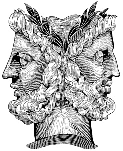
Computational Linguistics: Past, Present and Future

Mark Johnson
Department of Computing
Macquarie University

Australian Language Technology Association
December 2015

Tension between Science and Engineering



- *Engineering applications* (Natural Language Processing):
  - machine translation
  - speech recognition (automatic transcription)
  - information extraction and summarisation
  - *human-computer interaction* (e.g., question-answering)
- *Scientific side* (Computational Linguistics):
  - computation is the *manipulation of meaning-bearing symbols* in ways that respect their meaning
  - studies language comprehension, production and *acquisition* as *computational processes*

Why *computational* linguistics?

- Computers have revolutionised many areas of science
- Language is *computational* in a way that e.g., geology or gastroenterology aren't
  - *computation* is the manipulation of meaning-bearing symbols in ways that respect their meaning

  ⇒ *computation* is a *process*

⇒ Computational linguistics can contribute to scientific study of linguistic *processes*
  - *psycholinguistics*, which studies *human sentence comprehension and production*
  - *language acquisition*, which studies *how human children learn language*
  - *neurolinguistics*, which studies *how language is instantiated in the brain*

Outline

The Past

The Present

The Future

Conclusion

Machine Translation

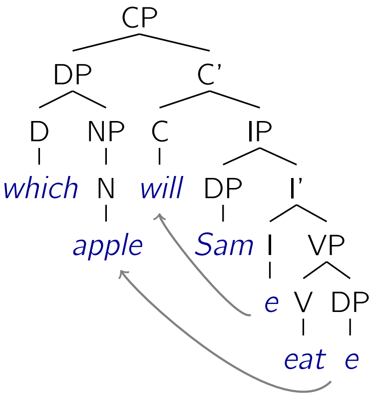
Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography — methods which I believe succeed even when one does not know what language has been coded — one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography.

When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

– Warren Weaver (1947)

The Cognitive Revolution

- The mind as a computer
- Chomsky's *generative grammars*
  - finite number of rules generate an infinite number of sentences
  - conflict between symbolic and statistical approaches



- Provided first formal descriptions of e.g., English auxiliary system

  Could Sam have been eating an apple?

Montague and Compositional Semantics

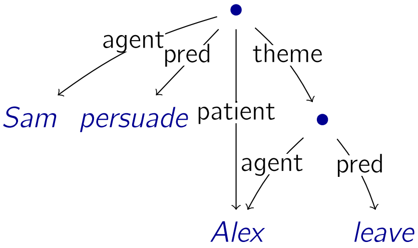
- Compositional semantics: the meaning of a phrase is a function of the meanings of its parts
- Montague extended *lambda calculus* to explain:
  - *quantification:* e.g., *A woman gives birth to a child every minute in India. We have to find her and stop her.*
  - *temporal expressions:* e.g., *The temperature is 90 and rising.*

$\Rightarrow$ Division of labour in computational linguistics:

  - linguists figure out the grammar of a language
  - computational linguists implement the grammar

Unification grammars

- Linguistic theories designed to be computationally tractable
- Syntactic structure encoded in *directed acyclic graphs*
- Parsing consists of *unifying* attribute-value structures



Sam persuaded Alex to leave

Why were manually-crafted grammars abandoned?

- Can construct grammars for any particular sentence or construction, so *why were manually-crafted grammars abandoned?*
- *Dilemma of coverage and ambiguity*
  - *Broad coverage* and *robustness* $\Rightarrow$ add more syntactic rules
  - $\Rightarrow$ *Ambiguity explosion:* thousands of syntactic parses for ordinary sentences
- *All dressed up but no place to go ...*
  - the parsers produced detailed linguistic analyses of tense, quantifier scopes, etc., we had no way to use
- *Grammaticality* is central to linguistic theory, but it's not important for a language understanding system
  - goal is to recover the speaker's intended meaning, whether or not sentence is grammatical

Outline

The Past

The Present

The Future

Conclusion

“All our models are wrong . . . ”

> *Remember that all models are wrong;*
>
> *The practical question is how wrong do they have to be to not be useful.*
>
> – George E. P. Box and Norman R. Draper

- One big surprise: how *useful* very simple models can be
  - especially if you train them on large amounts of data
- Don’t worry about “true” model: find simple models that are “right enough” to be useful

Statistical Inference and Big Data

- Simple statistical models often perform better than more complex non-statistical systems
  - HMM-based speech recognition, then word-based machine translation
- Probabilities provide a systematic way of integrating unreliable, possibly conflicting information
- In the 1990s we discovered how to build probabilistic variants of virtually any linguistic theory

$\Rightarrow$ *no principled conflict between rich structure and probabilities*

Probabilistic approaches avoid coverage/ambiguity dilemma

- Probabilistic grammars can avoid the dilemma by:
  - massively *over-generating* (e.g., grammar generates all possible trees for all possible strings)
  - using probabilities to *distinguish more plausible from less plausible analyses*
- Every string gets an analysis $\Rightarrow$ robust
- Probabilities can guide parsing process $\Rightarrow$ ambiguity not fatal
- Grammars are inferred from *manually-constructed* treebanks

  $\Rightarrow$ linguistic insights still necessary
  - tree-banking is a *more economical* way of building a parser

“Capturing a generalisation” vs. “Covering a generalisation”

- Goal of science is improved *understanding of phenomena* being studied
- Linguistics aims to *capture the generalisation* that explains a set of constructions
  - example: *subject-verb agreement*
    she talks / they talk
- In engineering work, it suffices to *cover the generalisation*:
  - adding subject-verb agreement to reranking parser *does not affect f-score*
  - parser already includes *head-to-head POS dependencies*
  - because the subject is a dependent of head verb, these *cover subject-verb agreement*

Mobile computing and the explosion in NLP

- Classic internet search is about as bad as can be for NLP
  - the queries are too short for parsing to help
  - the documents to retrieve are so long that "bag of words" methods work as well as any
  - but a major advance in semantics or discourse might change this (Deep Learning?)
- *Mobile computing* changes this completely
  - users likely to post complex requests if we can make speech recognition work well enough
  - mobile devices require short targeted responses
- Computational linguistics will be just a minor part of the apps of the future
  - these will be important enough to *demand custom technology*

  $\Rightarrow$ NLP may fracture into multiple separate disciplines

Outline

Prediction is very difficult, especially about the future

– Niels Bohr

- My main prediction for the future:
  Computational linguistics will be so successful that in the future
  it may fracture into many subdisciplines
  - sufficient funding that machine translation, document analysis, etc., will become fields in their own right
  - Computational Linguistics may survive as a service discipline, like statistics

Standards for natural language processing

- *Standards* play a crucial role in most engineering efforts because they *let us reuse the same solution for many different problems*
- There are *advantages* and *costs* to standardisation
- Penn treebank parsing is becoming a de facto standard
  - \+ often easier to use an existing PTB parser even if it isn't ideal for your task
  - \+ several fairly well engineered relatively interchangable implementations
  - – but for specialised tasks (e.g., IR, MT, SR) more specialised parsing tools are appropriate
- *Standard data formats* are what is usually meant by standards
  - what about the data content?

When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one.

– Vladimir Vapnik

What are the problems our methods reliably work on?

- Can a CRF reliably identify *Earnings per Share* in financial documents?
- Structural engineers have handbooks listing performance characteristics of different materials
  - MIT became famous by quantifying how long it takes to sterilise tin cans

Predicting system performance

- Need to be able to *accurately cost* new projects
  - so we can tell client "it will cost \$X to get Y% accuracy"

$\Rightarrow$ Predict system performance without investing large amounts of resources
  - pilot experiments
  - statistical power estimates (used e.g., to design medical experiments)

- Similiar principles apply to corpus design
  - how much data do we need, e.g., to train a parser to 90% f-score?
  - "more data is better" is *not* a good answer here!

Metrics and evaluation

- Quantitative testing and evaluation is *absolutely central* to an engineering effort
- No reason for "one size fits all"
  - major tasks typically have *multiple objectives* (e.g., at least X% precision, Y% recall, no more than Z% failure)

  $\Rightarrow$ multi-objective optimisation (?)
- Evaluation metric can be closely related to system's *business objective*

Contributing to a wider scientific enterprise

- Claim: a lot of what counts as progress in our field is often only loosely related to science
  - increasing f-score is often not a scientific contribution
  - but *how you did it* may be a scientific contribution

How can computational models contribute to scientific theory?

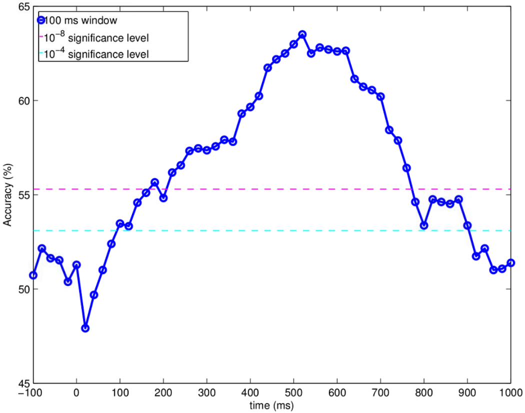
- Very hard to demonstrate that humans use a particular algorithm
  - not clear if neural computation is at all like current algorithms
  - how does computational complexity relate to psychological complexity?
    - lower probabilities $\Rightarrow$ slower processing, but why? (Levy)
- Marr's *3 levels of description* of a computational process
  - physical or implementational level
  - algorithmic and representational level
  - computational or informational level
- Major open problem: *how is hierarchical structure (trees) neurally represented?*

Computational neurolinguistics and "mind reading"

- *Magnetoencephalography* (MEG) uses superconducting sensors to detect magnetic fields generated by electrical currents in the brain
  - excellent temporal resolution, good spatial resolution
- "Mind reading": train classifiers to predict the experimental stimulus the subject is experiencing
- Use MEG signal to predict which word subject is hearing
- An L1-regularised logistic regression classifier can *distinguish the stimulus word with 65% accuracy*
  - the neuroscientists *don't care about classification accuracy* as long as it is *significantly above chance*

See: Bachrach, Haxby, Mitchell, Murphy

Classification accuracy versus time



- Although usually viewed as a 400msec response, *classifier predicts stimulus word* from 200msec post stimulus onset

⇒ Classifier provides information about *time course of language processing*

Sparse feature selection for localising neural responses

- Identifying the regions involved with language is very important e.g., for neurosurgery
- Our features are spatio-temporal regions of the brain
- L1 regularisation produces a *sparse model*, which identifies the spatio-temporal regions where the neural response to predicted variable differs

Localising the neural response

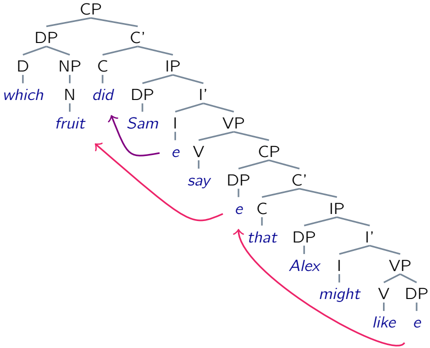
log unigram frequency

number of parser operations

- Both unigram frequency and number of parser operations are related to neural activity in the left anterior temporal lobe
- The number of parser operations is also related to neural activity in the left inferior frontal gyrus

How words and phrases compose in the brain

- Use "mind reading" to discover when and where words and phrases can be decoded during sentence comprehension
- Theories of syntax make different predictions about how words and phrases compose to form sentences



- Compare predictions about activation conventional syntactic theory, CCG and RNNs

How should we evaluate our work?

- *The goals of a scientific field may be very different to our usual goals*
  - I think this is common in real-world engineering problems too
- In a deployed engineering application, performance is critical
  - does it achieve the desired goal? (ultimately: does it achieve business objective?)
  - system performance, rather than the ideas involved, are what matters
- In scientific research, "success" is understanding the phenomenon being studied
  - ideally, evaluate work by how it advances our understanding
  - I suspect our scientific theories *lack key insights*

  $\Rightarrow$ too early to worry excessively about optimising performance (?)

What are we trying to do?

- Build a *unified model of all of language*
    - “pave it and put up a parking lot”
- Construct many different models for the different aspects of language and language processing
    - islands in the Pacific Ocean
    - perhaps we can build bridges between some of them?

See: van Benthem

A birds-eye view of computational linguistics

- The currently dominant reduction:

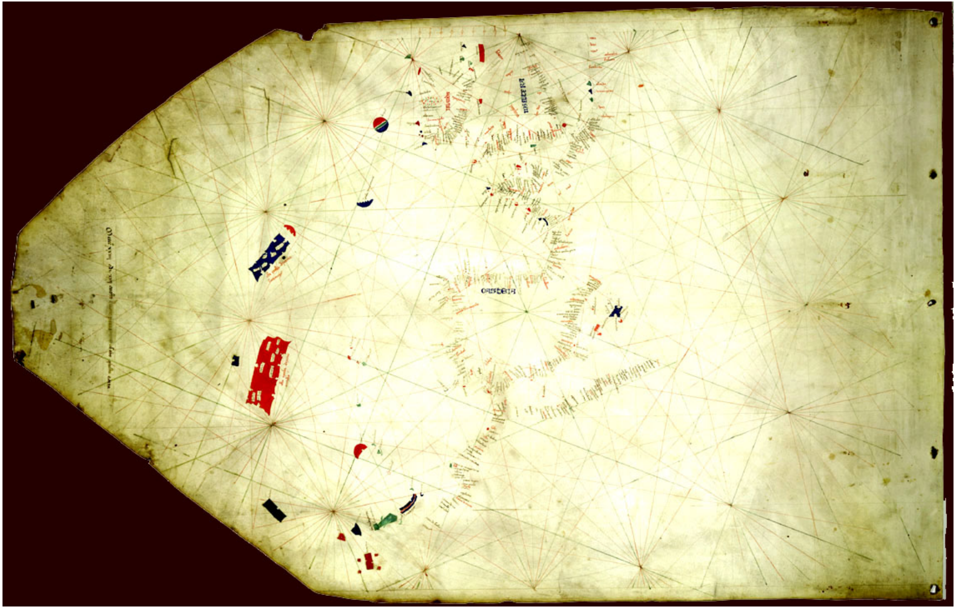
  Natural language problem
  $\Rightarrow$ Machine learning problem
  $\Rightarrow$ Statistical estimation problem
  $\Rightarrow$ Optimisation problem
- What might disrupt this?
  - "bolt from the blue" (e.g., Deep Learning, new discoveries in neuroscience (?))
  - statistical methods not based on optimisation, e.g., spectral methods, moment matching
- Perhaps we should concentrate on NL $\Rightarrow$ ML reduction, as this is where our community's strengths lie

Lessons from the history of science

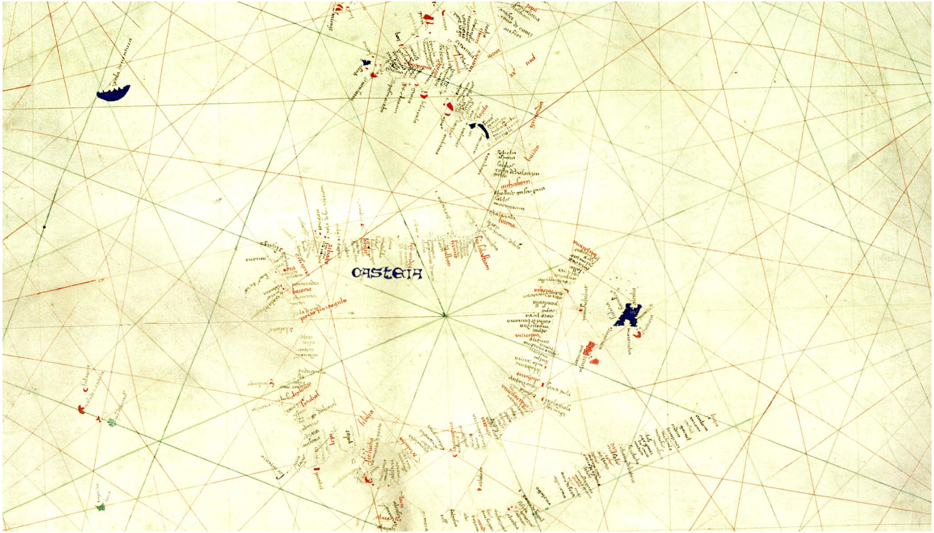
- Engineering has preceeded science in other areas as well
  - *Thermodynamics* and *statistical mechanics* took decades to develop after the steam engine
- Science isn't a story of continual progress
  - most ideas are wrong
  - Isaac Newton studied *alchemy* as well as gravitation
    - *transmutation* inspired his theory of optics
- The history of *maps and charts* is an interesting story about the interaction between academic research and practical "engineering" concerns

Psalter Mappa Mundi (1225?)

Portolan chart circa 1424

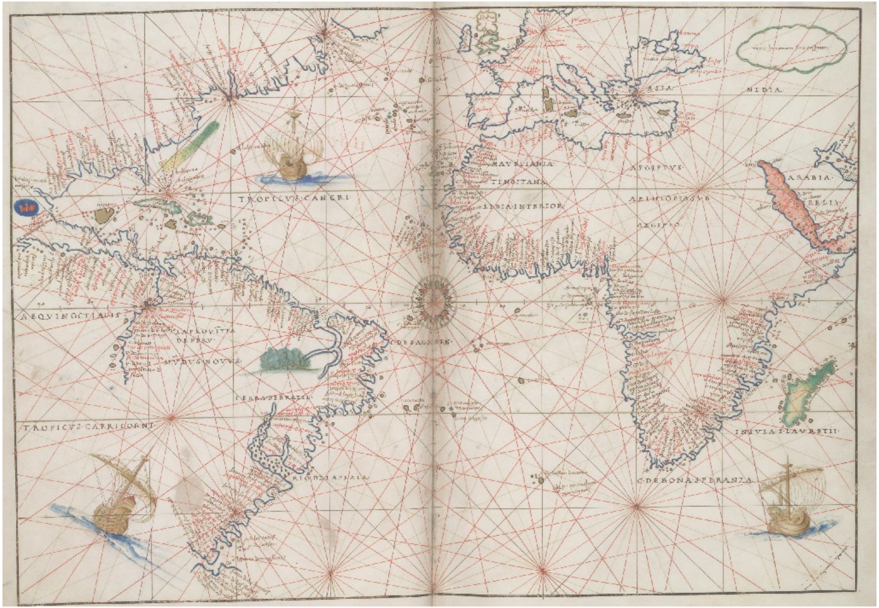


Portolan chart circa 1424 (center)



Waldseemüller 1507, after Ptolemy

Battista Agnese portolan chart circa 1550



Mercator 1569

... back to computational linguistics

- Be wary of analogies from the history of science!
  - we only remember the successes
- May wind up achieving something very different to what you expected
- Cartography and geography benefited from both the academic and Portolan traditions
- Geography turned out to be about brute empirical facts
  - geology and plate tectonics, rather than divinity and theology
- Mathematics (geometry and trigonometry) turned out to be essential
- Even wrong ideas can be important
  - the cosmographic tradition survives in celestial navigation

Outline

The Past

The Present

The Future

Conclusion

Where do we go from here?

- Expanding number of engineering and scientific applications
  - computational linguistics is one component of larger projects
  - will there be a *separate* field of computational linguistics in 50 years?
- Goals of scientific fields are often very different to those of CL
  - "covering generalisations" vs. "capturing generalisations"
  - CL is most relevant to the study of linguistic *processes*, e.g., psycholinguistics, language acquisition and neurolinguistics
  - other criteria are often more important than accuracy

Advice for beginning researchers

- "Keep your eyes on the prize"
    - focus on an important goal
    - be clear about *what you want to achieve* and *why you want to achieve it*
- The best researchers
    - can plot a path from where we are today to where they want to be
    - can *make what they do today contribute to their long-term goals*
    - adapt their research plans as new evidence comes in

Half the money I spend on advertising is wasted. The problem is: I don't know which half.

– John Wanamaker

Science advances one funeral at a time.

– Max Plank

We are recruiting!

- We're recruiting *post-docs* and *PhD students* for *academic* and *industrial research postions* who have skills in machine learning, statistical modelling and computational linguistics
- Contact **Mark.Johnson@mq.edu.au** for more information