HG8003 Technologically Speaking: The intersection of language and technology.

Overview of NLP, Main Issues

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 1
Location: LT8

HG8003 (2014)
Introduction

➤ How technology affects our use of language

➤ How technology enables us to study and process language in new ways.

➤ Representing Language (wk 1-6)
  ➢ Different Modalities (text, speech, . . .)
  ➢ Representing and Finding Meaning (implicit, explicit)
  ➢ The Semantic Web, Markup and Reputation

➤ Machine Translation (wk 7-12)
  Parsing, Generation, Transfer, Word Sense Disambiguation
Goals

➢ Gain understanding into:
  ➢ Representing, transmitting and transforming language.
  ➢ Parsing
  ➢ Generation
  ➢ Text Mining
  ➢ Machine Translation

➢ Know why language processing is so difficult interesting

➢ Know what the current state of the art is
Personal Introduction: Francis Bond

- BA in Japanese and Mathematics
- BEng in Power and Control
- PhD on “Determiners and Number in English contrasted with Japanese, as exemplified in Machine Translation”
  - Japanese - English/Malay Machine Translation
  - Japanese corpus, grammar and ontology (Hinoki)
  - Japanese - English, Chinese Machine Translation
  - Japanese WordNet
- 2009– NTU
Personal Introduction: Wang Shan (王珊)

➤ PhD on *Semantics of Event Nouns*
The Hong Kong Polytechnic University

➤ Visiting Scientist
Brandeis University, USA

➤ Research Fellow
NTU

➤ Homepage: https://sites.google.com/site/wangshanlanguages/

➤ Email: star8882008@qq.com
Administrivia

Coordinator  Francis Bond <bond@ieee.org> !<fcbond@ntu.edu.sg>

Lecture Time  Thursdays 16:30–19:30

Venue  NS — LT8

*  No tutorials
Assessment

- **Continuous Assessment** (40%)
  - Two quizzes (20% each, multiple choice)

- **Exam** (60%)
  - Multiple-choice
  - Shortish answers
A Note on Asymmetry

➤ This course has few lecturers and many students
   ➤ If you ask me something that takes only minute
   ➤ Consider that multiplied by 150
   ➤ 1 minute each → 2.5 hours

⇒ I will get grumpy if you ask a question that has already been answered

➤ I put all information up on the HG8003 page.
   ➤ Check there first
   ➤ Then ask other students
   ➤ Then email me: bond@ieee.org with HG8003 in the subject line
# Schedule

<table>
<thead>
<tr>
<th>Lec.</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>01-16</td>
<td>Introduction, Organization: Overview of NLP; Main Issues</td>
</tr>
<tr>
<td>2</td>
<td>01-23</td>
<td>Representing Language</td>
</tr>
<tr>
<td>3</td>
<td>02-06</td>
<td>Representing Meaning</td>
</tr>
<tr>
<td>4</td>
<td>02-13</td>
<td>Words, Lexicons and Ontologies</td>
</tr>
<tr>
<td>5</td>
<td>02-20</td>
<td>Text Mining and Knowledge Acquisition</td>
</tr>
<tr>
<td>6</td>
<td>02-27</td>
<td>Structured Text and the Semantic Web</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Quiz</strong></td>
</tr>
<tr>
<td>7</td>
<td>03-13</td>
<td>Citation, Reputation and PageRank</td>
</tr>
<tr>
<td>8</td>
<td>03-20</td>
<td>Introduction to MT, Empirical NLP</td>
</tr>
<tr>
<td>9</td>
<td>03-27</td>
<td>Analysis, Tagging, Parsing and Generation</td>
</tr>
<tr>
<td>10</td>
<td>Video</td>
<td>Statistical and Example-based MT</td>
</tr>
<tr>
<td>11</td>
<td>04-03</td>
<td>Transfer and Word Sense Disambiguation</td>
</tr>
<tr>
<td>12</td>
<td>04-10</td>
<td>Review and Conclusions</td>
</tr>
<tr>
<td>Exam</td>
<td>05-06</td>
<td>17:00</td>
</tr>
</tbody>
</table>

> Video week 10
Complimentary Courses

➤ HG8001 Language Puzzle — more linguistics, more on language use

➤ HG2051 Language and the Computer — solving NLP problems with Python: introduces both programming and linguistics

Main Issues

➢ Why is natural language interesting, difficult, complex and ambiguous?
  ➢ How do we resolve this ambiguity?

➢ Layers of linguistic analysis

➢ Computational Linguistics history, an overview, current successes

➢ Machine Translation
Consider robots

- They can walk (now run, and maybe even climb stairs)
- They can play chess
- They can play football
- They can recognize people by face

- They can’t hold a conversation
  - Language turns out to be very difficult
Querying a knowledge base

User query: Has my order number 4291 been shipped yet?

Database: 

<table>
<thead>
<tr>
<th>Order number</th>
<th>Date ordered</th>
<th>Date shipped</th>
</tr>
</thead>
<tbody>
<tr>
<td>4290</td>
<td>2/2/02</td>
<td>2/2/02</td>
</tr>
<tr>
<td>4291</td>
<td>2/2/02</td>
<td>2/2/02</td>
</tr>
<tr>
<td>4292</td>
<td>2/2/02</td>
<td></td>
</tr>
</tbody>
</table>

DB QUERY: order(number=4291, date_shipped=?)

RESPONSE: Order number 4291 was shipped on 2/2/02
What’s the difficulty?

Similar strings mean different things, different strings mean the same thing:

1. How fast is the 505G?
2. How fast will my 505G arrive?
3. Please tell me when I can expect the 505G I ordered.

Syntactic Ambiguity:

- Do you sell Sony laptops and disk drives?
- Do you sell (Sony (laptops and disk drives))?
- Do you sell (Sony laptops) and disk drives?

Lexical Ambiguity:

- 505G (Sony Vaio PCG 505G); LG 505G; RSQ DVD-505G
What’s in a name?

➢ Francis Bond
   Standard

➢ Bond, Francis
   Alphabetic order

➢ F. Bond
   Short

➢ F. C. Bond
   Short and uniquer

➢ フランシス ボンド furansisu bondo
   Japanese reversed

➢ 凡士 フランシス bondo furansisu
   Japanese normal (with Ateji)

➢ F・ボンド F·bo ndo
   Weird
Why it matters

» Hakodate Sushi sends a birthday discount (good sushi, but a bit expensive)

» 10% off

» Eight people got them at Professor Bond’s house

» Who were they?
Overview of NLP, Main Issues
Overview of NLP, Main Issues
Francis Bond

Overview of NLP, Main Issues
Overview of NLP, Main Issues
フランシスボンド
ボンドフランスシス

お誕生日おめでとうございます。
季節限定商品ぞくぞく入荷しております。
期間限定ですのでお早めに☆
有効期限は本状到着から2009年6月8日までとさせて頂きます。
【返付先】函館市場 奈良県奈良学園市
〒630-0112

CArTE POSTALE

料金後納郵便
函館市場

18817 1857

TEL 0742-40-4851

お誕生日おめでとうございます
本状を必ずご持参下さい。

他のダイレクトメール類似・複写物類似の送付、
株主優待カードとの併用はできません。
凡土フランシス

〒630-0112
奈良県生駒市鹿ノ台東2-6-5

CARTE POSTALE

凡土フランシス様

50612-62-090425-28723395 143

料金後払い郵便

お誕生日おめでとうございます。
季節限定商品ぞくぞく入荷しております。期間限定ですのでお早めに☆
有効期限は本状到着から2009年6月8日までとさせて頂きます。
【振付先】函館市場 奈良県近藤市場
奈良県奈良市中幸町6丁目7-7
TEL:0742-48-4851

お誕生日おめでとうございます

上記店舗にてご利用いただけます。
本状必ずご持参下さい。

函館市場

他のダイレクトメール・印刷・懸賞券・
株主優待コードとの併用はできません。
Who are these eight people?

- Francis Bond
- Francis BOND
- Francis Bond
- Francis BOND
- フランシス・ボンド
- フランシスボンド
- ボンドフランシス
- 凡土フランシス

They are all me! And I ate a lot of sushi that month.
Data Cleansing

- Detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

- Identifying incomplete, incorrect, inaccurate, irrelevant parts of the data and then replacing, modifying or deleting this dirty data.

≈ data scrubbing; ⊇ duplicate elimination

- This is the DB equivalent of parsing and disambiguation

  linking words to their meanings
How to clean up these?

Francis Bond
Francis BOND
Francis Bond
Francis BOND

➢ Word Order
➢ Segmentation
➢ Upper/Lower Case
➢ Chinese Characters, Katakana, Latin

Need to know something about names and scripts
Names are really, really difficult

➤ First (Middle)* (Last)?
  *Brian Peter George St. Jean le Baptiste de la Salle Eno*
  *Pablo* Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Ruiz y *Picasso*

➤ Last First (Middle)* *Ng Bee Chin*

➤ First (patronymic) *Osama bin Laden*

➤ *Willy; the the*

➤ the artist formerly known as prince
What to do?

➢ Impose Order: (nice if you can)

<table>
<thead>
<tr>
<th>First Name</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle Name</td>
<td></td>
</tr>
<tr>
<td>Last Name</td>
<td></td>
</tr>
</tbody>
</table>

➢ Allow anything: (where NLP comes in)

| Name |  |

➢ Allow anything: (plus a sort value)

<table>
<thead>
<tr>
<th>Name</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Last Name</td>
<td></td>
</tr>
</tbody>
</table>
Encoding Issues

A Name Too Unusual for the Chinese Government

China's Public Security Bureau plans to replace handwritten identity cards, like this one below, with computer-readable ones. However, the government's computers can read only 32,252 of the roughly 55,000 Chinese characters, forcing citizens with uncommon names to change them.

Trade Offs in Representing Names

➢ Simple Representation
  ➢ Easy to process
  ➢ Covers most cases
  ➢ Robust

➢ Complex Representation
  ➢ Can cover more cases
  ➢ Hard to process

➢ For Cheng, UNICODE would be enough (U+299E2).

➢ The full case of all names is open ended like Language
Linguistic Analysis
Layers of Linguistic Analysis

1. Phonetics & Phonology
2. Morphology
3. Syntax
4. Semantics
5. Pragmatics
1. Phonetics & Phonology

The study of **language sounds**, how they are physically formed; systems of discrete sounds, e.g. languages’ syllable structure.

- **disconnect**
- “It is easy to recognize speech.”
- “It is easy to wreck a nice beach.”
2. Morphology

The study of the sub-word units of meaning.

\texttt{dis|connect}  \\
“not” “attach”

Even more necessary in some other languages, e.g. Turkish:

(1) \texttt{uygarlastiramadiklarimizdanmissinizcasina}  \\
uyg\texttt{ar las tir ama dik lar imiz dan mis siniz casina}  \\
(behaving) as \texttt{if you are among those whom we could not civilize}

Also includes segmentation.
3. Syntax

The study of the structural relationships between words.

(2) *People saw her duck.*
4. Semantics

The study of the literal (truth conditional) meaning.

Structural Semantics:

(3) see(people, duck_i: PAST) poss(duck_i, pronoun:[3RD, SG, FEM]: PAST)

(4) see(people, duck_j: PAST) duck_j(pronoun:[3RD, SG, FEM])

(5) saw(people, duck_i) poss(duck_i, pronoun:[3RD, SG, FEM])

Lexical Semantics:

(6) people ⊂ entity (people is-a/kind-of entity)

(7) see ⊂ perceive

(8) saw ⊂ cut

(9) duck_i ⊂ bird

(10) duck_j ⊂ move
5. Pragmatics

The study of meaning in context.

➢ Which people?

➢ What duck?

➢ Why did you say that?

➢ What does it imply?
  ➢ Do you know who I am?
Linguistic Rules

E.g. Morphology

To make an English word plural, add “s”

➣ dog → dogs

⊗ baby → babies

⊗ dish → dishes

⊗ goose → geese

⊗ child → children

⊗ fish → fish (!)
Better rules

➢ To make a word plural
  ➢ look in a list of exceptions (several thousands)
  ➢ elsif it ends in y remove y add ies
  ➢ elsif it ends in s add es
  ➢ else add s

➢ Many sub-regularities, but not worth calculating
  ➢ focus → foci

➢ Note: not always the final word!
  attorney general → attorneys general
  this makes life hard
Inherent Ambiguity in Syntax

(11) Fed raises interest rates 0.5% in effort to control inflation

NY Times headline 17 May 2000
Some of the ambiguities? (33 parses!)
Two tasks

➤ Assign an interpretation to a sentence (Parsing, Syntax)
  ➤ DG, (P)CFG, HPSG, LFG, CCG, …
  ➤ Induced grammars
  ➤ ngrams

➤ Select the most suitable interpretation (Ranking, Interpretation)
  ➤ Semantic preferences and associations
  ➤ Stochastic models

➤ Both problems are hard

➤ The second seems to be harder
Ambiguous Headlines

(12) Miners refuse to work after death
(13) Stolen painting found by tree
(14) Milk drinkers are turning to powder
(15) Drunk gets nine months in violin case
(16) Panda mating fails; Veterinarian takes over
(17) Astronaut takes blame for gas in space craft
(18) Grandmother of eight makes hole in one
(19) Lack of brains hinders research
(20) Iraqi Head Seeks Arms
(21) Juvenile Court to Try Shooting Defendant
(22) Teacher Strikes Idle Kids
(23) British Left Waffles on Falkland Islands
(24) Ban on Nude Dancing on Governor’s Desk
What is grammatical and what isn’t?

(25)  Colorless green ideas sleep furiously.
(26)  *Furiously sleep ideas green colorless.  (Chomsky, 1957)

It can only be the thought of verdure to come, which prompts us in the autumn to buy these dormant white lumps of vegetable matter covered by a brown papery skin, and lovingly to plant them and care for them. It is a marvel to me that under this cover they are labouring unseen at such a rate within to give us the sudden awesome beauty of spring flowering bulbs. While winter reigns the earth reposes but these colourless green ideas sleep furiously. C. M Street (1985)

Using a simple probabilistic model (based only on the probability of a word occurring given the two proceeding words) Pereira (2000) showed that $P(25) \approx P(26) \times 200,000$. 
Language Evolves

➢ Morphology

➢ We learn new words all the time: bioterrorism, cyberstalker, infotainment, energy bar, parkour, tweet, crowdsourcing

➢ Part-of-speech

➢ Historically: kind and sort were always nouns:

(27) I knowe that sorte of men ryght well. [1560]

➢ Now also used as degree modifiers:

(28) I’m sort of hungry. [Present]
(29) It sort o’ stirs one up to hear about old times. [1833]
Natural Language Processing is hard because

Natural language is:

- highly ambiguous at all levels
- complex and subtle
- fuzzy, probabilistic
- interpretation involves combining evidence
- involves reasoning about the world
- embedded in a social system of people interacting
  * persuading, insulting and amusing them
  * changing over time
Probabilistic Models of Language

To handle this ambiguity and to integrate evidence from multiple levels we turn to:

The tools of probability:

- Bayesian Classifiers (not rules)
- Hidden Markov Models (not Deterministic Finite Automatons)
- Probabilistic Context Free Grammars
- Ranking Models
- ... other tools of Machine Learning, AI, Statistics
Natural Language Processing
Natural Language Processing (NLP) studies how to get computers to do useful things with natural languages:

* Most commonly **Natural Language Understanding**
* The complementary task is **Natural Language Generation**

NLP draws on research in Linguistics, Theoretical Computer Science, Artificial Intelligence, Mathematics and Statistics, Psychology, Cognitive Science, etc.
What & Where is NLP

➢ Goals can be very far-reaching
  ➢ True text understanding
  ➢ Reasoning and decision-making from text
  ➢ Real-time spoken dialog

➢ Or very down-to-earth
  ➢ Searching the Web
  ➢ Context-sensitive spelling correction
  ➢ Analyzing reading-level or authorship statistically
  ➢ Extracting company names and locations from news articles.
State of the Art

These days, the down-to-earth research predominates, as NLP becomes increasingly practical, focused on performing measurably useful tasks now.

Although language is complex, and ambiguity is pervasive, NLP can also be surprisingly easy sometimes: — rough text features often do half the job.

However, richer and richer models gradually bring us back to full-fledged linguistic analysis, but with graceful defaults (backing off to simpler models). Stochastic models are useful to guess when we don’t know (much of the time).
Linguistics

Linguistics is the study of natural languages:

- Understanding this naturally-occurring phenomenon.
- Structure, meaning, how acquired, differences and commonalities across languages.

Linguistics draws on research in Natural Language Processing, Theoretical Computer Science, Artificial Intelligence, Mathematics and Statistics, Psychology, Cognitive Science, etc.
Some brief history: 1950s

➤ Early computational linguistics (CL) on machines less powerful than pocket calculators.

➤ Foundational work on automata, formal languages, probabilities and information theory.

➤ First speech systems (Davis et al, Bell Labs).

➤ MT heavily funded by military, but basically just word substitution programs.

➤ Little understanding of natural language syntax, semantics, pragmatics.
Some brief history: 1960s

- Alvey report (1966) ends funding for MT in America
  the lack of real results realized

- ELIZA (MIT): Fraudulent NLP in a simple pattern matcher psychotherapist
  - It’s true, I am unhappy.
  - Do you think coming here will make you not to be unhappy?
  - I need some help; that much is certain.
  - What would it mean to you if you got some help?
  - Perhaps I could learn to get along with my mother.
  - Tell me more about your family.

- Early corpora: Brown Corpus (Kudera and Francis)
Some brief history: 1970s


- Could interpret questions, statements commands.
  - Which cube is sitting on the table?
  - The large green one which supports the red pyramid.
  - Is there a large block behind the pyramid?
  - Yes, three of them. A large red one, a large green cube, and the blue one.
  - Put a small one onto the green cube with supports a pyramid.
  - OK.
Some brief history: 1980s

- Procedural → Declarative (including logic programming)
- Separation of processing (parser) from description of linguistic knowledge.
- Representations of meaning: procedural semantics (SHRDLU), semantic nets (Schank), logic (perceived as answer; finally applicable to real languages (Montague))
- Large scale Knowledge Resources (Lenat and Cyc)
- Working MT in limited domains (METEO)
- Input Method Editors for Japanese, Chinese
Some brief history: 1990s

➢ Resurgence of finite-state methods for NLP: in practice they are incredibly effective.

➢ Speech recognition becomes widely usable.

➢ Large amounts of digital text become widely available and reorient the field. The Web.

➢ Resurgence of probabilistic/statistical and machine learning methods, led by a few centers, especially IBM (speech, parsing, Candide MT system), often replacing logic for reasoning.

➢ Recognition of ambiguity as key problem.

➢ Commercial rule-based MT (Systran, Atlas, . . . )
Some brief history: 2000s

A bit early to tell! But maybe:

- Continued surge in probability, Bayesian methods of evidence combination, and joint inference.
- Emphasis on meaning and knowledge representation.
- Emphasis on discourse and dialog.
- Strong integration of techniques, and levels: bringing together statistical NLP and sophisticated linguistic representations.
- Increased emphasis on unsupervised learning.
- Increase emphasis on social approaches (wikis and blogs)
General comments

➢ Even ‘simple’ applications might need complex knowledge sources

➢ Applications cannot be 100% perfect

➢ Applications that are < 100% perfect can be useful

➢ Aids to humans are easier than replacements for humans

➢ NLP interfaces compete with non-language approaches

➢ Shallow processing on arbitrary input or deep processing on narrow domains

➢ Limited domain systems require extensive and expensive expertise to port
Example Applications of NLP

➢ Input Methods for CJK

➢ Indexing and information retrieval: Google normalization of look up terms association nets (page rank)

➢ Spelling and Grammar correction: MS Word

➢ Aggregation and Clustering: Google News

➢ Information Extraction: CiteSeer, Google Scholar

➢ Crossword Solving: One Across

➢ Automated Phone Menus: All over the place
Machine Translation
Machine Translation: the Great Task

➢ The demand for Machine Translation

➢ Problems
  ➢ Linguistic
  ➢ Technical
  ➢ Interface

➢ Kinds of Machine Translation
  ➢ Rule-based (Knowledge-based): Transfer, Interlingual
  ➢ Data-driven: Example-based, Statistical

➢ Successful and Unsuccessful Applications

➢ The Future
Increased Demand

➤ Growing amount of cross-lingual communication
  ➤ A tenth of the U.N. Budget
  ➤ Global Economy requires interlingual communication
  ➤ Easy access to other languages over the Internet

➤ Large amounts of machine readable text
  ➤ Increase in the use of computers
  ➤ Improvement of scanners and speech-to-text systems

➤ A desire for quick translation
Linguistic Background

➢ No settled linguistic theory exists
  ➢ Can’t just implement it
  ➢ Non-core phenomena are very common
    often neglected by mainstream linguist research

➢ Translation is AI complete.
  ➢ Requires full knowledge of the world.

(30) 3階にオフィスがある
3 kai ni offisu ga aru
3 floor on office NOM have

There {is an office|are offices} on the third floor.

➢ Often requires specialist domain knowledge
➢ Even humans make mistakes
What should the output be?
I like words

- syntactic trees?
  \[(S \ (NP \ I) \ (VP \ (V \ like) \ (NP \ (N \ words))))\]
- semantic logical forms?
  \[like(speaker,word+PL)\]
- pragmatic speech acts?
  Speaker wants hearer to believe that speaker believes that
  like(speaker,word+PL)
- whatever is useful?
  watashi-wa kotoba-ga suki-da

How to model an infinite set of expressions?

What should the basic units of translation be?
Transfer — equivalents?

➤ Category changes: postwar → nach dem Krieg “after the war”

➤ Lexical gaps: wear → haku “wear below waist”
   kiru “wear above waist”
   kaburu “wear on the head”

➤ Head switching:

(31) I swam across the river
(32) J’ai traversé le fleuve en nageant
      I crossed the river by swimming
Transfer — mismatches

“The differences in languages lie not in what you can say, but rather what you must”
Roman Jakobson

- number
- definiteness
- gender
- politeness
- evidentiality
Transfer — discourse

➢ Different discourse order in Japanese and American stock market reports
  Conclusion at the end vs the beginning

➢ Differing conventional implicatures
  -te-mo ii “conditional” is much less positive than you may

➢ *Must you go, can’t you stay?* (in middle class English)
  ⇒ go home at once!

➢ Some work on speech acts in the Verbmbobil project

⊗ All to often ignored entirely
Technical Limitations

➢ Problems of Economy

➢ Memory Limitations (8Gb/sentence)
➢ Speed Problems (one hour/sentence)
   Some recent improvements in parallel processing

➢ Problems of Consistency

➢ Increased lexical choice leads to less consistency
➢ Large systems are often hard to predict

➢ The need for more information
Knowledge Acquisition

- Unknown words: 
  *Yahoo*, *sidewalk*, *togs*

- Unknown senses:
  (satellite) *footprint*, (system) *daemon*

- Unknown relationships:
  *Machine translation is easy, NOT!*

- Partially solved by:
  - Domain Specific Lexicons (and rules)
Interface

➢ OCR (Optical Character Recognition)

➢ Speech-to-Text
  ➢ Almost always impoverished
    no prosody, no spelling, no Chinese characters
  ➢ Is frequently wrong
    wreck a nice peach vs recognize speech

➢ Text
  ➢ Must often be cleaned
    correct spelling errors, loose fancy fonts
  ➢ May have useful structural mark up
    list header, list item
Rule based approaches to MT

➤ Transfer Based: \( n(n - 1) \)

Ja → En; Ja → No

➤ Commercial systems: SYSTRAN, METAL, L&H etc

➤ Research systems: ALT-J/E, Verbmobil, Logon, OpenTrad

➤ Interlingua: \( 2n \)

Ja → Inter; Inter → En, No

➤ Multilingual systems: Eurotra, CCIC, UNL, OntoSem

There is a convergence in real life, with an interlingual core.

“The first "live" translation occupied a 4Mb Microvax running Ultrix and C-Prolog for a complete weekend some time in early 1987. The sentence, translated from English into Danish, was "Japan makes computers".”

(On Eurotra, from Wikipedia)
Data-driven approaches to MT

➢ Memory-based translation:
   ➢ Translation Memories
     Very popular as an aid

➢ Example Based:
   ➢ Kyoto University: Nagao et al.; Kurohashi et al.
   ➢ ATR: TDMT

➢ Statistical Translation:
   ➢ IBM: Brown et al.; Google, Moses, . . .
   ➢ Automatic acquisition of lexico-grammar
   ➢ Current systems use simple model and lots of data
Data-driven approaches are popular

- They use the computer’s strengths
  - heavy computation over lots of data

- Results peak quickly

- Language is quirky ⇒ there are lots of exceptions
  - it is easier to pick them up from data than introspection
  - especially if you don’t speak the language

- Automatic evaluation favors statistical methods
Successful Applications

➢ Dissemination
  ➢ Producing information and making it widely available

➢ Browsing

➢ Machine Aided Translation
Dissemination: Controlled language

➤ Narrow Domain:
➤ Canada: Meteo
➤ NTT: ALTFLASH
➤ Travel aids
➤ Military tools

➤ Controlled Language:
➤ CMU: KANT
   Control languages
➤ Many manuals
Browsing

➢ Internet access

➢ Google, SYSTRAN, Pensee, Babelfish and many more
Machine Aided Translation

➢ Translation memory

➢ Dictionary look up/construction

➢ Automatic glossing

➢ Writing Assistance
Unsuccessful Applications

Fully Automatic High Quality Translation

But we still keep trying . . .
Spinoffs

➢ Automatic Proof-reading

➢ Writing Assistants
  Spell Checkers, Grammar Checkers

➢ Text-to-Speech

➢ Text-to-Braille

➢ Transliteration
References

➤ Websites and Mailing Lists

➤ linguist-list <www.linguistlist.org>
➤ acl-anthology <http://aclweb.org/anthology-new/>
➤ cmp-lg archives <http://xxx.lanl.gov/list/cs.CL/recent>

➤ Conferences

➤ COLING, ACL, ANLP, ...
➤ Theoretical and Methodological Issues in Machine Translation
➤ Machine Translation Summit
➤ Association for Machine Translation in the Americas
Journals

- Computational Linguistics
- Language Resources and Evaluation
- Computer Speech and Language
- Natural Language Engineering
- Machine Translation
- Journal of Natural Language Processing (English and Japanese)
Recogngments and Disclaimer

These slides contain material from Ann Copestake, Chris Manning, Andrew McCallum and Jason Eisner. English text was parsed with the English Resource Grammar.

Warning:

➢ This course can be fairly dry (I will try fit in some jokes :)
  ➢ I will try to make it interactive: beware — this means you!

➢ There is very little extra reading

➢ You will have trouble if you don’t come to class

➢ No ducks were harmed in making these slides