# Language Technology: Research and Development

Science and Research

Sara Stymne

Uppsala University

Based on slides from Joakim Nivre

# Research and Development

"Research and experimental development (R&D) comprise creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications." (OECD, 2002)

# Research and Development

“Research and experimental development (R&D) comprise creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications.” (OECD, 2002)

- Research – new knowledge
- Development – applied knowledge (cf. engineering)

# Research and Development

"Research and experimental development (R&D) comprise creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications." (OECD, 2002)

- Research – new knowledge
- Development – applied knowledge (cf. engineering)

# A Very Short History of (Western) Science

- Philosophy as a precursor of modern science
  - Antiquity: natural philosophy, Aristotle (600–300 BC)
  - Middle ages: scholastic philosophy (1100–1500)
- The scientific revolution (1500–1750)
  - Copernicus, Kepler, Galileo, Newton
  - Observation and experimentation
  - Mathematical models of physical phenomena
- Modern science (1900–):
  - Revolution in physics (relativity theory, quantum mechanics)
  - Explosion of new scientific disciplines
  - Natural, social and cultural sciences (arts, humanities)
  - Computational linguistics (1950s)

# Philosophy of Science

- Study of scientific methods
  - What distinguishes science from pseudo-science?
  - What is the nature of scientific reasoning?
  - What is a scientific explanation?
  - How does science make progress?

# Deduction and Induction

- Deductive inference

  All computational linguists are smart.
  Ann is a computational linguist.

  ---

  Therefore, Ann is smart.

  - Conclusion follows logically from premises
  - Characteristic of mathematical proofs

- Inductive inference

  All computational linguists I have met are smart.

  ---

  Therefore, all computational linguists are smart.

  - Conclusion does not follow logically from premises
  - Characteristic of empirical science (and everyday reasoning)

# Induction in Science

- Newton's law of universal gravitation (1686)
  - Every point mass in the universe attracts every other point mass with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them.
- Fleming's discovery of penicillin (1928)
  - Penicillium mold kills bacteria.
- Dürkheim's study of suicide (1897)
  - Suicide rates are higher in men than women.

# Hume's Problem of Induction

- Induction presupposes "uniformity of nature"
- How can we rationally justify this assumption?
  - By deduction – safe but impossible
  - By induction – more plausible but circular
- Conclusion:
  - The principle of induction cannot be rationally justified!

David Hume
(1711–1776)

# Verification and Falsification

Karl Popper
(1902–1994)

- Logical empiricism/positivism:
  - Scientific claims must be verifiable
  - Theories are verified inductively
  - Prefer the most probable of competing theories
  - Observations are objective and logically prior to theories
- Popper's alternative:
  - Scientific claims must be falsifiable
  - Theories are falsified deductively
  - Prefer the least probable of competing theories
  - Observations are theory-laden but must be replicable

# The Hypothetico-Deductive Method

- Universal claims can be falsified (but not verified) deductively:

  Bob is a computational linguist.
  Bob is not smart.

  ---

  Therefore, not all computational linguists are smart.

  "No amount of experimentation can ever prove me right;
  a single experiment can prove me wrong" (Einstein)

- Given hypothesis H with consequence C:
  - If C does not agree with observations, H is rejected (falsified)
  - Else H is provisionally accepted (corroborated)
- Science:
  - Progress through repeated testing, falsification, revision
  - Knowledge fundamentally uncertain ("current best theory")

# Inference to the Best Explanation (IBE)

- Another non-deductive inference type

  A window has been broken.
  A valuable painting is missing.

  ---

  A thief broke the window and took the painting.

  - Conclusion does not follow logically from premises
  - Alternative explanations are possible
- The principle of parsimony:
  - Prefer a simpler explanation (theory) over a more complex one
  - Darwin's theory of evolution
  - How can this principle be rationally justified?
  - Is IBE a form of induction (or the other way round)?

# Probabilistic Reasoning

- Laws and theories involving the notion of probability
  - Every gene has a 50% chance of being inherited (genetics)
  - Suicide rates are higher in men than women (sociology)
  - 90% of all lung cancers are caused by smoking (medicine)
- Inductive inference:

  80% of all computational linguists I have met are smart.

  ---

  Therefore, 80% of all computational linguists are smart.

- Deductive inference:

  80% of all computational linguists are smart.
  Ann is a computational linguist.

  ---

  Therefore, Ann has an 80% chance of being smart.

# Scientific Explanation

Carl G. Hempel
(1905–1997)

- Structured like an argument:
  - A set of premises (explanans)
  - A conclusion (explanandum)

Why did the metal rod expand?

---

All metal objects expand when their temperature increases.
Fire increases the temperature of objects.
The metal rod was placed in the fire.

---

Therefore, the rod expanded.

- Hempel's covering law model of explanation:
  - Conclusion follows logically from premises (deduction)
  - Premises are true and include at least one general law

# Problems with the Covering Law Model

- The problem of symmetry

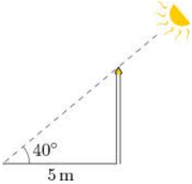
  Why is the shadow 5 meters long?

  ---

  Light travels in straight lines.
  Laws of trigonometry.
  Flagpole is 4.2 meters high.
  Angle of evelation of the sun is 40°.

  ---

  Therefore, the shadow is 5 meters long.



# Problems with the Covering Law Model

- The problem of symmetry

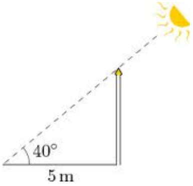
Why is the flagpole 4.2 meters high?

---

Light travels in straight lines.
Laws of trigonometry.
Shadow is 5 meters long.
Angle of evelation of the sun is $40^\circ$.

---

Therefore, the flagpole is 4.2 meters high.



# Problems with the Covering Law Model

- The problem of irrelevance

Why didn't the man become pregnant?

---

Anyone who takes birth control pills will not get pregnant.
The man took birth control pills.

---

Therefore, the man did not get pregnant.

- The problem of probabilistic laws

Why did the man get lung cancer?

---

90% of all lung cancers are caused by smoking.
The man was smoking.

---

Therefore, the man got lung cancer.

# Problems with the Covering Law Model

- The problem of irrelevance

  Why didn't the man become pregnant?

  ---

  Anyone who takes birth control pills will not get pregnant.
  The man took birth control pills.

  ---

  Therefore, the man did not get pregnant.

- The problem of probabilistic laws

  Why did the man get lung cancer?

  ---

  90% of all lung cancers are caused by smoking.
  The man was smoking.

  ---

  Therefore, his lung cancer was probably caused by smoking.

# Scientific Change

Thomas Kuhn
(1922–1996)

- Traditional view:
  - Science advances in a cumulative fashion
- Kuhn's notion of paradigm (normal science)
  - A set of shared theoretical assumptions
  - A set of accepted problems and methods ("puzzle solving")
- Scientific revolutions
  - Accumulation of anomalies lead to crisis and revolution
  - Old paradigm abandoned only if new paradigm available
  - Copernicus, Darwin, Einstein

# Beyond Natural Sciences

- Hermeneutics
  - Natural sciences seek explanation
    Why? = What caused it to happen?
  - Social/human sciences seek understanding
    Why? = Why did the agents bring it about?
  - Causality vs. Meaning
- Design science
  - Sciences of the artificial
  - Constructs, models, methods, instantiations
  - Truth vs. Utility
- Is there a universal scientific method?

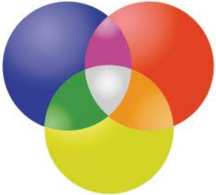
Hans-Georg Gadamer
(1900–2002)

Herbert Simon
(1916–2001)

# Research Ethics

- Traditional view:
  - Scientific knowledge is neither good nor bad per se
  - But scientific knowledge can be used unethically
  - Where does the responsibility of scientists begin and end?
- Ethical considerations in research activities:
  - Experimentation with humans or animals
  - Intellectual dishonesty (fabrication of data, plagiarism)
  - Discrimination and harrassment
  - Many disciplines have specific ethical guidelines

# The Name of our Field



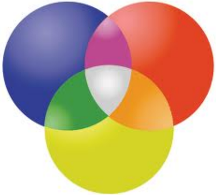
Computational Linguistics (CL)

Natural Language Processing (NLP)

[Human] Language Technology ([H]LT)

[Natural] Language Engineering ([N]LE)

# The Name of our Field



Computational Linguistics (CL)

- Study of natural language from a computational perspective

Natural Language Processing (NLP)

- Study of computational models for processing natural language

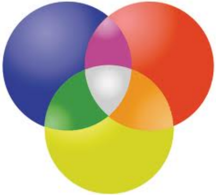
[Human] Language Technology ([H]LT)

- Development and evaluation of applications based on CL/NLP

[Natural] Language Engineering ([N]LE)

- Same as [H]LT but obsolete?

# The Name of our Field



Computational Linguistics (CL)

Natural Language Processing (NLP)

Often used more or less synonymously!

[Human] Language Technology ([H]LT)

[Natural] Language Engineering ([N]LE)

# An Interdisciplinary Field

Linguistics

- Theory, language description, data analysis (annotation)

Computer science

- Theory, data models, algorithms, software technology

Mathematics

- Theory, abstract models, analytic and numerical methods

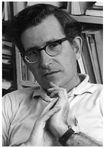
Statistics

- Theory, statistical learning and inference, data analysis

# Linguistics

F. de Saussure
(1857–1913)

L. Bloomfield
(1887–1949)



N. Chomsky
(1928–)

- Structuralist linguistics (1915–1960)
  - Language as a network of relations (phonology, morphology)
  - Inductive discovery procedures
- Generative grammar (1960–)
  - Language as a generative system (syntax)
  - Deductive formal systems (formal language theory)
  - NLP systems based on linguistic theories

# Linguistics

- Recent trends (1990–):
  - Language processing (psycholinguistics, neurolinguistics)
  - Strong empiricist movement (corpus linguistics)
  - NLP systems based on linguistically annotated data
- Theoretical and computational linguistics have diverged

  Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous? (Workshop at EACL 2009)

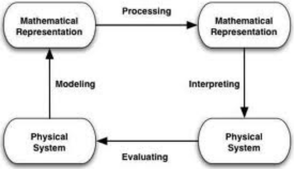
# Computer Science

Alan Turing
(1912–1954)

Herbert Simon and John Newell
(1916–2001) (1927–1992)

- Theoretical computer science
  - Turing machines and computability (Church-Turing thesis)
  - Algorithm and complexity theory (cf. formal language theory)
- Artificial Intelligence
  - Early work on symbolic logic-based systems (GOFAI)
  - Trend towards machine learning and sub-symbolic systems
  - Parallel development in natural language processing

# Mathematics



- Mathematical model
  - Description of real-world system using mathematical concepts
  - Formed by abstraction over real-world system
  - Provide computable solutions to problems
  - Solutions interpreted and evaluated in the real world
- Mathematical modeling fundamental to (many) science(s)

# Mathematics

- Real-world language technology problem:
  - Syntactic parsing: sentence ⇒ syntactic structure
  - No precise definition of relation from inputs to outputs
  - At best annotated data samples (treebanks)
- Mathematical model:
  - Probabilistic context-free grammar $G$

$$T^* = \operatorname*{argmax}_{T:yield(S)=T} P_G(T)$$

  - $T^*$ can be computed exactly in the model
  - $T^*$ may or may not give a solution to the real problem
- How do we determine whether a model is good or bad?

# Statistics

Probability theory

- Mathematical theory of uncertainty

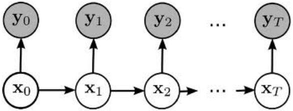
Descriptive statistics

- Methods for summarizing information in large data sets

Statistical inference

- Methods for generalizing from samples to populations

# Statistics



- Probability theory
  - Framework for mathematical modeling
  - Standard models: HMM, PCFG, Naive Bayes
- Descriptive statistics
  - Summary statistics in exploratory empirical studies
  - Evaluation metrics in experiments (accuracy, precision, recall)
- Statistical inference
  - Estimation of model parameters (machine learning)
  - Hypothesis testing about systems (evaluation)

# Language Technology R&D

Sections in Transactions of the ACL (TACL):

- Theoretical research
- Empirical research
- Applications and tools
- Resources and evaluation

# Language Technology R&D

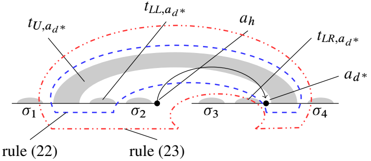
Sections in Transactions of the ACL (TACL):

- Theoretical research – deductive approach
- Empirical research – inductive approach
- Applications and tools – design and construction
- Resources and evaluation – data and method

# Theoretical Research

- Formal theories of language and computation
- Studies of models and algorithms in themselves
- Claims justified by formal argument (deductive proofs)
- Often implicit relation to real-world problems and data

# Theoretical Research



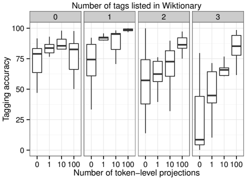
Satta, G. and Kuhlmann, M. (2013)
Efficient Parsing for Head-Split Dependency Trees.
*Transactions of the Association for Computational Linguistics* 1, 267–278.

- Contribution:
  - Parsing algorithms for non-projective deendency trees
  - Added constraints reduce complexity from $O(n^7)$ to $O(n^5)$
- Approach:
  - Formal description of algorithms
  - Proofs of correctness and complexity
  - No implementation or experiments
  - Empirical analysis of coverage after adding constraints

# Empirical Research

- Empirical studies of language and computation
- Studies of models and algorithms applied to data
- Claims justified by experiments and statistical inference
- Explicit relation to real-world problems and data

# Empirical Research



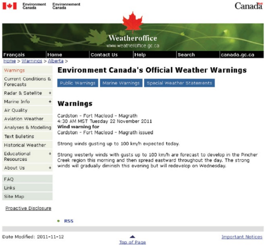
Täckström, O., Das, D., Petrov, S., McDonald, R. and Nivre, J. (2013)
Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging.
*Transactions of the Association for Computational Linguistics* 1, 1–12.

- Contribution:
  - Latent variable CRFs for unsupervised part-of-speech tagging
  - Learning from both type and token constraints
- Approach:
  - Formal description of mathematical model
  - Statistical inference for learning and evaluation
  - Multilingual data sets used in experiments

# Applications and Tools

- Design and construction of LT systems
- Primarily end-to-end applications (user-oriented)
- Claims often justified by proven experience
- May include experimental evaluation or user study

# Applications and Tools



Environment Canada's Official Weather Warnings

Warnings

Cardston - Fort Macleod - Magrath
4:30 AM MST Tuesday 22 November 2011
**Wind warning for**
Cardston - Fort Macleod - Magrath issued

Strong winds gusting up to 100 km/h expected today.

Strong westerly winds with gusts up to 100 km/h are forecast to develop in the Pincher Creek region this morning and then spread eastward throughout the day. The strong winds will gradually diminish this evening but will redevelop on Wednesday.

Date Modified: 2011-11-12

Gotti, F., Langlais, P. and Lapalme, G. (2014)
Designing a Machine Translation System for Canadian Weather Warnings:
A Case Study. *Natural Language Engineering* 20(3): 399–433.

- Contribution:
  - In-depth description of design and application development
  - Extensive evaluation in the context of application (real users)
- Approach:
  - Case study – concrete instance in context
  - Semi-formal system description (flowcharts, examples)
  - Statistical inference for evaluation

# Resources and Evaluation

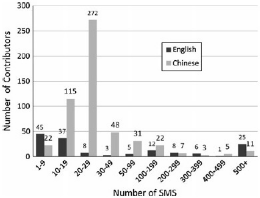
Resources

- Collection and annotation of data (for learning and evaluation)
- Design and construction of knowledge bases (grammars, lexica)

Evaluation

- Protocols for (empirical) evaluation
  - Intrinsic evaluation – task performance
  - Extrinsic evaluation – effect on end-to-end application
- Methodological considerations:
  - Selection of test data (sampling)
  - Evaluation metrics (intrinsic, extrinsic)
  - Significance testing (statistical inference)

# Resources and Evaluation



Chen, T. and Kan, M.-Y. (2013)
Creating a Live, Public Short Message Service Corpus:
The NUS SMS Corpus. *Language Resources and Evaluation* 47:299–335.

- Contribution:
  - Free SMS corpus in English and Chinese ($>$ 70,000 msgs)
  - Discussion of methodological considerations
- Approach:
  - Crowdsourcing using mobile phone apps
  - Automatic anonymization using regular expressions
  - Linguistic annotation as future plans

# Language Technology as a Science

- Scientific reasoning
  - Deduction common in theoretical research
  - Induction underlies machine learning and statistical evaluation
  - Inference to the best explanation in experimental studies
- Scientific explanation
  - Explanations based on general laws are rare
  - Explanations based on statistical generalizations are the norm
- Reproducibility/replicability
  - Important in theory but problematic in practice
  - Initiatives to publish data and software with papers

    Fokkens et al. (2013) Offspring from Reproduction Problems: What Replication Failure Teaches Us. In *Proceedings of ACL*.

    Reprolang 2020

# Ethics in Language Technology

- Increasing attention in the (larger) community
- Some issues raised by Hovy and Spruit:
    - Exclusion – data bias
    - Overgeneralization – modeling bias
    - Dual-use problems
- First Workshop on Ethics in NLP held in 2017
- First workshop on NLP for Positive Impact held in 2021 (by NLP for social good initiative)

# Coming up

- Research groups: web site
- Read the articles for seminar 1
- Check if you are presenting an article at seminar 1
- Debate session tomorrow, 10-12 in Turing
- First literature seminar: Wednesday September 7
- Take home exam: September 8-14

# Take-home exam

- Covers the Okasha book
- Handed out: September 8
- Deadline: September 14, 23:59
- Anonymous, so do not write your name, only your Ladok code
- Studium used for handing out and submitting
- Registered students have been signed up to Ladok

# Literature seminars

- 3 articles to read per seminar
- One person repsonsible for presenting each article
  - short summary (MAX 2 minutes)
  - main points, strengths, problems, difficulties
  - points for discussion
- Everyone is expected to have read all articles and to contribute to all discussions!
- Bring the articles to the seminar (on paper or electronically)
- Campus only (unless exception granted)
- If absent: write a short report

# Reminder deadlines etc.

- All course deadlines are strict!
- Hand in in Studium at the latest 23.59. Then it closes.
- Backup deadlines specified on the course web page (not recommended!)

# Reminder deadlines etc.

- All course deadlines are strict!
- Hand in in Studium at the latest 23.59. Then it closes.
- Backup deadlines specified on the course web page (not recommended!)
- If you cannot respect a deadline due to **extraordinary** circumstances, discuss this with your teacher well before the deadline. No exceptions will be given after the deadline!

# Reminder deadlines etc.

- All course deadlines are strict!
- Hand in in Studium at the latest 23.59. Then it closes.
- Backup deadlines specified on the course web page (not recommended!)
- If you cannot respect a deadline due to **extraordinary** circumstances, discuss this with your teacher well before the deadline. No exceptions will be given after the deadline!
- Take home exam:
  - Individual examination
  - No cooperation