Text Analysis and Topic Models

Wray Buntine
National ICT Australia (NICTA)
Australian National University (ANU)

MLSS, Oct., 2010

Part I

Motivation and Background: Language and Document Analysis

What a good Statistical NLP Course Needs

Apart from the usual CS background (algorithms, data structures, coding, *etc.*):

- prerequisites or coverage of information theory, and computational probability theory;
- theory of context free grammars, normal forms, parsing theory,*etc.*;
- programming tools: Java for tools, Java & Python for experiments.

None of this is presented here!

Outline

Outline

We do a review of the analysis of formal natural language (not a formal analysis of natural language).

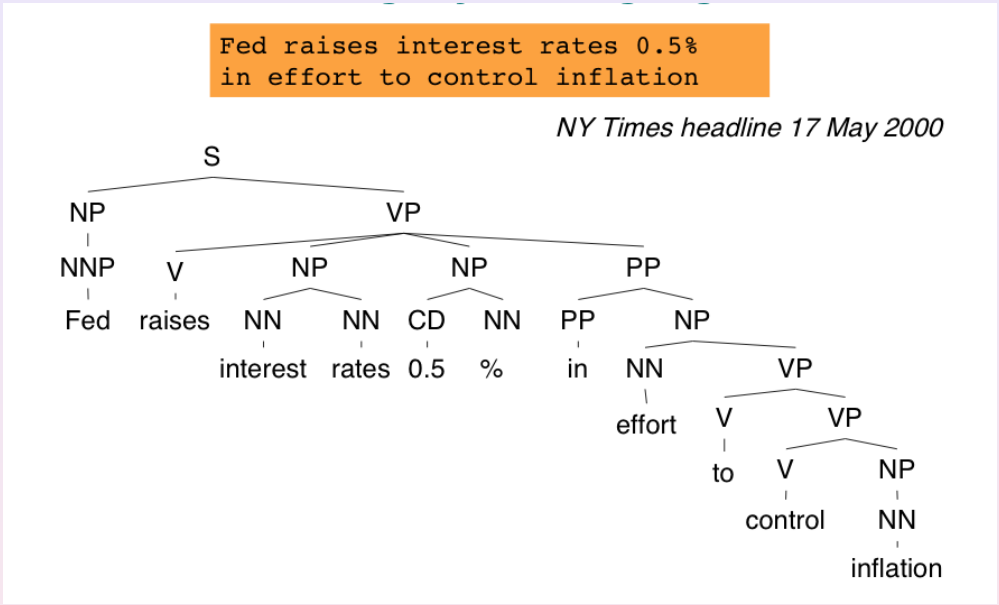
1. Formal Natural Language
   - NLP Processing and Ambiguity
   - Words
   - Parsing
2. Document Processing
3. Document Analysis

What is Formal Natural Language

- Formal language is taught in schools (*e.g.*, grammar schools) with correct grammar, punctuation and spelling.
- Most books, more traditional print media, formal business communication, and newspapers use this.
- But errors exist even in the *The Times* and *The New York Times*.
- In contrast, informal language is found in email, people's web pages, chat groups, and "trendy" print media.

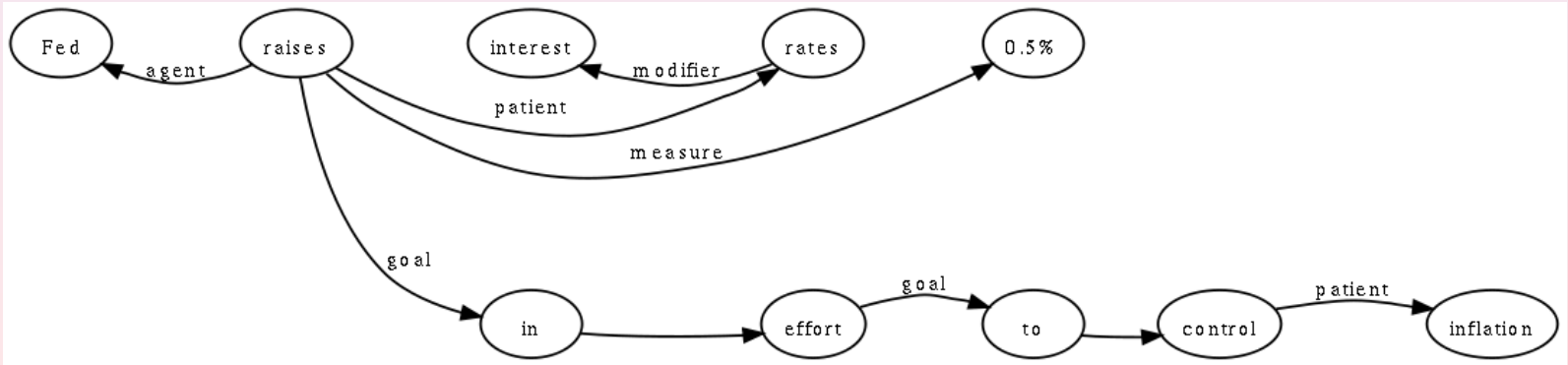
Outline

Analysing Language



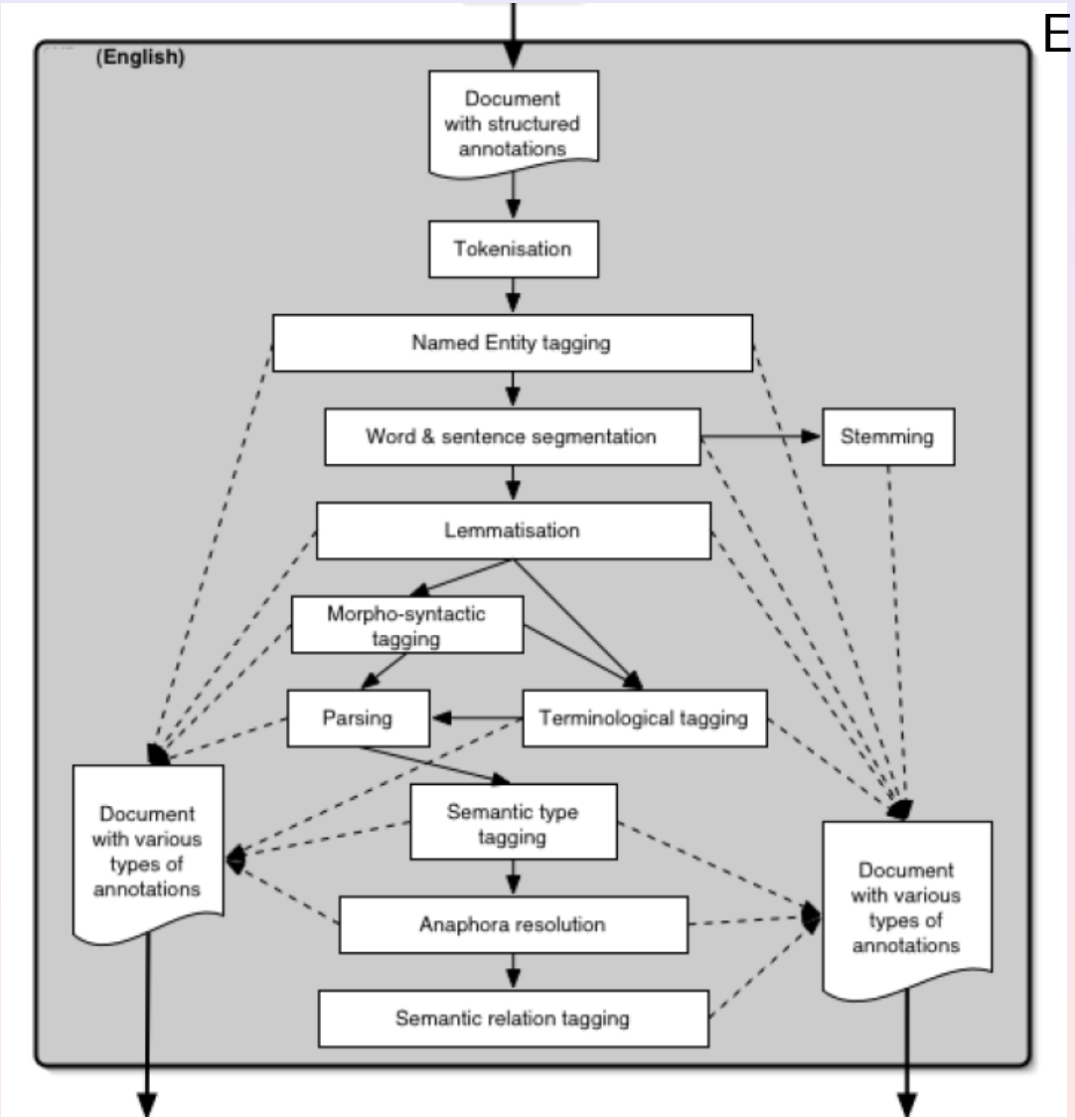
Example from *McCallum's NLP course*

- Left, a traditional parse tree showing constitutent phrases.
- Below, a dependency graph showing *semantic roles*.



Traditional NLP Processing

Full processing pipeline might look like this for English.



- Typical accuracies for various stages might be 90-98%.
- But it can drop down to 60% for the later semantic analysis.
- Errors earlier on magnify later.
- Recent research propagates uncertainty and alternatives along with the linguistic results.

Common Tasks in NLP

Tokenisation: breaking text up into basic tokens such as word, symbol or punctuation.

Chunking: detecting parts in a sentence that correspond to some unit such as "noun phrase" or "named entity".

Part-of-speech tagging: detecting the part-of-speech of words or tokens.

Named entity recognition: detecting proper names.

Parsing: building a tree or graph that fully assigns roles/parts-of-speech to words, and their inter-relationships.

Semantic role labelling: assigning roles such as "actor", "agent", "instrument" to phrases.

NLP in Chinese

Input
A Chinese sentence
我弟弟要买两个足球。
My brother wants to buy two balls.

Output (the word and POS sequence)
我/r (my) 弟弟/n (brother) 要/v (want)
买/v (buy) 两/m (two) 个/q (classifier)
足球/n (football) 。/w (period)

- Tokenisation (segmenting words) is very difficult.
- Easier in Japanese[1] because their foreign words use separate phonetic alphabets.
- Little morphology used.

[1] Japanese writing is based on traditional Chinese, the precursor to modern Simplified Chinese.

NLP in Arabic

القاهرة هي أكبر مدينة أفريقية والأكثر سكاناً في أفريقيا والشرق الأوسط .وهي محافظة مدينة، أي أنها محافظة تشغل كامل مساحتها مدينة واحدة، وفي نفس الوقت مدينة كبيرة تشكل محافظة بذاتها. وبالرغم من كونها كمدينة هي الأكبر إلا أنها تعد من أصغر محافظات مصر كمحافظة .

- Here is part of an article in Arabic about Cairo.
- Underlined words are ambiguous due to lack of vowels.
- Red parts are attached prefixes (like English prepositions "on", "of").
  - Turkic, Finnish, and some archaic Indo-European languages use suffixes similarly. Dative cases in Germanic are remnants of this aspect of language.
- Note Arabic and Hebrew share general features, their scripts can be traced to versions of Aramaic.
  - Many Asian and European alphabets are derived from Phoenician, a precursor to Aramaic, but they also have vowels. Phoenician itself was influenced by Egyptian hieratic, Egypt's alphabetic simplification of Egyptian hieroglyphics. Hieroglyphics is closer to Chinese in concept.

NLP in Arabic, cont.

درست	***darasat***	she studied (feminine)
درست	***darrasat***	she taught (feminine)
درست	***durisat***	it was studied (feminine)
درست	***durrisat***	it was taught (feminine)
درست	***darastu***	i studied
درست	***darrastu***	i taught
درست	***duristu***	i was studied
درست	***durristu***	i was taught
درست	***darasta***	you studied (masculine)
درست	***darrasta***	you taught (masculine)
درست	***durista***	you were studied (masculine)
درست	***durrista***	you were taught (masculine)
درست	***darasti***	you studied (feminine)
درست	***darrasti***	you taught (feminine)
درست	***duristi***	you were studied (feminine)
درست	***durristi***	you were taught (feminine)

- Has a fairly rich morphology (i.e., modification of words to match case).
- Vowels not included in alphabet.

NLP in Arabic, cont.

Prefixes: some English prepositions are translated to prefixes in Arabic.

بالدرس	***be*ddars**	With/In the lesson
للدرس	***le*ddars**	For/To the lesson
كالدرس	***ka*ddars**	As the lesson
فالدرس	***fa*ddars**	Then the lesson
فالدرس	***fi*ddars**	In the lesson

Lack of vowels: ambiguity due to lack of vowels in Hebrew

ספק SAFEK = doubt
ספק SAFAK = clapped
ספק SIPEK = provided
ספק SUPAK = has been provided
ספק SAPAK = provider

Agglutinating and Compounding

English: I am in the cafe too.

Finnish: On kahvilassahan.

Finnish, an *agglutinating language* like Mongolian and Turkish, can express four English words in one! The translation:

$$On_{I\ am}\ kahvi_{coffee} la_{place} ssa_{in} han_{emphasis}\ .$$

This makes statistical machine translation very difficult. For instance, only the base word "kahvila" will be in any dictionary.

English: dog food

Finnish: koirarouka

On the other hand, detecting *compound words* is much easier:

$$koira_{dog} food_{rouka}$$

Translation Difficulties

Some languages represent names differently, especially those originating outside of the Latin based alphabets.

Code	Language	Translation
EN	English	Saddam Hussein
LV	Latvian	Sadams Huseins
HU	Hungarian	Szaddám Huszein
ET	Estonian	Saddäm Husayn

Language Ambiguities

An unnamed high-performance commercial parser made the following analysis of a sentence from Reuters Newswire in 1996.

Clothes made of [hemp and smoking paraphernalia]$_{\text{phrase}}$ were on sale.

The correct analysis is:

[Clothes made of hemp]$_{\text{phrase}}$ and [smoking paraphernalia]$_{\text{phrase}}$ were on s

This misinterpretation is a common semantic problem with current parsing technology.

Language Ambiguities, cont.

- New$_{\text{adjective}}$ York Tennis Club$_{\text{name}}$ opening today. versus
  New York Tennis Club$_{\text{name}}$ opening today.
- He worked at Yahoo!$_{\text{sentence}}$ Tuesday.$_{\text{sentence}}$ versus
  He worked at Yahoo!$_{\text{name}}$ Tuesday.$_{\text{sentence}}$
- Stolen painting found by tree$_{\text{location}}$. versus
  Stolen painting found by tree$_{\text{actor}}$.
- Iraqi head$_{\text{body part}}$ seeks arms$_{\text{body part}}$. versus
  Iraqi head$_{\text{politician}}$ seeks arms$_{\text{weapons}}$.

Language Ambiguities, cont.

- Ambiguities arise in all processing steps, due to the tokenisation done, the identification of proper names, the part of speech assigned, the parse, or the semantic role assigned.
- All languages have particular versions of the ambiguity problem. *e.g.*, standard Arabic and Hebrew don't represent vowels in their text!

We resolve ambiguity by appeal to *distributional semantics*, that the meaning of a word is given by its distribution with the words surrounding it, its context.

Handling of ambiguity generally requires that intermediate processing carry uncertainty, for instance, by using latent variables in statistical methods.

Outline

Word Classes (dictionary version of part of speech)

Part of speech	Function	Examples
Verb	action or state	(to) be, have, do, like, work, sing, can, must
Noun	thing or person	pen, dog, work, music, town, London, John
Adjective	describes a noun	a/an, 69, some, good, big, red, well, interesting
Adverb	describes a verb, adjective or adverb	quickly, silently, well, badly, very, really
Pronoun	replaces a noun	I, you, he, she, some
Preposition	links a noun to another word	to, at, after, on, but
Conjunction	joins clauses or sentences or words	and, but, when, because
Interjection	short exclamation, can be in sentence	oh!, ouch!, hi!

Word Forms

Morpheme: Is a semantically meaningful part of a word.

Inflection: A version of the word within the one word class by adding a grammatical morpheme. "walk" to "walks", "walking", and "walked".

Lemma: The base word form without inflections, but no change in word class. "walking" lemmatizers back to "walk", but "redness" (N) does not lemmatise to "red" (A).

Derivation: Adding grammatical morphemes to change the word class. "appoint" (V) to "appointee" (N), "clue" (N) to "clueless" (A). Uses "-ation", "-ness", "-ly" *etc.*

Stemming: Primitive version of lemmatization that strips off grammatical morphemes naively, usually in a context free manner.

Open versus Closed: Nouns, verbs, adjectives, adverbs are considered *open* word classes that continually admit new entries.

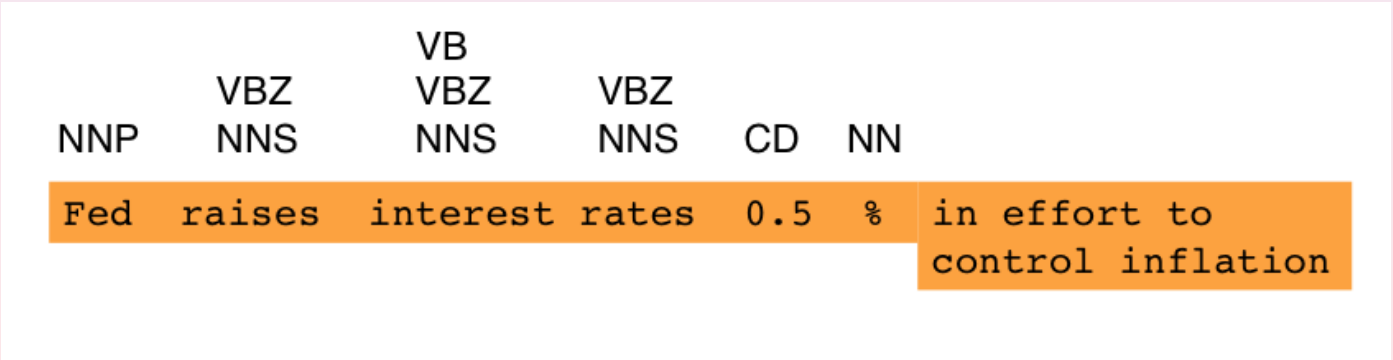
Parts of Speech (computational)

Example parts of speech from the Tagging Guidelines for the Penn Treebank.

POS	Function	Examples
CC	coordinating conjunction	and, but, either
CD	cardinal number	three, 27
DT	determiner	a, the, those
IN	preposition or subordinating conjunction	out, of, into, by
JJ	adjective	good, tall
JJS	adjective, superlative	best, tallest
MD	modal	he *can* swim
NN	noun, singular or mass	the *ice* is cold
NNS	noun plural	the *iceblocks* are cold
PDT	predeterminer	*all* the boys
SYM	symbol	$, %
VBD	verb, past tense	swam, walked
...	...	...

Parts of Speech (computational version), cont.

- For computational analysis, more detail over the 8 word classes is needed in order to capture inflections and variations supporting a parse.
- With just candidate POS for each word, many different parses can exist. McCallum's initial example is shown again below.



Collocations

Small, usually contiguous, sequence of word that behaves semantically like a single word: "hot dog", "with respect to", "home page", "fourth quarter", "run down",

- Meaning of a collocation is different to the meaning of its parts.
  - The collocation cannot be modified easily without changing the meaning: "kicked the bucket" versus "kicked the tub", "the bucket was kicked".
  - We identify collocations by appeal to distributional semantics.
- Related: multi-word expression/unit, compound, idiom.
- In some languages, collocations replaced by compounds (words are joined with no space or hyphen).
- Important for parsing, dictionaries, terminology extraction, ...

Outline

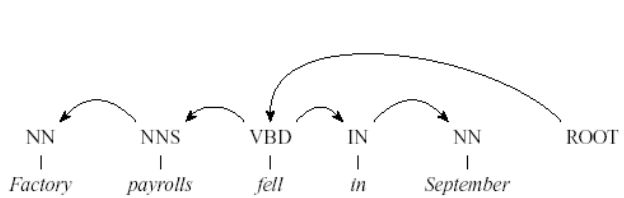
Constituents

A word or a group of words that functions as a single unit within a hierarchical structure.

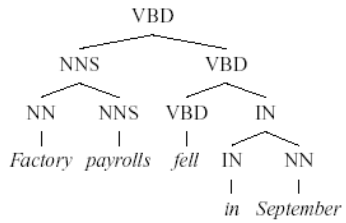
e.g. noun phrase, prepositional phrase, collocation, *etc.*

- Often can be replaced by a single pronoun and the enclosing sentence is still gramatically valid.
- Serve as a valid answer to some question.
  e.g., How did you get to work? By train.
- Admits standard syntactic manipulations.
  e.g., can be joined with another using "and", can be moved elsewhere in the sentence as a unit.
- Building a parse tree involves building the complete set of constituents for a sentence.

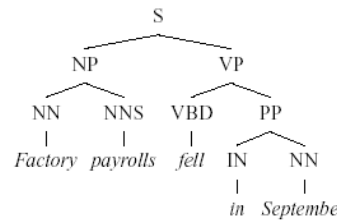
Parsing



(a) Classical Dependency Structure



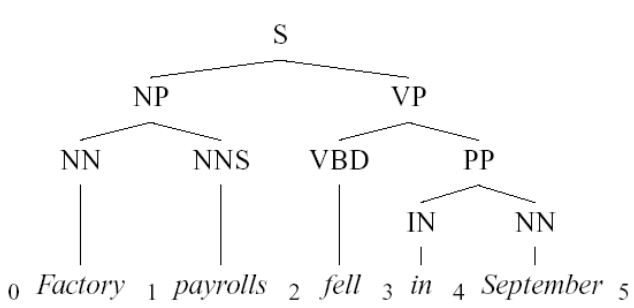
(b) Dependency Structure as CF Tree



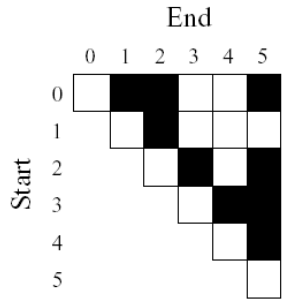
(c) CFG Structure

- Sometimes we want a dependency tree showing syntactic or semantic relationships, as in (a).
  - Usually, we want the relationships labelled.
    e.g. arc from "fell" to "in" labelled with *time*, arc from "fell" to "payrolls" labelled with *patient*.
- Some formal linguistic theory develops a parse tree, in this case a Context Free Grammar (CFG) is used in (c).
- Figure shows a derivation of the parse tree from the dependency tree.

Shallow Parsing



(a)



(b)

Span	Label	Constituent	Context
⟨0,5⟩	S	NN NNS VBD IN NN	◇ – ◇
⟨0,2⟩	NP	NN NNS	◇ – VBD
⟨2,5⟩	VP	VBD IN NN	NNS – ◇
⟨3,5⟩	PP	IN NN	VBD – ◇
⟨0,1⟩	NN	NN	◇ – NNS
⟨1,2⟩	NNS	NNS	NN – VBD
⟨2,3⟩	VBD	VBD	NNS – IN
⟨3,4⟩	IN	IN	VBD – NN
⟨4,5⟩	NN	NNS	IN – ◇

(c)

1: (a) Example parse tree with (b) its associated bracketing and (c) the yields and contexts for each constituent span

- A full parse yields many subtrees or constituents, labelled verb phrase (VP), prepositional phrase (PP), *etc.*
- We can also note the labels of a particular type (*e.g.*, all NPs), and build a classifier that recognises just that type.
- Recognising the start and end of a particular type of constituent is called **shallow parsing** or **chunking**.
- Parsing can also be represented as a structured classification problem, recognising the best coherent set of constituents.

Case Frames

- Case frames give the functional characteristics of a verb, the number of arguments, and their syntactic cases.
- Roles are the argument types in a single position for a verb relation, *e.g.*, agent, actor, instrument, ...
- Example case frames with roles.
  - `actor` "buy" `object` (syntactic)
  - person/organisation "buy" `thing` (semantic)
  - `agent` "fix" `thing`
  - `animate-object` "walks"
- Various databases collect case frames with more or less semantics: FrameNet, PropBank, VerbNet.
- Allows mapping of verb syntax to semantics.

Outline

We look beyond the text content to consider applications of document processing.

1. Formal Natural Language
2. Document Processing
   - Language in the Electronic Age
   - Information Warfare
   - Why Analyse Documents
3. Document Analysis

Processing of Documents

- Documents have a structure with text, links to other documents, citations to publications, images, indexes, and so forth.
- Why do we care about documents?
- What applications can be made?

Outline

Informal Language

Text messages: My smmr hols wr CWOT. B4, we used 2go2 NY 2C my bro, his GF & thr 3 :- kids FTF. ILNY, it's a gr8 plc.

IRC Chat:
`Meta-man:` NLP is a little tricky to do over IRC
`Dan_26:` I see no diff
`galamud:` I'm not pissed! I'm flattered! I mean, er... =)
`Meta-man:` hold that thought ...to your checkbook :]
`JonathanA:` HAH! LOL

Web Page Structure

- Web pages have complicated structures and *genre*, more so than traditional documents (letters, books, *etc.*).
- Example genres: product page, personal home page, FAQ, news item, blog, corporate data sheet, ...
- Much of the content will be template content shared across many similar pages.
- No standard guidelines, so must determine heuristically.

Linguistic Resources

- A large number of different resources now becoming available, due to the Internet and digitisation.
- Included: gazetteers, dictionaries, tagged text (tagged with POS, name entity types, *etc.*), word sense data, case frame and semantic role data (*i.e.*, for verbs), collocations, aligned translations.
- Tagged and marked up linguistic resources are the hardest to get, but are the ones most needed for supervised statistical NLP.

Availability of linguistic resources is a key determining factor in the success of statistical NLP projects.

Unsupervised (or semi-supervised) approaches to statistical NLP are most needed.

Outline

The Internet Society

- Primary school students have internet component in coursework, are given internet search tasks as assignments.
- Internet news and blogs have overtaken newspapers as primary information source; the business models still developing.
- E-government, business and consumer e-services booming.
- Search and internet-based multimedia now a significant form of entertainment.

  e.g. 8 year-old boy with keywords "dinosaur", "meteor".

The Internet Society, cont.

- Advertising on specialist websites, on particular keyword searches, or on your email based on its content, is well focussed.
- Targeted advertising through the web, for instance Google AdSense, is considered the best value for money for advertising.
- Major industry companies track "green" websites and blogs for potential environmental scandals.

Document analysis has taken on a new life due to the internet. Business, government and consumer ramifications still unfolding.

Information Warfare

Definition: "the use and management of information in pursuit of a competitive advantage over an opponent."

- Email spam, link spam, *etc.* Whole websites are now fabricated with fake content in the effort by spammers.
- "More than half of Americans say US news organizations are politically biased, inaccurate, and don't care ...," *Pew Research Center on "news"* (Aug. 2007)
  - "Poll respondents who use the Internet as their main source of news – roughly one quarter of all Americans – were even harsher with their criticism."
  - 80% of the watchers of FOX news had one or more major misconceptions over Iraq war, compared with only 23% for PBS/NPR, *WorldPublicOpinion.ORG survey* (Oct. 2003)

It's an information war out there on the internet (between consumers, companies, not-for-profits, voters, parties, news publishers, ...).

Outline

Bioinformatics: Medline

- *PubMed* is the most popular database in Biology, and the main database MedLine has over 16 million entries.
  - entries are abstracts and metadata in (*MedLine format*, *XML format*, ...
  - 2,000-4,000 new entries/day from 5000 journals in 37 languages.
- The abstract databases are searchable using free text and contolled vocabularies, such as *MeSH* terms.

Tasks in MedLine

- The MeSH terms are generally entered by users and not thorough. Thus subject-specific searching patchy.
- Named entities (genes, proteins) have many different versions so it is difficult to search for them.
- Same problems apply to many technical information resources, such as patent databases.

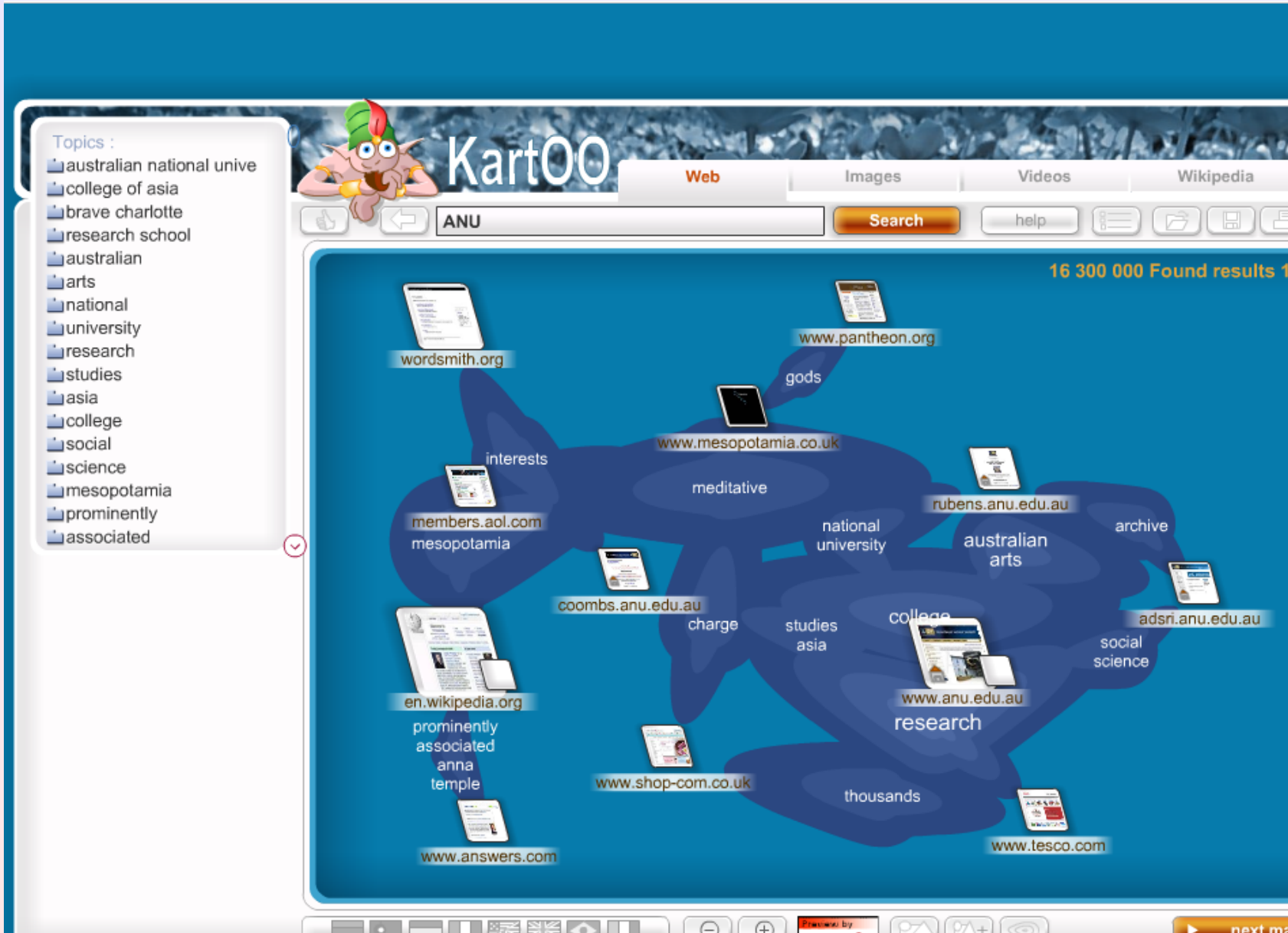
European Media Monitor: NewsExplorer

- Developed at the European Commission's Joint Research Center (JRC) in Italy. Online at `http://press.jrc.it/`.
- Completely automated:
  - automatically generate daily news summaries, and provides a *daily briefing*,
  - collect and cluster news events, and *news personalities*,
  - provide geographical, *theme* and time summaries,
  - cross-lingual capabilities.
- Uses relatively simple NLP and SML technology cleverly.
- Widely regarded within the EU Commission.

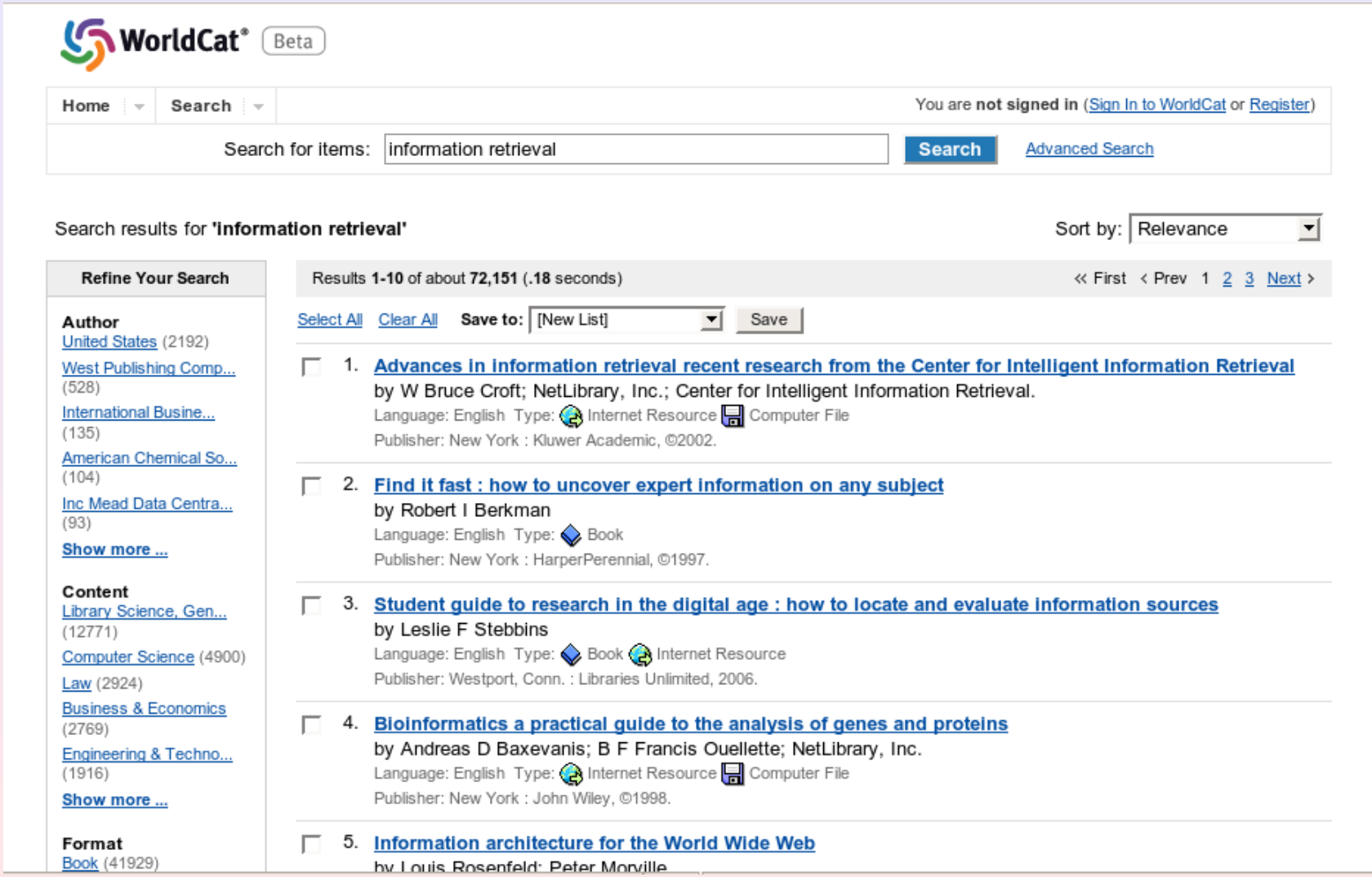
Advanced Search Engines

- Clustering output to give a dynamic snapshot of the area, such as *Clusty*.
- Providing a stronger typing of content in terms of area, keyword, genre, document type, such as *Exalead*
- Subject specific areas such as *academic search*, product search and *library catalogue search*.

Advanced Search Engines: Visualisation



World Wide Library



Patent Search: PatentLens

- Started out as a *patent search engine* for Bioinformatics to support patent packaging.
- Software is open source, but largely developed in-house at *Cambia* based on `Lucene`, the Java IR system.
- Many specific facilities to support patents (organisation/company matching, cross-nation support, gene name search ...).
- The patent landscape is changing, see *Open Invention Network*.

Social Bookmarks: Del.icio.us

- *Del.icio.us* is one of the best known social bookmarking sites.
- Uses tagging to provide higher-weighted keywords.
- Uses social bookmarks to get popularity/"authority" for pages.
- Purchased by Yahoo in 2005.

Opinion: their search returns best pages on fairly general topic areas, *e.g.* *information retrieval*, (*i.e.*., but not "home page" or "lost page" search).

Business Applications

Intelligence: information from the web about consumer trends and opinions, and about competitors.

Summaries: executive reports and overviews based on a large collection of documents input.

Intranet support: search and browse, personalisation, categorization, document management.

Administration: eGovernment and electronic document processing.

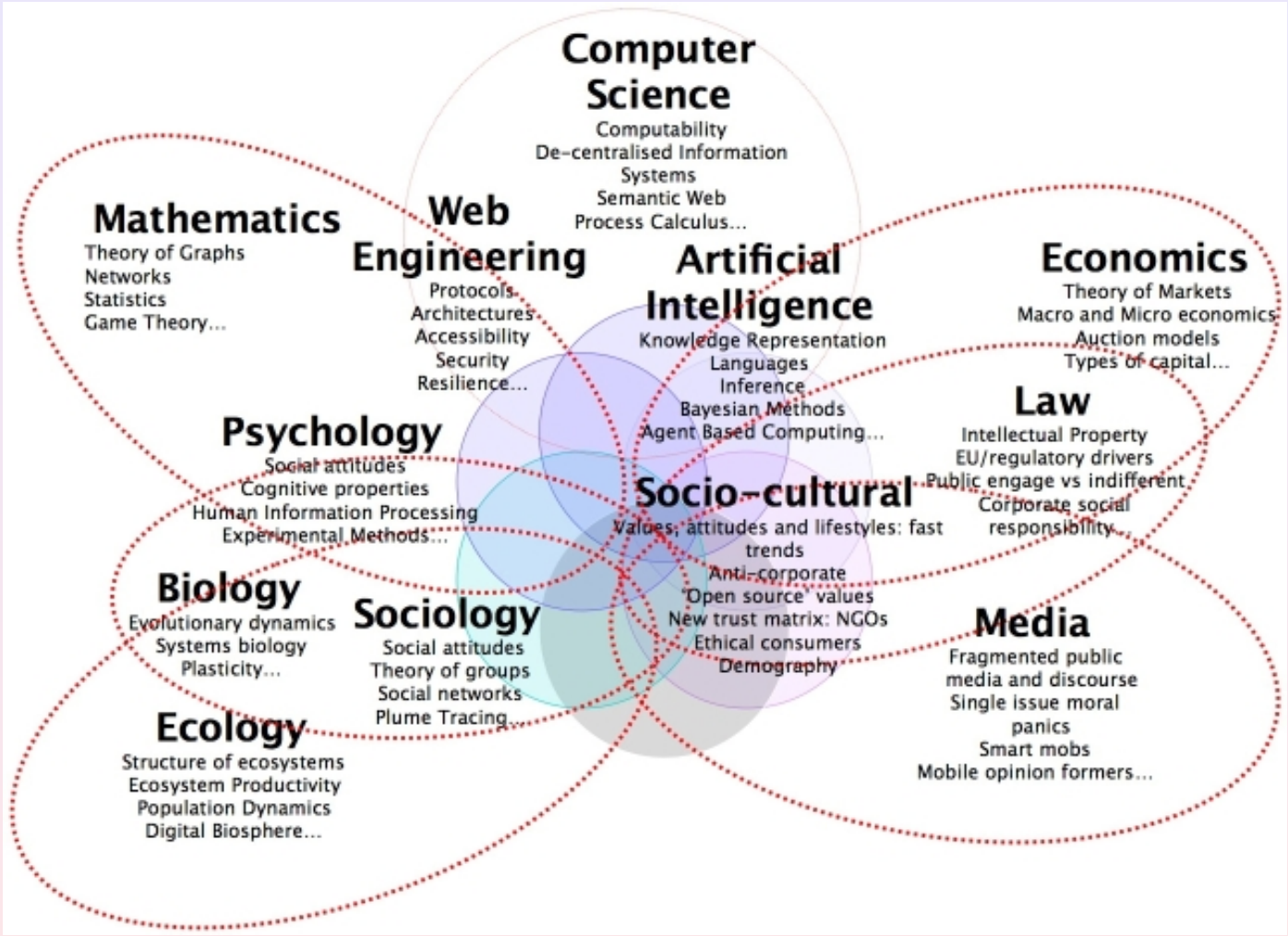
Advertising: many aspects of advertising now running online.

Outline

We sketch out the field of document analysis, with major emphasis on text.

1. Formal Natural Language
2. Document Processing
3. Document Analysis
   - Representation
   - Resources
   - Other Areas

Web Science



From *Web Science*.

Outline

Linguistic Representation

Linguistic aspects:

- basic representations presented previously: morpheme, token, word class, part-of-speech, lemma, collocation, term, named entity, constituent, phrase, parse tree, case frame, semantic role, dependency graph;
- transformations and default processing steps between them;
- differences for different languages;
- sources of ambiguity.

It is important to understand the linguists viewpoints, and their whys and wherefores.

Computational Representation

Computational aspects for the text in documents:

- data formats such as XML and its support tools and representations such as Schema, XQuery, ...;
- data structures and manipulation such as trees, graphs, regular expressions, FSA, ...;
- character processing, UTF8, simplified Chinese, Latin, ...

All of these aspects make a scripting language like Python (or Perl) the best platform for beginning statistical NLP.

Meaning Representation

The layers of processing for the text in documents.

Character level: characters $\longrightarrow$ tokens sentences $\longrightarrow$ paragraphs $\longrightarrow$ documents.

Syntactic level: morphemes $\longrightarrow$ lemmas and parts of speech $\longrightarrow$ collocations, terms and named entities $\longrightarrow$ constituents, phrases $\longrightarrow$ sentences.

Semantic level: case frames and semantic roles, dependencies, topic modelling, genre.

The three levels tend to interact, and the various stages in each level interact as well.

Outline

Part of Speech Data

- Human annotators have taken, say, 20Mb of Wall Street Journal text and carefully assigned POS to tokens.
- There can be some difficulty in assigning POS:
  - "She stepped off/IN the train." `versus` "She pulled off/RP the trick."
  - "We need an armed/JJ guard." *versus* "Armed/VBD with only a knife, ..."
  - "There/EX was a party in progress there/RB."
- POS data laborious to construct, but very useful for statistical methods.

Most parsers don't require POS tagging beforehand. It is generally done as a pre-processing step for information extraction. or shallow parsing.

Computer Dictionary: CELEX

- CELEX is the Dutch Centre for Lexical Information.
- Provides CDROM with lexical information for English, German and Dutch, called *CELEX2*. Available from LDC.
- Contains orthography (spelling), phonology (sound), morphology (internal structure of words), syntax, and frequency for both lemmas and word-forms.
- Provided for 50,000 lemmata.

Headword	Pronunciation	Morphology	Cl	Type	Freq
celebrant	"sE-lI-br@nt	((celebrate),(ant))	N	sing	6
cellarages	"sE-l@-rIdZIs	((cellar),(age),(s))	N	plu	0
cellular	"sEl-jU-l@r*	((cell),(ular))	A	pos	21

Computer Thesaurus: WordNet

- Developed at Princeton University under the direction of psychology professor George A. Miller from 1985 on.
- Contains over 150,000 words or collocations, *e.g.* see *make*, *red*, *text*.
- Words in a network with link types corresponding to:

hypernym:	generalisation,
hyponym:	specialisation,
holonym:	has as a part,
meronym:	is a part of,
antonym:	contrasting or opposite,
derivationally related:	"textual" is for "text",
word senses:	different semantic use cases identified,
case frames:	case frames for verbs.

- Available free (with an "unencumbered license"), and lots of supporting software.

Gazetteers

- Term originally applies to geographic name databases that might contain auxiliary data such as type (mountain, town, river, *etc.*), location, parent state, *etc.*
- Sometimes extended in NLP to apply to other specialised databases of proper names.
- Proper names treated differently in NLP because:
  - they behave as single tokens and don't inflect,
  - generally are marked with first letter uppercase,
  - are the greatest source of new or unknown words in text, and are not usually in dictionaries.

Good gazetteers and dictionaries are critical for performance in any specialised domain.

Linguistic Data Consortium

- *LDC* is an open consortium initially funded by ARPA.
- Wide *variety of data* including speech and transcripts, news and transcripts, language resources, annotated and parsed data.
- Includes the famous Penn Treebank which has POS tagging and parse trees for some news sources.
- Includes the Google 5-gram data (frequencies for contiguous sequences of 5 words as they occur in internet text.

Major Software

GATE: A long-time leader, Java platform from Univ. of Sheffield, provides for pipelining, and default components and plugins.

UIMA: Open source pipeline/component platform supports distributed processing, but no specific tools, via IBM.

Lingpipe: Good commercial tools with "free for non-commercial" license.

OpenNLP: Good open source tools, in Java.

Other: Many individual tools for parsing, stemming, entity extraction, *etc.*, most often in Java, older ones sometimes in C or available as libraries.

Outline

Important Issues

We've looked at applications, representation and linguistic resources, what about:

Software: many open source tools exist of varying quality, though some of the best tools are commercial and expensive.

Evaluation: a myriad of evaluation tracks exist for every aspect, and these generate some important data sets and resources.

Algorithms: space and time complexity, *etc.*

Statistical prerequisites: the field has prodigious users and creators of statistical techniques.

Recognised Problems

Information retrieval (IR): given query words, retrieve relevant parts from a document collection.

Question answering (QA): similar to IR but return an answer.

Document summarisation: taking a small set of documents on a given theme and preparing a short summary or executive brief.

Topic detection and tracking (TDT): tracking topics, and discovering new ones in information streams.

Semantic web annotation: annotating documents with appropriate semantic mark-up.

Classification: categorising documents into topic hierarchies, or creating hierarchies suited for a collection.

Genre identification: predicting the genre type.

Sentiment analysis: predicting the sentiment (negative, satisfied, happy, ...) of a blog or chat participant or commentary.

Recognised Problems, cont.

Document structure analysis: identifying the parts of a web page or document such as title, index, advertising, body, *etc.*

Linguistic resource development: tagging of text with parse structures, POS, semantic roles, name entities, *etc.*, and development of dictionaries, gazetteers, case frames, *etc.*, especially in specialised subjects.

Recommendation: from user characteristics and prior selections, make recommendations, such as collaborative filtering.

Ranking: given candidate responses for a recommendation or retrieval task, do the fine grained ranking.

Cleaning up Wikipedia: the Wikipedia would be an amazing linguistic resource if only, ...

Recognised Problems, cont.

Machine translation (MT): automatically convert text to another language,

Cross language IR (CLIR): from queries in one language probe document collection in another.

Email spam detection: recognising spam email.

Trust and authority: measures of document/author quality in terms authority and trust based on content, links, citation, history, *etc.*

Communities: analysis and identification of online communities.

Video and Image X: most of the above applied to video and images.

Outline

And so ends Part 1. Next we look at specific problems and algorithms.