

COMP348 — Document Processing and the Semantic Web

Week 10 Lecture 2: Information Extraction

Diego Mollá

COMP348 2008H1

Abstract

We will address the issues related with information extraction, especially from the aspect of rule-based approaches.

Update May 14, 2008

Contents

1 Information Extraction	1
2 Generic Architecture	4

Some Useful Extra Reading

- D. Appelt and D. Israel. Introduction to Information Extraction Technology (1999) <http://www.ai.sri.com/~appelt/ie-tutorial>
- Hobbs et al. FASTUS (1997) <http://www.ai.sri.com/natural-language/projects/fastus-schabes.html>
- R. Grisham, Information Extraction, In: R. Mitkov (ed), *The Oxford Handbook of Computational Linguistics*, pp. 545–559, 2003.

1 Information Extraction

The Motivation for Information Extraction

Observations

- Most of the information is contained in text in human languages and not in databases or similar structured formats.
 - Estimate: about 85%
- Most of new information is now stored in digital form.
 - Estimate: about 92%

Conclusion

You're missing out on a lot of good stuff if you can't get answers from all that digital information written in human languages.

What Information Extraction is About

The Problem

Extract well-defined pieces of information (for example named entities or events) from collections of documents

The Goal

To populate a template or database

Particularities

- Typically, most of the information in a document is ignored.
- IE can be contrasted with earlier goals of building story understanding systems, where broad and deep coverage is needed.

An Example Document

From MUC-4

San Salvador, 19 Apr 89 (ACAN-EFE) – [TEXT] Salvadoran President-elect Alfredo Cristiani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of the crime.

...

Garcia Alvarado, 56, was killed when a bomb placed by urban guerrillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador.

...

Vice President-elect Francisco Merino said that when the attorney general's car stopped at a light on a street in downtown San Salvador, an individual placed a bomb on the roof of the armored vehicle.

...

According to the police and Garcia Alvarado's driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.

A Corresponding Filled Template

Incident: Date	19 Apr 89
Incident: Location	El Salvador: San Salvador (CITY)
Incident: Type	Bombing
Perpetrator: Individual ID	urban guerrillas
Perpetrator: Organization ID	FMLN
Perpetrator: Confidence	Suspected or Accused by Authorities: FMLN
Physical Target: Description	vehicle
Physical Target: Effect	Some Damage: vehicle
Human Target: Name	Roberto Garcia Alvarado
Human Target: Description	attorney general: Roberto Garcia Alvarado driver bodyguards
Human Target: Effect	Death: Roberto Garcia Alvarado No Injury: driver Injury: bodyguards

Information Extraction vs NL Understanding

- We don't care about subtleties in author's intentions
- We don't need to be able to answer general questions about the text
- We do need to be able to extract specific predetermined information from the text
- We can settle for a less expressive representation of the 'meaning' of the document (i.e., templates)

Target Applications

- Converting unstructured texts to databases
- Providing input to summarization systems
- Creating indexes for IR systems

The Message Understanding Conferences

MUC

- A set of tasks related with information extraction
- Competition-based focusing on measurable evaluations
- Fostered research in IE and related tasks

Timeline

- MUC-1 (1987); MUC-2 (1989)
 - Naval operations messages
- MUC-3 (1991); MUC-4 (1992)
 - Terrorism in latin american countries
- MUC-5 (1993)
 - Joint ventures and microelectronics domain
- MUC-6 (1995)
 - News articles on management changes
- MUC-7 (1998)
 - Satellite launch reports

Benefits of MUCs

- A common task for everyone
- Development of large corpora with associated "key templates"
- Set format for the templates, methods of automatically scoring program output to hand-created key templates
- Methods for evaluating system performance

Evaluation

Precision Correct Slots/Slots Filled

Recall Correct Slots/Total Possible Correct Slots

F Measure A Weighted combination of Precision and Recall

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- For complex IE tasks, people agree on slot fills in the 60-80% range
- Current state-of-the-art:
 - the F = 0.60 barrier
 - systems achieve about 60% of human performance

2 Generic Architecture

The General Approach

1. Locate sentences or fragments that contain relevant information
2. Ignore information that is not relevant
3. Extract the information
4. Output information in a predetermined form

A System Architecture [Appelt and Israel]

1. Tokenisation
2. Morphological and Lexical Processing
3. Syntactic Analysis
4. Domain Analysis

A More Detailed System Architecture [Hobbs]

1. Text Zoning
2. Preprocessing
3. Filtering
4. Preparsing
5. Parsing
6. Fragment Combination
7. Semantic Interpretation
8. Lexical Disambiguation
9. Coreference Resolution
10. Template Generation

Text Zoning

- Turns a text into a set of useful text segments
 - Headers, paragraphs, clusters of paragraphs, tables, and other useful parts
 - May be ‘topic’-based, using cue words or statistics
 - Depends on the structures of the texts in the domain of application.
- Discards unwanted segments of the text
 - Mail headers, signature blocks . . .
- Can make use of explicit logically-oriented markup
 - HTML, XML, SGML
 - Word’s Rich Text Format (maybe), LaTeX
- In the absence of a logically-oriented markup, use typographic information
 - Centered blocks
 - Paragraph breaks
- Low level markup (PostScript, PDF) not so useful
- Subsequent processing may focus on specific zones
 - Searches for information about the date of a message can be restricted to the mailer headers
 - Information in tables can be handled by special purpose code
- Text-zoning code is usually very specific to the kinds of text being handled

Preprocessing

- Carries out tokenisation and sentence segmentation
- Attributes may be added to the words (e.g. PoS)
- Named entities may be identified
- Techniques are required for unknown words
- Spelling correction also carried out at this point
- Techniques used:
 - Lexical lookup, perhaps in conjunction with morphological analysis
 - Part of speech tagging
 - Finite-state pattern-matching for recognizing and normalizing basic entities
 - Standard spelling correction techniques
 - A variety of heuristics for handling unknown words

Filtering

- Throws away sentences considered to be irrelevant
 - No point in expending machine cycles on sentences which are not important to the task
- Speed vs accuracy trade-off: cheaper (ie faster) heuristics for determining relevance may make more mistakes.
- Relevance may be zone-dependent

Preparsing

- Observation:
 - in going from a sequence of words to a parse tree, some structures can be identified more reliably than others.
- Examples:
 - noun groups: “*the six dead terrorists* in the vehicle were . . .”
 - appositives: “*John Bull, the forty-year old CEO*, said . . .”
 - some prepositional phrases: “*the CEO of the company* said”.
- Typically uses finite state grammars and special word lists.

Parsing

- Takes as input a sequence of lexical items and small-scale structures built by the parser
- Produces as output a set of parse tree fragments, corresponding to subsentential units
 - There is no need to find a full parse
- Many parsing techniques are available:
 - *Chart parsing* is a popular choice

Fragment Combination

- Takes as input a set of parse tree fragments derived from a sentence
- Tries to combine the fragments into a representation for the entire sentence
- Generally based on heuristics:
 - Overcomes the problems of not having a rich enough syntactic analysis for the entire sentence
 - Domain-based heuristics much faster, especially for the long sentences found in real text

Semantic Interpretation

- Generates a *semantic structure* or *logical form* or *event frame* from a parse tree or a collection of parse tree fragments.
 - Participants in a sentence
 - Explicit relationships between the participants (agent, patient, location, instrument, etc)
- Traditionally done with heuristics, currently there is a move for using statistical approaches
 - Semantic Role Labelling

FrameNet

- <http://framenet.icsi.berkeley.edu/>
- [*Cook* Matilde] fried [*Food* the catfish] [*Heating-instrument* in a heavy iron skillet].
- [*Item* Colgate's stock] rose [*Difference* \$3.64] [*Final-value* to \$49.94].

Lexical Disambiguation

- Turns a semantic structure with general or ambiguous predicates into a semantic structure with specific, unambiguous predicates
- This task may be carried out in a number of different places in a system
- In restricted domains this may not be an issue — the ‘one sense per document’ assumption

Coreference Resolution

- Identifies different descriptions of the same entity in different parts of the text and relates them in some way
- A range of anaphoric relationships may need to be dealt with:
 - Identity** (different ways of referring to the same thing):
 - “*Bill Gates ... he ... Microsoft's founder ...*”
 - Meronymy** (part-of relationships between entities)
 - “*A new program ... the documentation is weak ...*”
 - Event Reference** – “*the murder of the civilians was a new development ...*”.
- Techniques:
 - Number and gender agreement for pronouns:
 - * “Bill Gates met with Esther Dyson ... *she* later stated ...”
 - Semantic consistency based on taxonomic information:
 - * “Toyota Motor Corp ... the Japanese automaker”
 - Some notion of ‘focus’:
 - * Pronouns typically refer to something mentioned in the previous sentence

Template Generation

- Derives final output templates from the semantic structures
- Carries out low-level formatting and normalisation of data
- Usually implemented via domain-dependent heuristics

Take-home Messages

Essential Skills — for a passing grade

- Explain the goal of IE
- Compare IE with IR
- Describe the general approach to IE

Additional Skills — for credit or above

- Describe each IE stage: inputs and outputs, and issues

What's Next

Week 10

- Lecture 3: Named Entity Recognition