Information Extraction (IE)

Fadi Biadsy

CS4705

Oct 30, 2008

Information Extraction (IE) -- Task

- Idea: ‘extract’ or tag particular types of information from arbitrary text or transcribed speech

Named Entity Tagger

- Identify types and boundaries of named entity
- For example:
  - Alexander Mackenzie , (January 28, 1822 - April 17, 1892), a building contractor and writer, was the second Prime Minister of Canada from ….

  ➔ <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 ….

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 casualty in the accidents.

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

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

Approaches

1. Statistical Sequence Labeling
2. Supervised
3. Semi-Supervised and Bootstrapping

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 approach 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 two classifiers:
  1. A binary classifier:
     - Given a span of text and two entities
     - Decide if there is a relationship between these two entities.
  2. A classifier is trained to determine the types of relations exist between the entities

- **Features:**
  - Types of two named entities
  - Bag-of-words
  - …

- Example:
  - A rented **SUV** went out of control on Sunday, causing the death of **seven** people in Brooklyn
  - Relation: Type=Accident, Vehicle Type= SUV, causality = 7, weather = ?

- **Pros and Cons?**

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>"
- **<u>Matching Techniques</u>**
  - **Exact matching**
    - Pros and Cons?
  - **Flexible matching** (e.g., [X] was .* killed .* by [Y])
    - Pros and Cons?

Pattern Matching

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

Semi-Supervised Approach

AutoSlog-TS (Riloff, 1996)

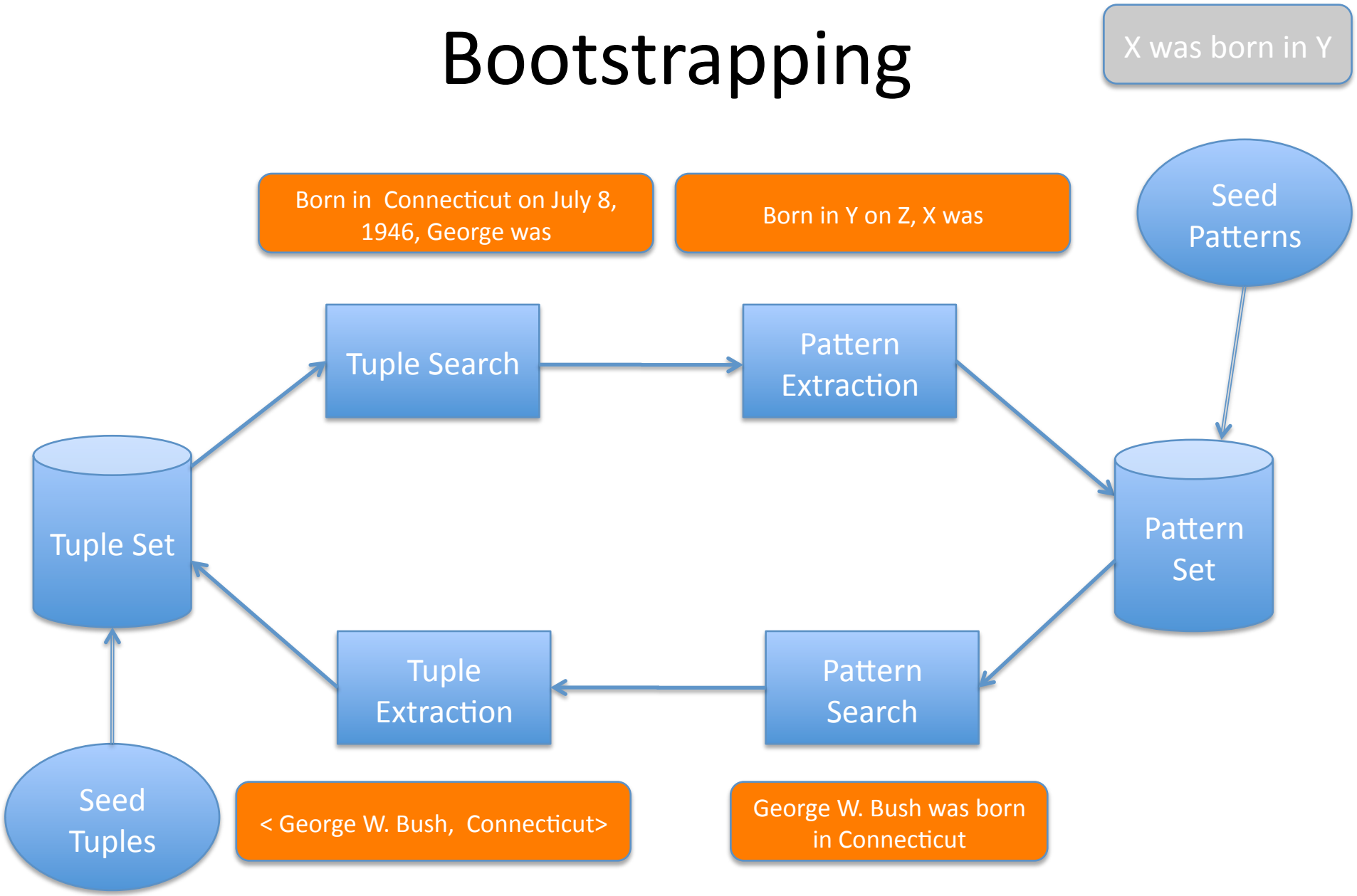
- **MUC-4 task:** extract information about terrorist events in Latin America.
- Two corpora:
  1) Domain-dependent corpus that contains relevant information
  2) A set of irrelevant documents
- Algorithm:
  1. Using some heuristic rules, all patterns are extracted from both corpora. For example:

     Rule: <Subj> passive-verb

     ➔ <Subj> was murdered

     ➔ <Subj> was called
  2. Pattern Ranking: The output patterns are then ranked by frequency of their occurrences in corpus1 / corpus2.
  3. Filter out the patterns by hand

Bootstrapping



TASK 12: (DARPA – GALE year 2)
PRODUCE A BIOGRAPHY OF [PERON].

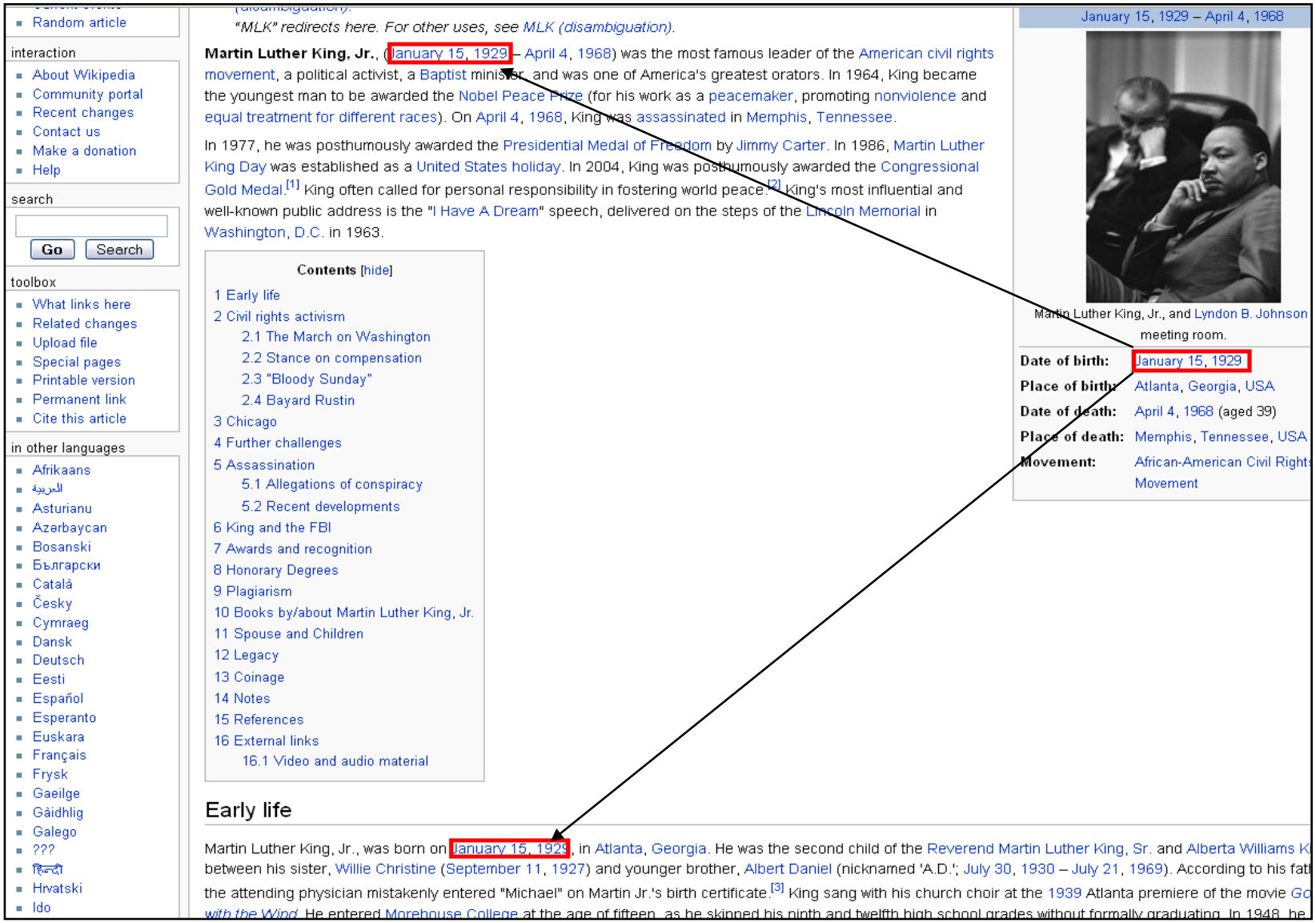
1. Name(s), aliases:
2. *Date of Birth or Current Age:
3. *Date of Death:
4. *Place of Birth:
5. *Place of Death:
6. Cause of Death:
7. Religion (Affiliations):
8. Known locations and dates:
9. Last known address:
10. Previous domiciles:
11. Ethnic or tribal affiliations:
12. Immediate family members
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, we handle each slot independently using bootstrapping to learn IE patterns.

- To improve the recall, we utilize a biographical-sentence classifier.

Biography patterns from Wikipedia



- Random article

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

search

Go Search

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this article

in other languages
- Afrikaans
- العربية
- Asturianu
- Azərbaycan
- Bosanski
- Български
- 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.

Contents [hide]

January 15, 1929 – April 4, 1968

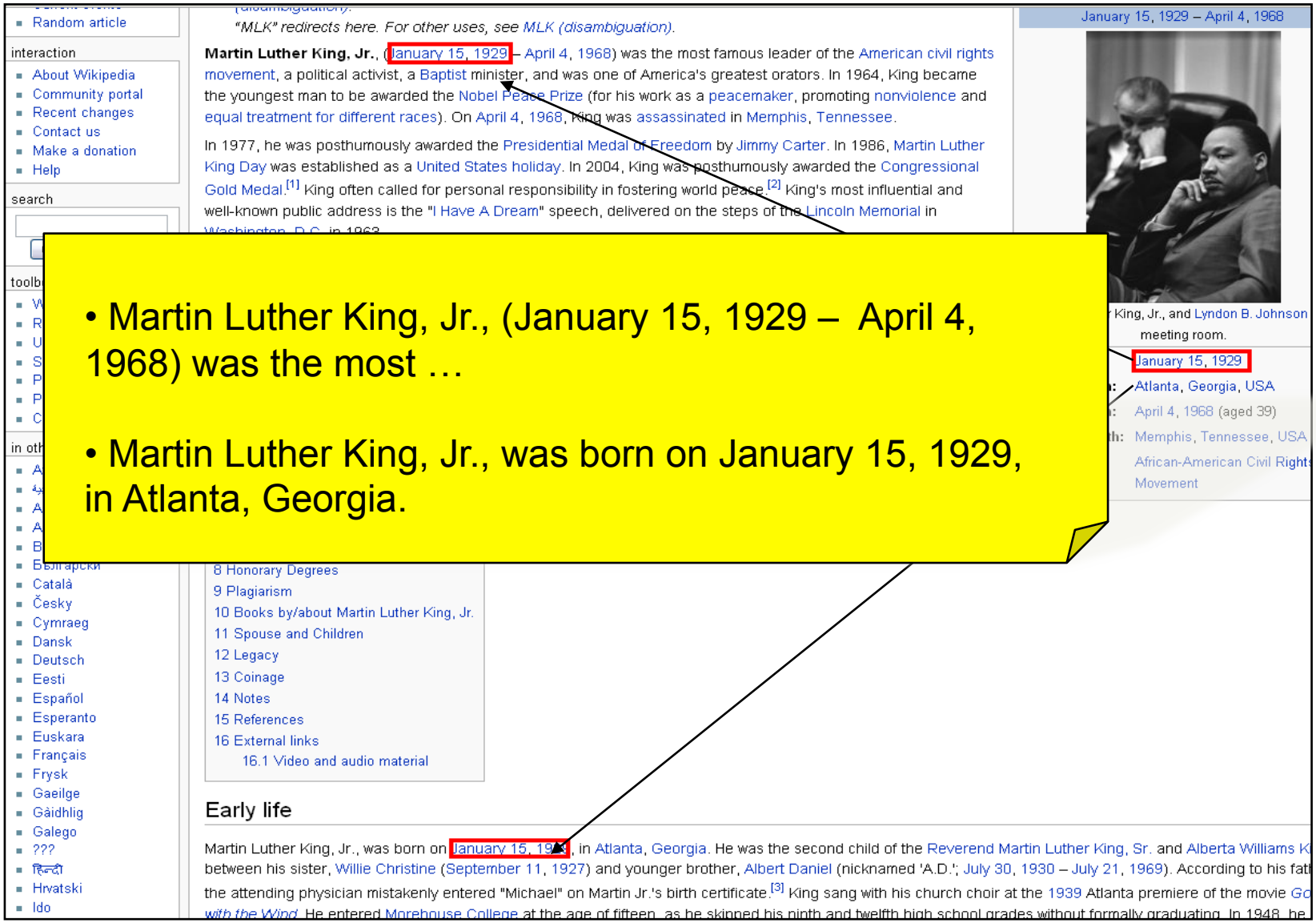
Martin Luther King, Jr., and Lyndon B. Johnson meeting room.

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 K 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 fat 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 with the Wind*. He entered Morehouse College at the age of fifteen, as he skipped his ninth and twelfth high school grades without formally graduating. In 1948, he

Biography patterns from Wikipedia



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:

Web Images Video News Maps more »

Arafat "civil engineering" Search Advanced Search Preferences

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 ...
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:

Web Images Video News Maps more »

site:en.wikipedia.org ~graduated with a bache Search Advanced Search Preferences

Web Results **1**

[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 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!

Bouncing between Wikipedia and Google

- This approach worked well for a few fields such as: *education, publication, Immediate family members, and Party or other organization affiliations*
- Did not provide good patterns for some of the fields, such as: *Religion, Ethnic or tribal affiliations, and Previous domiciles)*, we got a lot of noise
- Why the bouncing idea is better than using only one corpus?
- Non of the patterns match? Back-off strategy…

Biographical-Sentence Classifier

(Biadsy, et al., 2008)

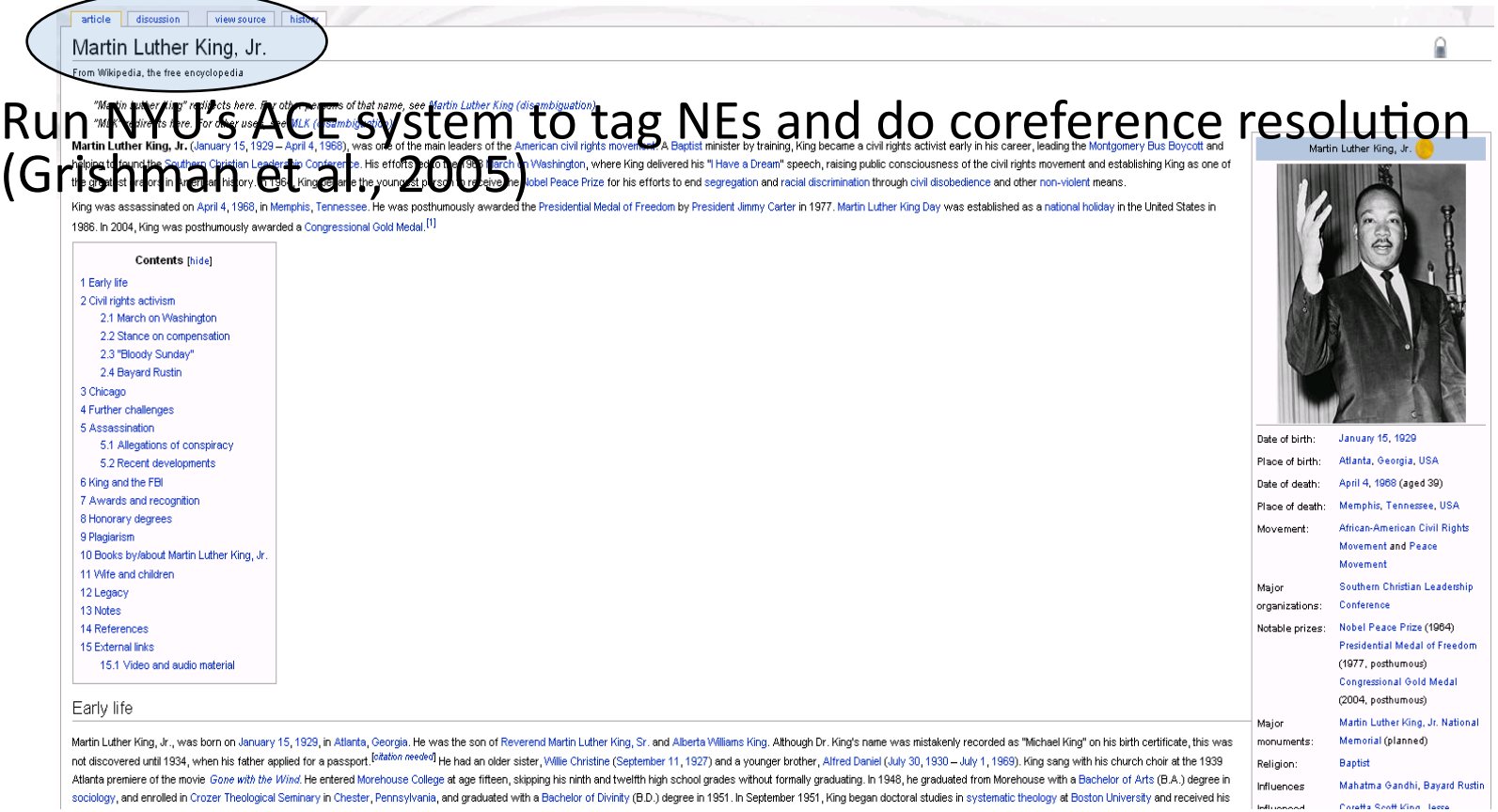
- Train a binary classifier to identify biographical sentences
- Manually annotating a large corpus of biographical and non-biographical information (e.g., Zhou et al., 2004) is labor intensive
- Our approach: collect biographical and non-biographical corpora automatically

Training Data – Biographical Corpus from Wikipedia

- Utilize Wikipedia biographies
- Extract 17K biographies from the xml version of Wikipedia
- Apply simple text processing techniques to clean up the text

Constructing the Biographical Corpus

1. Identify the subject of each biography



2. Run NYU's ACE system to tag NEs and do coreference resolution (Grishman et al., 2005)

Constructing the Biographical Corpus

3. Replace each **NE** by its tag type and subtype

In **September 1951**, **King** began his doctoral studies In theology at **Boston University.**

In **[TIMEX]** , **[PER_ Individual]** began [TARGET_HIS] doctoral studies In theology at **[ORG_Educational]** .

Constructing the Biographical Corpus

3. Replace each NE by its tag type and subtype

4. Non-pronominal referring expression that is coreferential with the target person is replaced by **[TARGET_PER]**

In September 1951, **King** began his doctoral studies In theology at Boston University.

In **[TIMEX]** , **[TARGET_PER]** began **[TARGET_HIS]** doctoral studies In theology at **[ORG_Educational]** .

Constructing the Biographical Corpus

3. Replace each NE by its tag type and subtype

4. Non-pronominal referring expression that is coreferential with the target person is replaced by [TARGET_PER]

5. Every pronoun *P* that refers to the target person is replaced by **[TARGET_*P*]**, where P is the pronoun replaced

In September 1951, King began his doctoral studies In theology at Boston University.

In **[TIMEX]** , **[TARGET_PER]** began **[TARGET_HIS]** doctoral studies In theology at **[ORG_Educational]** .

Constructing the Biographical Corpus

3. Replace each NE by its tag type and subtype

4. Non-pronominal referring expressions that are coreferential with the target person are replaced by [TARGET_PER]

5. Every pronoun *P* that refers to the target person is replaced by [TARGET_*P*], where P is the pronoun replaced

6. Sentences containing no reference to the target person are removed

In September 1951, King began his doctoral studies In theology at Boston University.

In **[TIMEX]** , **[TARGET_PER]** began **[TARGET_HIS]** doctoral studies In theology at **[ORG_Educational] .**

Constructing the Non-Biographical Corpus

- English newswire articles in TDT4 used to represent non-biographical sentences

1. Run NYU's ACE system on each article
2. Select a PERSON NE mention at **random** from all NEs in article to represent the target person
3. Exclude sentences with no reference to this target
4. Replace referring expressions and NEs as in biography corpus

Biographical-Sentence Classifier

- Train a classifier on the biographical and non-biographical corpora
  - Biographical corpus:
    - 30,002 sentences from Wikipedia
    - 2,108 sentences held out for testing
  - Non-Biographical corpus:
    - 23,424 sentences from TDT4
    - 2,108 sentences held out for testing

Biographical-Sentence Classifier

- Features:
  - Frequency of 1-2-3 grams of class-based/lexical, e.g.:
    - [TARGET_PER] was born
    - [TARGET_HER] husband was
    - [TARGET_PER] said
  - Frequency of 1-2 grams of POS
- Chi-square for feature selection

Classification Results

- Experimented with three types of classifiers:

Classifier	Accuracy	F-Meassure
SVM	87.6%	0.87
M. Naïve Bayes (MNB)	84.1%	0.84
C4.5	81.8%	0.82

- Note: Classifiers provide a confidence score for each classified sample

Thank you