Information Extraction

NER

What is NER?

Approaches

Gazetteers

Rule-based methods

NER via ML

Relation Extraction

What is RE?

RE Methods

Rule-based methods

Machine learning

Other approaches

Web Search & Text Analysis

Information Extraction

Karin Verspoor

04 April 2017

THE UNIVERSITY OF MELBOURNE

What is Information Extraction?

Information Extraction is the task of identifying core entities, relationships, and events in a text.

There are two common subtasks:

- Named Entity Recognition: finding mentions of particular entity types (e.g., Person, Place, Company name)
- Relation/Event Extraction: finding specific events or relationships that involve the entities.

Typically the objective is not full *natural language understanding*, but to find specific, pre-defined types of information in the text.

- Extraction aims to fill a given "template" which defines the structure of the information to be extracted.
- Domain-specific
- (Recently, there have been efforts to build "open" information extraction systems which do not pre-define domain or relation types.)

MUC3: Kidnapping

TST-1-MUC3-0080

BOGOTA, 3 APR 90 (INRAVISION TELEVISION CADENA 1) - [REPORT] [JORGE ALONSO SIERRA VALENCIA] [TEXT] LIBERAL SENATOR FEDERICO ESTRADA VELEZ WAS KIDNAPPED ON 3 APRIL AT THE CORNER OF 60TH AND 48TH STREETS IN WESTERN MEDELLIN, ONLY 100 METERS FROM A METROPOLITAN POLICE CAI [IMMEDIATE ATTENTION CENTER]. THE ANTIOQUIA DEPARTMENT LIB- ERAL PARTY LEADER HAD LEFT HIS HOUSE WITHOUT ANY BODYGUARDS ONLY MINUTES EARLIER. AS WE WAITED FOR THE TRAFFIC LIGHT TO CHANGE, THREE HEAVILY ARMED MEN FORCED HIM TO GET OUT OF HIS CAR AND INTO A BLUE RENAULT.

HOURS LATER, THROUGH ANONYMOUS TELEPHONE CALLS TO THE METROPOLITAN POLICE AND TO THE MEDIA, THE EXTRADITABLES CLAIMED RESPONSIBILITY FOR THE KIDNAPPING. IN THE CALLS, THEY ANNOUNCED THAT THEY WILL RE- LEASE THE SENATOR WITH A NEW MESSAGE FOR THE NATIONAL GOVERNMENT. LAST WEEK, FEDERICO ESTRADA HAD REJECTED TALKS BETWEEN THE GOVERNMENT AND THE DRUG TRAFFICKERS.

TST-1-MUC3-0080

0	MESSAGE ID	TST1-MUC3-0080
1	TEMPLATE ID	1
2	DATE OF INCIDENT	03 APR 90
3	TYPE OF INCIDENT	KIDNAPPING
4	CATEGORY OF INCIDENT	TERRORIST ACT
5	PERPETRATOR: ID OF INDIV(S)	“THREE HEAVILY ARMED MEN”
6	PERPETRATOR: ID OF ORG(S)	“THE EXTRADITABLES”
7	PERPETRATOR: CONFIDENCE	CLAIMED OR ADMITTED: “THE EXTRADITABLES”
8	PHYSICAL TARGET: ID(S)	*
9	PHYSICAL TARGET: TOTAL NUM	*
10	PHYSICAL TARGET: TYPE(S)	*
11	HUMAN TARGET: ID(S)	“FEDERICO ESTRADA VELEZ” (“LIBERAL SENATOR”)
12	HUMAN TARGET: TOTAL NUM	1
13	HUMAN TARGET: TYPE(S)	GOVERNMENT OFFICIAL: “FEDERICO ESTRADA VELEZ”
14	TARGET: FOREIGN NATION(S)	–
15	INSTRUMENT: TYPE(S)	*
16	LOCATION OF INCIDENT	COLOMBIA: MEDELLIN (CITY)
17	EFFECT ON PHYSICAL TARGETS	*
18	EFFECT ON HUMAN TARGETS	*

MUC5 text

<DOC>
<DOCNO> 0592 </DOCNO>
<DD> NOVEMBER 24, 1989, FRIDAY </DD>
<SO>Copyright (c) 1989 Jiji Press Ltd.;</SO>
<TXT>
BRIDGESTONE SPORTS CO. SAID FRIDAY IT HAS SET UP A JOINT VENTURE IN TAIWAN WITH A LOCAL CONCERN AND A JAPANESE TRADING HOUSE TO PRODUCE GOLF CLUBS TO BE SHIPPED TO JAPAN.

THE JOINT VENTURE, BRIDGESTONE SPORTS TAIWAN CO., CAPITALIZED AT 20 MILLION NEW TAI- WAN DOLLARS, WILL START PRODUCTION IN JANUARY 1990 WITH PRODUCTION OF 20,000 IRON AND METAL WOOD CLUBS A MONTH. THE MONTHLY OUTPUT WILL BE LATER RAISED TO 55,000 UNITS, BRIDGESTON SPORTS OFFICIALS SAID.

THE NEW COMPANY, BASED IN KAOHSIUNG, SOUTHERN TAIWAN, IS OWNED 75 PCT BY BRIDGE- STONE SPORTS, 15 PCT BY UNION PRECISION CASTING CO. OF TAIWAN AND THE REMAINDER BY TAGA CO., A COMPANY ACTIVE IN TRADING WITH TAIWAN, THE OFFICIALS SAID. BRIDGESTONE SPORTS HAS SO FAR BEEN ENTRUSTING PRODUCTION OF GOLF CLUBS PARTS WITH UNION PRECISION CASTING AND OTHER TAIWAN COMPANIES.

WITH THE ESTABLISHMENT OF THE TAIWAN UNIT, THE JAPANESE SPORTS GOODS MAKER PLANS TO INCREASE PRODUCTION OF LUXURY CLUBS IN JAPAN.
</TXT>
</DOC>

MUC5 template

```
<TEMPLATE-0592-1> :=
   DOC NR: 0592
   DOC DATE: 241189
   DOCUMENT SOURCE: "Jiji Press Ltd."
   CONTENT: <TIE UP RELATIONSHIP-0592-1>
<TIE UP RELATIONSHIP-0592-1>:=
   TIE-UP STATUS: EXISTING
   ENTITY: <ENTITY-0592-1>
   JOINT VENTURE CO:<ENTITY-0592-4>
   OWNERSHIP: <OWNERSHIP-0592-1>
   ACTIVITY:<ACTIVITY-0592-1>
<ENTITY-0592-1>:=
   NAME: BRIDGESTONE SPORTS CO
   ALIASES: "BRIDGESTONE SPORTS"
          "BRIDGESTON SPORTS"
   NATIONALITY: Japan (COUNTRY)
   TYPE: COMPANY
   ENTITY RELATIONSHIP:<ENTITY RELATIONSHIP-0592-1>
<OWNERSHIP-0592-1>:=
   OWNED: <ENTITY-0592-4>
   TOTAL-CAPITALIZATION: 20000000
   TWD OWNERSHIP-%: (<ENTITY-0592-1> 75)
```

```
<ENTITY-0592-4>:=
   NAME: BRIDGESTONE SPORTS TAIWAN CO
   ALIASES: "UNION PRECISION CASTING"
          "BRIDGESTON SPORTS"
   LOCATION: "KAOHSIUNG" (UNKNOWN) Taiwan (COUNTRY)
   TYPE: COMPANY
   ENTITY RELATIONSHIP:<ENTITY RELATIONSHIP-0592-1>
<ENTITY RELATIONSHIP-0592-1>:=
   ENTITY1: <ENTITY-0592-1>
   ENTITY2: <ENTITY-0592-4>
   REL OF ENTITY2 TO ENTITY1: CHILD
   STATUS: CURRENT
<ACTIVITY-0592-1>:=
   INDUSTRY: <INDUSTRY-0592-1>
   ACTIVITY-SITE: (Taiwan (COUNTRY) <ENTITY-0592-4>)
   START TIME: <TIME-0592-1>
<TIME-0592-1>:=
   DURING: 0190
<INDUSTRY-0592-1>:=
   INDUSTRY-TYPE: PRODUCTION
   PRODUCT/SERVICE: (CODE 39 "20,000 IRON AND METAL WOOD")
```

Named Entity Recognition

NE types:

- ENAMEX (type= person, organisation, location)
- TIMEX (type= time, date)
- NUMEX (type= money, percent)

NE markup with subtypes:

<ENAMEX TYPE='PERSON'>Flavel Donne</ENAMEX> is an analyst with <ENAMEX TYPE='ORGANIZATION'>General Trends</ENAMEX>, which has been based in <ENAMEX TYPE='LOCATION'>Little Spring</ENAMEX> since <TIMEX>July 1998</TIMEX>.

Reflective Activity: Named Entity Recognition

- How might you approach Named Entity Recognition? (what kinds of methods can you think of?)
- What challenges might you run into with these methods?

Easy vs Hard Named Entity Categories

ENAMEX is harder, more context dependent than TIMEX and NUMEX:

- Is *Granada* a COMPANY or a LOCATION?
- Is *Washington* a PERSON or a LOCATION?
- Is *Arthur Anderson* a PERSON or an ORGANISATION?

Named Entity Recognition approaches

- Gazetteers (fixed list containing names of a certain type, e.g. countries, last names, titles, state names, rivers...)
- Manually written rules and regular expressions
  - Rules about mid initials, postfixes, titles
  - Acronyms: Hewlett Packard Inc. $\rightarrow$ HP
  - Patterns: "president of <company>"
    matches *executive vice president of Hupplewhite*

Gazetteers

Gazetteer of full names impossible and not useful, as both first and last names can occur on their own

Last name gazetteer impractical

- Almost infinite set of name patterns possible: last names are productive (1.5M surnames in US alone)
- Overlap with common nouns/verbs/adjectives
  - First 2 pages of Cambridge phone book include 237 names
  - Of those, 6 (2.5%) are common nouns: Abbey, Abbot, Acres, Afford, Airs, Alabaster

First name gazetteer less impractical, but still not foolproof

- First names can be surprising, eg. MUC-7 example: “Llennel Evangelista”
- First names are productive, eg. Moon Unit Zappa, Apple Paltrow . . .

Names are tricky

Overlap with common nouns:

- River and Rain Phoenix, Moon Unit Zappa, Apple Paltrow
- "Virtue names": Grace (134), Joy (390), Charity (480), Chastity (983), Constance, Destiny
- "Month names": June, April, May
- "Flower names": Rose, Daisy, Lily, Erica, Iris . . .

From US Social Security Administration's list of most popular girls' names in 1990, with rank:

Amber (16), Crystal (41), Jordan (59), Jade (224), Summer (291), Ruby (300), Diamond (450), Infant (455), Precious (472), Genesis (528), Paris (573), Princess (771), Heaven (902), Baby (924) . . .

- Additional problem: non-English names alliterated into English; variant spellings
- Complicated name patterns with titles: Sammy Davis Jr, HRH The Prince of Wales, Dr. John T. Maxwell III

Name type ambiguity

Ambiguity of name types: Columbia (Org.) vs. (British) Columbia (Location) vs. Columbia (Space shuttle)

- Company names often use common nouns ("Next", "Boots", "Thinking Machines". . .) and can occur in variations ("Peter Stuyvesant", "Stuyvesant")
- Coordination problems/ left boundary problems:
  - One or two entities in: "China International Trust and Investment Corp invests $2m in . . . "?
  - Unknown word at beginning of potential name: in or out?
    Suspended Ceilings Inc vs Yesterday Ceilings Inc
    Mason, Daily and Partners vs. Unfortunately, Daily and Partners

Experiments show: simple gazetteers fine for locations (90%P/80%R) but not for person and organisations (80%P/50%R)

High-precision rules

Mix regular expression patterns with words recognised from gazetteers (e.g., PROF = profession; REL = relative)

Rule	Assign	Example
`(Xx+)+ (is|,) a? JJ* PROF`	PERS	Yuri Gromov, a former director
`(Xx+)+ is? a? JJ* REL`	PERS	John White is beloved brother
`(Xx+)+ himself`	PERS	White himself
`(Xx+)+, DD+ ,`	PERS	White, 33,
`share? in (Xx+)+`	ORG	shares in Trinity Motors
`(Xx+)+ Inc.`	ORG	Hummingbird Inc.
`PROF (of|at|with) (Xx+)+`	ORG	director of Trinity Motors
`(Xx+)+ (region|area)`	LOC	Lower Beribidjan area

NER as a machine learning task

Can we frame NER as a machine learning task?

What if we treat it as a word tagging task, like POS Tagging?

- not immediately obvious, as some entities are multi-word
- one solution is to change the model to handle multi-word observations
  - hidden semi-Markov models
  - semi-Conditional Random Field models
- . . . Simple approach: map to a token-based tagset

NER as token tagging task: In/Out (IOB) sequence labelling

(Known variously as IOB or BIO)
BIO labelling trick: Apply a tag to each word

- B = begin entity
- I = inside (continuing) entity
- O = outside, non-entity

Tony/B-PERSON Abbott/I-PERSON has/O declared/O the/O GP/O co-payment/O as/O “/O dead/O , /O buried/O and/O cremated/O ”/O after/O it/O was/O finally/O dumped/O on/O Tuesday/B-DATE ./O

Analysis

- allows for adjacent entities(e.g., B-PERSON B-PERSON)
- expands the tag set by a small factor, some impact on efficiency and learning quality

BIO Tag Sequence learning

How can we learn a BIO tag sequence?

- In exactly the same way we learn a POS tag sequence!
- N-gram models, Hidden Markov Models, Maximum Entropy Markov Models
- Conditional Random Fields

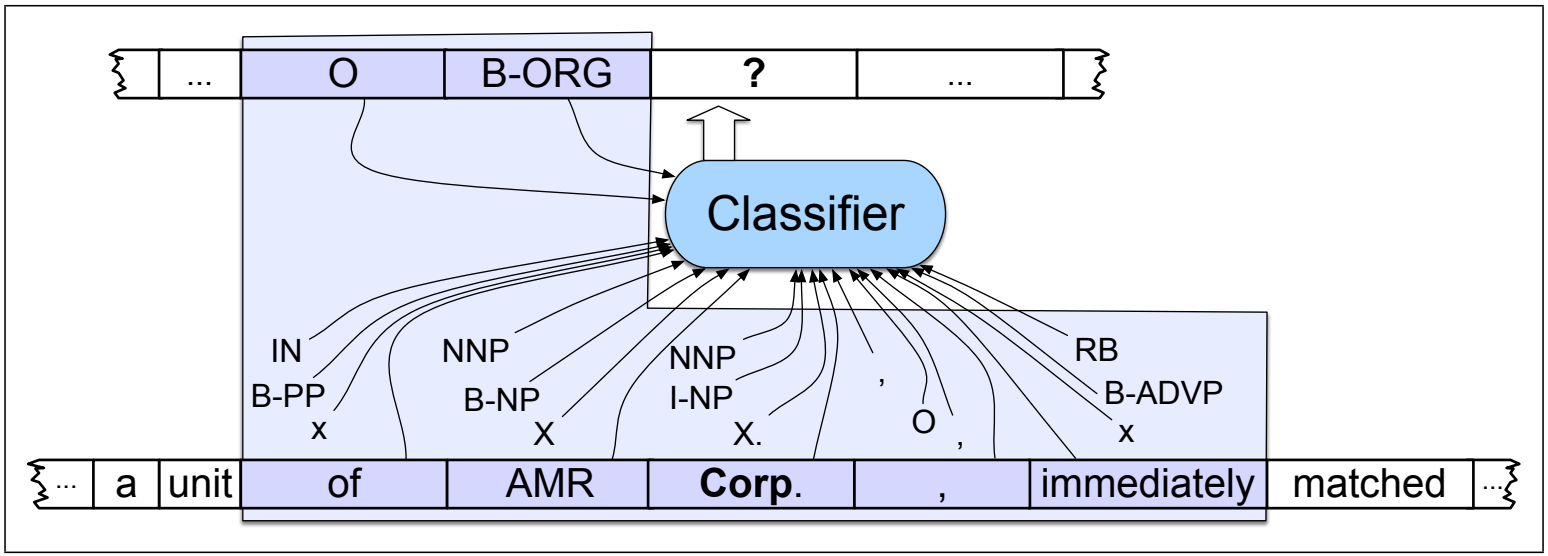


Figure 21.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

Introduction to Relation Extraction

Web Search & Text Analysis
Karin Verspoor
Information Extraction
NER
What is NER?
Approaches
Gazetteers
Rule-based methods
NER via ML
Relation Extraction
What is RE?
RE Methods
Rule-based methods
Machine learning
Other approaches

We introduced the idea of Relation Extraction above as

- Relation/Event Extraction: finding specific events or relationships that involve the entities.

We also saw some examples from the early MUC competitions, which included some very complex event representations.

Let's focus on the simplified task of extracting *relation triples*. These have the form (for a binary relationship):

Predicate (Argument1, Argument2)

For instance, to express the core relational information in *Julian is a lecturer of the subject Web Search & Text Analysis*, we might formally write:

lecturer ('Julian', 'Web Search & Text Analysis')

Why Relation Extraction?

Relation Extraction allows us to summarise the core relationships that are expressed in a text.

This has applications in

- Transforming "unstructured" information into "structured", relational information
  - This in turn allows us to store information extracted from text in databases in a more computable form.
  - (Example: populating a job candidate experience database by analysing CVs)
  - The information can more easily be manipulated for subsequent retrieval or analysis.
- "InfoBox" summaries
- Using relational information extracted from text in other predictive applications (more next week)
- Question answering
- Supporting reasoning and inference – *"connecting the dots"*

Who was the last President before Clinton to visit Vietnam?

Central African Republic president visits Vietnam
May 19, 2009

VietNamNet Bridge -- The President of the Central African Republic, Francois Bozize Yangouvonda, began a five-day official visit to Vietnam on May 18 at the invitation of State President Nguyen Minh Triet.

Who was the last President before Clinton to visit Vietnam?

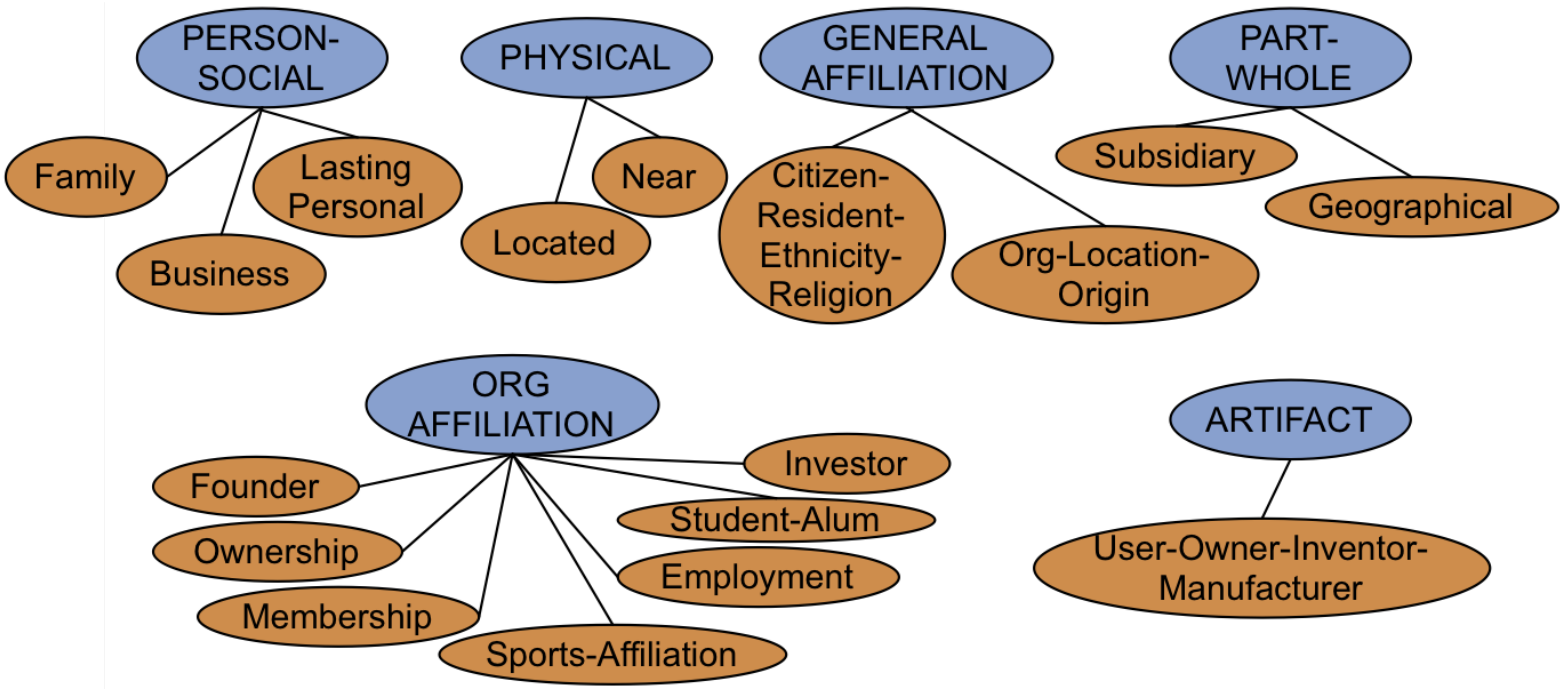
Thanks to massive digital archiving of news, we might find the answer.

Bill Clinton - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Bill_Clinton
As **president**, **Clinton** presided over the longest period of peacetime ... **After** a failed health care reform attempt, the Republican Party won control of ... a budget surplus between the years 1998 and 2000, the **last** three years of **Clinton's** presidency. ..... **Clinton's** November 2000 **visit** to **Vietnam** was the first by a U.S. **President** ...

BBC News | ASIA-PACIFIC | Clinton's Vietnam visit
news.bbc.co.uk/2/hi/asia-pacific/1025169.stm
16 Nov 2000 – **President** Bill **Clinton** is the first US head of state to **visit Vietnam** since the end of the ... His four-day **visit**, which starts later on Thursday, is also his **last** ... the **president** said at a Veteran's Day ceremony shortly **before** the **visit**.

Clinton Makes Historic Visit to Vietnam - ABC News
abcnews.go.com › International
Richard Nixon was the **last** serving **president** to **visit** the region when he traveled ... But several hours **before** the **president** arrived, the **Clinton** charm had already ...

Automated Content Extraction shared task



ACE relation examples

Physical-Located PER-GPE

- He was in Tennessee

Part-Whole-Subsidiary ORG-ORG

- XYZ, the parent company of ABC

Person-Social-Family PER-PER

- John's wife Yoko

Org-AFF-Founder PER-ORG

- Steve Jobs, co-founder of Apple

Relations: 'Encyclopedic' or hierarchical knowledge

- IS-A (hypernym): subsumption between classes
  `Giraffe IS-A ruminant IS-A ungulate IS-A mammal IS-A vertebrate IS-A animal ...`
- Instance-of: relation between individual and class
  `San Francisco instance-of city`
- IS-PART (meronym): part/whole relations between classes
  `wheel IS-PART car`

RE: Cues from language

`Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.`

What is Gelidium?

How do you know?

Intuition from Hearst (1992): "Automatic Acquisition of Hyponyms from Large Text Corpora"

RE Methods: Hand-written patterns

As with NER rules, use indicative linguistic patterns to identify relations.

(Yes! Regular Expressions to the rescue!)

E.g., for hyponyms (plus some patterns for dealing with lists):

`Y such as X ((, X)* (, and|or) X)`	The bow lute, such as the Bambara ndang
`such Y as X`	such authors as Herrick, Goldsmith, and Shakespeare
`X or other Y`	bruises, wounds, broken bones or other injuries
`X and other Y`	temples, treasuries, and other important civic buildings
`Y(,)? including X`	common-law countries, including Canada and England
`Y, especially X`	European countries, especially France, England, and Spain

Relations have constraints

Relations often hold between specific entity types.

- located-in (ORGANIZATION, LOCATION)
- founded (PERSON, ORGANIZATION)
- cures (DRUG, DISEASE)

. . . So let's look for pairs of co-occurring named entities.

Many relations can hold between two given NEs

Drug — Disease

Drug -cause- Disease
Drug -treat- Disease
Drug -prevent- Disease
Drug -cure- Disease

Person — Organisation

Person -founder- Organisation
Person -employee- Organisation
Person -member- Organisation
Person -investor- Organisation
Person -president- Organisation

NE-based patterns

We can define patterns in terms of previously identified named entities.

```
>>> IN = re.compile(r'.*\bin\b(?!\b.+ing)')
>>> for doc in nltk.corpus.ieer.parsed_docs('NYT_19980315'):
...     for rel in nltk.sem.extract_rels('ORG', 'LOC', doc,
...                                      corpus='ieer', pattern = IN):
...         print(nltk.sem.rtuple(rel))

[ORG: 'WHYY'] 'in' [LOC: 'Philadelphia']
[ORG: 'McGlashan &AMP; Sarrail'] 'firm in' [LOC: 'San Mateo']
[ORG: 'Freedom Forum'] 'in' [LOC: 'Arlington']
```

NE-based patterns

Who holds what office in what organisation?

`PERSON, POSITION of ORG`
George Marshall, Secretary of State of the United States

`PERSON (named|appointed|chosen|etc.) PERSON Prep? POSITION`
Truman appointed Marshall Secretary of State

`PERSON [be]? (named|appointed|etc.) Prep? ORG POSITION`
George Marshall was named US Secretary of State

Rules: Pros and Cons

Pros

- Generally high-precision
- Can be defined to capture domain-specific regularities and interpretations

Cons

- Generally low-recall
- Costly and time-consuming to develop by hand

Supervised classification of relations: Annotated Data

Given annotated training data, we can learn a relation classifier.

- Define the relations of interest.
- Define and annotate the entity types relevant to those relations.
- Annotate the relevant Named Entities (automatically or manually).
- Annotate the relations between the entities.

We can train a classifier with one of many algorithms (Naïve Bayes, Maximum Entropy, SVM . . .).

Supervised classification of relations: Approach

- Given a pair of NEs
- Classify the relation between them

(Usually sentence-bound; i.e., we only consider pairs of NEs within a sentence.)

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

Is there a relation between these two NEs? What is it?

Features for a RE classifier

American Airlines$_{M1}$, a unit of AMR, immediately matched the move, spokesman Tim Wagner$_{M2}$ said.

- Headwords of the NE mentions
  (Airlines, Wagner, combination: Airlines-Wagner)
- Bag of words, bigrams in the NEs
  ({American, Airlines, Tim, Wagner, American Airlines, Tim Wagner})
- Context: preceding/following words
  e.g., for M2: (spokesman, said)
- Bag of words between the NEs
  ({a, AMR, of, immediately, matched, move, spokesman, the, unit})
- NE type features
  M1: ORG, M2: PERSON, concatenation: ORG-PERSON
- Phrasal structure of NE
  e.g., Proper Noun, Pronoun *(he, she, it)*, NP *(the company)*

Richer Linguistic Features for RE classification

Features characterising the path between the Named Entities

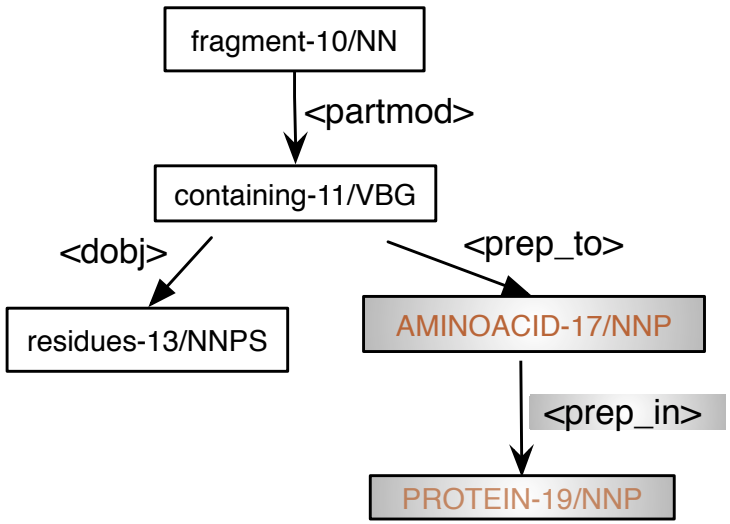
- Syntactic chunk sequence
  NP NP PP VP NP NP
- Path in syntax tree
  NP $\uparrow$ NP $\uparrow$ S $\uparrow$ S $\downarrow$ NP
- Dependency path (words)
  Airlines matched Wagner said
- Dependency graph kernel
  - shortest path vs. all paths
  - generalisation from words to POS tags to NE classes

Dependency graph patterns

Original Sentence: The present studies demonstrate the presence of a native fragment containing 14 residues from Ile16 to ***Trp29*** in ***alpha-chymotrypsin*** that binds ...

Tokenized & Entity Tagged Sentence: The present studies demonstrate the presence of a native fragment containing 14 residues from Ile16 to AMINOACID in PROTEIN that binds ...

Dependency Graph of Partial Sentence:



Protein-Residue Association Rule: (highlighted in the above graph)
Protein: PROTEIN-19/NNP, AMINOACID: AMINOACID-17/NNP <==
prep_in(PROTEIN-19/NNP, AMINOACID-17/NNP)

Supervised RE classification, pros and cons

Pros

- More robust than hand-developed rules
- Higher Recall

Cons

- Requires large amounts of annotated data
- Domain-dependent and genre-dependent

Bootstrapping for learning relations

- Given seed examples
- Given a few high-precision patterns
- Use seeds to populate a relation iteratively
- e.g., given IS-A(apple, fruit) find all ways *apple* and *fruit* are connected in sentences $\rightarrow$
- Infer new patterns
- Find new entities connected with those patterns
- repeat

Alternative approaches

- Semi-supervised: distant supervision
- Unsupervised: “open” information extraction, capture relations based on syntactic connections

Summary

- We introduced the NLP task of *information extraction* from text, including *named entity recognition* and *relation extraction*.
- Information Extraction is a key component of "structuring unstructured data" – identifying key entities and relations between them.
- We discussed some challenges of extracting named entities and relations between them.
- We saw that gazetteer-based methods, rule-based methods and learning-based methods have all been attempted for NER & Relation Extraction.
- We learned how to treat NER as a sequence labelling (tagging) task.
- We saw how to approach Relation Extraction as a classification task.
- We learned that constraints on relations can be harnessed both for higher precision, and for bootstrapping.

Further reading

- Chapter 21 (Information Extraction), 3rd Edition, Jurafsky & Martin, *Speech and Language Processing*
- NLTK Chapter 7
  `http://www.nltk.org/book/ch07.html`
- LingPipe Tutorial on NER
  `http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html`
- GATE JAPE pattern engine:
  `https://gate.ac.uk/sale/tao/splitch8.html`
- On kernels: Bunescu, R. C., & Mooney, R. J. (2005a). A shortest path dependency kernel for relation extraction. HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 724731). Vancouver, British Columbia, Canada: Association for Computational Linguistics.
  `www.cs.utexas.edu/~ml/papers/spk-emnlp-05.pdf`