Natural Language Processing and Applications

Information Extraction & Retrieval

“Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information upon it.”
- Samuel Johnson

Dr. Mark Lee
mgl@cs.bham.ac.uk
http://www.cs.bham.ac.uk/~mgl/nlpa/

Lecture Overview

Part 1
Information Extraction

Part 2
Information Retrieval

IE versus IR

- Information Retrieval
  - Assume information is already conveniently present as a document
  - Query + Collection of documents -> Relevant documents
  - Varied Applications:
    - Web search
    - Library systems
    - Spam filtering

- Information Extraction
  - We have a document (or several million), we need information!
  - Documents are not a convenient form for information (for later processing)
  - Applications
    - Question Answering Systems
    - Data Mining etc.

IE Tasks

- Information Extraction typically involves either tagging entities or filling in templates.
- Two Sample Tasks
  - Named Entity Recognition
  - Identify and Classify All Named Entities in a document
  - Event Extraction
  - Extract information on all events in a document

- IR: Information between documents
- IE: Information within documents
Wall Street Journal Text

- J.P. Bolduc, vice chairman of W.R. Grace & Co., which holds a 83.4% interest in this energy-services company, was elected a director. He succeeds Terrence D. Daniels, formerly a W.R. Grace vice chairman, who resigned. W.R. Grace holds three of Grace Energy’s seven board seats.

- Named Entity Recognition
  - Tag All Named Entities
  - Classify all Entities as Person, Place, Organisation.

Software Engineering & IE

- IE is an exemplar of rapid prototyping
  - Quickly develop a baseline classifier which might capture 80% of the information
  - Add to the prototype heuristics for special cases to boost accuracy

- Named Entity Recognition
  Large number of Regular patterns expressed as Finite State Transducers
  Capitalized-word+ “Corp” -> Organization
  “Mr.” Capitalized Word+ -> Person

  Such rules need to be ordered by preference.
  - (Typically rule which captures most words is preferred)
  - We do this by building a large network of FSTs

Corpora & other resources

- Developing a high quality classifier requires a systematic approach
  - Large Annotated Corpus
  - Programme to compare results of gold standard corpus and classifier

- Annotation requires
  - Hard Work
  - High Degree of Inter-annotator agreement (measure in Kappa scores etc.)

- Gazetteers – large lists of names
  (we’ll typically use e-copies of phone directories of the right nationality and date)

- Rules to cover aliases
  - John Smith -> Mr. Smith -> John

Machine Learning

- Building a high performance Named Entity Recognizer is hard work
- Multiple Languages and multi-domains represent significant challenges

- We can use machine learning to help here
  - Sentence = W₁, W₂, ..., Wₙ
  - four tags = start_of_name, end_of_name, middle_of_name, not_name
  - P(tag₁ | Wₙ) probability of tag₁ given a word

- We can also use n-gram models to capture immediate context
- Use dynamic programming to select most likely sequence of tags
- Additional Features (caps, punctuation, partial parsing?)

In practice the trick is knowing which features to ignore.
Performance

- Named Entity Recognition is regarded as "done"
- NER was introduced as an evaluation task in MUC6 (1996)
  - 95% is possible with a professionally developed NER
  - Slightly lower accuracy if developed on a different domain corpus to the application.
- NER since NER relies on FST technology it's also fast
- Fast enough to be part of other applications or as a stage of preprocessing
- NER technology for Information Retrieval (Winchester & Lee, 2001)
- (also some work on summarization)

Sentence Splitting

- So far we have assumed:
  - Sentences are clearly marked as such.
  - Punctuation is clearly used and unambiguous
  - Full stops mark the end of sentences, commas (etc.) mark clause boundaries.
- Today, Dr. Mark Lee lectured a class on Natural Language Processing (N.L.P.). It was sponsored by a Rich Company Inc. and Another Rich Company Inc.
- ‘Wo-ho!’ said the coachman. "So, then! One more pull and you’re at the top and be damned to you, for I have had trouble enough to get you to it! - Joe!" (From A Tale of Two Cities)
- 90% of full stops are sentence boundaries in the Brown Corpus
- How do we deal with the remaining 10%?

Heuristics

**General Case**
- Sentence boundaries occur at one of ".", "?" or "!

**Exceptions**
- "So men, women, children etc. will be given safety jackets” in mixed case text, periods followed by whitespace followed by a lower case letters are not sentence boundaries
- "Most cases are benign however .5 % of all cases are terminal”
  - Periods followed by a digit with no whitespace are not sentence boundaries
- "Mr. White runs the company”
  - Periods preceded by titles “Mr/Dr/Ms/etc.” followed by whitespace and then a capital letter are not sentence boundaries.
- www.google.com
  - Periods surrounded by letters are probably web addresses or acronyms etc.

Further Problems

- Automatic Speech Recognition
  - Periods are not usually recognised
  - Neither is letter case
  - GOOD EVENING GIANNI VERSACE ONE OF THE WORLDS LEARNING FASHION DESIGNERS HAS BEEN MURDERED IN MIAMI POLICE SAY THAT IT WAS WAS A PLANNED KILLING CARRIED OUT LIKE AN EXECUTION
- Partial Solutions
  - Probability of Words beginning sentences
  - Pauses in the speech signal
  - NER (Names often start sentences)
  - Rapid prototyping (but Sentence Splitting is still tricky in practice)
Word Segmentation

- Word Segmentation in Chinese & Japanese
  “How do you say Octopus in Japanese?”
  ri-wen zhang-yu zen-me shuo (or)
  ri wen-zhang yu zen-me shuo
  “How do you say Japan essay fish?”

Sproat et al. (1996)
Most likely segmentation is the one which contains the set of most likely words.
Each word has a probability \( P(W) \) derived from a corpus
\[
P(S1) = P(w_1) \times P(w_2) \times \ldots \times P(w_n)
\]
\[
P(S2) = P(w_1) \times P(w_2) \times \ldots \times P(w_n)
\]
Use Dynamic Programming to pick maximal probability
Word Lexicon is a FST for each word (see lecture 2)

Event Extraction

- J.P. Bolduc, vice chairman of W.R. Grace & Co., which holds a 83.4% interest in this energy-services company, was elected a director. He succeeds Terrence D. Daniels, formerly a W.R. Grace vice chairman, who resigned. W.P. Grace holds three of Grace Energy’s seven board seats.

- Event Extraction
Suppose we want to scan newspaper text and extract all cases of executives starting or leaving management positions.
- Person:
- Company:
- Start/Leave job:

Event Recognisers

- In the same way we approached NER
- Regular expressions associated with templates
  capitalized-word+1 appointed capitalized-word+2 "as" "president"
  Person: 2
  Company: 1
  Start/Leave: Start

- This doesn’t get us very far.
  IBM appointed Harriet Smith as President
  Harriet Smith was appointed president by IBM
  Harriet Smith, who was appointed president by IBM
  IBM, the famous computer company, appointed Harriet Smith as director of sales
- We could add more regular expressions but isn’t this missing the point?

Partial Parsing

- Some problems can be dealt with by recognising constituent structure and then writing regular expressions for partial structure rather than words
- In particular, names and noun phrases can be useful

- This is Bottom Up Parsing
- Robustness is required
- Shallow disambiguation
IE development

- Information Extraction remains a specialised area of NLP
  - Personal Expertise
  - “Craft”
- Several Workbenches exist
  - GATE (General Architecture for Text Extraction)
- Some Machine Learning Techniques
  - Hand Annotated Corpora
  - Standard Stat. NLP techniques
  - The big question is always what information to learn from.

Applications

- Much of IE is hand crafted
  - Applications are therefore costly to develop & maintain
  - The “better than google” effect (or “just F'ing Google it”)
- IE in Medical Records
  - Hospitals generate HUGE amounts of textual data
  - Electronic data is becoming standard
  - Medical research relies on analysis of patient care and outcomes
- Military/Intelligence Applications
  - Email Analysis
  - Newswire summarization etc.

Click to add title

Time for a Break!