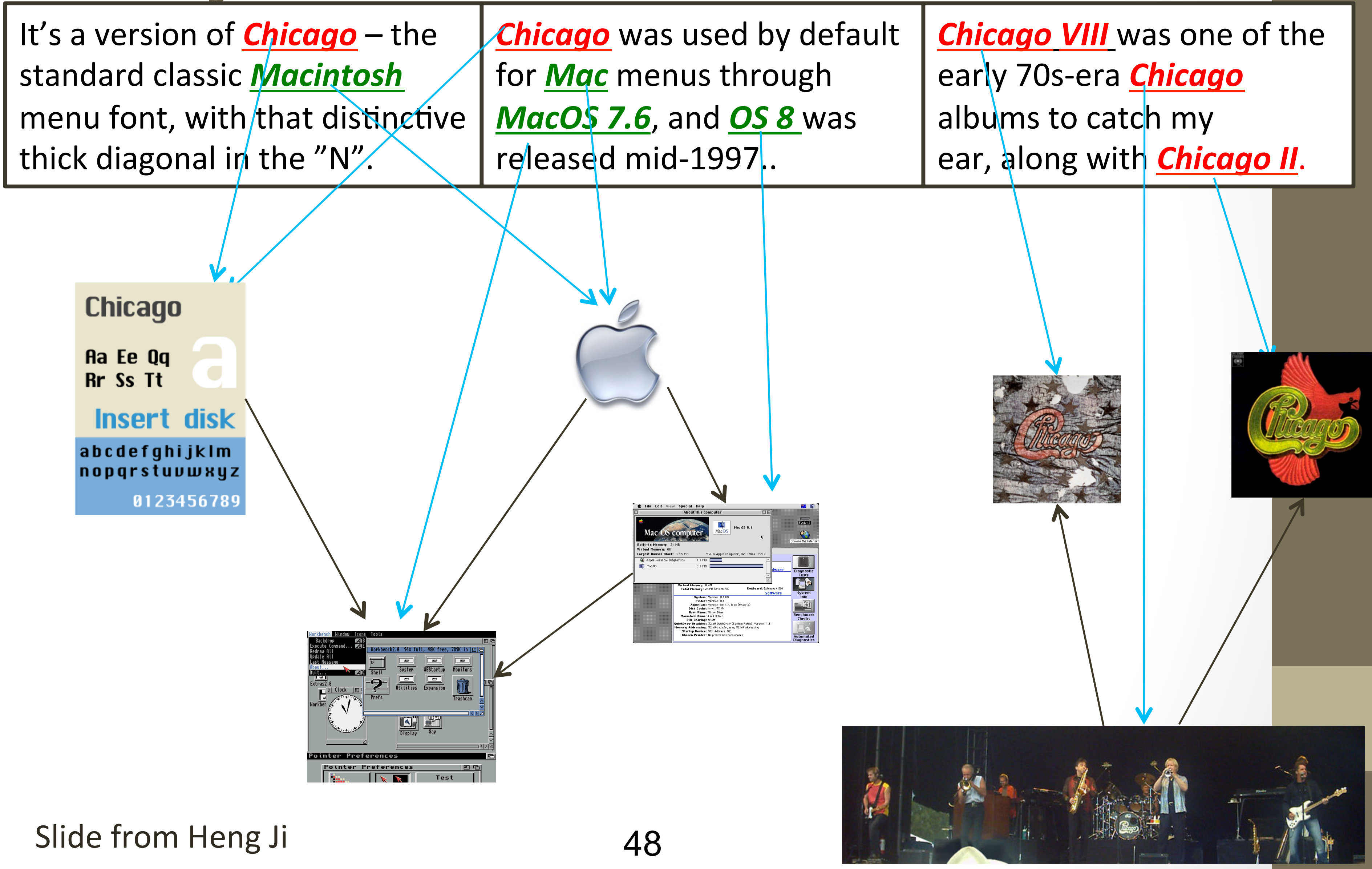
The "Reference" Collection has Structure

It's a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the "N".	***Chicago*** was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.



Slide from Heng Ji

Analysis of Information Networks



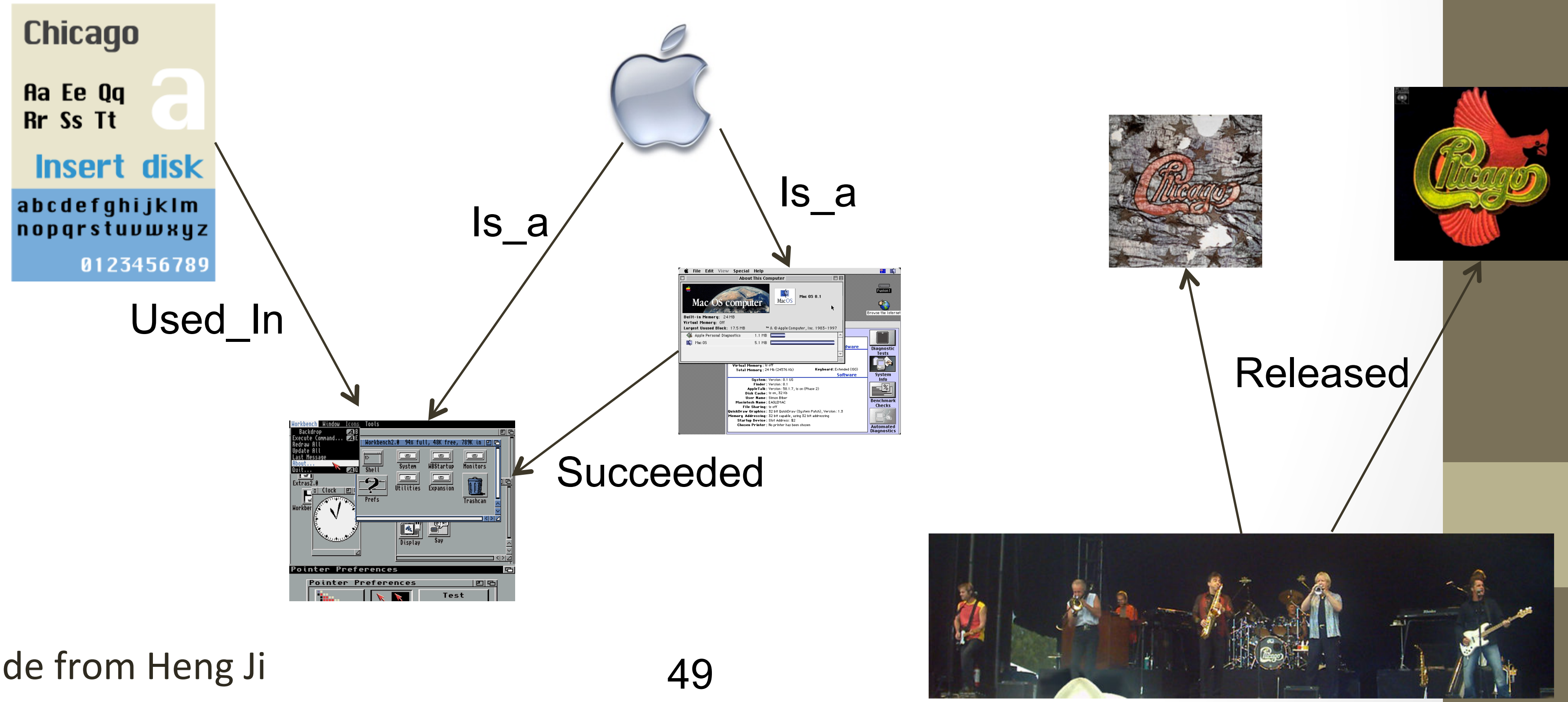
It’s a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the ”N”.

Chicago was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..

Chicago VIII was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.

Slide from Heng Ji

Here – Wikipedia as a knowledge resource .... bu can use other resources



Used_In

Is_a

Is_a

Succeeded

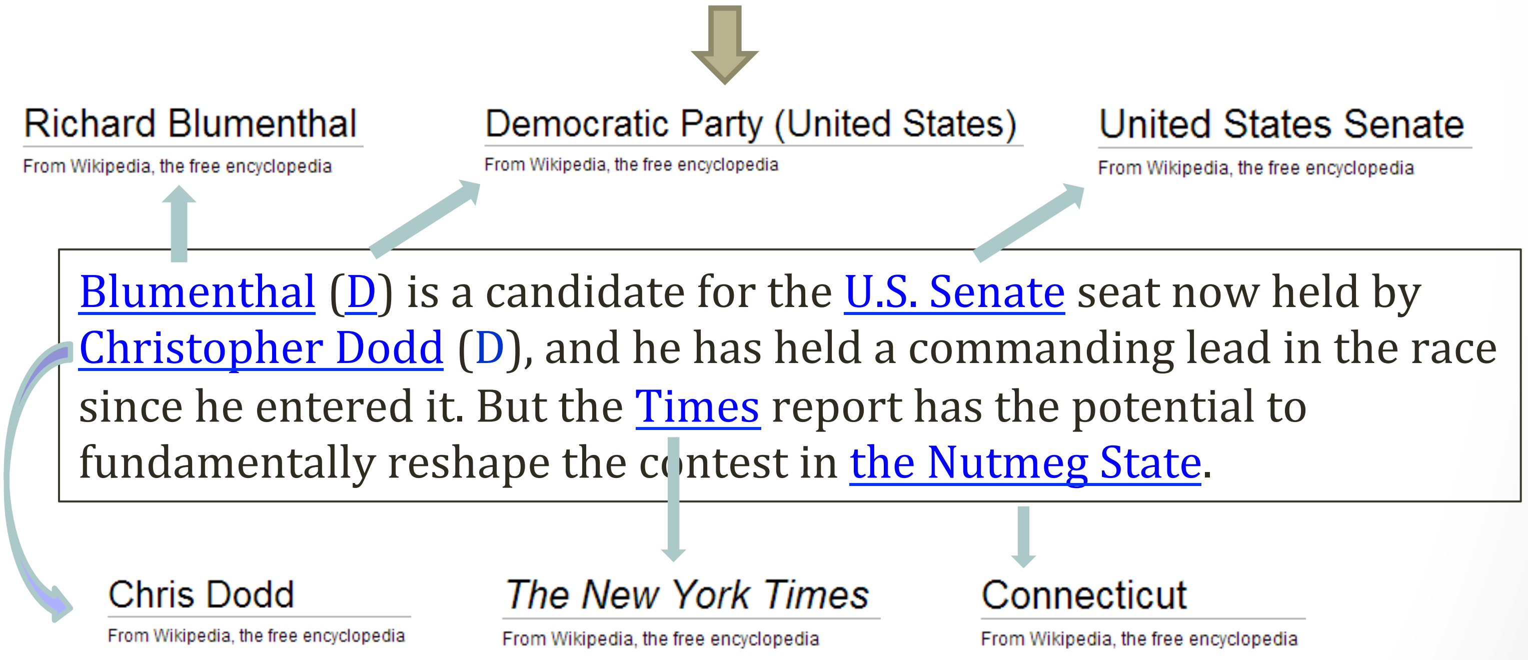
Released

Slide from Heng Ji

Wikification: The Reference Problem

Cycles of Knowledge: Grounding for/ using Knowledge

Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.



Richard Blumenthal
From Wikipedia, the free encyclopedia

Democratic Party (United States)
From Wikipedia, the free encyclopedia

United States Senate
From Wikipedia, the free encyclopedia

Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

Chris Dodd
From Wikipedia, the free encyclopedia

The New York Times
From Wikipedia, the free encyclopedia

Connecticut
From Wikipedia, the free encyclopedia

Slide from Heng Ji

Task Definition

- A formal definition of the task consists of:

1. A definition of the **mentions** (concepts, entities) to highlight

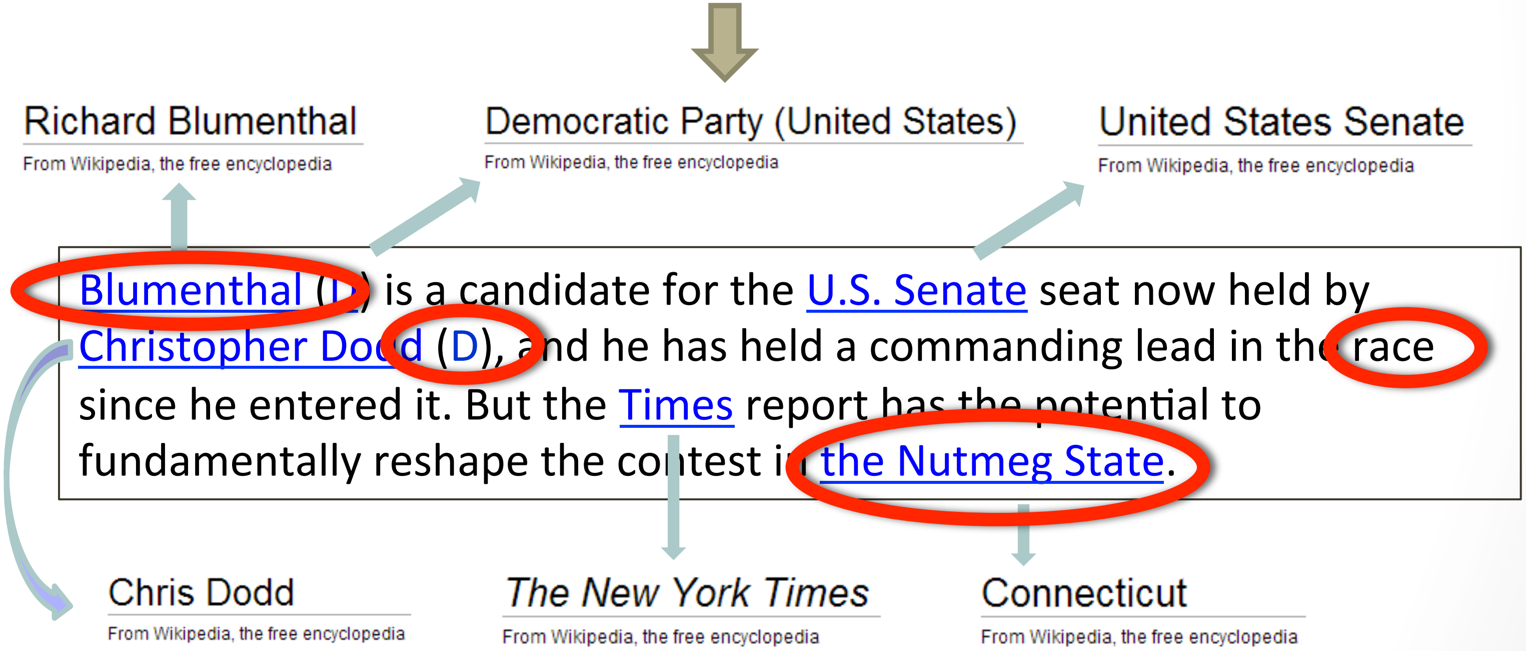
2. Determining the target encyclopedic resource (**KB**)

3. Defining what to point to in the KB (**title**)

Slide from Heng Ji

Examples of Mentions (1)

Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.



Richard Blumenthal
From Wikipedia, the free encyclopedia

Democratic Party (United States)
From Wikipedia, the free encyclopedia

United States Senate
From Wikipedia, the free encyclopedia

Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

Chris Dodd
From Wikipedia, the free encyclopedia

The New York Times
From Wikipedia, the free encyclopedia

Connecticut
From Wikipedia, the free encyclopedia

Slide from Heng Ji

Neural Approach to Entity Linking (Wikification)

Gupta, Singh and Roth, EMNLP 2017

- Learns a dense, unified representation of entities
    - Encodes semantic and background knowledge from multiple sources
    - An encoder for each source of information
    - Entity embeddings learned to be similar to encodings
- Only uses indirect supervision from Wikipedia/Freebase
- Can incorporate new entities without retraining existing representations
- http://cogcomp.org/papers/GuptaSiRo17.pdf

Jointly Embedding Entity Information

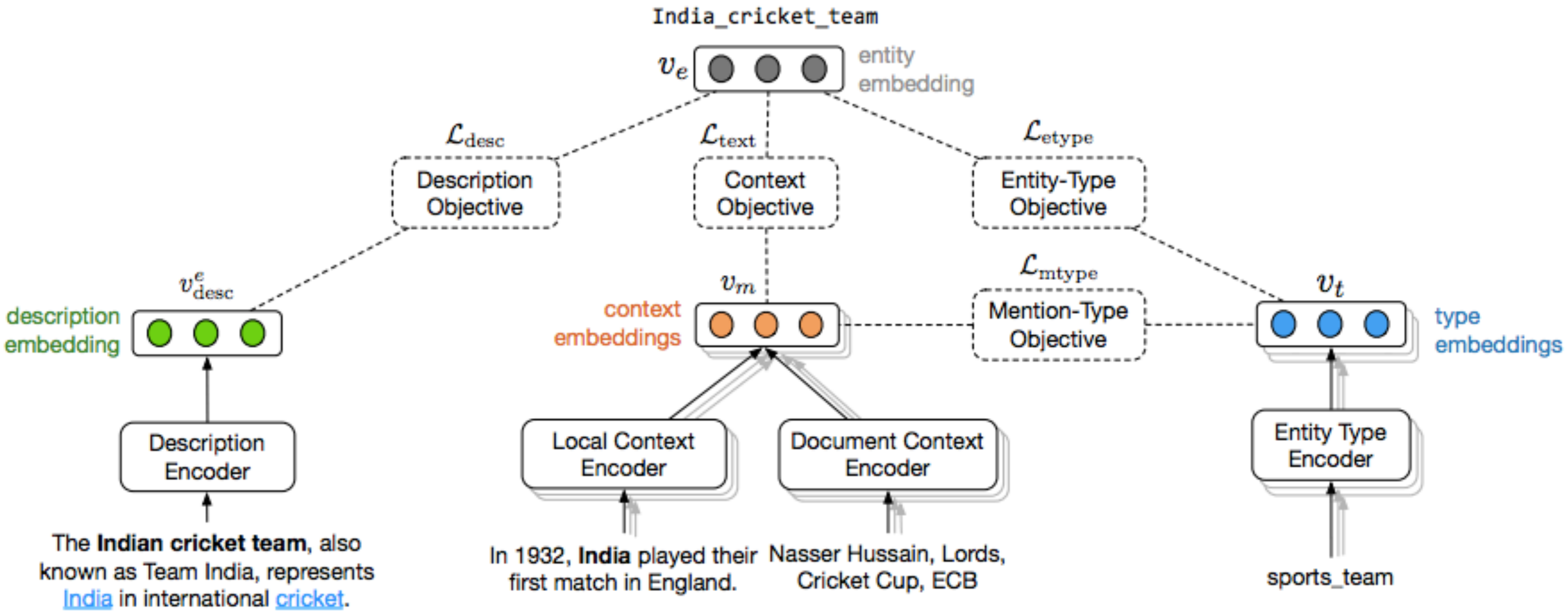


Figure 1: **Overview of the Model** (§ 3): Each entity has a Wikipedia description, linked mentions in Wikipedia (only one shown), and fine-grained types from Freebase (only one shown). We encode local and document-level mention contexts (§ 3.1), entity-description (§ 3.2), and fine-grained entity-types (§ 3.3 & § 3.4). Joint optimization (§ 3.5) over these provides the unified entity representations $\{v_e\}$.

Jointly Embedding Entity Information

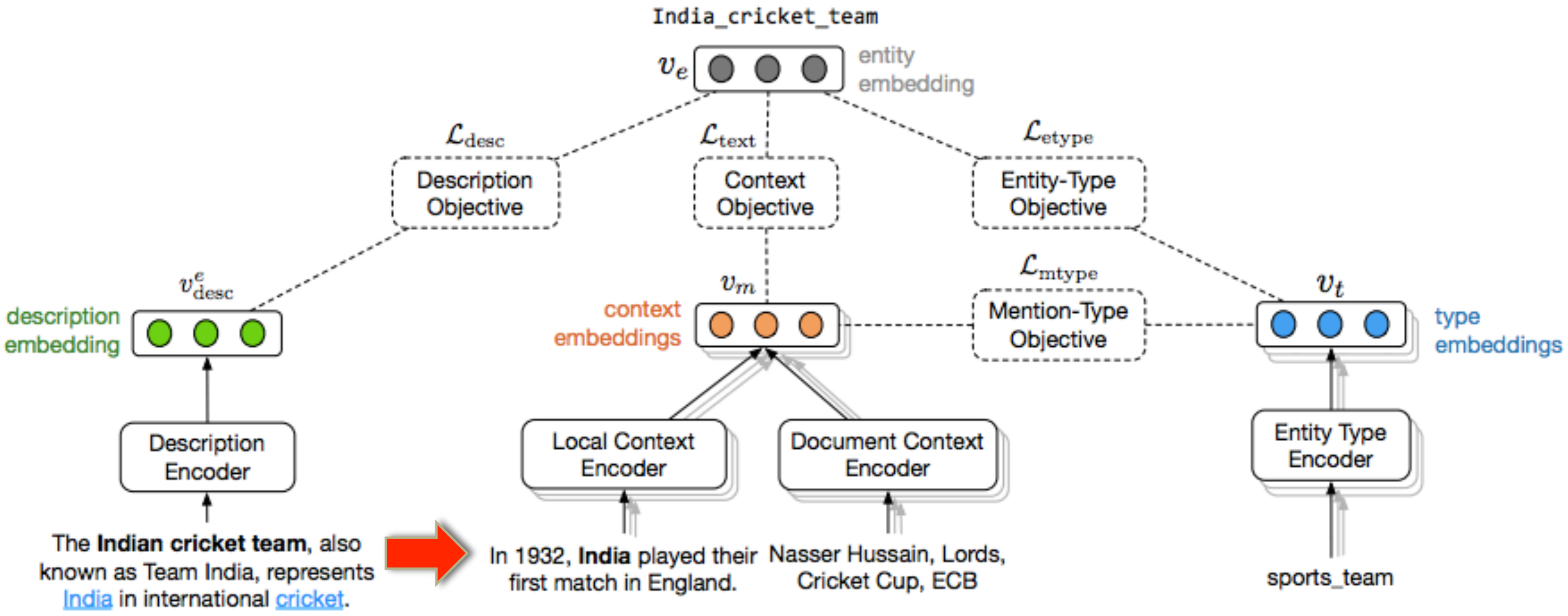

India_cricket_team
v_e
entity embedding
$\mathcal{L}_{\text{desc}}$
$\mathcal{L}_{\text{text}}$
$\mathcal{L}_{\text{etype}}$
Description Objective
Context Objective
Entity-Type Objective
v^e_{desc}
v_m
$\mathcal{L}_{\text{mtype}}$
v_t
description embedding
context embeddings
Mention-Type Objective
type embeddings
Description Encoder
Local Context Encoder
Document Context Encoder
Entity Type Encoder
The **Indian cricket team**, also known as Team India, represents India in international cricket.
In 1932, **India** played their first match in England.
Nasser Hussain, Lords, Cricket Cup, ECB
sports_team

Figure 1: **Overview of the Model** (§ 3): Each entity has a Wikipedia description, linked mentions in Wikipedia (only one shown), and fine-grained types from Freebase (only one shown). We encode local and document-level mention contexts (§ 3.1), entity-description (§ 3.2), and fine-grained entity-types (§ 3.3 & § 3.4). Joint optimization (§ 3.5) over these provides the unified entity representations $\{v_e\}$.

Look at Wikipedia

- Entity description: https://en.wikipedia.org/wiki/India_national_cricket_team

Encoding the mention context

In 1932, India played their first game in England.

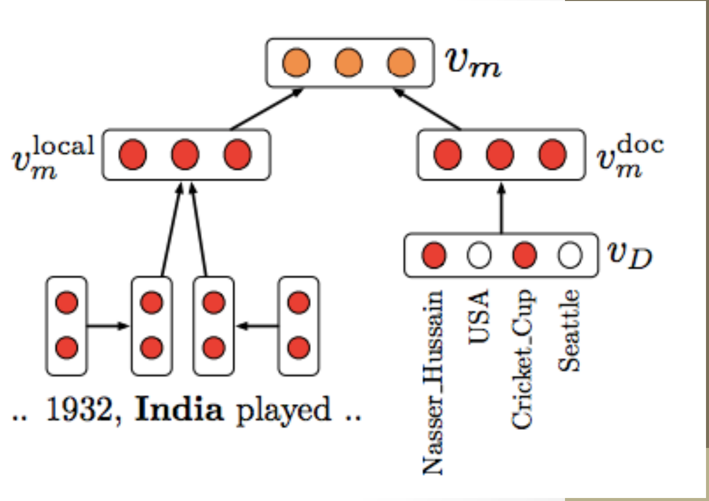
- Example mention contains two mentions: "*India*" and "*England*"
- Aim to disambiguate "*India*" to the team
  - **Local context**: "*played*" and "*match*"
  - **Document context**: to identify the sport
- Preserve the semantics: "*England*" should not match to a team

Local Context

- Given mention m in sentence: {w_1, …, m, …. w_N}
- Left LSTM applied to w_1…m -> $\overrightarrow{h^l_m}$
- Right LSTM applied to m …. w_N -> $\overleftarrow{h^r_m}$
- $\overrightarrow{h^l_m}; \overleftarrow{h^r_m}$ concatenated and passed through a single layer feed forward network

Document Context Encoder

- Bag of mentions vector:
  - USA, Pearl Jam, Nasser Hassain
- Compressed to a low dimensional representation using a single layer feed forward neural network
- Combine local and document representations to get a mention level encoding using concatenation and feed through a single layer feed forward network



$$v_m \in \mathbb{R}^d.$$

Encoding Entity Description D

- Embed each word of the Wikipedia description as a d-dimensional vector

- Encode as a fixed vector using a CNN:

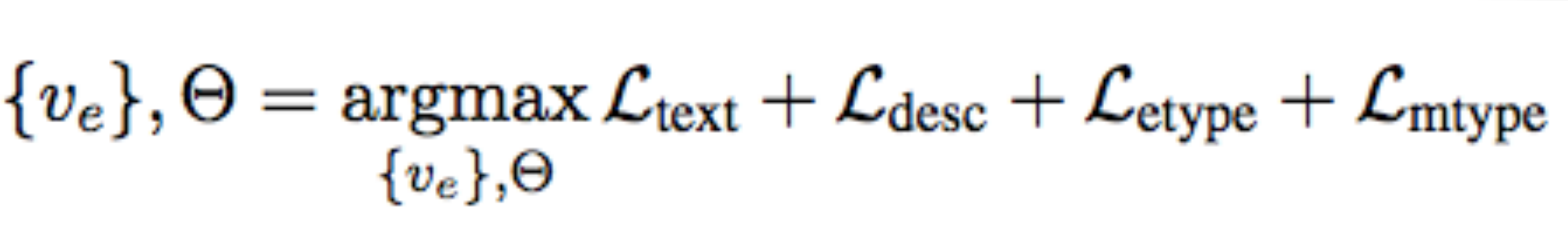
$$v_{\text{desc}}^{e} \in \mathbb{R}^{d}$$

Learning the Type Representation

- Embed type T in Freebase
- Each entity can have multiple types
- Jointly learn entity and type representations

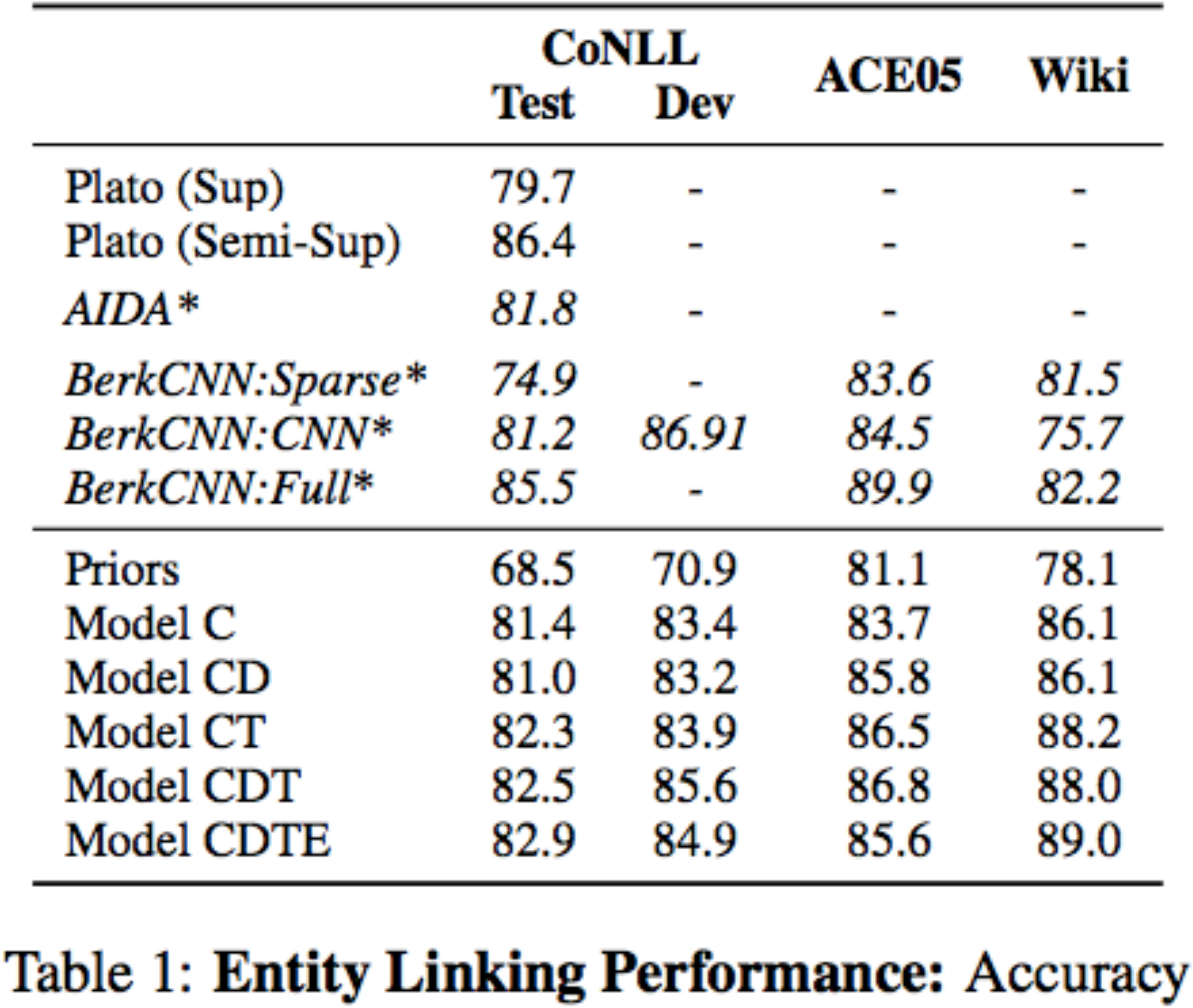
Learning Unified Entity Representations

- Separate models for entity mentions, entity descriptions, type descriptions
- To learn the different entity representations and their parameters, jointly maximize the total objective



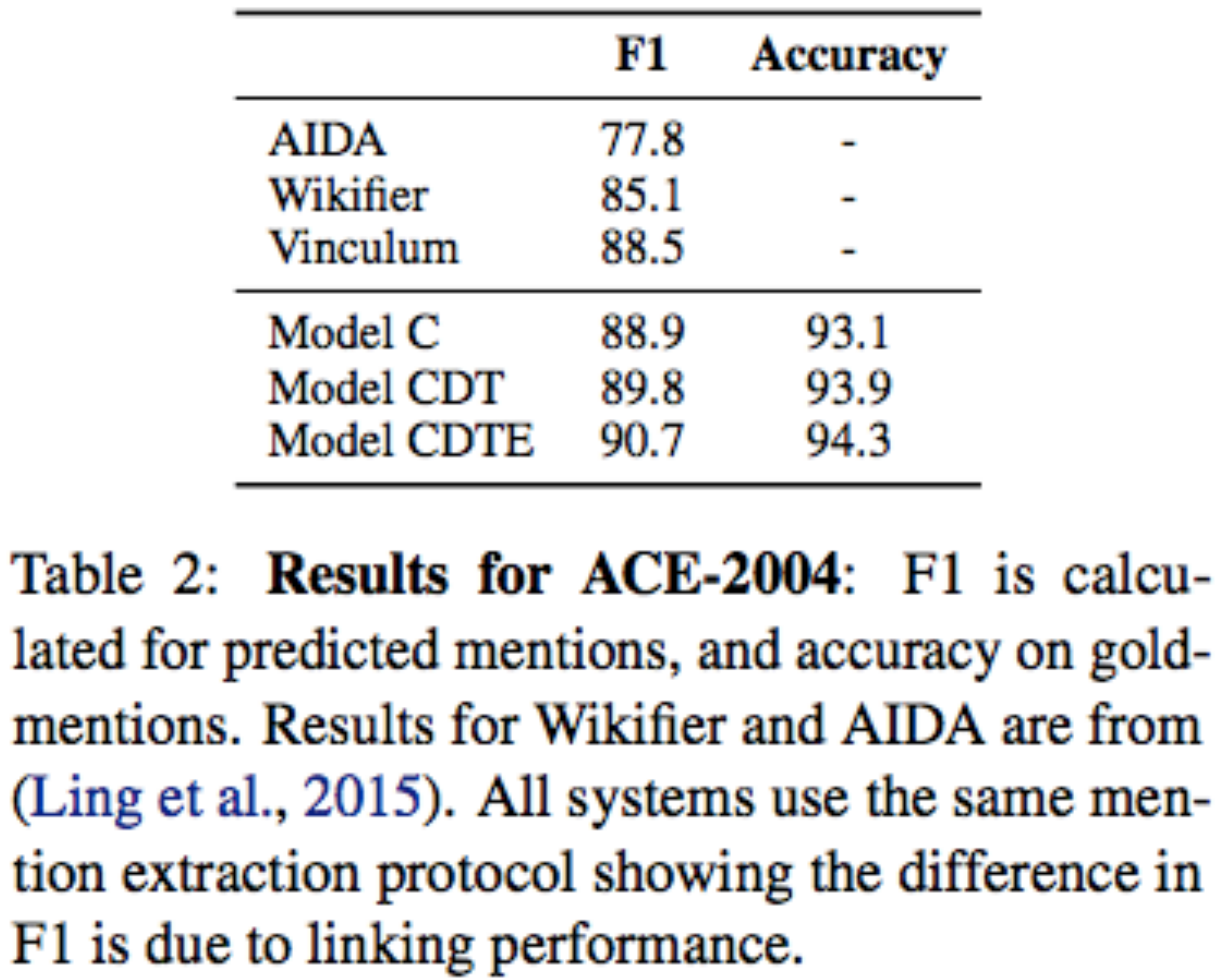
$$\{v_e\}, \Theta = \underset{\{v_e\},\Theta}{\operatorname{argmax}} \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{desc}} + \mathcal{L}_{\text{etype}} + \mathcal{L}_{\text{mtype}}$$

where v_e are the set of entity representations and θ are the parameters



	CoNLL Test	 Dev	ACE05	Wiki
Plato (Sup)	79.7	-	-	-
Plato (Semi-Sup)	86.4	-	-	-
*AIDA**	*81.8*	-	-	-
*BerkCNN:Sparse**	*74.9*	-	*83.6*	*81.5*
*BerkCNN:CNN**	*81.2*	*86.91*	*84.5*	*75.7*
*BerkCNN:Full**	*85.5*	-	*89.9*	*82.2*
Priors	68.5	70.9	81.1	78.1
Model C	81.4	83.4	83.7	86.1
Model CD	81.0	83.2	85.8	86.1
Model CT	82.3	83.9	86.5	88.2
Model CDT	82.5	85.6	86.8	88.0
Model CDTE	82.9	84.9	85.6	89.0

Table 1: **Entity Linking Performance:** Accuracy



	F1	Accuracy
AIDA	77.8	-
Wikifier	85.1	-
Vinculum	88.5	-
Model C	88.9	93.1
Model CDT	89.8	93.9
Model CDTE	90.7	94.3

Table 2: **Results for ACE-2004**: F1 is calculated for predicted mentions, and accuracy on gold-mentions. Results for Wikifier and AIDA are from (Ling et al., 2015). All systems use the same mention extraction protocol showing the difference in F1 is due to linking performance.

Looking forward

- More languages: 3000!
- Multi-media
- Streaming mode
- No more training data
- Context-aware, living