Chapter VI: Information Extraction

Information Retrieval & Data Mining
Universität des Saarlandes, Saarbrücken
Wintersemester 2013/14
Chapter VI: Information Extraction

VI.1 Motivation & Applications
   Knowledge Queries, Entities & Relations, RDF(S), SPARQL

VI.2 Natural Language Processing Basics
   Part-of-Speech Tagging, Dependency Parsing, Word Sense Tagging

VI.3 Rule-Based Information Extraction
   Wrapper Induction

VI.4 Learning-Based Information Extraction
   Hidden Markov Models, Conditional Random Fields

VI.5 Named Entity Reconciliation
   Fellegi-Sunter Model

VI.6 Knowledge Base Construction & Open IE
   SNOWBALL, YAGO, TextRunner, NELL
VI.1 Motivation & Applications

• Beyond keywords as queries and documents as retrieval units
  • extract entities and annotate text documents or web pages (e.g., named entity recognition)
  • find instances of semantic classes (e.g., not yet known in WordNet)
  • extract facts (relations among entities) from text documents or web pages (e.g., Wikipedia) to automatically populate ontology/knowledge base
  • answer questions by analyzing natural language and translating it into machine-processable format

• Technologies:
  • Lexicon lookups (name dictionaries, geo gazetteers, etc.)
  • NLP (PoS tagging, chunking/parsing, semantic role labeling, etc.)
  • Pattern matching & rule learning (regular expressions, FSAs)
  • Statistical learning (HMMs, CRFs, etc.)
  • Text mining
Google Knowledge Graph

Albert Einstein - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Albert_Einstein
Albert Einstein (ˈaɪl伯t ˈaɪnstaɪn; German: [ˈalbert ˈʔaɪnʃtaɪn] (listen); 14 March 1879 – 18 April 1955) was a German-born theoretical physicist who... Hans Albert Einstein - Eduard Einstein - Mileva Maric - Elsa Einstein

Albert Einstein – Wikipedia
dw.wikipedia.org/wiki/Albert_Einstein ▶ Translate this page
Albert Einstein († 14. März 1879 in Ulm; † 18. April 1955 in Princeton, New Jersey) war ein theoretischer Physiker. Seine Forchungen zur Struktur von Materie, ...
Relativitätstheorie - Ulm - Zionismus - Thomas Harvey

Albert Einstein - Biographical - Nobelprie.org
www.nobelprie.org/nobel_prizes/physics/laurates/.../einstein-bio.html ▶
Albert Einstein - Biographical. Albert Einstein was born at Ulm, in Württemberg, Germany, on March 14, 1879. Six weeks later the family moved to Munich, where... News for albert einstein

Providence digital-content company chosen for online publishing aspects of Albert Einstein collection
The Providence Journal - 17 hours ago
PROVIDENCE – As Princeton University Press works to publish "The Collected Papers of Albert Einstein," it has selected a Providence ...

Maybe he’s a relative? Caterpillar bears incredible resemblance to Albert E... Daily Mail - 3 days ago
Congress proves Albert Einstein’s definition of insanity
San Jose Mercury News - 1 day ago

Einstein Archives Online
www.alberteinstein.info/ ▶
The homepage of the repository of the personal papers of the great scientist, humanist and Jew, Albert Einstein.

http://www.google.com
Freebase

http://www.freebase.com
DBpedia

http://www.dbpedia.org
The Linked Data Project

- As of 2011:
  - 295 sources
  - 32 billion RDF triples
  - 504 million links

http://www.linkeddata.org
Semantic Web

- **Semantic Web** [Berners-Lee ’01] is an extension of the World Wide Web to make its contents *interpretable for machines*

- **World Wide Web Consortium (W3C) Semantic Web standards**
  - **Unified Resource Identifier (URI)** to uniquely identify abstract or physical resources
  - **Resource Description Framework (RDF)** to describe properties of abstract or physical resources
  - **Resource Description Framework Schema (RDF/S)** to describe schemata
  - **Web Ontology Language (OWL)** to describe ontologies
  - **SPARQL Protocol and Query Language (SPARQL)** to formulate queries
Unified Resource Identifier

- **Unified Resource Identifier** (URI) is a string of characters that uniquely identifies an abstract or physical resource

  - `http://www.bbc.co.uk/music/artists/67f66c07-6e61-4026-ade5-7e782fad3a5d`
  - `http://www.musicbrainz.org/artist/67f66c07-6e61-4026-ade5-7e782fad3a5d`

- `http://www.host.org/pub/bands?query=FF#albums`

  - **scheme** (e.g., http, ftp, urn) determines interpretation of URI
  - **authority** indicates who is responsible for the resource (e.g., a host)
  - **path** provides hierarchical information for identifying the resource
  - **query** provides non-hierarchical information for identifying the resource
  - **fragment** refers to a specific part of the resource
• **Resource Description Framework** (RDF) provides a data model to describe properties of resources (identified by their URI)

• **RDF statements** are (S, P, O) triples consisting of a **subject** (URI), a **predicate** (a URI), and an **object** (a URI or literal)

• **Example**: Dave Grohl is a member of Foo Fighters

http://dbtune.org/musicbrainz/page/artist/67f66c07-6e61-4026-ade5-7e782fad3a5d (S)
http://xmlns.com/foaf/spec/20100809.html#member (P)
http://dbtune.org/musicbrainz/page/artist/4d5f891d-9bce-45ae-ad86-912dd27252fa (O)
RDF (cont’d)

- RDF triples form a **RDF graph**

- **Namespaces** represent *common URI prefixes* and allow for a more compact representation of RDF data

  ```
  @prefix a: http://allaboutmusic.org/
  a:Foo_Fighters a:member a:Dave_Grohl
  a:Foo_Fighters a:member a:Pat_Smear
  ```

- **RDF/N3** as one possible text representation of RDF data

  ```
  @prefix a: http://allaboutmusic.org
  a:Foo_Fighters a:member a:Dave_Grohl
  a:Foo_Fighters a:member a:Pat_Smear
  ```
SPARQL

- **SPARQL Protocol and Query Language** (SPARQL) is a **query language** for the Semantic Web standardized by the W3C

- SPARQL has a **SQL-inspired syntax** to define **graph patterns** and retrieves all matching subgraphs as query results

- **Example:**

  ```sql
  PREFIX a: <http://allmusic.org/>
  SELECT DISTINCT ?b, ?r, ?p WHERE {
    ?b a:hasMember ?p .
  }
  ORDER BY ?p
  ```
SPARQL

- SPARQL Protocol and Query Language (SPARQL) is a query language for the Semantic Web standardized by the W3C.

- SPARQL has a SQL-inspired syntax to define graph patterns and retrieves all matching subgraphs as query results.

- Example:

```sparql
PREFIX a: <http://allmusic.org/>

SELECT DISTINCT ?b, ?r, ?p WHERE {
  ?b a:hasMember ?p .
}
ORDER BY ?p
```

Query:

Result:

- Foo Fighers
  - hasMember: Nirvana
- Nate Mendel
  - livesIn: Seattle
- Kurt Cobain
  - diedIn: Seattle
- Seattle
Who was German chancellor in 1992?

Result:
Helmut Kohl

Basic information:
- official position: Chancellor
- country: Germany
- political affiliation: Christian Democratic Union
- start date: 01. October 1982 (31 years 2 months 2 days ago)
- end date: 27. October 1998 (15 years 1 month 17 days ago)
- duration of leadership: 16 years 26 days

Sequence:
- Tuesday, November 22, 2005 to Tuesday, December 3, 2013
  - Angela Merkel
  (Christian Democratic Union)
- Tuesday, October 27, 1998 to Tuesday, November 22, 2005 (7 years)
  - Gerhard Schröder
  (Social Democratic Party of Germany)
- Friday, October 1, 1982 to Tuesday, October 27, 1998 (16 years)
  - Helmut Kohl
  (Christian Democratic Union)
- Thursday, May 16, 1974 to Friday, October 1, 1982 (8 years 5 months)
  - Helmut Schmidt
  (Social Democratic Party of Germany)
- Thursday, May 7, 1974 to Thursday, May 16, 1974 (94 days)
  - Walter Scheel (acting)
  (Free Democratic Party)

Personal information:
- full name: Helmut Michael Kohl
- date of birth: 03. April 1930 (age: 83 years)
- place of birth: Ludwigsafen, Rhineland-Palatinate, Germany

Timeline:
- 1940
- 1969
- 1980
- 2000

How was the weather last year?

Recorded weather for Saarbrucken, Germany:
- time range: 2012
- temperature: -21°C to 37°C (average low: 4°C | average high: 15°C)
- relative humidity: average 75%
- wind speed: average: 3 m/s (maximum: 15 m/s)

http://www.wolframalpha.com
Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (née Patzig) Planck. Planck studied at the Universities of Munich and Berlin, where his teachers included Kirchhoff and Helmholtz, and received his doctorate of philosophy at Munich in 1879. He was Privatdozent in Munich from 1880 to 1885, then Associate Professor of Theoretical Physics at Kiel until 1889, in which year he succeeded Kirchhoff as Professor at Berlin University, where he remained until his retirement in 1926. Afterwards he became President of the Kaiser Wilhelm Society for the Promotion of Science, a post he held until 1937. He was also a gifted pianist and is said to have at one time considered music as a career. Planck was twice married. Upon his appointment, in 1885, to Associate Professor in his native town Kiel he married a friend of his childhood, Marie Merck, who died in 1909. He remarried her cousin Marga von Hösslin. Three of his children died young, leaving him with two sons.
Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (née Patzig) Planck. Planck studied at the Universities of Munich and Berlin, where his teachers included Kirchhoff and Helmholtz, and received his doctorate of philosophy at Munich in 1879. He was Privatdozent in Munich from 1880 to 1885, then Associate Professor of Theoretical Physics at Kiel until 1889, in which year he succeeded Kirchhoff as Professor at Berlin University, where he remained until his retirement in 1926. Afterwards he became President of the Kaiser Wilhelm Society for the Promotion of Science, a post he held until 1937. He was also a gifted pianist and is said to have at one time considered music as a career.

Planck was twice married. Upon his appointment, in 1885, to Associate Professor in his native town Kiel he married a friend of his childhood, Marie Merck, who died in 1909. He remarried her cousin Marga von Hösslin. Three of his children died young, leaving him with two sons.
Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (née Patzig) Planck. Planck studied at the Universities of Munich and Berlin, where his teachers included Kirchhoff and Helmholtz, and received his doctorate of philosophy at Munich in 1879. He was Privatdozent in Munich from 1880 to 1885, then Associate Professor of Theoretical Physics at Kiel until 1889, in which year he succeeded Kirchhoff as Professor at Berlin University, where he remained until his retirement in 1926. Afterwards he became President of the Kaiser Wilhelm Society for the Promotion of Science, a post he held until 1937. He was also a gifted pianist and is said to have at one time considered music as a career. Planck was twice married. Upon his appointment, in 1885, to Associate Professor in his native town Kiel he married a friend of his childhood, Marie Merck, who died in 1909. He remarried her cousin Marga von Hösslin. Three of his children died young, leaving him with two sons.
Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (née Patzig) Planck. Planck studied at the Universities of Munich and Berlin, where his teachers included Kirchhoff and Helmholtz, and received his doctorate of philosophy at Munich in 1879. He was Privatdozent in Munich from 1880 to 1885, then Associate Professor of Theoretical Physics at Kiel until 1889, in which year he succeeded Kirchhoff as Professor at Berlin University, where he remained until his retirement in 1926. Afterwards he became President of the Kaiser Wilhelm Society for the Promotion of Science, a post he held until 1937. He was also a gifted pianist and is said to have at one time considered music as a career.

Planck was twice married. Upon his appointment, in 1885, to Associate Professor in his native town Kiel he married a friend of his childhood, Marie Merck, who died in 1909. He remarried her cousin Marga von Hösslin. Three of his children died young, leaving him with two sons.
Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (née Patzig) Planck. Planck studied at the Universities of Munich and Berlin, where his teachers included Kirchhoff and Helmholtz, and received his doctorate of philosophy at Munich in 1879. He was Privatdozent in Munich from 1880 to 1885, then Associate Professor of Theoretical Physics at Kiel until 1889, in which year he succeeded Kirchhoff as Professor at Berlin University, where he remained until his retirement in 1926. Afterwards he became President of the Kaiser Wilhelm Society for the Promotion of Science, a post he held until 1937. He was also a gifted pianist and is said to have at one time considered music as a career.

Planck was twice married. Upon his appointment, in 1885, to Associate Professor in his native town Kiel
he married a friend of his childhood, Marie Merck, who died in 1909. He remarried her cousin Marga von Hösslin.

Three of his children died young, leaving him with two sons.
IE for Knowledge Base Construction

{{Infobox_Scientist
| name = Max Planck
| birth_date = [[April 23]], [[1858]]
| birth_place = [[Kiel]], [[Germany]]
| death_date = [[October 4]], [[1947]]
| death_place = [[Göttingen]], [[Germany]]
| residence = [[Germany]]
| nationality = [[Germany|German]]
| field = [[Physicist]]
| work_institution = [[University of Kiel]]</br>[[Humboldt-Universität zu Berlin]]</br>[[Georg-August-Universität Göttingen]]
| alma_mater = [[Ludwig-Maximilians-Universität München]]
| doctoral_advisor = [[Philipp von Jolly]]
| doctoral_students = [[Gustav Ludwig Hertz]]
| known_for = [[Planck's constant]],
[[Quantum mechanics|quantum theory]]
| prizes = [[Nobel Prize in Physics]] (1918)

Categories: 1858 births | 1947 deaths | German Nobel laureates | German physicists | Members of the Pontifical Academy of Sciences | Members of the Prussian Academy of Sciences | Nobel laureates in Physics | Recipients of the Copley Medal | People from Kiel | People from the Province of Schleswig-Holstein | Quantum physicists | Recipients of the Pour le Mérite (civil class) | Theoretical physicists | Thermodynamicists | University of Munich alumni | University of Munich faculty | Humboldt University of Berlin alumni | Humboldt University of Berlin faculty | University of Kiel faculty | Columbia University faculty | German Christians | Religion and science | Fellows of the Leopoldina

Max Planck

Born
April 23, 1858
Kiel, Holstein
Died
October 4, 1947 (aged 89)
Göttingen, West Germany
Nationality
German
Fields
Physica
Institutions
University of Kiel
University of Berlin
University of Göttingen
Kaiser-Wilhelm-Gesellschaft
Alma mater
Ludwig-Maximilians-Universität München
Doctoral advisor
Alexander von Brill
Doctoral students
Gustav Ludwig Hertz
Erich Krebschmann
Walther Meßinger
Walther Schottky
Max von Laue
Max Abraham
Moritz Schlick
Walther Bothe
Julius Edgar Lilienfeld
Known for
Planck's constant
Planck postulate
Planck's law of black body radiation
Notable awards
Nobel Prize in Physics (1918)
Religious stance
Protestant[1]
Notes
He is the father of Erwin Planck who was hanged in 1945 by the Gestapo for his part in the July 20 plot.
NLP-Based IE on the Web

Che Guevara

From Wikipedia, the free encyclopedia.

Ernesto Rafael Guevara de la Serna (June 14, 1928 [1] October 9, 1967), commonly known as Che Guevara or el Che, was an Argentine-born Marxist revolutionary and Cuban guerrilla leader. Guevara was a member of Fidel Castro's 26th of July Movement that seized power in Cuba in 1959. After serving in various important posts in the new government, Guevara left Cuba in 1965 with the hope of fomenting revolutions in other countries, first in the Congo-Kinshasa (currently the Democratic Republic of the Congo) and later in Bolivia, where he was captured in a CIA-organized military operation. It is believed by some that the CIA wished to keep Guevara alive for interrogation but, after his capture in the Yuro ravine, he died at the hands of the Bolivian Army in La Higuera near Vallegrande on October 9, 1967. Testimony by various individuals who were participants in, or
NLP-Based IE on the Web

Ernesto "Che" Guevara (Spanish pronunciation: [eˈɾes'to ñeˈβara]; 15 June 1928 – 9 October 1967), commonly known as Che or simply Che, was an Argentine Marxist revolutionary, physician, author, guerrilla leader, diplomat, and military theorist. A major figure of the Cuban Revolution, his stylized visage has become a ubiquitous countercultural symbol of rebellion and global iconography within popular culture.

As a young medical student, Guevara traveled throughout South America and was radicalized by the poverty, hunger, and disease he witnessed. His burgeoning desire to help overthrow what he saw as the capitalist exploitation of Latin America by the United States prompted his involvement in Guatemala's social reforms under President Jacobo Árbenz, whose eventual CIA-assisted overthrow at the behest of the United Fruit Company solidified Guevara's political ideology.

[9] In 1959, in Mexico City, he met Raúl and Fidel Castro, and later to Cuba aboard the yacht Granma, with the intention of overthrowing US-backed Cuban dictator Fulgencio Batista. Guevara soon rose to prominence among the insurgents, was promoted to second-in-command, and played a pivotal role in the victorious two-year guerrilla campaign that deposed the Batista regime. Following the Cuban Revolution, Guevara performed a number of key roles in the new government. These included reviewing the appeals and firing squads for those convicted as war criminals during the revolutionary tribunal.

[10] Instituting agrarian land reform as minister of industries, helping spearhead a successful nationwide literacy campaign, serving as both national bank president and instructional director for Cuba's armed forces, and traversing the globe as a diplomat on behalf of Cuban socialism. Such positions also allowed him to play a central role in training the military forces that repelled the Bay of Pigs Invasion and bringing the Soviet nuclear-armed ballistic missiles to Cuba which precipitated the 1962 Cuban Missile Crisis.

[11] Additionally, he was a prolific writer and artist, composing a seminal manual on guerrilla warfare, along with a best-selling memoir about his youthful continental motorcycle journey. His experiences and studying of Marxism-Leninism led him to posit that the Third World's underdevelopment and dependency was an intrinsic result of imperialism, neocolonialism, and monopoly capitalism, with the only remedy being proletarian internationalism and world revolution. Guevara left Cuba in 1965 to foment revolution abroad, first unsuccessfully in Congo Kinshasa and later in Bolivia, where he was captured by CIA-assisted Bolivian forces and summarily executed.

Guevara remains both a revered and reviled historical figure, polarized in the collective imagination in a multitude of biographies, memoirs, essays, documentaries, songs, and films. As a result of his well-documented martyrdom, poetic invocations for class struggle, and desire to create the consciousness of a "new man" driven by moral rather than material incentives, he has evolved into a quintessential icon of various leftist-inspired movements. Time magazine named him one of the 100 most influential people of the 19th century.

While an Alberto Korda photograph of him entitled Guerrillero Heroico (shown), was cited by the Maryland Institute College of Art as "the most famous photograph in the world."
Extracting Structured Records from the Deep Web
Extracting Structured Records from the Deep Web

Mining the Web: Analysis of Hypertext and Semi-Structured Data (The Morgan Kaufmann Series in Data Management Systems) (Hardcover)

by Soumen Chakrabarti

List Price: $62.82

Price: $62.82

& this item ships for FREE with Super Saver Shipping.
Extracting Structured Records from the Deep Web

Extracted Record:
Title: Mining the Web
Author: Soumen Chakrabarti
Hardcover: 344 pages
Publisher: Morgan Kaufmann
Language: English
Jeopardy!

A big U.S. city with two airports, one named after a World War II hero, and one named after a World War II battle field.
Jeopardy!

A big U.S. city with two airports, one named after a World War II hero, and one named after a World War II battle field.
Jeopardy!

A big U.S. city with two airports, one named after a World War II hero, and one named after a World War II battlefield.

History [edit]

World War II [edit]

See also: Illinois World War II Army Airfields

The airport was constructed in 1942–43 as a manufacturing plant for Douglas C-54s during World War II.[12] The site was chosen for its proximity to the city and transportation.[12] The two million square foot (180,000 m²) factory needed easy access to the workforce of the nation's then-second-largest city, as well as its extensive railroad infrastructure. Orchard Place was a small nearby farming community.[12]

Douglas Company's contract ended in 1945 and though plans were proposed to build commercial aircraft, the company ultimately chose to concentrate production on the west coast. With the departure of Douglas, the airport took the name Orchard Field Airport, the source of its three-letter IATA code ORD.

In 1945, the facility was chosen by the city of Chicago as the site for a facility to meet future aviation demands. Matthew Laflin Rockwell (1915–1988) was the director of planning for the U.S. Army Corps of Engineers and responsible for the site selection and design of O'Hare International Airport. He was the great grandson of Matthew Laflin, a founder and pioneer of Chicago.

In 1949, the airport was renamed "O'Hare International Airport" to honor Edward O'Hare, the U.S. Navy's first flying ace and Medal of Honor recipient in World War II. Its IATA code, "ORD", remained unchanged, however, resulting in the infrequent case of an airport's three-letter designation bearing no connection to the airport name or metropolitan area.
Jeopardy!

A big U.S. city with two airports, one named after a World War II hero, and one named after a World War II battle field.
Structured Knowledge Queries

A big U.S. city with two airports, one named after a World War II hero, and one named after a World War II battlefield.

```
SELECT DISTINCT ?c WHERE {
  ?c type City . ?c locatedIn USA .
  ?a1 type Airport . ?a2 type Airport .
  ?a1 locatedIn ?c . ?a2 locatedIn ?c .
}
```

• Use manually curated templates for mapping sentence patterns to structured queries

• Focus on factoid and list questions
William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia" inspired this author's most famous novel.

This town is known as "Sin City" & its downtown is "Glitter Gulch".

As of 2010, this is the only former Yugoslav republic in the EU.

99 cents got me a 4-pack of Ytterlig coasters from this Swedish chain.

- **Full details:** [Ferrucci et al. ’10] [Ferrucci et al. ’12]
More IE Applications

• **Comparison shopping & recommendation portals**
  (e.g., consumer electronics, used cars, real estate, pharmacy, etc.)

• **Business analytics** on customer dossiers, financial reports, etc.
  (e.g., how was company X performing in the last 5 years?)

• Market/customer, PR impact, and **media coverage analysis**
  (e.g., how are our products perceived by teenagers?)

• **Job brokering** (applications/resumes, job offers)
  (e.g., how well does the candidate match our desired profile?)

• **Knowledge management** in consulting companies
  (e.g., do we have experience on retail in Brasil?)

• **Knowledge extraction** from scientific literature
  (e.g., which HIV drugs have been found ineffective recently?)
IE Viewpoints and Approaches

- IE as learning (restricted) **wrappers/regular expressions** (wrapping pages with common structure from Deep Web)

- IE as learning **relations** (rules for identifying instances of $n$-ary relation)

- IE as learning **fact boundaries**

- IE as learning **text segmentation** (HMMs, etc.)

- IE as learning **contextual patterns**

- IE as **natural-language analysis** (NLP methods)

- IE as **large-scale text mining** for knowledge acquisition (combinations of tools including web queries)
IE Viewpoints and Approaches

**Lexicons**

Abraham Lincoln was born in Kentucky.

• **Source:** [Cohen ’03]

**Classify Pre-Segmented Candidates**

Abraham Lincoln was born in Kentucky.

**Sliding Window (+ Classifier)**

Abraham Lincoln was born in Kentucky.

Try alternate window sizes:

**Boundary Models**

Abraham Lincoln was born in Kentucky.

**Finite State Machines**

Abraham Lincoln was born in Kentucky.

Most likely state sequence?

**Context Free Grammars**

Abraham Lincoln was born in Kentucky.

Most likely parse?
IE Quality Assessment

• Fix IE task (e.g., extract all book records from bookseller website)

• Manually extract all correct records

• Use standard IR effectiveness measures
  • precision, (relative) recall, F1 measure, etc.
  • statistical tests w/ confidence intervals for precision, recall, etc.
    based on a sample of manually inspected records

• Benchmark settings:
  • MUC (Message Understanding Conference), discontinued
  • ACE (Automatic Content Extraction) (http://www.nist.gov/speech/tests/ace/)
  • TAC (Text Analysis Conference) (http://www.nist.gov/tac/)
  • …
Additional Literature for VI.1


- **H. Cunningham**: *Information Extraction*, Encyclopedia of Language and Linguistics, 2005


- **D. Ferrucci**: *Introduction to “This is Watson”*, IBM Journal of Research and Development, 56(3):1, 2012

- **S. Sarawagi**: *Information Extraction*, Foundations & Trends in Databases 1(3), 2008
VI.2 Natural Language Processing Basics

• **Tokenization** of input documents into
  • **meaningful input units** (e.g., NL sentences, tables, lists, etc.)
  • **input tokens** (e.g., words, phrases, semantic sequences)
  • **token features** (e.g., position in document, capitalization, length, etc.)

• **Linguistic preprocessing** of input documents
  • **part-of-speech tagging** maps words to their grammatical role
  • **chunk parsing** maps a sentence to labeled segments
  • **dependency parsing** identifies logically connected segments

• Both are **important preprocessing steps** for many IE tasks
Part-of-Speech Tagging

- **Part-of-Speech (PoS) tagging** maps each word (group) to its grammatical role (e.g., noun, verb, adjective, determiner, etc.)

- Often uses **Hidden Markov Models** trained on large corpora

- **PoS Tags (Penn Treebank):**
  - CD: cardinal number
  - DT: determiner
  - EX: existential *there*
  - JJ: adjective
  - NN: noun
  - POS: possessive ending
  - PRP: personal pronoun
  - RB: adverb
  - VB: verb, base form
  - WDT: *wh*-determiner (*which*, …)
  - WP: *wh*-pronoun (*who*, *whom*, …)

- Example: *The/DT bright/JJ student/NN who/WP works/VBZ hard/RB will/MĐ pass/VB all/DT exams/NNS*

[Link to Penn Treebank tags:](http://www.lsi.upc.edu/~nlp/SVMTool/PennTreebank.html)
Word Sense Tagging/Disambiguation

- **Idea**: Tag each word with its word sense (meaning, concept) by mapping to a thesaurus/ontology/lexicon such as WordNet

- **Typical approach**:
  - Form context $con(w)$ of word $w$ in sentence (or passage)
  - Form context $con(s)$ of candidate sense (e.g., using the corresponding WordNet synset, gloss, neighboring concepts, etc.)
  - Assign $w$ to $s$ with highest similarity between $con(s)$ and $con(w)$ or highest likelihood of $con(s)$ generating $con(w)$
  - Incorporate prior, i.e., relative frequency of senses for same word
  - **Joint disambiguation**: map multiple words to their most likely meaning (taking into account semantic coherence, compactness)

- **Benchmark initiative**: http://www.senseval.org
Deep Parsing for Constituent Trees

• Construct syntax-based parse tree of sentence constituents

  • **Non-deterministic context-free grammars** (natural ambiguity)
  • **Probabilistic context-free grammars** (likely vs. unlikely parse trees)

• Extensions and variations:

  • **lexical parser**: enhanced with lexical dependencies (e.g., only specific verbs can be followed by two noun phrases)
  • **chunk parser**: simplified to detect only phrase boundaries
Dependency Parsing

• Reveal dependencies between logically connected segments

(ROOT
  (S
    (NP (DT The) (JJ bright) (NN student))
    (SBAR
      (WHNP (WP who))
      (S
        (VP (VBZ works)
          (ADVP (RB hard ))))))
  (VP (MD will)
    (VP (VB pass)
      (NP (DT all) (NNS exams))))))

Typed dependencies:

  det(student-3, The-1)
  amod(student-3, bright-2)
  nsubj(passes-7, student-3)
  nsubj(works-5, who-4)
  rcmode(student-3, works-5)
  advmod(works-5, hard-6)
  root(ROOT-0, passes-7)
  det(exams-9, all-8)
  dobj(passes-7, exams-9)

• Stanford Dependencies:

  nsubj nominal subject
  rel relative
  dobj direct object
  det determiner
  amod adjectival modifier
  rcmode relative clause modifier
  acomp adjectival complement
  advmod adverbial modifier

Named Entity Recognition (NER)

- Identify mentions of **named entities** (e.g., persons, locations, organizations, dates, etc.)
- Runs text through part-of-speech tagging or probabilistic parsing
- Uses dictionaries to validate/falsify candidate entities
- Does not disambiguate candidate entities

- **Example:** *Bayern Munich with their captain Philipp Lahm lost the final in Munich on May 19 2012*
Coreference Resolution (Anaphor Resolution)

• Connect **pronouns** etc. to **subject/object** of previous sentence.

• **Example**: *Diego Maradona was soccer player of the year. He is also known as the hand of god.*
Semantic Role Labeling (SRL)

• Identify **semantic types** of events or \( n \)-ary relations based on taxonomy (e.g., FrameNet, VerbNet, PropBank)

• Fill **components** of \( n \)-ary tuples (semantic roles, slots of frames)

• **Example:** *Thompson is understood to be accused of importing heroin into the United States*

    <event>
    <type> drug-smuggling </type>
    <destination> <country>United States</country> </destination>
    <source> unknown </source>
    <perpetrator> <person> Thompson </person> </perpetrator>
    <drug> heroin </drug>
    </event>
FrameNet Representation for SRL

Smuggling

Definition:

The words in this frame describe situations in which the Perpetrator secretly takes Goods into or out of a country or other area which are prohibited by law or on which one has not paid the required duty.

FEs:

Core:

Source (Goal) 
Goal is the location the Goods end up in.

Goal
SemType
Goal

Source (Goods) 
The FE Goods is anything (including labor, time, or legal rights) that can be a country.

Goods (Goal)

Path (Goal)
The path refers to (a part of) the ground the Goods travel over or to a country.

Path (Event)

Perpetrator (Event)
The person (or other agent) that illegally takes the Goods into or out of a country.

Perpetrator (Event)

Source (Source) 
The Source is the location the Goods occupy initially before change of location.

Source (Event)

SemType

Source

Non-Core:

Duration (Core)
The amount of time for which a state holds or a process is ongoing.

Event (Core)
The unlawful movement of Goods.

Duration
Event

Frequency (Event)
The number of times that a smuggling event occurs.

Frequency

Initiator (Event)
A description of the Perpetrator not covered by more specific FE(s), including secondary effects (such as, etc.), and general descriptions comparing events (the same way). In most cases, it indicates relevant characteristics of a Perpetrator that also affect the action (presumptively, willfully, deliberately, expertly, rashly).

Initiator

Initiates (Drawable) portable marijuana into the prison

Associated (Event)

Object (Event)

Property (Event)
The action that the Perpetrator is trying to accomplish by the act of smuggling.

Property

We smuggle you in here and sell it but...

Time (Event)
The reason for which an event occurs.

Time

Where the event takes place.

Related (Event)


Source: http://framenet.icsi.berkeley.edu/
PropNet Representation for SRL

• Large collection of annotated newspaper articles; roles are simpler (more generic) than FrameNet

Arg0, Arg1, Arg2, … and ArgM with modifiers

- **LOC**: location
- **ADV**: general purpose
- **MOD**: modal verb
- **TMP**: time
- **MNR**: manner
- **EXT**: extent
- **NEG**: negation marker
- **CAU**: cause
- **PNC**: purpose
- **DIR**: direction

• **Example**: Revenue edged up 3.4% to $904 million from $874 million in last year’s third quarter

[Arg0: Revenue] *increased* [Arg2-EXT: by 3.4%] [Arg4: to $904 million] [Arg3: from $874 million] [ArgM-TMP: in last year’s third quarter]

• **Source**: [http://verbs.colorado.edu/~mpalmer/projects/ace.html](http://verbs.colorado.edu/~mpalmer/projects/ace.html)
Stanford CoreNLP

• **Stanford CoreNLP Tools**
  
  • implemented in Java
  
  • wrappers for Python, Ruby, Perl, etc.
  
  • part-of-speech tagging
  
  • dependency parsing
  
  • coreference resolution
  
  • named entity recognition
  
  • sentiment analysis
  
  • models for English, Arabic, Chinese, French, German


http://nlp.stanford.edu:8080/corenlp/
NLTK

- Natural Language Toolkit
  - implemented in Python
  - part-of-speech tagging
  - dependency parsing
  - named entity recognition
  - sentiment analysis
  - models for English, Chinese, and Spanish

- Link: http://nltk.org

http://text-processing.com/demo/tag/
Additional Literature for VI.2

- **C. Manning and H. Schütze**: *Foundations of Statistical Natural Language Processing*, MIT Press, 2000