Introducing Document Analysis

Wray Buntine
Monash University
http://topicmodels.org

August 17, 2015
Background, Career


1990 → US → 2002 → Finland → 2007 → Australia

Now: 2014 → MONASH University
Information Technology
Background, Research

- Bayesian machine learning
- inductive logic programming and statistical relational learning
- Bayesian decision trees (inspired Breiman’s “bagging”)
- model averaging and ensembles
- graphical models for machine learning
- machine learning in text, documents and search
- topic models
- non-parametric Bayesian methods
Research: Discrete Bayesian Non-Parametrics

- using Discrete Bayesian Non-Parametrics:
  - Pitman-Yor processes
  - Poisson processes techniques
- on complex discrete structures and networks
- for data like
  - FreeBase
  - WordNet, parse trees
  - citation networks with abstracts

Applied to Machine Learning by folks like Lancelot James (HKUST) and Yee Whye Teh (Oxford).
Research: Non-Parametric Topic Models

vanilla LDA

non-parametric LDA with burstiness
Aaron Li: First CUDA NP-LDA, 2012
Outline

Motivation

Formal Natural Language

Document Processing

Document Analysis
Information Overload

- 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?
“Big Data” is a hot topic in the business world these days. But there’s a subset of this broad field that has yet to take a turn in the spotlight. It’s called “text mining,” and you’re probably going to be hearing a lot more about it over the coming months and years. Basically, text mining is the process of combing through countless pages of plain-language digitized text to find useful information that’s been hiding in plain sight. First developed—as a labor-intensive manual discipline—in the 1980s, text mining has become ever more efficient as computing power has increased. Relevant today to many businesses, the practice nonetheless brings with it as much potential for conflict as opportunity. Which is why we’re going to be hearing more about it.

*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

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>corporate</td>
<td>expert finding, document summarisation</td>
</tr>
<tr>
<td>health care</td>
<td>disease monitoring, nursing notes analysis</td>
</tr>
<tr>
<td>government</td>
<td>records classification, FOI requests, service delivery, policy research</td>
</tr>
<tr>
<td>insurance</td>
<td>problem identification from claims, fraud detection</td>
</tr>
<tr>
<td>legal</td>
<td>search and discovery</td>
</tr>
<tr>
<td>oil and gas</td>
<td>analysis of maintenance and repair logs</td>
</tr>
<tr>
<td>retail</td>
<td>brand or product analysis, customer retention</td>
</tr>
<tr>
<td>security</td>
<td>log analysis</td>
</tr>
</tbody>
</table>
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.
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 (properly) here!
We do a review of the analysis of formal natural language (not a formal analysis of 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 a student.
I have been a doctor for 5 years.
I am ill.
I was sick on Monday.
I have been ill for the last 2 days.

married.
I was married in 2001.
I 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

Motivation

Formal Natural Language

NLP Processing and Ambiguity

Words

Parsing

Overview

Document Processing

Document Analysis
## NLP Steps

<table>
<thead>
<tr>
<th>NLP Step</th>
<th>Example</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentence segmentation</strong></td>
<td>Frank met the president. He said: “Hi! What’s up – Mr. President?”</td>
<td>Pivotal’s Data Science blog</td>
</tr>
<tr>
<td><strong>Tokenization</strong></td>
<td>My phone tries to change ‘eating’ to ‘dating’.</td>
<td></td>
</tr>
<tr>
<td><strong>Stemming/lemmatization</strong></td>
<td>eating, ate, eat</td>
<td></td>
</tr>
<tr>
<td><strong>Part-of-Speech tagging</strong></td>
<td>If you build it, he will come</td>
<td></td>
</tr>
<tr>
<td><strong>Parsing</strong></td>
<td>Jon and Frank went into a bar.</td>
<td></td>
</tr>
<tr>
<td><strong>Named entity recognition</strong></td>
<td>Let’s meet John in DC at 6pm.</td>
<td></td>
</tr>
<tr>
<td><strong>Co-reference resolution</strong></td>
<td>John drank a beer. He thought it was warm.</td>
<td></td>
</tr>
</tbody>
</table>
Analysing Language

Example from McCallum’s NLP course

- Left, a traditional parse tree showing constituent 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.
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.

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

<table>
<thead>
<tr>
<th>drastu</th>
<th>darasat</th>
<th>she studied (feminine)</th>
</tr>
</thead>
<tbody>
<tr>
<td>drastu</td>
<td>darrasat</td>
<td>she taught (feminine)</td>
</tr>
<tr>
<td>drastu</td>
<td>durisat</td>
<td>it was studied (feminine)</td>
</tr>
<tr>
<td>drastu</td>
<td>darrisat</td>
<td>it was taught (feminine)</td>
</tr>
<tr>
<td>drastu</td>
<td>darastu</td>
<td>i studied</td>
</tr>
<tr>
<td>drastu</td>
<td>darrastu</td>
<td>i taught</td>
</tr>
<tr>
<td>drastu</td>
<td>duristu</td>
<td>i was studied</td>
</tr>
<tr>
<td>drastu</td>
<td>durrastu</td>
<td>i was taught</td>
</tr>
<tr>
<td>drastu</td>
<td>darasta</td>
<td>you studied (masculine)</td>
</tr>
<tr>
<td>drastu</td>
<td>durrasta</td>
<td>you taught (masculine)</td>
</tr>
<tr>
<td>drastu</td>
<td>durista</td>
<td>you were studied (masculine)</td>
</tr>
<tr>
<td>drastu</td>
<td>durrasta</td>
<td>you were taught (masculine)</td>
</tr>
<tr>
<td>drastu</td>
<td>darasti</td>
<td>you studied (feminine)</td>
</tr>
<tr>
<td>drastu</td>
<td>darrasti</td>
<td>you taught (feminine)</td>
</tr>
<tr>
<td>drastu</td>
<td>duristi</td>
<td>you were studied (feminine)</td>
</tr>
<tr>
<td>drastu</td>
<td>durristi</td>
<td>you were taught (feminine)</td>
</tr>
</tbody>
</table>
NLP in Arabic, cont.

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

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Arabic</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>با للدرس</td>
<td>بيددرس</td>
<td>With/In the lesson</td>
</tr>
<tr>
<td>للدرس</td>
<td>ليددرس</td>
<td>For/To the lesson</td>
</tr>
<tr>
<td>كالدرس</td>
<td>كيددرس</td>
<td>As the lesson</td>
</tr>
<tr>
<td>فالدرس</td>
<td>ففيددرس</td>
<td>Then the lesson</td>
</tr>
<tr>
<td>فالدرس</td>
<td>فيددرس</td>
<td>In the lesson</td>
</tr>
</tbody>
</table>

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

SAFEK = doubt
SAFAK = clapped
SIPEK = provided
SUPAK = has been provided
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:

\[
\text{On}\_{\text{am}}\ \text{kahvi}_{\text{coffee}}\ \text{la}_{\text{place}}\ \text{ssa}_{\text{in}}\ \text{han}_{\text{emphasis}}\ .
\]

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:

\[
\text{koira}_{\text{dog}}\text{rouka}_{\text{food}}
\]
Translation Difficulties

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

<table>
<thead>
<tr>
<th>Code</th>
<th>Language</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>English</td>
<td>Saddam Hussein</td>
</tr>
<tr>
<td>LV</td>
<td>Latvian</td>
<td>Sadams Huseins</td>
</tr>
<tr>
<td>HU</td>
<td>Hungarian</td>
<td>Szaddám Huszein</td>
</tr>
<tr>
<td>ET</td>
<td>Estonian</td>
<td>Saddäm Husayn</td>
</tr>
</tbody>
</table>
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 were on sale.

The correct analysis is:

Clothes made of hemp and smoking paraphernalia were on sale.

This misinterpretation is a common semantic problem with current parsing technology.
Language Ambiguities, cont.

- **New** adjective **York Tennis Club** name opening today. versus **New York Tennis Club** name opening today.

- He worked at **Yahoo!** sentence Tuesday. versus He worked at **Yahoo!** name Tuesday.

- Stolen painting found by **tree** location. versus Stolen painting found by **tree** actor.

- Iraqi **head** body part seeks **arms** body part. versus Iraqi **head** politician seeks **arms** 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 carry uncertainty, for instance, by using latent variables in statistical methods.
Outline

Motivation

Formal Natural Language
  NLP Processing and Ambiguity
  Words
  Parsing
  Overview

Document Processing

Document Analysis
Parts of Speech

**adverb**
A word that describes a verb, an adjective, or another adverb and tells where, when, how, or to what extent.
- James strolled arrogantly into the dance.

**verb**
A word that shows action or state of being.
- Lola raced breathlessly down the court.

**noun**
A word that names a person, place, or thing.
- The artist nervously searched the museum for his painting.

**adjective**
A word that describes or gives more information about a noun or pronoun.
- The pudgy pook quickly devoured the greasy bacon.

**pronoun**
A word used in place of a noun.
- She would like it better with sprinkles, whipped cream, and a cherry.

**conjunction**
A word that connects words or groups of words.
- I enjoy rock, hip-hop, and jazz, but not classical music.

**interjection**
A word that expresses surprise or strong feeling.
- Wow! That was the best movie I’ve ever seen.

**preposition**
A word that shows the relationship of a noun or a pronoun to another word.
- The clue is hidden in the book between the pages.
# Word Classes (dictionary version of part of speech)

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

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

<table>
<thead>
<tr>
<th>Fed</th>
<th>raises</th>
<th>interest rates</th>
<th>0.5 %</th>
<th>in effort to control inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP</td>
<td>VBZ</td>
<td>NNS</td>
<td>NNS</td>
<td>CD</td>
</tr>
</tbody>
</table>
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

Motivation

Formal Natural Language
  NLP Processing and Ambiguity
  Words
  Parsing

Overview

Document Processing

Document Analysis
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 grammatically 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 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

- 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

Motivation

Formal Natural Language
  NLP Processing and Ambiguity
  Words
  Parsing
  Overview

Document Processing

Document Analysis
Levels of Linguistic Structure

Acoustic signal → phonetics → light 1

Phones

phonology \{light 1, dark 1\} → /l/

Phonemes or Letter-Strings

US English /p/+/θ/+/r/+/k/ → pork

Australian English /p/+/θ/+/k/ → pork

lexicon

Morphemes

morphology dis- + loyal + -ly → disloyally

Words Blair is Bush's poodle

syntax

Phrases, Sentences

Blair is Bush's poodle

semantics

Meaning out of context

reference, pragmatics

derogatoryStatement(subservient(agent(Tony_Blair), object(George_W_Bush))))

Meaning in context

from Bill Wilson's unit Introduction to Natural Language Processing at UNSW
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, Jurafsky, 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!
Document segmentation:
Collocation recognition:
Topic models:
Twitter clustering: non-parametric hierarchical Bayesian methods
Sentiment analysis:
Language models: deep neural networks

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

Motivation

Formal Natural Language

Document Processing

Language in the Electronic Age

Why Analyse Documents

Document Analysis
Social Media

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

Text messages: My smmr hols wr CWOT. B4, we used 2 go2 NY 2 C
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.”
▶ “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

<table>
<thead>
<tr>
<th>Availability of linguistic resources is a key determining factor in the success of statistical NLP projects.</th>
</tr>
</thead>
</table>

**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

Motivation

Formal Natural Language

Document Processing

Language in the Electronic Age

Why Analyse Documents

Document Analysis
Using Information

Real Time Citizen Journalism in Mumbai Terrorist Attacks

.............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 controlled 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.


Advertising: many aspects of advertising now running online.
Outline

Motivation

Formal Natural Language

Document Processing

Document Analysis
  Representation
  Resources
  Other Areas

We sketch out the field of document analysis, with major emphasis on text.
From Web Science.
Outline

Motivation

Formal Natural Language

Document Processing

Document Analysis

Representation

Resources

Other Areas
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 $\rightarrow$ tokens $\rightarrow$ sentences $\rightarrow$ paragraphs $\rightarrow$ documents.

**Syntactic level:** morphemes $\rightarrow$ lemmas and parts of speech $\rightarrow$ collocations, terms and named entities $\rightarrow$ constituents, phrases $\rightarrow$ 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

Motivation

Formal Natural Language

Document Processing

Document Analysis

Representation

Resources

Other Areas
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.
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.

<table>
<thead>
<tr>
<th>Headword</th>
<th>Pronunciation</th>
<th>Morphology</th>
<th>Cl</th>
<th>Type</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>celebrant</td>
<td>&quot;sE-ll-br@nt&quot;</td>
<td>((celebrate),(ant))</td>
<td>N</td>
<td>sing</td>
<td>6</td>
</tr>
<tr>
<td>cellarages</td>
<td>&quot;sE-l@-rldZls&quot;</td>
<td>((cellar),(age),(s))</td>
<td>N</td>
<td>plu</td>
<td>0</td>
</tr>
<tr>
<td>cellular</td>
<td>&quot;sEl-jU-l@r*&quot;</td>
<td>((cell),(ular))</td>
<td>A</td>
<td>pos</td>
<td>21</td>
</tr>
</tbody>
</table>
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:** 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.
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 of 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

Motivation

Formal Natural Language

Document Processing

Document Analysis