Information Extraction

CS4705
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 ...
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.

<table>
<thead>
<tr>
<th>Vehicle Type</th>
<th># accidents</th>
<th># casualties</th>
<th>Weather</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUV</td>
<td>1200</td>
<td>190</td>
<td>Rainy</td>
</tr>
<tr>
<td>Trucks</td>
<td>200</td>
<td>20</td>
<td>Sunny</td>
</tr>
</tbody>
</table>
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

- Statistical sequence labeling
- Supervised
- Semi-supervised and bootstrapping
<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 two classifiers:
  - 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 = ?

- 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>”

Matching Techniques
- 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, not scalable
Semi–supervised approach
AutoSlog–TS (Riloff 1996)

- MUC–4 task: extract information about terrorist events in Latin America
- Two corpora:
  - Domain–dependent corpus that contains relevant information
  - A set of irrelevant documents
- Algorithm:
  1. Using heuristics, 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 the frequency of their occurrences in corpus1/corpus2
  3. Filter out the patterns by hand
Bootstrapping

- Born in Connecticut on July 8, 1946, George was born in Connecticut.
- Born in Y on Z, X was born in Y.
- Seed Patterns
- Tuple Set
  - Tuple Search
  - Tuple Extraction
- Pattern Set
  - Pattern Extraction
  - Pattern Search
- Seed Tuples
  - <George W. Bush, Connecticut>
  - George W. Bush was born in Connecticut.
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

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 King. He died from a bullet fired from a sniper at the age of 39. A significant event in his life was the March on Washington on August 28, 1963, where he delivered his famous "I Have a Dream" speech. King was assassinated on April 4, 1968, in Memphis, Tennessee. His life and work continue to inspire people around the world.
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 1954, 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 1966, Martin Luther King Day was established as a United States holiday. In 2004, King was posthumously awarded the Congressional Gold Medal. King often called for personal responsibility in fostering world peace. 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.

- 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.
- \(\Rightarrow\) `<Target_Person>`, was born on `<Target_Date>`
- \(\Rightarrow\) `<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:
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:
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 ...

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. ...

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 ...

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 ...

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 ...

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). ...

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_PER] said
  – Frequency of 1-2 grams of POS

• Chi-square for feature selection
Classification Results

- Experimented with three types of classifiers:

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>87.6%</td>
<td>0.87</td>
</tr>
<tr>
<td>M. Naïve Bayes (MNB)</td>
<td>84.1%</td>
<td>0.84</td>
</tr>
<tr>
<td>C4.5</td>
<td>81.8%</td>
<td>0.82</td>
</tr>
</tbody>
</table>

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