Information Extraction
Lecture 10 – Ontological and Open IE

CIS, LMU München
Winter Semester 2015-2016

Dr. Alexander Fraser, CIS
Administravia

• Suggested Klausur date is in the last week of the Vorlesung (the week before Fasching)
  • Klausur: February 3rd
  • There will be a review for the Klausur on Wed January 27th

• NEW: there is a conflict with a different course, I will look into this
• Before I start on Ontological IE, two topics I wanted to briefly talk about today:
  • Semantic Role Labeling
  • Wikification
Syntactic Parsing and Relation Extraction

• We saw in the previous two lectures that syntactic features are useful for relation extraction (and event extraction)

• For instance...
Parse Features for Relation Extraction

*American Airlines*, a unit of AMR, immediately matched the move, spokesman *Tim Wagner* said.

Mention 1

- Base syntactic chunk sequence from one to the other
  
  **NP**  **NP**  **PP**  **VP**  **NP**  **NP**

- Constituent path through the tree from one to the other
  
  **NP**  **↑**  **NP**  **↑**  **S**  **↑**  **S**  **↓**  **NP**

- Dependency path
  
  *Airlines*  *matched*  *Wagner*  *said*
Semantic Role Labeling

- A generalization beyond syntactic parsing is Semantic Role Labeling (often abbreviated to SRL)
- Here the idea is to identify the arguments to a verb
  - So this can capture the same information as, e.g., a dependency parse
  - It should be clear that this will be useful in IE
- But the difference is that the arguments are captured in terms of their semantic function rather than their syntactic function
Subcategorization Frame

• Consider the sentences:
  • The man was bitten by the dog
  • The dog bit the man

• In terms of the verb and the subcategorized arguments, there is no difference here

• In Semantic Role Labeling, these will have the same representation!

• Consider also:
  • The man was bitten.
Semantic Role Labeling

Example from Kozhevnikov and Titov

List of SRL tools (see also the comments):
Last Word: Training Data

• The critical problem for statistical approaches is labeled training data

• There are two mature data sets for training semantic role labelers for English
  
  • **Framenet** is the one that may be more useful for many IE purposes (but **Propbank** is also interesting)

• There has been some work on projecting these two resources to other languages using machine translation techniques
  
  • E.g., for German, the "Salsa" project at Uni SB
Wikification

• Wikification is the problem of automatically annotating entities in free text with their (English) Wikipedia page
• Let's start with motivation...
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.
Wikification: Motivation

• Dealing with Ambiguity of Natural Language
  o Mentions of entities and concepts could have multiple meanings

• Dealing with Variability of Natural Language
  o A given concept could be expressed in many ways

• Wikification addresses these two issues in a specific way:

• The Reference Problem
  o What is meant by this concept? (WSD + Grounding)
  o More than just co-reference (within and across documents)
In the last two lectures, we discussed how to extract relations and events from text

- We looked in detail at relations expressed in a single sentence
- Event extraction captures relations which are often expressed at either the sentence or at the document level (i.e., in multiple sentences)
  - Consider the CMU Seminar task – the task is to extract events (seminars), with speaker, location, start time and end time

Today we will discuss updating a knowledge base with the extracted relations or events
- This is called "Ontological IE"
Ontologies

An **ontology** is a consistent knowledge base without redundancy

<table>
<thead>
<tr>
<th>Person</th>
<th>Nationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>German</td>
</tr>
<tr>
<td>Merkel</td>
<td>Germany</td>
</tr>
<tr>
<td>A. Merkel</td>
<td>French</td>
</tr>
</tbody>
</table>

- Every entity appears only with exactly the same name
- There are no semantic contradictions

Slide from Suchanek
Ontological Information Extraction (IE) aims to create or extend an ontology.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>citizenOf</td>
<td>Germany</td>
</tr>
</tbody>
</table>

Angela Merkel is the German chancellor.... Merkel was born in Germany...

...A. Merkel has French nationality...

<table>
<thead>
<tr>
<th>Person</th>
<th>Nationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>German</td>
</tr>
<tr>
<td>Merkel</td>
<td>Germany</td>
</tr>
<tr>
<td>A. Merkel</td>
<td>French</td>
</tr>
</tbody>
</table>
Ontological IE Challenges

Challenge 1:
Map names to names that are already known

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>citizenOf</td>
<td>Germany</td>
</tr>
</tbody>
</table>

Merkel
Angie
A. Merkel
Ontological IE Challenges

Challenge 2:
Be sure to map the names to the right known names

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>citizenOf</td>
<td>Germany</td>
</tr>
<tr>
<td>Una Merkel</td>
<td>citizenOf</td>
<td>USA</td>
</tr>
</tbody>
</table>

Merkel is great!
Ontological IE Challenges

Challenge 3: Map to known relationships

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>citizenOf</td>
<td>Germany</td>
</tr>
</tbody>
</table>

... has nationality ...
... has citizenship ...
... is citizen of ...
Ontological IE Challenges

Challenge 4:
Take care of consistency

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>citizenOf</td>
<td>Germany</td>
</tr>
</tbody>
</table>

Angela Merkel is French...
A **triple** (in the sense of ontologies) is a tuple of an entity, a relation name, and another entity:

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>citizenOf</td>
<td>Germany</td>
</tr>
</tbody>
</table>

Most ontological IE approaches produce triples as output. This decreases the variance in schema.
Triples

A triple can be represented in multiple forms:

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>citizenOf</td>
<td>Germany</td>
</tr>
</tbody>
</table>

\[<\text{Angela Merkel}, \text{citizenOf}, \text{Germany}>\]

Slide from Suchanek
Example: Elvis in YAGO
• Let's talk about ontological IE using extraction from Wikipedia as an example

• Then we will go on to open IE, which uses similar ideas to extract from all the text on the web!
Why is Wikipedia good for information extraction?
• It is a huge, but homogenous resource
  (more homogenous than the Web)
• It is considered authoritative
  (more authoritative than a random Web page)
• It is well-structured with infoboxes and categories
• It provides a wealth of meta information
  (inter article links, inter language links, user discussion,...)
Ontological IE from Wikipedia

Wikipedia is a free online encyclopedia
- 3.4 million articles in English
- 16 million articles in dozens of languages

Every article is (should be) unique
=> We get a set of unique entities
that cover numerous areas of interest

Angela_Merkel
Una_Merkel
Germany
Theory_of_Relativity
<table>
<thead>
<tr>
<th>Background information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Birth name</strong></td>
</tr>
<tr>
<td><strong>Born</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Died</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Genres</strong></td>
</tr>
<tr>
<td><strong>Occupations</strong></td>
</tr>
<tr>
<td><strong>Instruments</strong></td>
</tr>
<tr>
<td><strong>Years active</strong></td>
</tr>
<tr>
<td><strong>Labels</strong></td>
</tr>
<tr>
<td><strong>Associated acts</strong></td>
</tr>
<tr>
<td><strong>Website</strong></td>
</tr>
</tbody>
</table>

| Birth_name = Elvis Aaron Presley |
| Born = {{Birth date | 1935 | 1 | 8}}<br />[[Tupelo, Mississippi | Tupelo]] |
IE from Wikipedia

Elvis Presley

~Infobox~
Born: 1935
...

Categories: Rock singers

Exploit Infoboxes

bornOnDate = 1935 (hello regexes!)

born = 1935

Slide from Suchanek
Blah blah blub fasel (do not read this, better listen to the talk) blah blah Elvis blub (you are still reading this) blah Elvis blah blub later became astronaut blah

Categories: Rock singers

Exploit conceptual categories

Exploit Infoboxes

Rock Singer

born: 1935

type: Rock Singer

IE from Wikipedia
Consistency Checks

Check uniqueness of functional arguments
Check domains and ranges of relations
Check type coherence
Ontological IE from Wikipedia

YAGO
- 3m entities, 28m facts
- focus on precision 95% (automatic checking of facts)
  [http://yago-knowledge.org](http://yago-knowledge.org)

DBpedia
- 3.4m entities
- 1b facts (also from non-English Wikipedia)
- large community
  [http://dbpedia.org](http://dbpedia.org)

Freebase
- Community project on top of Wikipedia (bought by Google, but still open)
  [http://freebase.com](http://freebase.com)
- Now integrated into Wikidata!!!
Ontological IE by Reasoning

Recap: The challenges:
- deliver canonic relations
- deliver canonic entities
- deliver consistent facts

Idea: These problems are interleaved, solve all of them together.

Elvis was born in 1935

died in, was killed in

Elvis, Elvis Presley, The King

born (Elvis, 1970)
born (Elvis, 1935)
Using Reasoning

Ontology

Documents
Elvis was born in 1935

Consistency Rules
birthdate < deathdate

First Order Logic

type(Elvis_Presley,singer)
subclassof(singer,person)
...

appears("Elvis","was born in","1935")
...

means("Elvis",Elvis_Presley,0.8)
means("Elvis",Elvis_Costello,0.2)
...

born(X,Y) & died(X,Z) => Y<Z
appears(A,P,B) & R(A,B)
  => expresses(P,R)
appears(A,P,B) & expresses(P,R)
  => R(A,B)
...

SOFIE system

Slide from Suchanek
Ontological IE by Reasoning

Reasoning-based approaches use logical rules to extract knowledge from natural language documents.

Current approaches use either
- Weighted MAX SAT
- or Datalog
- or Markov Logic

Input:
- often an ontology
- manually designed rules

Condition:
- homogeneous corpus helps
Ontological IE Summary

Ontological Information Extraction (IE) tries to create or extend an ontology through information extraction.

Current hot approaches:
• extraction from Wikipedia
• reasoning-based approaches
• integrating uncertainty
Open Information Extraction

Open Information Extraction/Machine Reading aims at information extraction from the entire Web.

Vision of Open Information Extraction:
- the system runs perpetually, constantly gathering new information
- the system creates meaning on its own from the gathered data
- the system learns and becomes more intelligent, i.e. better at gathering information
Open Information Extraction

Open Information Extraction/Machine Reading aims at information extraction from the entire Web.

Rationale for Open Information Extraction:
• We do not need to care for every single sentence, but just for the ones we understand
• The size of the Web generates redundancy
• The size of the Web can generate synergies
KnowItAll &Co

KnowItAll, KnowItNow and TextRunner are projects at the University of Washington (in Seattle, WA).

Egyptians built pyramids 400
Americans built pyramids 20
...

Valuable common sense knowledge (if filtered)

http://www.cs.washington.edu/research/textrunner/
TextRunner took .80 seconds.

Retrieved 391 results for Predicate containing "built" and Argument 2 containing "pyramids"

Grouping results by predicate. Group by: argument 2 | argument 1

**built** - 159 results

Egyptians (297), aliens (71), Pharaohs (40), built the pyramids
Egyptians (26), Khufu (18), Maya (9), 30 more... built the Great Pyramid
Imhotep (8), Pharaoh Zoser (4), Egyptians (2), King Djoser (2) built the Step Pyramid
two symbols of life (4), 6th dynasty kings (3), King Sneferu (3), Snefru (3) built two large Pyramids
Egyptians (8) built the Great Pyramids
ancient Egyptians (6) built more than 90 royal pyramids
colonial silver city of Taxco (3), Explore (2) built the gigantic pyramids of the Sun
Central America (2), part of Mexico (2) built great cities, temples and pyramids

http://www.cs.washington.edu/research/textrunner/
“Read the Web” is a project at the Carnegie Mellon University in Pittsburgh, PA.

Krzewski coaches the Blue Devils.

Krzewski  Blue Angels
Miller    Red Angels

sports coach \neq scientist

Type Check
If I coach, am I a coach?

http://rtw.ml.cmu.edu/rtw/

Slide from Suchanek
Open IE: Read the Web

- **arthropod** (100.0%)
  - Seed
    - CPL @156 (100.0%) on 30-sep-2010
      - "hind wings of _" "invertebrates, such as _" _ swarm from" "other insects, including _" _ do not eat wood" "many legs as _" _ produce such " "complete metamorphosis" "I do n't see anymore _" "ants, so _" "insecticide for such insects as _" _ are the only insects "red imported _" "insects like _" "social insects, such as _" "arthropods include _" "insect pests including _" "meaty foods like _" "arthropods include _" "insect pests, such as _" "other insects such as _" "insects, in particular _" "release a pheromone like _" "many insects, including _" _ are social insects" "insect pests such as _" _ are arthropods, including _" _ are beneficial insects _"
    - SEAL @151 (50.0%) on 26-sep-2010
      - kateretes (Seed)
      - mosquito (Seed)
      - peppered_moth (Seed)
      - sap_beetle (Seed)
      - tettigoniidae (Seed)
      - triatoma_protracta (Seed)
      - honeylocust_spider_mite
      - grape_flea_beetle
      - blueberry_leaf_beetle
      - sugarcane_moth_borer
      - psychoda_moth_flies
      - bagworm_moth
      - carpenterworm_moths
      - leafcurl_plum_aphid
      - merchant_grain_beetle

http://rtw.ml.cmu.edu/rtw/
Open Information Extraction

Open Information Extraction/Machine Reading aims at information extraction from the entire Web.

Main hot projects
- TextRunner (University of Washington)
- Read the Web (Carnegie Mellon)
- Prospera/SOFIE (Max-Planck Informatics Saarbrücken)

Input
- The Web
- Read the Web: Manual rules
- Read the Web: initial ontology

Conditions
- none
• Slide sources
  – Many of the slides today on Ontological IE and Open IE are from Fabian Suchanek (Télécom ParisTech)
  – See the web page I mentioned for a list of semantic role labelers
  – Some of the Wikification slides are from Dan Roth's tutorial, this is highly recommended
• Thank you for your attention!