Introducing Document Analysis

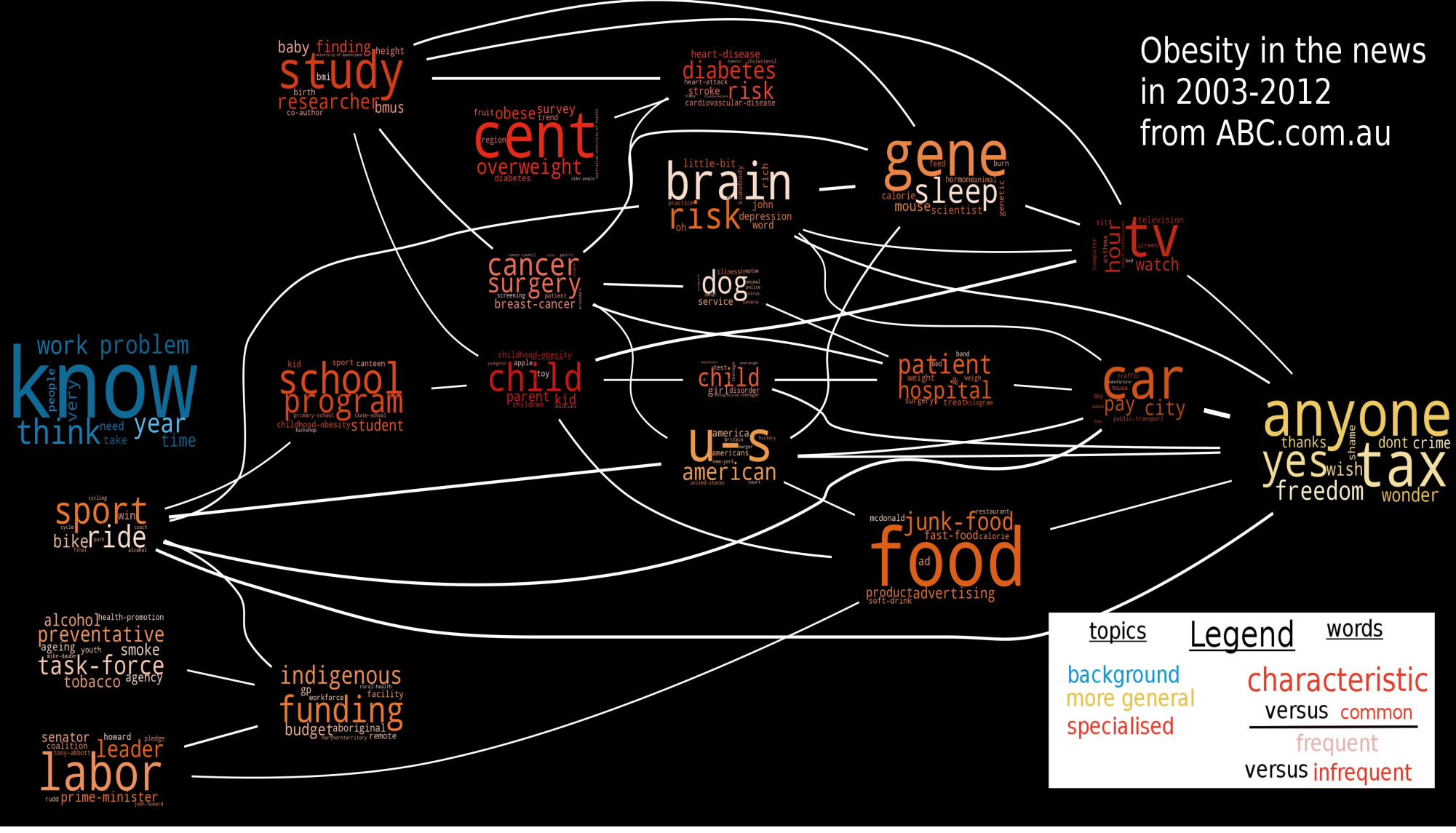
Wray Buntine
Monash University

http://topicmodels.org ⟵ get the slides from here

April 4, 2016

Outline

A word about me.



topic models are used to find themes or "soft" groupings in document collections, social networks, bibliographic networks, or other semi-structured content

Research: Discrete Bayesian Non-Parametrics

- using Discrete Bayesian Non-Parametrics:
  - hierarchical Pitman-Yor processes
  - Poisson process techniques
  - on complex discrete structures and networks
- for data like:
  - FreeBase, WordNet, parse trees
  - social netorks, blogs and recommendations
  - citation networks with abstracts
  - matrix and tensor factorisation
- non-parametric topic modelling software, `hca`
  - at `MLOSS.org` and `GitHub`
  - multicore, coded in C

Outline

Why work with documents and text.

Information Overload

 langwitches' photostream @ Flickr

- need tools to help us: organize, search, summarise and understand information
- field information access serves this purpose

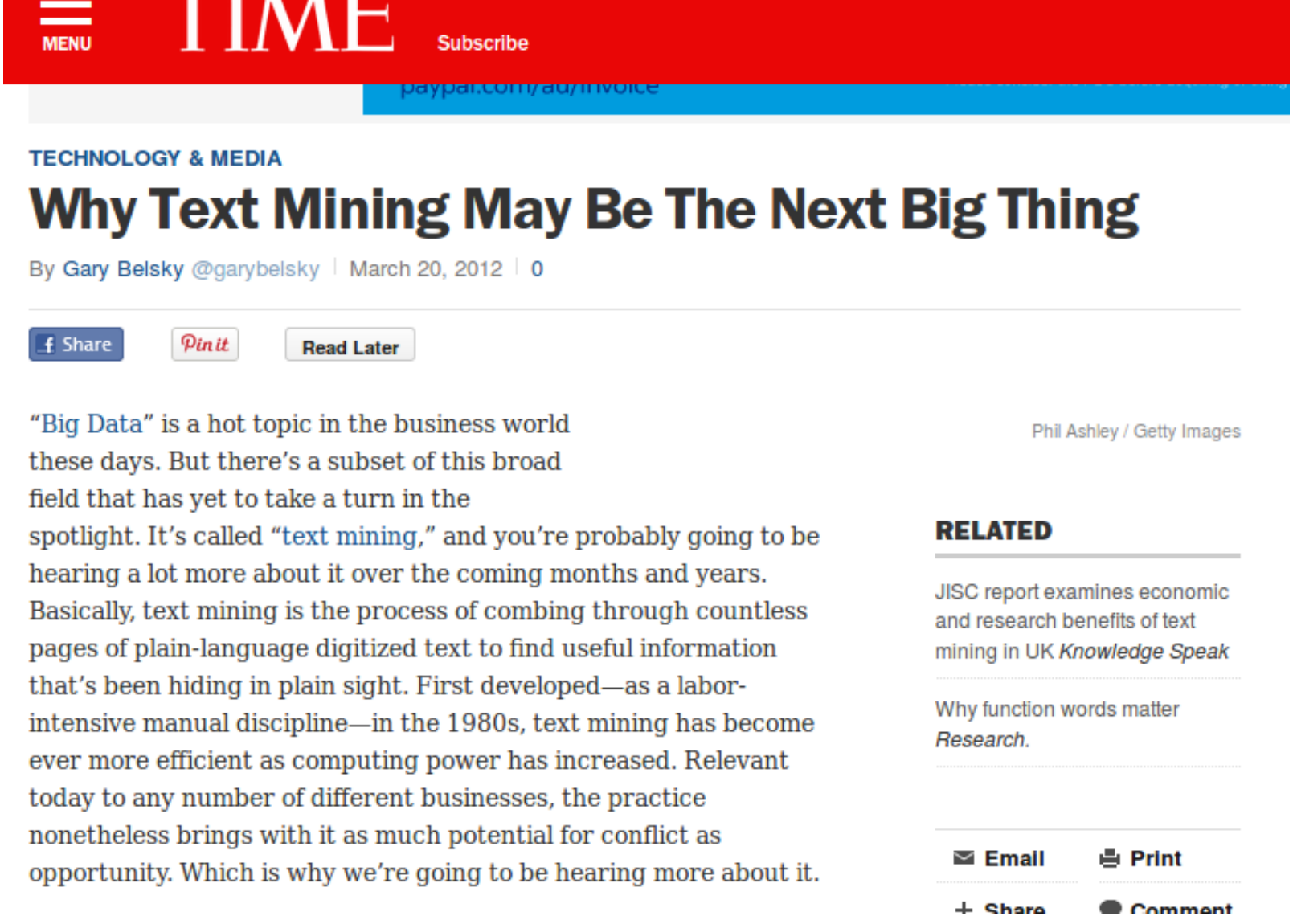
Information Warfare

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

- Email spam, link spam, *etc.*
  - Whole websites are fabricated with fake content.
  - Spammers using social networks *to personalise attacks*.
- trust in information on the web is being damaged by people *"paid by companies to post comments"* (Dec. 2011).
- *Pew Research Center* (Sept. 2011) says people think
  - "news organizations tend to favor one side," ...
  - "are often influenced by powerful people and organizations"

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

How much of the internet world is text or semi-structured content?



Time magazine claims 80% of Corporate/Government content is text.

Examples of Text Analysis

- the document is about 'immigration' or 'sales tax'
- the webpage mentions a particular product
- the tweet describes a problem with a product
- the author of the blog post is likely a Labour Party voter
- the email contains bullying language
- the review gives positive remarks about a hotel
- from news reports this person is likely the same as this other

Example Tasks

corporate	expert finding, document summarisation
health care	disease monitoring, nursing notes analysis
government	records classification, FOI requests, service delivery, policy research
insurance	problem identification from claims, fraud detection
legal	search and discovery
oil and gas	analysis of maintenance and repair logs
retail	brand or product analysis, customer retention
security	log analysis

Demonstrations

Nonsense paragraph for testing:

> *Sir Nigel Shadbolt, Professor of Artificial Intelligence at Southampton University, happily believes in the power of open data. With the venerable Sir Tim Berners-Lee, he persuaded two UK Prime Ministers of the importance of letting us all get our hands on information that's been willfully collected about us by the government and other organisations. The enormous potential for corruption and distortion can be avoided.*

Demos:

- *OpenCalais* or *Semantria* for tagging
- the *Stanford Parser*

What People Want from Document Analysis

- formal name and key phrase (“colocation”) recognition
- sentiment and emotion analysis
- classification, theme and topic analysis
- prediction, in many ways
- information retrieval, in many ways
- summarisation
- custom analysis

What a good Documents or Statistical NLP Course Needs

Apart from the usual computer science 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;
- document, text and internet standards.
- deep neural networks, non-parametric Bayesian statistics

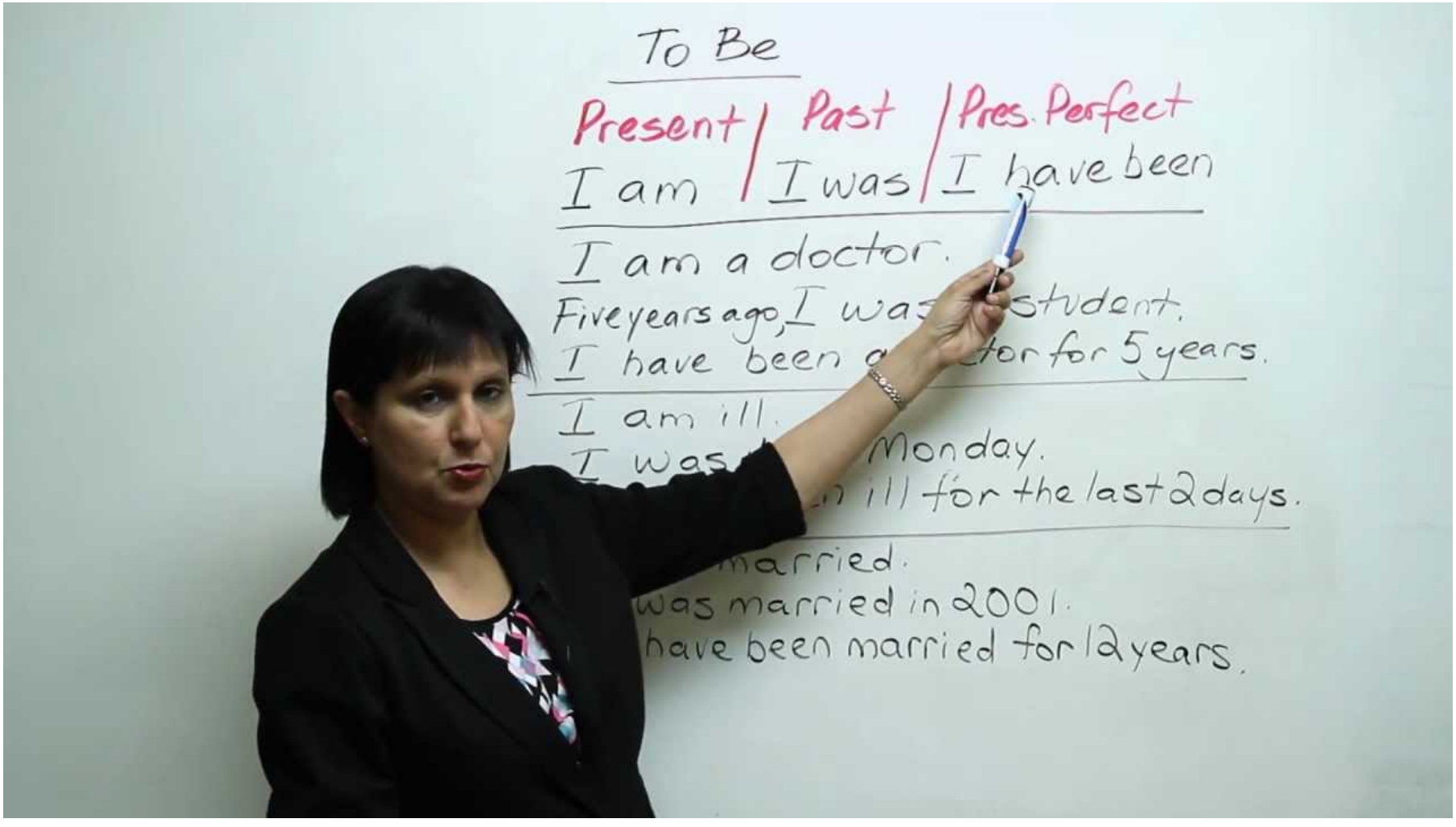
None of this is presented here!

Outline

We do a review of the analysis of formal natural language.

Formal Natural Language

Grammar


To Be
Present | Past | Pres. Perfect
I am | I was | I have been
I am a doctor.
Five years ago, I was student.
I have been tor for 5 years.
I am ill.
I was Monday.
ill for the last 2 days.
married.
was married in 2001.
have been married for 12 years.

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* (and other *newspapers of record*)
- In contrast, informal language is found in email, people's web pages, chat groups, and "trendy" print media.

Outline

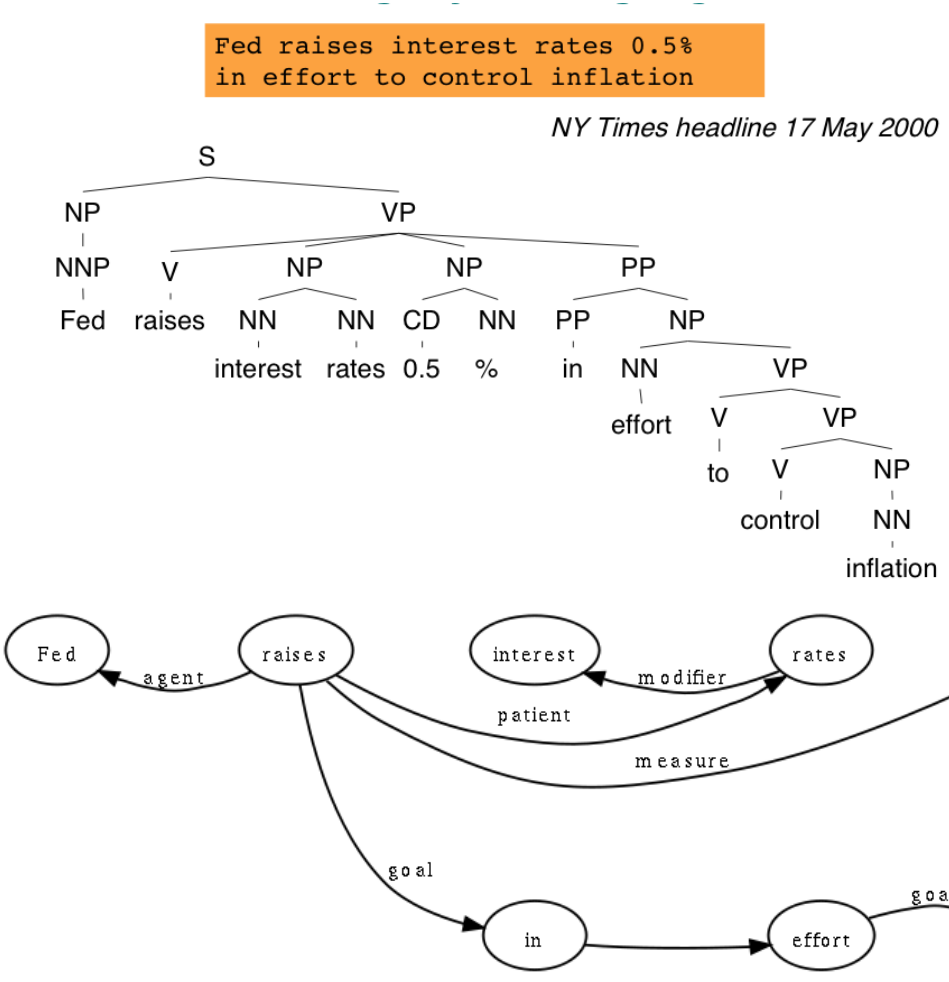
NLP Steps

Sentence segmentation ***Identify sentence boundaries***	Frank met the president. He said: "Hi! What's up – Mr. President?"	**Sentence 1:** Frank met the president. **Sentence 2:** He said: "Hi What's up – Mr. President?"
Tokenization ***Identify word boundaries***	My phone tries to change 'eating' to 'dating'. #hateautocorrect	[My] [phone] [tries] [to] [change] ['] [eating] ['] [to] ['] [dating] ['] [.] [#hateautocorrect]
Stemming/lemmatization	eating, ate, eat	eat, eat, eat
Part-of-Speech tagging	If you build it, he will come	If you build it , he will come IN PRP VBP PRP , PRP MD VB
Parsing	Jon and Frank went into a bar.	(S (NP (NP John) and (NP Frank)) (VP went (PP into (NP a bar))))
Named entity recognition	Let's meet John in DC at 6pm.	Let's meet **John** in **DC** at **6pm**. Pers Loc Time
Co-reference resolution	John drank a beer. He thought it was warm.	**John** drank a **beer**. **He** thought **it** was warm.

Source *Pivotal's Data Science blog*

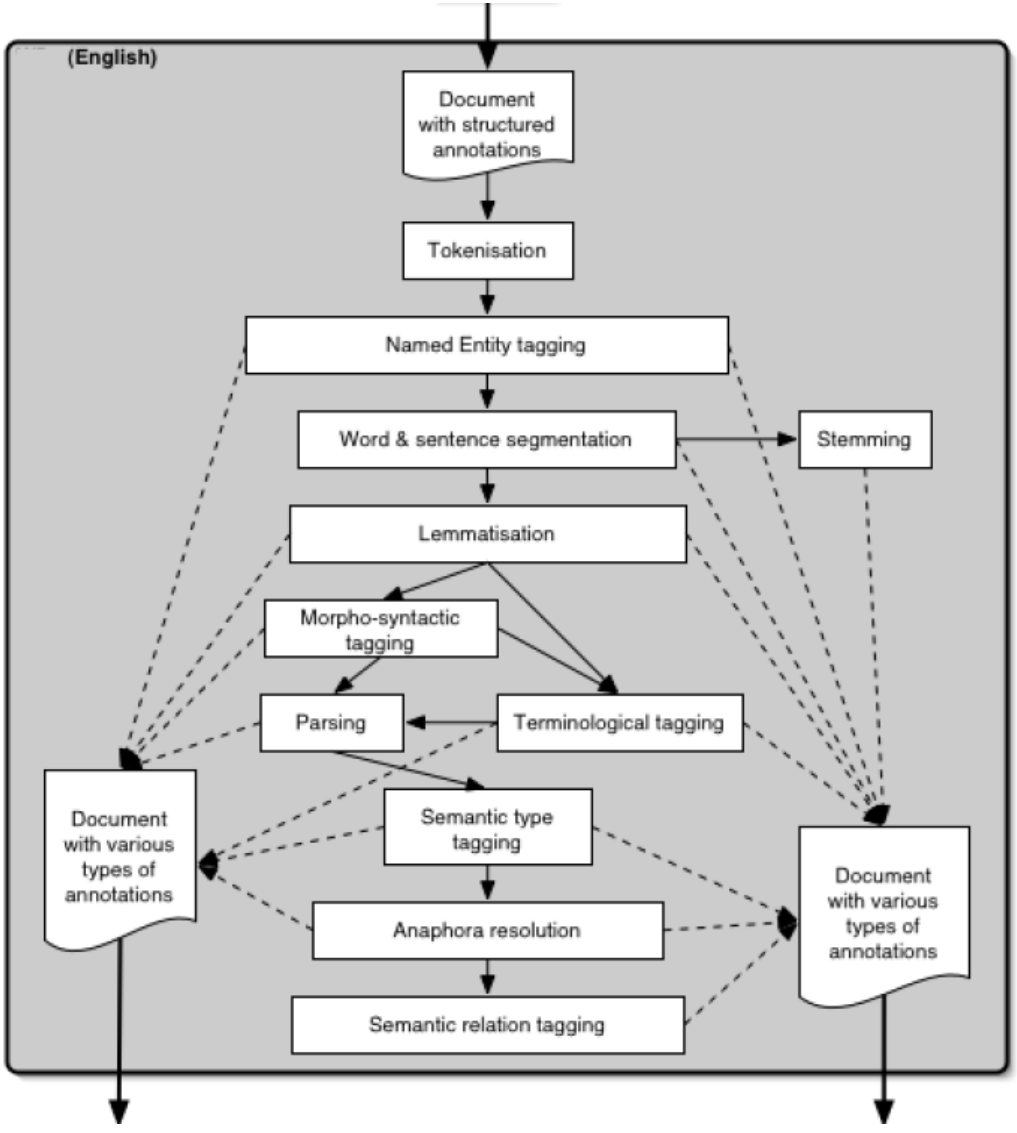
Analysing Language

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 writing 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:

$$koira_{dog} rouka_{food}$$

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 sal

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.

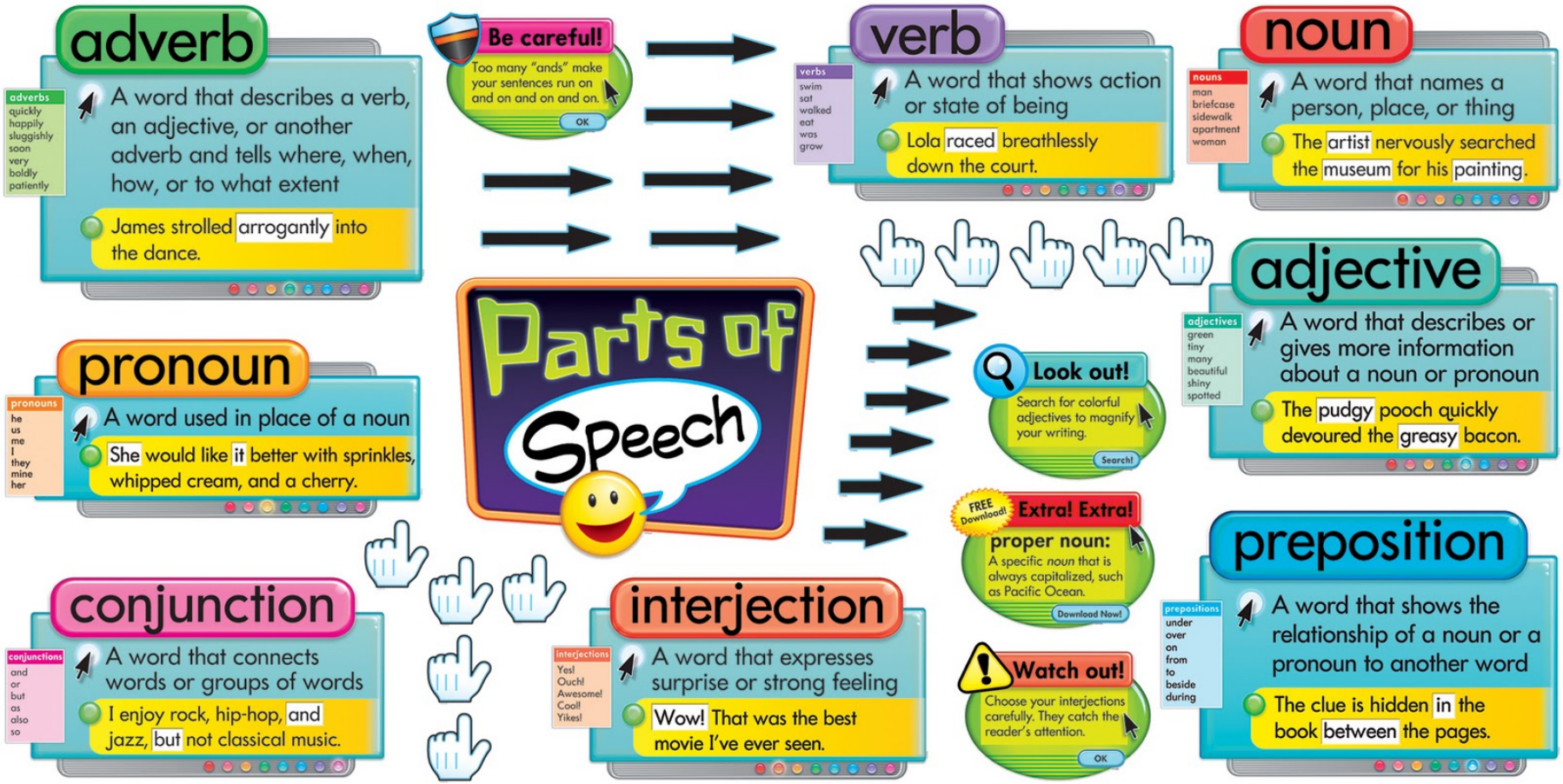
- Ambiguities arise in all processing steps.
- All languages have particular versions of the ambiguity problem.

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 manages uncertainty, for instance, by using latent variables in statistical methods.

Outline

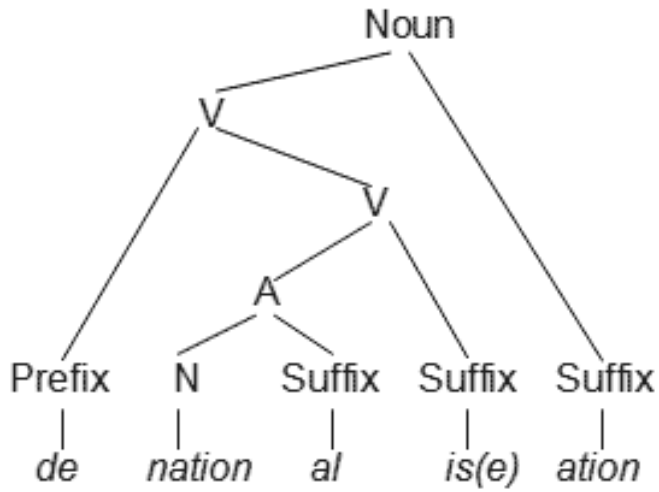
Parts of Speech



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!

Morphology



Source: http://artsfaculty.auckland.ac.nz

- handy when we do not know a word
- or haven't seen enough of the word to infer its semantics

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.

	VBZ	VB VBZ	VBZ			
NNP	NNS	NNS	NNS	CD	NN	
Fed	raises	interest	rates	0.5	%	in effort to control inflation

Collocations

e.g. “hot dog", “with respect to", “home page", “fourth quarter", “run down",

- meaning of collocation different to meaning of its parts:
  - cannot be modified easily without changing the meaning:
    - “kicked the bucket" versus “kicked the tub", “the bucket was kicked"
  - identify collocations by distributional semantics.
- Related: multi-word expression/unit, compound, idiom.
- In some languages, collocations replaced by compounds: “dog food” versus “koirarouka” (Finnish)
- Important for parsing, dictionaries, terminology extraction, ...

Outline

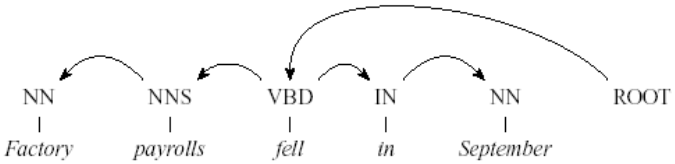
Constituents

Constituents ::= a group of words that functions as a single unit within a hierarchical structure

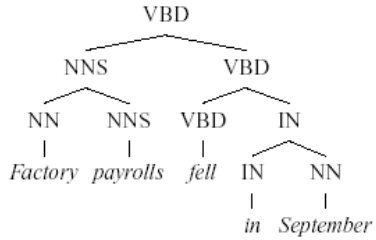
e.g. noun phrase, prepositional phrase, collocation, *etc.*

- 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.

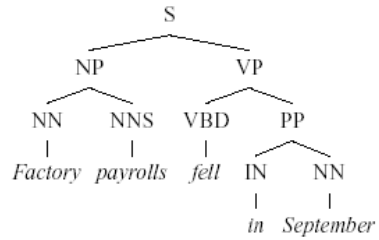
Parsing



(a) Classical Dependency Structure



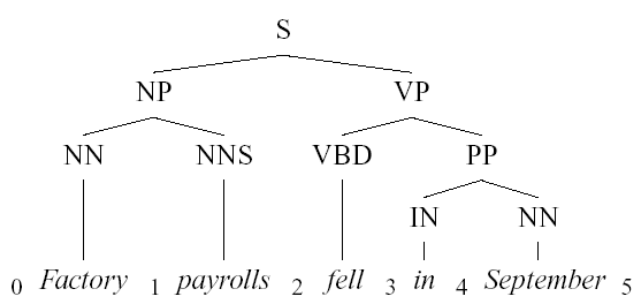
(b) Dependency Structure as CF Tree



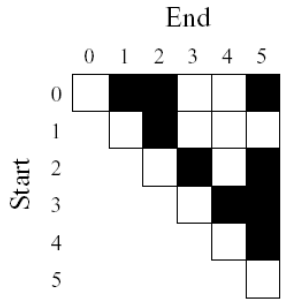
(c) CFG Structure

- a dependency tree, in (a), shows syntactic or semantic relationships
  - we want the relationships labelled.
    e.g. arc from “fell" to “in" labelled with *time*, arc from ”fell" to ”payrolls" labelled with *patient*.
- Context Free Grammar (CFG) gives a parse tree, in (c), using formal linguistic theory
- (b) 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

- full parse yields many subtrees or constituents, labelled verb phrase (VP), prepositional phrase (PP), *etc.*
- recognising the start and end of a particular type of constituent (without parsing) is called shallow parsing or chunking.
- parsing can also be represented as a structured classification problem, as coordinated shallow parsing

Case Frames

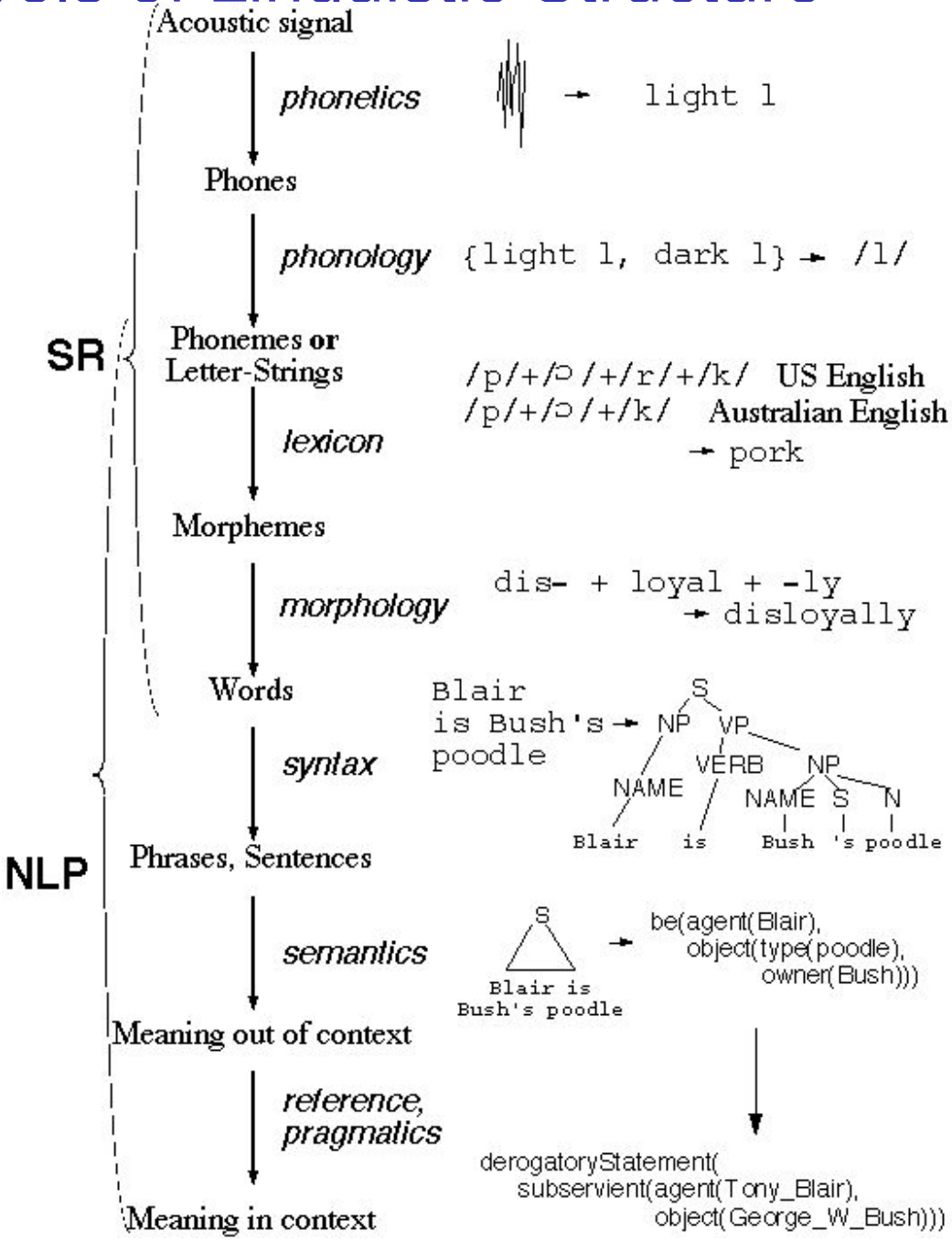
- Example case frames with roles.
  - `actor` “buy” `object` (syntactic)
  - person/organisation “buy” `thing` (semantic)
  - `agent` “fix” `thing`
  - `animate-object` “walks”
- allows mapping of verb syntax to semantics
- give the functional characteristics of a verb
- roles are the argument types in a single position

  e.g. agent, actor, instrument, ...
- various databases:

  e.g. FrameNet, PropBank, VerbNet.

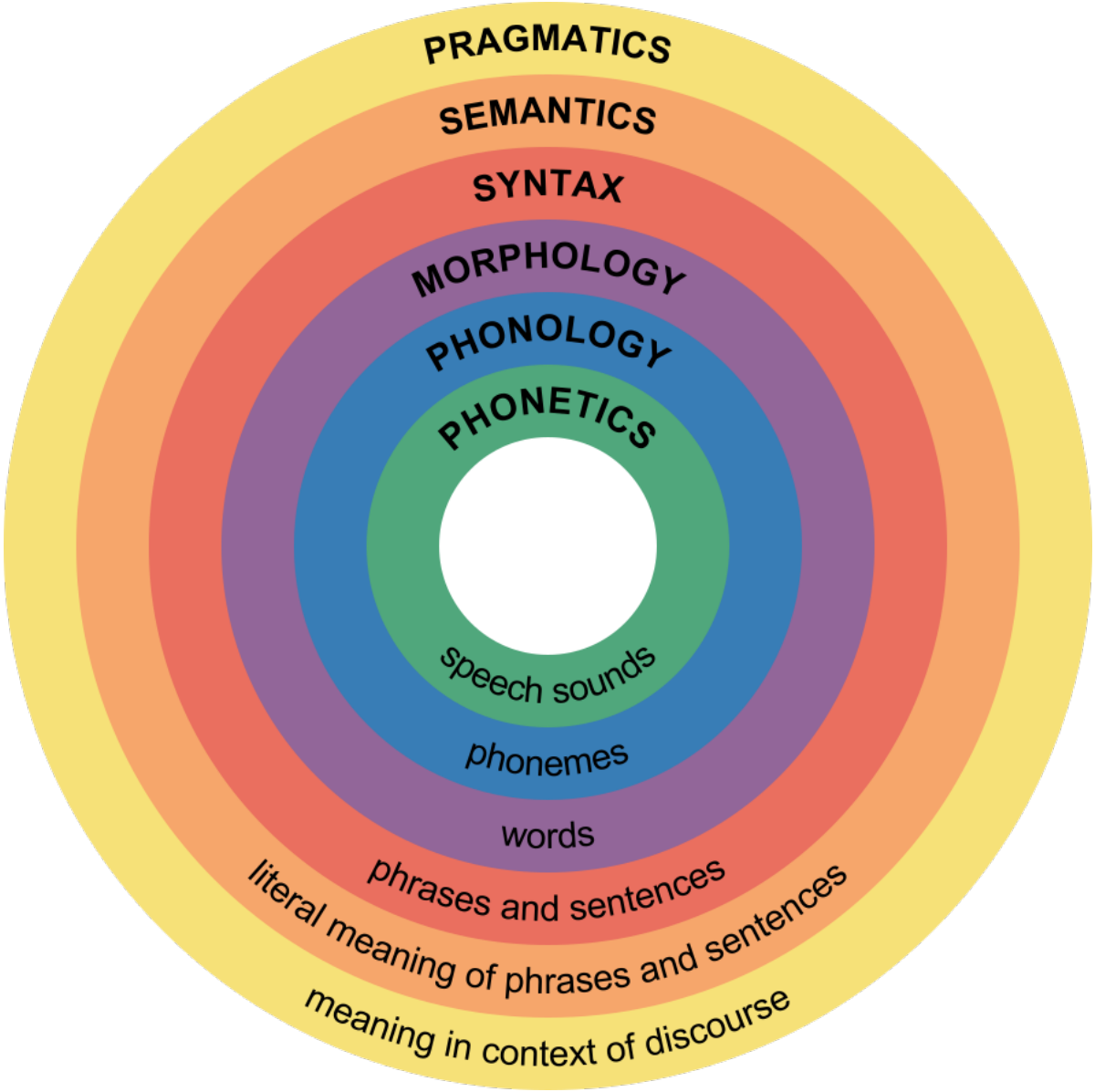
Outline

Levels of Linguistic Structure



from Bill Wilson's unit *Introduction to Natural Language Processing* at UNSW

Levels of Linguistic Structure, cont



by J.J. Thomas and K.A. Cook (Ed.) derivative work: McSush [Public domain], via Wikimedia Commons

History of NLP

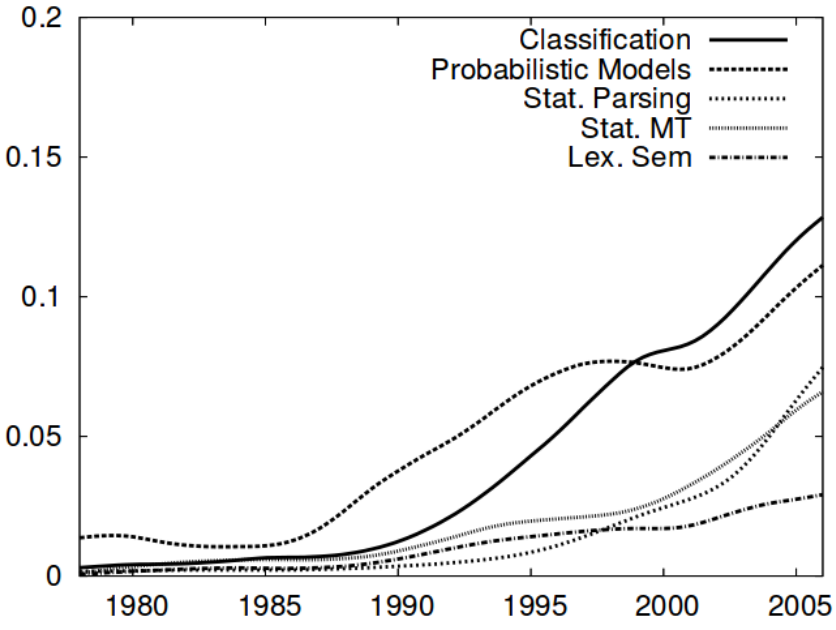


Figure 1: Topics in the ACL Anthology that show a strong recent increase in strength.

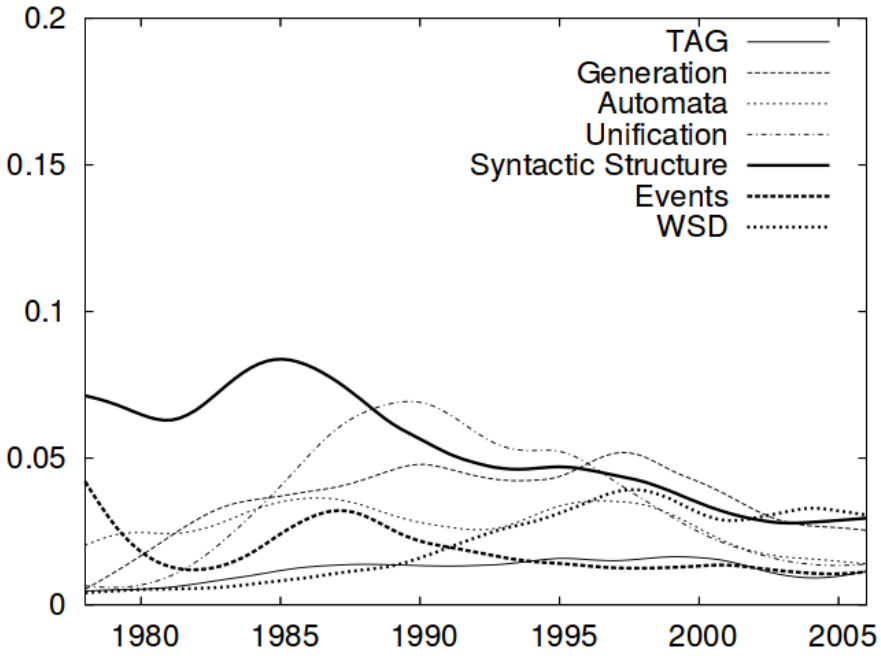


Figure 6: Peaked topics

"Studying the History of Ideas Using Topic Models" Hall, Jurafsaky, Manning, EMNLP 2008

State of the Art

Speech recognition: non-uniform internal-handcrafting Gaussian mixture model/Hidden Markov model (GMM-HMM) technology based on generative models of speech trained discriminatively

Machine translation: statistical language models and statistical translation models using “big data"

parsing: statistical parsing; deep neural networks; massively parallel probabilistic evidence-based architecture (IBM Watson)

Disclaimer: I am no expert in these, just my rough guess!

State of the Art, cont

Non-parametric hierarchical Bayesian methods: document segmentation, collocation recognition, topic models, twitter clustering.

Deep neural networks: sentiment analysis, topic models, language models, parsing.

Both deep neural networks and non-parametric hierarchical Bayesian models allow creative hierarchical structures with flexible estimation.

Outline

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

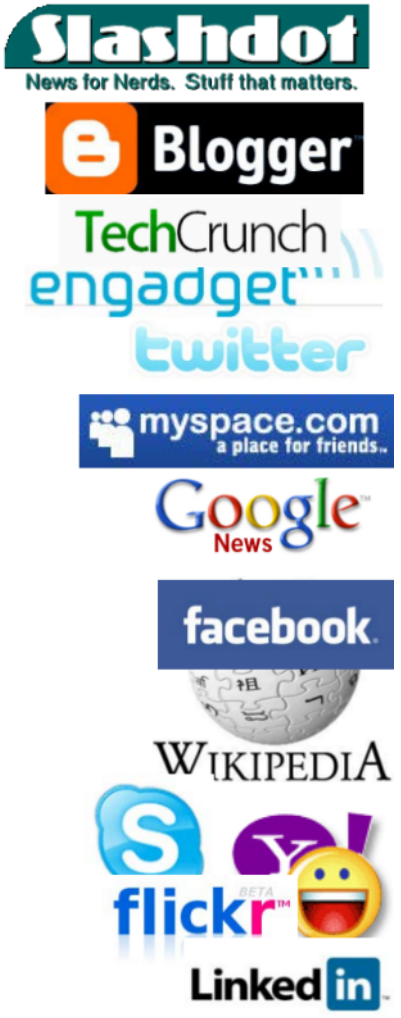
Processing of Documents

- 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

Social Media

- big data, rich text
- opinions, rants, reviews, gossip
- informal and formal language
- sentiment and style
- links and metadata
- linked data
- context and stucture



Informal Language

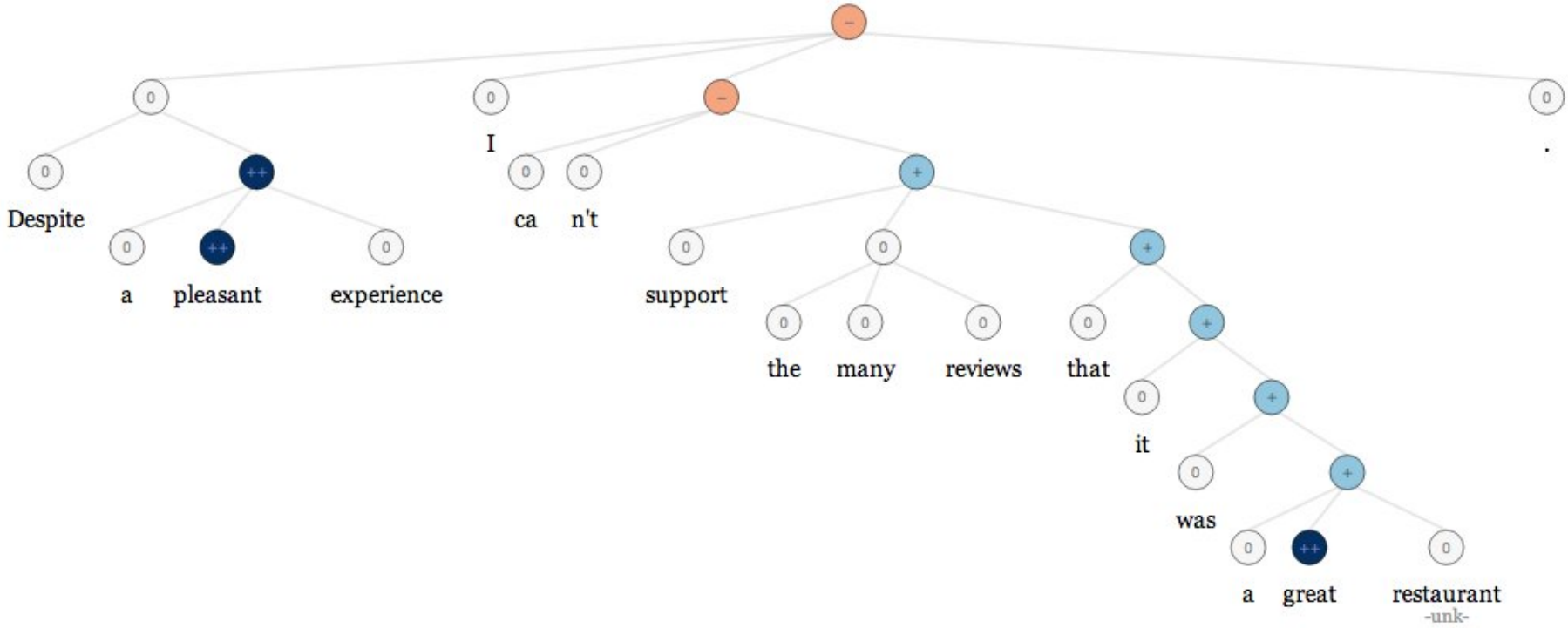
Text messages: My smmr hols wr CWOT. B4, we used 2go2 NY 2C my bro, his GF & thr 3 :- kids FTF. ILNY, it's a gr8 plc.

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

Emotive Language and Sentiment

- “The government will reduce interest rates.” *versus* “The government will slash interest rates.”
- “You are meticulous.” *versus* “You are nitpicking.”
- and sarcasm: “Thanks a lot, HR. I’m unable to access the payroll system!”

Sentiment: Example Parse



from Socher *et al.*, see *Deep Learning for Sentiment Analysis*

Web Page Structure

- web pages have more complicated structures and *genre* than traditional documents
- Genres:
  - product page
  - personal home page
  - FAQ
  - blog
- much of the content templated
- no standard formatting guidelines

Linguistic Resources

- large number of resources available, due to open data, crowdsourcing, and digitisation.
  - *e.g.* gazetteers, dictionaries, annotated text (tagged with POS, name entity types, *etc.*), semantic role data (*i.e.,* for verbs), collocations, aligned translations.
- but correctly annotated and marked up linguistic resources are the hardest to get

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

Unsupervised learning (or semi-supervised) for statistical NLP is most needed.

Linked Open Data (Semantic Web)

- data stored in RDF (triples) and XML interlinked via URIs
- integrated information extraction
- good structured content for semi-supervised training.

Outline

Using Information


NOVEMBER 29TH, 2008

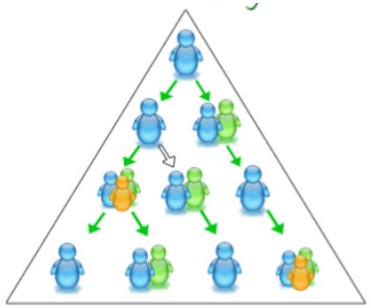
Real Time Citizen Journalism in Mumbai Terrorist Attacks

Welcome to Gauravonomics Blog! Subscribe to my combined feed in a feed reader or by e-mail and you'll never miss a single post. Thanks for visiting!

(The Mumbai terror attack has finally ended after more than 60 hours.

Even as I continue to track instances of citizen journalism in the Mumbai terror attack on this post, I'm trying to make sense of what happened in a work in progress case study and a Flickr set of screenshot on the role of social media in the Mumbai terror attack. I'm

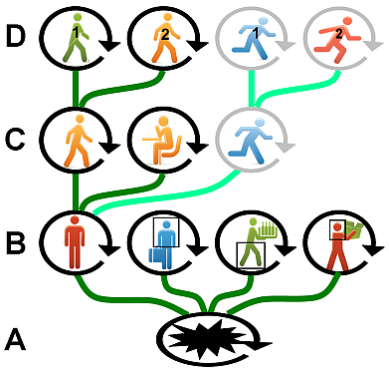
.....real-time citizen journalism



social media marketing .....

..... reputation management

behaviour analysis .....

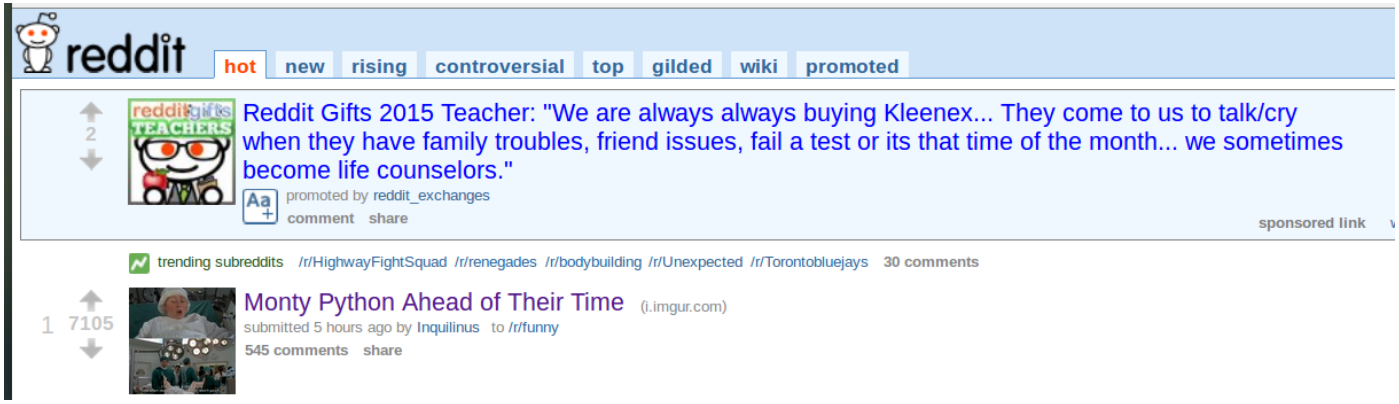


Bioinformatics: Medline

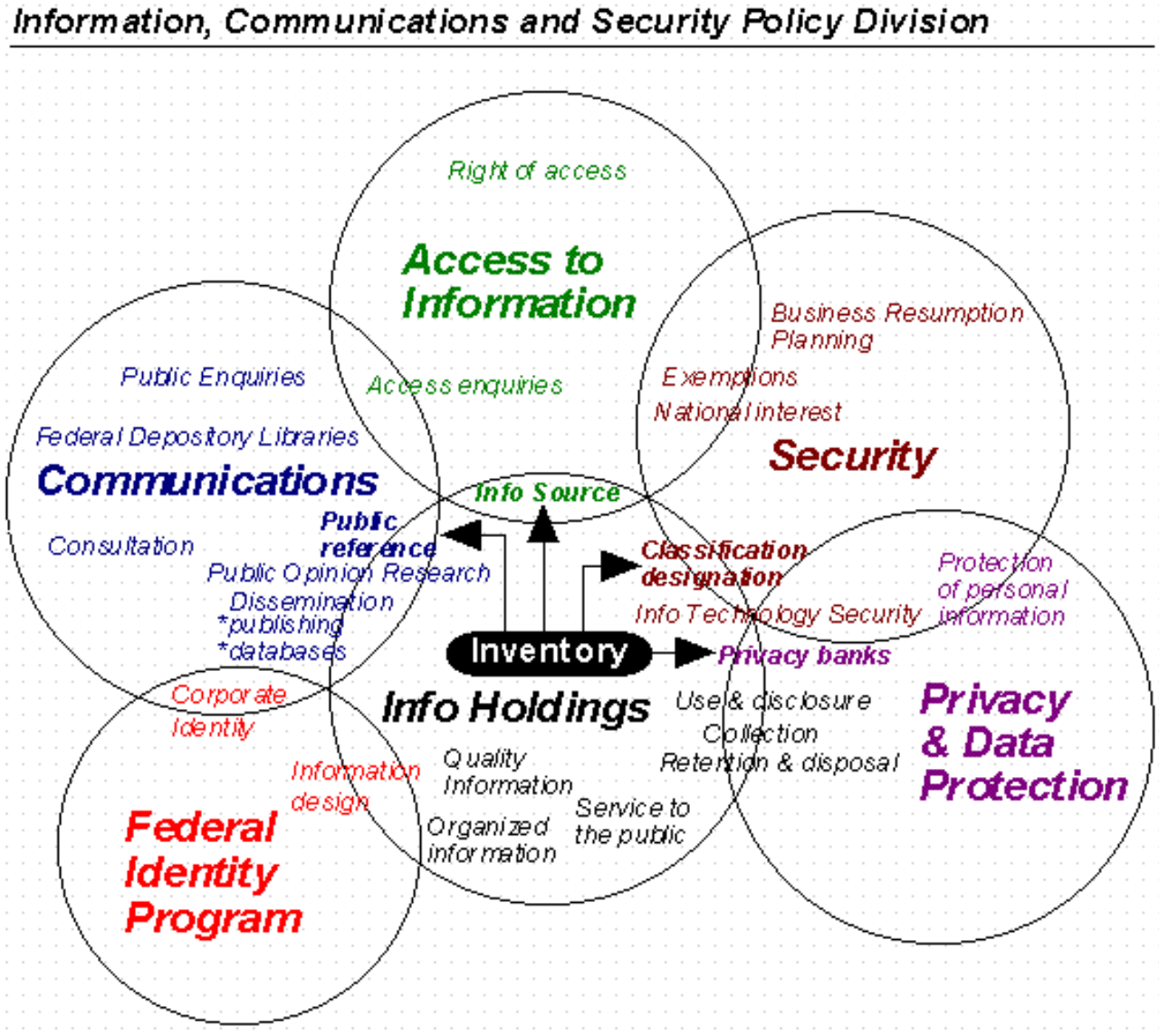
- *PubMed* is the most popular database in Biology, and the main database MedLine has over 16 million entries.
  - entries are abstracts and metadata in *abstract format* *MedLine format*, *XML format*, ...
  - 2,000-4,000 new entries/day from 5000 journals in 37 languages.
- The abstract databases are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

  e.g. *browser* and *text analysis*

Social Bookmarks: Reddit.com



- *Reddit* is one of the best known social bookmarking sites.
- can use tagging to provide higher-weighted keywords
- use social bookmarks to get popularity/“authority” for pages



Federal Government Information Policy, Canada
(all about processing documents)

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.

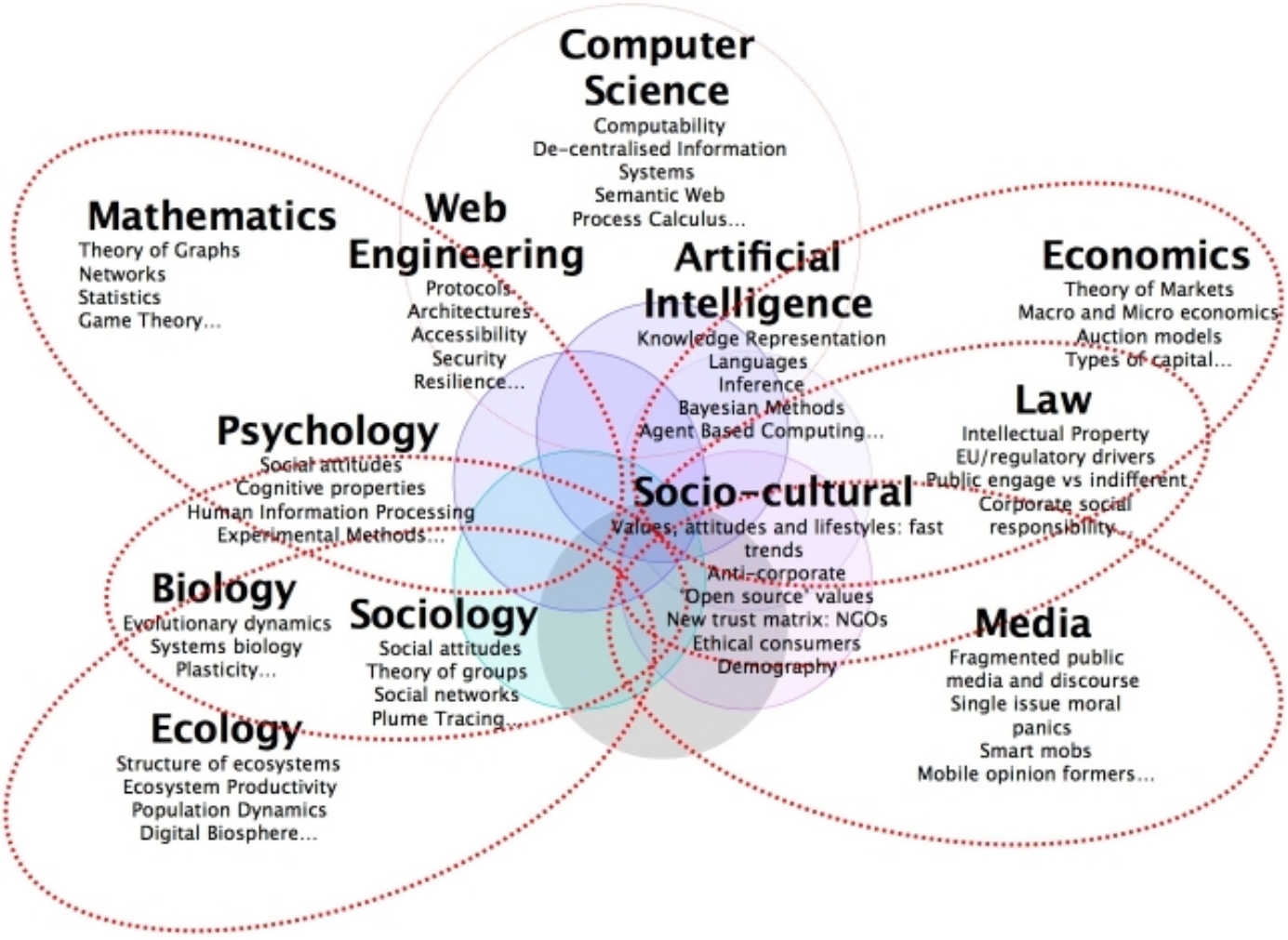
Administration: eGovernment and electronic document processing.

Advertising: many aspects of advertising now running online.

Outline

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

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 language like Python (also Java) 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.

Stanford CoreNLP[2]: Good open source tools, in Java, with sentiment and negation.

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

[2] I use this a lot!

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.

Favorite Websites

EventRegistry.Org: from Jozef Stefan Institute, amazing news analysis service

OpenCalais.com: by Thomson Reuters, bring structure to unstructured content, great demo

Semantria.com: another web-based tagging API. great demo

Idibon: new cloud-based NLP, TBD

Outline