Information Extraction

Text mining [7]
5 april 2005

Antal van den Bosch en
Walter Daelemans

http://ilk.uvt.nl/~antalb/textmining/

Partial Assignment I

- Better results when test text is of same type as type on which the parser is trained (generally)
  - Input news: WSJ > Biomint
  - Input bio: WSJ <= Biomint
- Some misconceptions
  - Shallow parser assigns base constituents (no nesting)!
  - Chunker and relation finder are the same in both parsers!

Partial Assignment I

- Some observations
  - Machine Learning based shallow parsers can be inconsistent
    - Mr. Bush (proper) vs. Mr. Fischer (common)
  - Limited percolation of errors from lower to higher levels
  - Test texts (20 sentences) in general too small for reliable results
- Assignment done well to very well by all!

What is Information Extraction?

- Input: unstructured text
- Output: structured information, fills pre-existing template (find salient information)
- Most often stored in database for futher processing (e.g. data mining)

What is information extraction NOT?

- Information retrieval (we need to extract info, not only find relevant documents)
- Text understanding (only specific parts of the text are interesting)
  - large corpora can be used
  - possible to score objectively

Applications of IE

- Can make information retrieval more precise
- Summarization of documents in well-defined subject areas
  - Even multilingual!
- Automatic generation of databases from text

Why is IE difficult?

- Many different ways of expressing the same information
  - BNC Holdings Inc named Ms G Torretta as its new chairman.
  - Nicholas Andrews was succeeded by Gina Torretta as chairman of BNC Holdings Inc.
  - Ms. Gina Torretta took the helm at BNC Holdings Inc.

Why is IE difficult?

- Information can be spread out over many sentences
  - After a long boardroom struggle, Mr Andrews stepped down as chairman of BNC Holdings Inc. He was succeeded by Ms Torretta.

Why is IE difficult?

- Language
  - E.g. compounds in Dutch and German
- Genre
  - Newspaper text vs. speech transcripts
- Text
  - Length
  - Non-text
- Task
  - Identifying entities
  - Finding attributes of entities
  - Finding relations between entities

Example

(From SAIC website on information extraction)

Example:

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

First mark entities…

Persons:

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Organisations:

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Locations:

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Artifacts:

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Dates:

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Templates derived:

http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

Events:

COMPANY-FORMATION_EVENT:

COMPANY:	La Jolla Genomatics
PRINCIPALS:	Fletcher Maddox Oliver Ambrose
DATE:	
CAPITAL:	

RELEASE-EVENT:

COMPANY:	La Jolla Genomatics
PRODUCT:	Geninfo
DATE:	June 1999
COST:	

MUCs

- Message Understanding Conference
- Started by DARPA (Defense Advance Research Project Agency)
- Information Extraction task
- Provided template for output
- Training data provided
- Evaluation by comparison with human performance
- Evaluation: recall, precision, F-score

Topics of MUC tasks

- MUC-1 (’87) and MUC-2 (’89):
  - Naval operations
- MUC-3 (’91) and MUC-4 (’92):
  - Terrorism
- MUC-5 (’95):
  - Joint ventures and electronics
- MUC-6 (’95):
  - Management changes
- MUC-7 (’97):
  - Space vehicle and missile launch

Other ‘good’ topics

- Patient records
- Medical information (e.g. genome project)
- Job adverts / CVs
- Entertainment
- Using WWW for own purposes

Evaluation

Evaluation is difficult

- Partial answers
- Humans don’t do too well: inter-annotator agreement of around 60 to 85 percent (Appelt and Israel)

Recall

- Recall = number of correct answers + 0.5 * number of partial answers / number of possible answers

Precision

- Precision = number of correct answers + 0.5 * number of partial answers / number of given answers

Two approaches

- Hand-crafting (aka knowledge engineering)
  - Rules are made by hand
  - Human experts are needed
- Machine learning
  - Rules are automatically derived from corpora
  - Annotated corpora are needed

Hand crafting

- Pro:
  - Tends to make good systems
  - Is not too hard
- Con:
  - Can take a lot of time and effort (money)
  - Makes changes to a system hard to accommodate
  - Experts needed

Machine learning

- Pro:
  - Portable to different domain
  - No expert needed (apart from annotation)
  - Realistic training corpus → real cases covered
- Con:
  - Data expensive
  - Retraining can take long, for minor changes

Which is best?

- Depends on task:
  - Stable / bound to change
  - Data available?
  - Experts cheap?
  - Easy problem?
- Possible to mix approaches! – Components can each have different architecture

System architecture

- Text analysis
  - Lexical analysis
  - Named entity recognition
  - Parsing
- Extraction
  - Pattern matching
  - Merging of facts
- Output template generation

Text analysis

- Lexical analysis
- Named entity recognition
- Parsing

Lexical analysis

- Tokenisation
- Part of speech tagging
- Lemmatization
- Word sense tagging

Named Entity Recognition

- Actors (people, organisations)
- Time expressions
- Numerical expressions (money, weights,…)
- Specific domain related names (eg drugs, diseases, genes,…)

Parsing

- Chunking
- Shallow parsing
- Full parsing

System architecture

- Text analysis
  - Lexical analysis
  - Named entity recognition
  - Parsing
- Extraction
  - Pattern matching
  - Merging of facts
- Output template generation

Extraction

- Extraction of simple facts, using pattern matching
- Integration of small facts into a bigger picture
  - Coreference resolution
  - Merging smaller chunks of knowledge together
  - Using domain / world knowledge / Inference

Pattern matching

- X is the CEO of Y.
- The drug X can be used to treat disease Y.
- X, the leading producer of Y, …

Coreference resolution

- Narrow definition: recognise noun phrases pointing at the same thing
- Broad definition:
  - Whole – part relations
  - Set – subset
  - Type – token

Coreference resolution (narrow definition)

- Names can differ
  - Microsoft, Microsoft Corp.
  - Mr. George W. Bush, George Bush, Bush Junior
- Definite noun phrases
  - The president of the United States, the Texan ex-governor
- Pronouns

Merging of smaller bits of knowledge

- Partial information is learned from different parts of a text / corpus
- They need to be merged to end up with full event / entity descriptions
- Different partially filled templates are matched, and merged if they agree on key points

Factoids

Text mining [7]
5 april 2005

Antal van den Bosch en
Walter Daelemans

http://ilk.uvt.nl/~antalb/textmining/

Overview

- Factoids
  - Goal
  - Factoid Memory Machine:
    - Functionality
    - Components
  - Integration
- Partial assignment 2
- End assignment information

Factoids

Wikipedia:

Factoid originally meant a wholly spurious "fact" invented to create or prolong public exposure or to manipulate public opinion and was coined by Norman Mailer in his 1973 biography of Marilyn Monroe. Mailer himself described a factoid as "facts which have no existence before appearing in a magazine or newspaper". Mailer created the word by combining the word "fact" and the ending "-oid" to mean "like a fact".

The term is sometimes now also used to mean a small piece of true but often valueless or insignificant information. This definition was popularized by the CNN Headline News TV channel which during the 1980s and 90s used to frequently include such a fact under the heading of the word "factoid" during newscasts.

Factoids (2)

- Here:
  - A relation between a named entity, a verb, and other named entities (such as places)
  - Found in factual documents, such as newspaper articles
  - Supposedly true(-ish)

Goal

- Question answering:
  - Simple who did what, where questions
  - Return the factoid as an answer
- Usable offline knowledge:
  - Pre-compiled fast QA retrieval instead of expensive on-line search
  - Bibliographical profiling, journalistic search; Quick listing of things a person did, places he/she went

Factoid Memory Machine

- Simple demo system
- http://ilk.uvt.nl/~factoids/index.php
- Illustration on Dutch newspaper archive
  - ILK Corpus, 1985-1998
  - Several newspapers
  - 123 M words

FMM: Functionality

- User gives query terms in four categories:
  - Who?
  - Did what?
  - Where?
  - Other terms?
- Query is matched to all stored factoids
- All matching factoids (AND) are returned

FMM: Functionality

- User gives query terms in four categories:
  - Who? subject person NEs
  - Did what? verbs
  - Where? location NEs
  - Other terms? anything
- Query is matched to all stored factoids
- All matching factoids (AND) are returned

FMM: Components

- Offline extraction of factoids:
  - Tokenizer
  - Named-entity recognizer (persons, locations)
  - Shallow parser (verbs, subjects)
- No problem when slow
- Problem with (in)accuracy
- 774139 factoids in 123 M words

Simplification

- Some syntactic structures are "preferred", "easy".
- E.g. subject-verb-rest sentence structure in Dutch.
- So: only take subject-initial sentences.

Example factoids

Simple markup:

```
18431 {person}Kohl | [HD]hield het op een
  simpele [HD]ruil van {nonname}Schleswig-
  Holstein tegen {city}Hessen , waar de
  {nonprofit}CDU vorig jaar de macht van de
  {nonprofit}SPD en de {nonprofit}Groenen
  [HD]overnam .
100347 {person}Richard {person}Nixon | [HD]is de
  enige {nonname}VS-president die tot aftreden
  [HD]werd [VC]gedwongen
```

FMM: Integration

- Technical modules:
  - PHP
  - mysql
- Operation:
  - mysql database stores all factoids
  - PHP reads in queries from forms,
  - PHP queries mysql for an AND of query terms with field specifications
  - PHP prints results in browser
- Not optimized for speed

FMM: Demo

- Go to http://ilk.uvt.nl/~factoids/index.php
- In beta stage; limited functionality:
  - No location questions
  - At least one word in "wie" field
- Example queries:
  - Paus bezoekt
  - Gullit verlaat
  - Thatcher ontkende

Partial assignment 2

- QA performance of Factoid Memory Machine
  - http://ilk.uvt.nl/~factoids/index.php
- Given 20 "who did what" queries *that return at least one factoid* (bias!),
  - Measure the precision (fraction of # correct answers / total # returned factoids) for each query.
  - When factoid is not an answer, determine the source of error.
- Present and discuss aggregated and averaged outcomes.
- Analyse the most serious sources of errors and suggest improvements to FMM.

Partial assignment 2 (2)

- Notes:
  - An answer is correct if the factoid has the queried person as a subject of the queried verb.
  - Don't count double answers for one document (it's a bug).
  - Send crashes and weird bugs to Antal.vdnBosch@uvt.nl

End assignment

- Write a scientific paper on a text mining topic
  - Literature overview, experiments with an existing system, a new machine-learning-based module, …
  - 5-8 pages
  - **May 31, 2005**
- Presentation on the topic
  - **May 10, 2005**
  - 15 minutes