

Natural Language Processing and Computational Linguistics: from Theory to Application

Professor Mark Johnson

Director of the Macquarie Centre for Language Sciences (CLaS)
Department of Computing

October 2012

Computational linguistics and Natural language processing

- *Computational linguistics* is a *scientific discipline* that studies *linguistic processes* from a *computational perspective*
 - ▶ language comprehension (computational psycho-linguistics)
 - ▶ language production
 - ▶ language acquisition
- *Natural language processing* is an *engineering discipline* that uses computers to do *useful things with language*
 - ▶ information retrieval
 - ▶ topic detection and document clustering
 - ▶ document summarisation
 - ▶ sentiment analysis
 - ▶ machine translation
 - ▶ speech recognition

Outline

A crash course in linguistics

Machine learning and data mining

Brief survey of NLP applications

Word segmentation and topic models

Conclusion

Phonetics, phonology and the lexicon

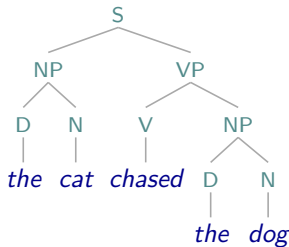
- *Phonetics* studies the *sounds* of a language
 - ▶ E.g., English *aspirates stop consonants* in certain positions
- *Phonology* studies the *distributional properties of these sounds*
 - ▶ E.g., the English noun plural is [s] following unvoiced segments and [z] following voiced segments
- A language has a *lexicon*, which lists for each word and morpheme
 - ▶ how it is pronounced (phonology)
 - ▶ what it means (semantics)
 - ▶ its distributional properties (morphology and syntax)

Learning the lexicon

- Speech does not have pauses in between words
 - ⇒ children have to *learn* how to segment utterances into words
- As part of an ARC project, we've built a computational model that performs word segmentation using:
 - ▶ utterance boundaries
 - ▶ non-linguistic context
 - ▶ *syllable structure* and *rhythmic patterns*
 - language-specific ⇒ *must be learned*
 - ▶ *other known words* and *longer-range linguistic context*
- With our model, we can
 - ▶ measure the contribution of each information source
 - ▶ predict the effect of changing the percept or changing the input

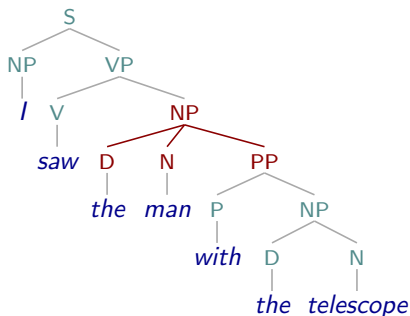
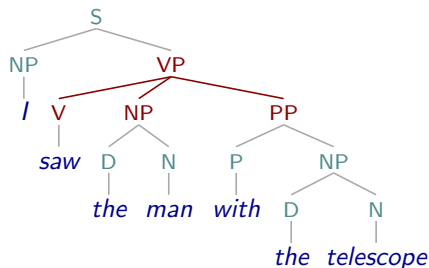
Morphology and syntax

- *rich hierarchical structure* is pervasive in language
- *morphology* studies the structure of words
 - ▶ E.g., *re+structur+ing*, *un+remark+able*
- *syntax* studies the ways words combine to form phrases and sentences
 - ▶ phrase structure helps identify *who did what to whom*



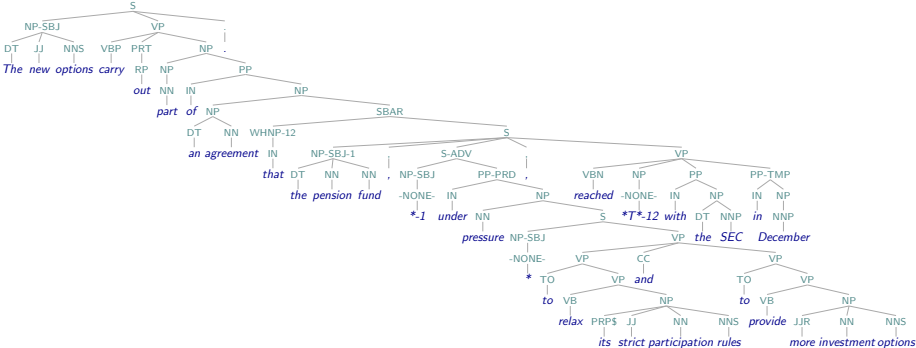
Parsing identifies phrase structure

- *Ambiguity* is pervasive in human languages



- Recover English phrase structure with *over 90% accuracy*
- We have an ARC project to parse running speech
 - ▶ coupled with a speech recogniser
 - ▶ our models are *robust* against *speech disfluencies*

Phrase structure of real sentences



Semantics and pragmatics

- *Semantics* studies the *meaning of words, phrases and sentences*
 - ▶ E.g., *I ate the oysters in/for an hour.*
 - ▶ E.g., *Who do you want to talk to \emptyset /him?*
- *Pragmatics* studies how we use language to *do things in the world*
 - ▶ E.g., *Can you pass the salt?*
 - ▶ E.g., a letter of recommendation:
Sam is punctual and extremely well-groomed.

Outline

A crash course in linguistics

Machine learning and data mining

Brief survey of NLP applications

Word segmentation and topic models

Conclusion

The “statistical revolution”

- Hand-written rule-based approach
 - ▶ linguist crafts patterns or rules to solve problem
 - ▶ complicated and expensive to construct
 - ▶ hand-written systems are often *brittle*
- Statistical “machine learning” approach
 - ▶ collect statistics from large corpora
 - ▶ combine a variety of information sources using machine-learning
 - ▶ statistical models tend to be easier to maintain and less brittle
- Statistical models of language are *scientifically interesting*
 - ▶ humans are very sensitive to statistical properties
 - ▶ statistical models make *quantitative predictions*

Machine learning vs. statistical analysis

- Machine learning and statistical analysis often use similar mathematical models
 - ▶ E.g., linear models, least squares, logistic regression
- The *goals* of statistical analysis and machine learning are different
- Statistical analysis:
 - ▶ goal is *hypothesis testing* or *identifying predictors*
 - E.g., *Does coffee cause cancer?*
 - ▶ size of data and number of factors is small (thousands)
- Machine learning:
 - ▶ goal is *prediction* by *generalising from examples*
 - E.g., *Will person X get cancer?*
 - ▶ size of data and number of factors can be huge (billions)
 - let learning algorithm decide which factors to use

Outline

A crash course in linguistics

Machine learning and data mining

Brief survey of NLP applications

Word segmentation and topic models

Conclusion

Named entity recognition

- *Named entity recognition* finds the named entities and their classes in a text

- ▶ Example:

Sam Spade bought 300 shares in Acme Corp in 2006

PERSON NUMBER CORPORATION TIME

Noun phrase coreference

- *Noun phrase coreference* tracks mentions to entities within documents
 - ▶ Example:
Julia Gillard met with the president of Indonesia yesterday. Ms. Gillard told him that she . . .
- *Cross-document coreference* identifies mentions to same entity in different documents
- We're doing this on speech data as part of our ARC project

Relation extraction

- *Relation extraction* mines texts to find instances of specific *relationships between named entities*

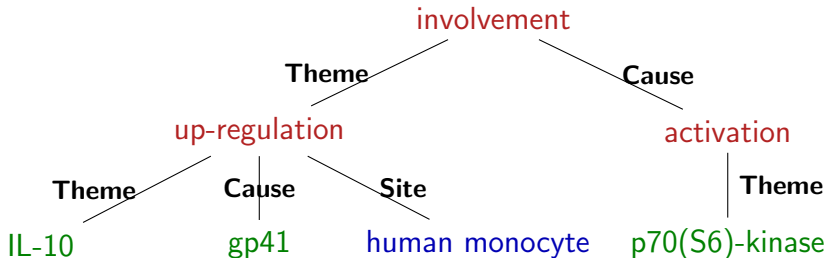
Name	Role	Organisation
<i>Steven Schwartz</i>	<i>Vice Chancellor</i>	<i>Macquarie University</i>
...

- Has been applied to mining bio-medical literature

Event extraction and role labelling

- *Event extraction* identifies the events described by a text
- *Role labelling* identifies “who did what to whom”
 - ▶ Example:

Involvement of *p70(S6)-kinase activation* in *IL-10 up-regulation* in *human monocytes* by *gp41* envelope protein of *human immunodeficiency virus type 1* ...



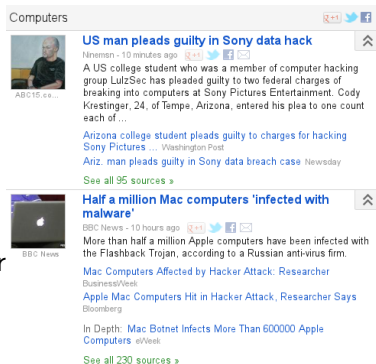
Opinion mining and sentiment analysis

- Used to analyse social media (Web 2.0)
- Classify message along a *subjective–objective scale*
- Identify *polarity* of message
 - ▶ in some genres, simple *keyword-based approaches* work well
 - ▶ but in others it's necessary to model *syntactic structure* as well
 - E.g., *I doubt she had a very good experience . . .*
- Often combined with
 - ▶ *topic modelling* to cluster messages with similar opinions
 - ▶ multi-document *summarisation* to present comprehensible results
- We're currently applying this to *financial announcements* with the CMCRC

Topic models for document processing

- Topic models *cluster documents into one or more topics*
 - ▶ usually *unsupervised* (i.e., topics aren't given in training data)
- Important for document analysis and information extraction
 - ▶ Example: clustering news stories for information retrieval
 - ▶ Example: tracking evolution of a research topic over time

Computers



US man pleads guilty in Sony data hack
NineMSN - 10 minutes ago

A US college student who was a member of computer hacking group LulzSec has pleaded guilty to two federal charges of breaking into computers at Sony Pictures Entertainment. Cody Krestinger, 24, of Tempe, Arizona, entered his plea to one count each of ...

[Arizona college student pleads guilty to charges for hacking Sony Pictures ...](#) Washington Post

[Ariz. man pleads guilty in Sony data breach case](#) Newsday

[See all 95 sources >](#)

Half a million Mac computers 'infected with malware'
BBC News - 10 hours ago

More than half a million Apple computers have been infected with the Flashback Trojan, according to a Russian anti-virus firm.

[Mac Computers Affected by Hacker Attack: Researcher](#) BusinessWeek

[Apple Mac Computers Hit in Hacker Attack, Researcher Says](#) Bloomberg

In Depth: [Mac Botnet Infects More Than 600000 Apple Computers](#) eWeek

[See all 230 sources >](#)

Topic modelling task

- Given just a *collection of documents*, simultaneously identify:
 - ▶ which topic(s) each document discusses
 - ▶ the words that are characteristic of each topic
- Example: TASA collection of *37,000 passages* on language and arts, social studies, health, sciences, etc.

Topic 247		Topic 5		Topic 43		Topic 56	
word	prob.	word	prob.	word	prob.	word	prob.
DRUGS	.069	RED	.202	MIND	.081	DOCTOR	.074
DRUG	.060	BLUE	.099	THOUGHT	.066	DR.	.063
MEDICINE	.027	GREEN	.096	REMEMBER	.064	PATIENT	.061
EFFECTS	.026	YELLOW	.073	MEMORY	.037	HOSPITAL	.049
BODY	.023	WHITE	.048	THINKING	.030	CARE	.046
MEDICINES	.019	COLOR	.048	PROFESSOR	.028	MEDICAL	.042
PAIN	.016	BRIGHT	.030	FELT	.025	NURSE	.031
PERSON	.016	COLORS	.029	REMEMBERED	.022	PATIENTS	.029
MARIJUANA	.014	ORANGE	.027	THOUGHTS	.020	DOCTORS	.028
LABEL	.012	BROWN	.027	FORGOTTEN	.020	HEALTH	.025
ALCOHOL	.012	PINK	.017	MOMENT	.020	MEDICINE	.017
DANGEROUS	.011	LOOK	.017	THINK	.019	NURSING	.017
ABUSE	.009	BLACK	.016	THING	.016	DENTAL	.015
EFFECT	.009	PURPLE	.015	WONDER	.014	NURSES	.013
UNKNOWN	.008	CROSS	.011	FORGET	.012	PHYSICIAN	.012
PILLS	.008	COLORS	.009	RECALL	.012	HOSPITALS	.011



Mixture versus admixture topic models

- In a *mixture model*, each document has a *single topic*
 - ▶ all words in the document come from this topic
- In *admixture models*, each document has a *distribution over topics*
 - ▶ a single document can have multiple topics (number of topics in a document controlled by prior)
 - ⇒ can capture more complex relationships between documents than a mixture model
- Both mixture and admixture topic models typically use a “*bag of words*” representation of a document

Example: documents from NIPS corpus

Annotating an unlabeled dataset is one of the bottlenecks in using supervised learning to build good predictive models. Getting a dataset labeled by experts can be expensive and time consuming. With the advent of crowdsourcing services ...

The task of recovering intrinsic images is to separate a given input image into its material-dependent properties, known as reflectance or albedo, and its light-dependent properties, such as shading, shadows, specular highlights, ...

In each trial of a standard visual short-term memory experiment, subjects are first presented with a display containing multiple items with simple features (e.g. colored squares) for a brief duration and then, after a delay interval, their memory for ...

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some ...

Example (cont): ignore function words

Annotating an unlabeled dataset is one of the bottlenecks in using supervised learning to build good predictive models. Getting a dataset labeled by experts can be expensive and time consuming. With the advent of crowdsourcing services ...

The task of recovering intrinsic images is to separate a given input image into its material-dependent properties, known as reflectance or albedo, and its light-dependent properties, such as shading, shadows, specular highlights, ...

In each trial of a standard visual short-term memory experiment, subjects are first presented with a display containing multiple items with simple features (e.g. colored squares) for a brief duration and then, after a delay interval, their memory for ...

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some ...

Example (cont): mixture topic model

Annotating an unlabeled dataset is one of the bottlenecks in using supervised learning to build good predictive models. Getting a dataset labeled by experts can be expensive and time consuming. With the advent of crowdsourcing services ...

The task of recovering intrinsic images is to separate a given input image into its material-dependent properties, known as reflectance or albedo, and its light-dependent properties, such as shading, shadows, specular highlights, ...

In each trial of a standard visual short-term memory experiment, subjects are first presented with a display containing multiple items with simple features (e.g. colored squares) for a brief duration and then, after a delay interval, their memory for ...

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some ...

Example (cont): admixture topic model

Annotating an unlabeled dataset is one of the bottlenecks in using supervised learning to build good predictive models. Getting a dataset labeled by experts can be expensive and time consuming. With the advent of crowdsourcing services ...

The task of recovering intrinsic images is to separate a given input image into its material-dependent properties, known as reflectance or albedo, and its light-dependent properties, such as shading, shadows, specular highlights, ...

In each trial of a standard visual short-term memory experiment, subjects are first presented with a display containing multiple items with simple features (e.g. colored squares) for a brief duration and then, after a delay interval, their memory for ...

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some ...

Finding topics in document collections

- If we're not told word–topic and document–topic mappings, this is an *unsupervised learning* problem
- Can be solved using *Bayesian inference* with “sparse prior”
 - ▶ most documents discuss few topics
 - ▶ most topics have a small vocabulary
- Simple iterative learning algorithm
 - ▶ randomly assign words to topics
 - ▶ repeat until converged:
 - assign topics to documents based on word-topic assignments
 - assign words to topics based on document-topic assignments
- *Nothing language-specific* \Rightarrow these models can be applied to other domains
 - ▶ search for a *hidden causes* in a sea of data

Outline

A crash course in linguistics

Machine learning and data mining

Brief survey of NLP applications

Word segmentation and topic models

Conclusion

Unsupervised word segmentation

- Input: phoneme sequences with sentence boundaries
- Task: identify *word boundaries*, and hence *words*

j △ u ▲ w △ a △ n △ t ▲ t △ u ▲ s △ i ▲ ð △ ə ▲ b △ u △ k

ju want tu si ðə bʊk

“you want to see the book”

- Ignoring phonology and morphology, this involves learning the pronunciations of the lexical items in the language

Mapping words to referents



- Input to learner:
 - ▶ word sequence: *Is that the pig?*
 - ▶ objects in nonlinguistic context: DOG, PIG
- Learning objectives:
 - ▶ identify utterance topic: PIG
 - ▶ identify word-topic mapping: *pig* \rightsquigarrow PIG

Word learning as a kind of topic modelling

- Learning to map words to referents is like topic modelling: we identify referential words by clustering them with their contexts
 - Word learning also involves finding *sequences* of phonemes (word pronunciations) that cluster with a context
 - Idea: apply the word learning model with words (rather than phonemes) as the basic units
- ⇒ An extended topic model that learns *topical multi-word expressions*
- Currently negotiating a contract with NICTA to develop this idea

Learning topical multi-word expressions

Annotating an unlabeled dataset is one of the bottlenecks in using supervised learning to build good predictive models. Getting a dataset labeled by experts can be expensive and time consuming. With the advent of crowdsourcing services ...

The task of recovering intrinsic images is to separate a given input image into its material-dependent properties, known as reflectance or albedo, and its light-dependent properties, such as shading, shadows, specular highlights, ...

In each trial of a standard visual short-term memory experiment, subjects are first presented with a display containing multiple items with simple features (e.g. colored squares) for a brief duration and then, after a delay interval, their memory for ...

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some ...

Outline

A crash course in linguistics

Machine learning and data mining

Brief survey of NLP applications

Word segmentation and topic models

Conclusion

Conclusion

- *Computational linguistics* studies language *production*, *comprehension* and *acquisition*
 - ▶ what information is used, and how is it used?
 - ▶ may give insights into language disorders, and suggest possible therapies
- *Natural language processing* uses computers to process speech and texts
 - ▶ *information retrieval, extraction and summarisation*
 - ▶ machine translation
 - ▶ human-computer interface
- Statistical models and machine learning play a central role in both
- Theory and practical applications interact in a productive way!