Computational linguistics
Where do we go from here?

Mark Johnson

Macquarie University

ACL 2012

Prediction is very difficult, especially about the future

– Niels Bohr

- My main prediction for the future:
  *Computational linguistics will be so successful that in 50 years …
  it may not exist as a discipline any more*
- I hope I'm wrong

The future looks good!

- 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 and discourse parsing might change this (?)
- *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*

“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 can leverage large amounts of data
- Don’t worry about “true” model: find simple models that are “right enough” to be useful

“What goes around, comes around”

- The *empiricist turn* in computational linguistics connects back to the very earliest work in the field
  - George Miller
  - Victor Ingve
  - Warren Weaver
- Time for a *rationalist revival*? (Ken Church)
- But it’s never the same the second time around . . .

Rationalism vs empiricism in engineering

- Rationalism vs empiricism is a deep and interesting intellectual question
- But if your goal is to "get something done" it doesn't really matter whether your system is "rationalist" or "empiricist" *as long as it works*
- Real question: what combination of software and data achieves your goals *as cheaply as possible*
  - often *a small amount of annotated data* is incredibly valuable
- From an engineering perspective, rationalism vs. empiricism becomes *a question of economics*
  - may depend only "accidental" properties, e.g., what annotated data is already available
  - an intermediate position (e.g., semi-supervised learning) may be best

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
  - I'm not sure these are important: if someone can use a parser, they can probably also write a Python wrapper to reformat the input and output (?)

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

MACQUARIE UNIVERSITY

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"

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

"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*

Where do we fit?

- *Computer science and machine learning:*
  - but CS and ML aren't obviously sciences
- *Artificial intelligence:*
- *Mathematics:*
- *Linguistics:*
- *Psychology:*
- *Cognitive science:*

Why *computational* linguistics?

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

⇒ Computational linguistics contributes theoretically 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*

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
- My guess: it's premature to focus on the algorithmic level
  - our algorithms (e.g., EM, MCMC, particle filters) are designed to be very general, but humans solve very specific problems
  - neural wetware probably constrains representations and algorithms in ways we don't understand
    - major open problem: *how is hierarchical structure (trees) neurally represented?*

Two case studies of computational linguistics with a scientific goal

- Unsupervised models of language acquisition
- Computational linguistics and neuroscience
- In both cases we'll see that the usual goals of computational linguistics (e.g., improving f-score) align badly with broader scientific goals

MACQUARIE UNIVERSITY

Unsupervised parsing and grammar induction is a strange task

- *Unsupervised parsing* and *grammar induction* study how a grammar and parses can be learnt from terminal strings alone
  - this is a hard problem: "plain" EM does really badly!
- Standard motivation for this work:
  - help us *understand human language acquisition*
  - inducing parsers for *under-resourced languages*
- *These are very different goals!*
  - very lightly supervised methods are almost certainly more economical for under-resourced languages
- Unsupervised parsing from POS-tagged sequences isn't a cognitively-realistic task
  - POS-tags only make sense as part of a grammar

Identifying information sources for language acquisition

- A computational model can identify *which information sources suffice* to do something
  - *word segmentation* is first step to *learning a lexicon*

    y △ u ▲ w △ a △ n △ t ▲ t △ u ▲ s △ i ▲ D △ e ▲ b △ U △ k

  - using distributional information and syllable structure *achieves about 90% token f-score*
- *Synergies in acquisition*:
  - learning word segmentation and syllable structure jointly learns both more accurately than learning each on its own
  - learning word ⇝ object mapping together with word segmentation improves word segmentation accuracy
- *“Animals don’t move on wheels”*

  – Tom Wasow

See: Fleck, Goldwater, Swingley and many others

MACQUARIE UNIVERSITY

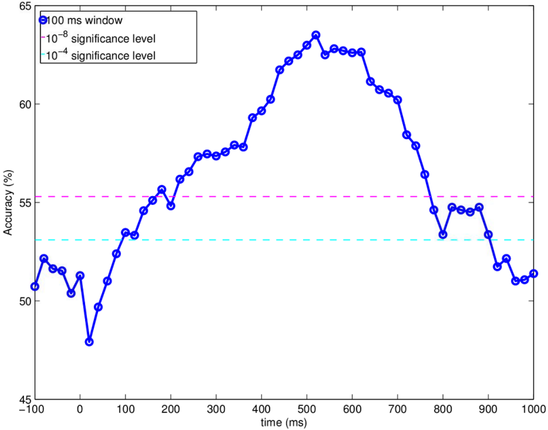
Computational neurolinguistics and "mind reading"

- *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 whether a context is "constraining" versus "non-constraining"

  constraining: Ruth has a necklace of glass *beads*
  non-constraining: Tom has been discussing the *beads*
- An L1-regularised logistic regression classifier can *predict context type 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 context type from 200msec post stimulus onset

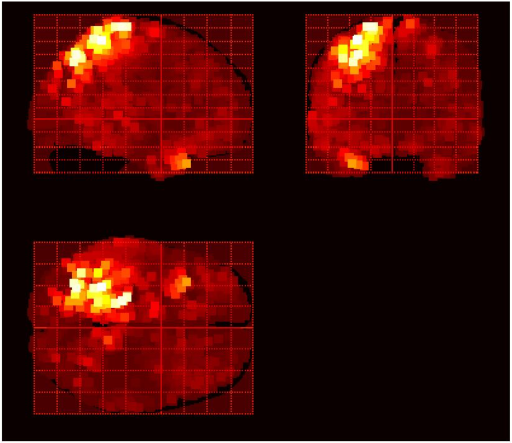
⇒ Classifier provides information about *time course of language processing*

MACQUARIE UNIVERSITY

Sparse feature selection for localising neural responses

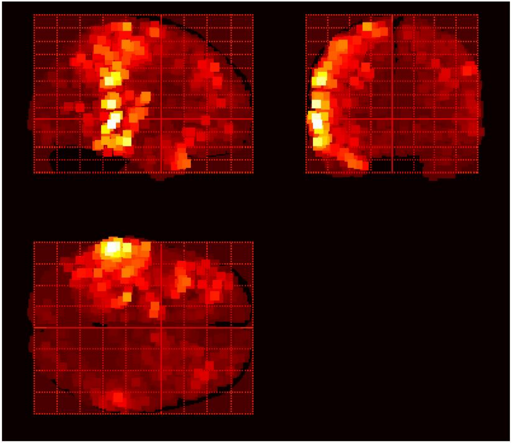
- Identifying the regions involved with language is very important e.g., for neurosurgery
- Our features are spatio-temporal regions of the brain
- L1 regularisation produces a *sparse model*, which identifies the spatio-temporal regions where the neural response to predicted variable differs

Predicting target word unigram probability



- Highest weight features in *parahippocampal gyrus*

Predicting target word conditional probability



- Highest weight features in *superior temporal gyrus*

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*

  ⇒ 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., discovery 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

Unification grammar

- Another reduction (dominant in 1980s–1990s):

  Natural language problem

  $\Rightarrow$ Logical problem in "feature logic"

  $\Rightarrow$ Satisfiability or deductive inference problem
- Whatever weaknesses this approach may have, it has developed a tight connection between CL and Linguistics, so it can be done!
- Most of their complex representations aren't relevant to statistical NLP (?)
- Maybe statistics will go this way too – all our students' students will learn is stochastic gradient ascent?

Lessons from the history of science

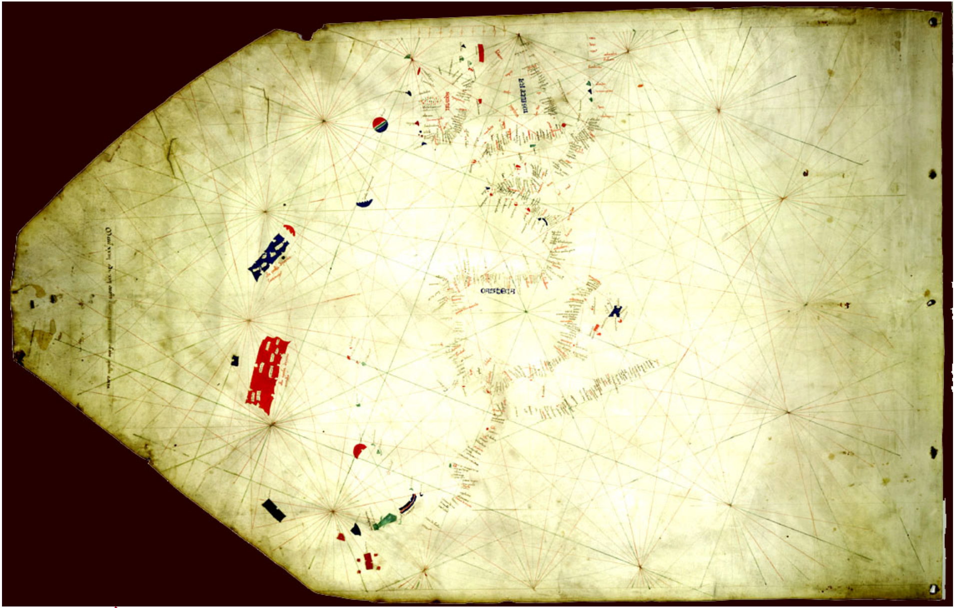
- Engineering has preceeded science in other areas as well
  - *Thermodynamics* and *statistical mechanics* took decades to develop after the steam engine
- Science isn't a story of continual progress
  - most ideas are wrong
  - Isaac Newton studied *alchemy* as well as gravitation
    - *transmutation* 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

MACQUARIE UNIVERSITY

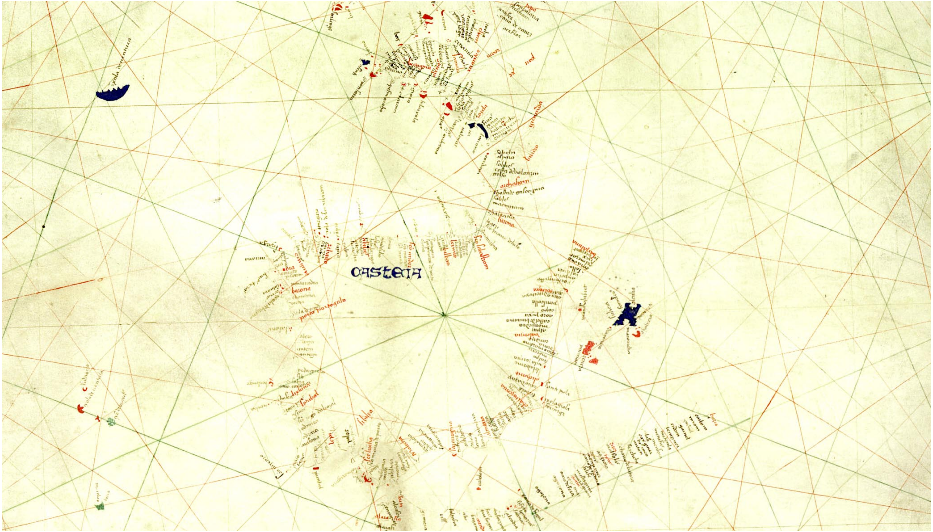
Psalter Mappa Mundi (1225?)

MACQUARIE UNIVERSITY

Portolan chart circa 1424

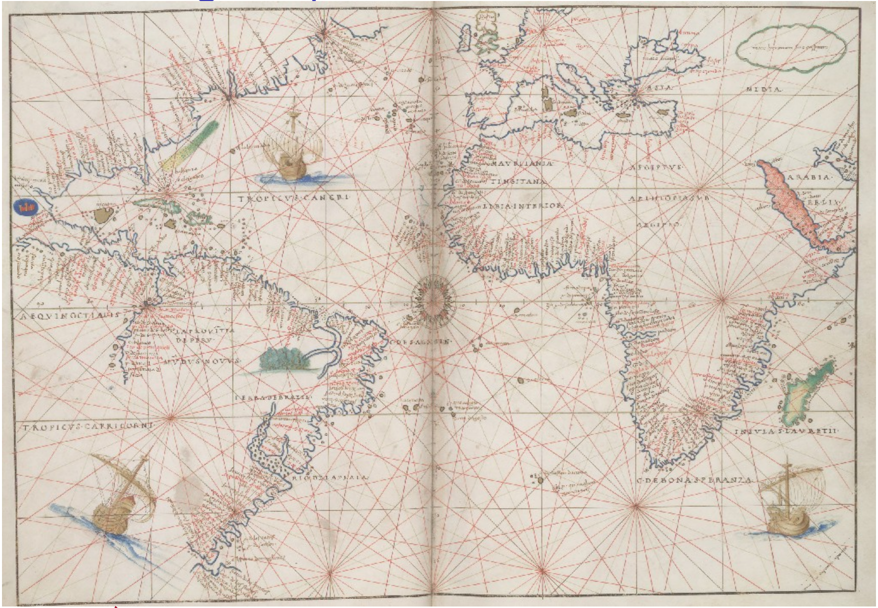


Portolan chart circa 1424 (center)



Waldseemüller 1507, after Ptolemy

Battista Agnese portolan chart circa 1550



MACQUARIE UNIVERSITY

Mercator 1569

MACQUARIE UNIVERSITY

… 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

Where do we go from here?

- Expanding number of engineering and scientific applications
  - computational linguistics will be just a component of a larger effort
  - should 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
  - computational models are most likely to help at Marr's computational (rather than algorithmic) level
  - computational models can help identify *information sources* used in linguistic processes, and *synergies* between linguistic processes
- Are there other ways of CL contributing to science?

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

Science advances one funeral at a time

– Max Plank

- Take everything I've said "with a grain of salt"
- But if you have an interesting idea, don't wait until I'm dead . . .

MACQUARIE UNIVERSITY