HG8003 Technologically Speaking: The intersection of language and technology.

Final Review and Conclusions

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

Lecture 12
Location: LT8
## 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>Quiz</td>
<td></td>
<td></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>Quiz</td>
<td></td>
<td></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
Overview of the Exam

Quiz 1: 20%; Quiz 2: 20%; Final Exam: 60%

Part A (50%): 50 multiple choice questions (like the quiz)

Part B (50%): 5 short questions (∼1 page each)
  - We will go through some sample questions today

Non-English examples: transliterate and gloss
  - 犬 inu “dog”
  - ayam “chicken”

Make your answers easy to read — help me give you marks
Review: Goals of this course

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

➢ Know why language processing is so difficult interesting

➢ Know what the current state of the art is

➢ Learn a little about best practice (evaluation)
Introduction
Review of Introduction

➢ Natural language is ambiguous and has a lot of variation

➢ We need to resolve this ambiguity for many tasks
  ➢ Humans are good at this task
  ➢ Machines find it hard

➢ Example of vagueness: Names
  ➢ Vary in Word order, Segmentation, Orthography, Case
    * ボンドフランシス
    * フランシスボンド
    * フランシス・ボンド
    * Francis・Bond
    * Francis・BOND
Layers of Linguistic Analysis

There are many layers of linguistic analysis

1. Phonetics & Phonology (sound)
2. Morphology (intra-word)
3. Syntax (grammar/structure)
4. Semantics (sentence meaning)
5. Pragmatics (contextual meaning)
Representing Language
Review of Representing Language

➤ Writing Systems

➤ Encodings

➤ Speech

➤ Bandwidth

Final Review and Conclusions
Three Major Writing Systems

➢ Alphabetic (Latin)
   ➢ one symbol for each consonant or vowel (simple sounds)
   ➢ Typically 20-30 base symbols (1 byte)

➢ Syllabic (Hiragana)
   ➢ one symbol for each syllable consonant+vowel (complex sounds)
   ➢ Typically 50-100 base symbols (1-2 bytes)

➢ Logographic (Hanzi)
   ➢ pictographs, ideographs (sound-meaning combinations)
   ➢ Typically 10,000+ symbols (2 bytes for most, 3 for all)
Encoding

➤ Need to map characters to bits

➤ More characters require more space

➤ Moving towards unicode for everything

➤ If you get the encoding wrong, it is gibberish
Speech

- Speech is an analog signal
  - considerable variation
  - no clear boundaries

- Hard to convert to symbols
  - single speaker trained models work OK
  - noisy speech is still an unsolved problem
## Speed is different for different modalities

Speed in words per minute (one word is 6 characters)
(English, computer science students, various studies)

<table>
<thead>
<tr>
<th>Activity</th>
<th>Speed (wpm)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>300</td>
<td>200 (proof reading)</td>
</tr>
<tr>
<td>Writing</td>
<td>31</td>
<td>21 (composing)</td>
</tr>
<tr>
<td>Speaking</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td>Hearing</td>
<td>150</td>
<td>210 (speeded up)</td>
</tr>
<tr>
<td>Typing</td>
<td>33</td>
<td>19 (composing)</td>
</tr>
</tbody>
</table>

➢ Reading >> Speaking/Hearing >> Typing/Writing

⇒ Speech for input
⇒ Text for output
Representing Meaning
Review of Representing Meaning

➢ Three ways of defining meaning
  ➢ Attributional (Compositional)
  ➢ Relational
  ➢ Distributional
Attributional Meaning

➢ Give a semantic description of word use in isolation of the categorisation of other lexical items
  ➢ definitions
  ➢ decompositional semantics (break down into primitives)

➢ Easy for humans to understand

➢ Hard to decide on sense boundaries (granularity: splitters vs. lumpers)

➢ Definitions are circular (the grounding problem)

➢ Hard to be consistent
Relational Meaning

- Capture correspondences between lexical items by way of a finite set of pre-defined semantic relations
- Captures many generalizations usefully
- Hard to make complete
- Leads to large, complex graphs
Distributional Meaning

- Capture word meanings as collections of contexts in which words appear
  - n-grams
  - syntactic relations
  - sentences
  - documents

- Good for synonymy, not so good for antonymy

- Computationally tractable
Some Relations in WordNet

**hyponyms:** Y is a hyponym of X if every Y is a (kind of) X
  
  \[
  \text{cat} \subset \text{animal}
  \]

**hypernyms:** Y is a hypernym of X if every X is a (kind of) Y

**meronym:** Y is a meronym of X if Y is a part of X
  
  \[
  \text{nose meronym face (part-of)}
  \]
  \[
  \text{wolf meronym pack (member-of)}
  \]

**holonym:** Y is a holonym of X if X is a part of Y

**antonym:** Y is an antonym of X if they are opposites
  
  \[
  \text{hot} \leftrightarrow \text{cold}
  \]
Why are dictionaries important?

➢ For humans
  ➢ find meaning of unknown words
  ➢ find more information about known words
  ➢ codify knowledge about word usage (glossaries)

➢ For machines
  ➢ store information about words
  ➢ link between text and knowledge
Words, Lexicons and Ontologies
Review of Words, Lexicons and Ontologies

- Storing information on machines allows us to manipulate it in many ways
- Information for humans can be made easier to search and validate
  - Machine Readable Dictionaries
- Information for machines must be made explicit
  - Dictionaries for various processors
  - Ontologies
- We can reuse knowledge to make new resources
**definition** (n) a concise explanation of the meaning of a word or phrase or symbol

➢ Headword: definition

➢ Part of Speech: n (noun)

➢ Definition:

➢ genus: explanation (class)

➢ differentia: concise; of the meaning of a word or phrase or symbol (meaning within that class)
Erin McKean’s TED Talk

➢ Redefining the dictionary (by Erin McKean; TED Talk 2007)
  (http://blog.ted.com/2007/08/30/redefining_the/)

➢ Dictionaries still don’t cover all words
many, many new words are undefined
as many as one per book?

➢ We need to define these words in context

➢ On-line dictionaries allow us to do this without space limitations
  ➢ Dictionaries can describe usage with real examples
Ontology Example (WordNet)

Synset 06744396-n: definition

Def: ‘a concise explanation of the meaning of a word or phrase or symbol.’

Hype: account
Hypo: redefinition, explicit definition, recursive definition, stipulative definition, contextual definition, ostensive definition, dictionary definition

SUMO: = equivalentContentInstance

Has-Part: genus
Has-Part: differentia
What is an Ontology?

- A set of statements in a formal language that describes/conceptualizes knowledge in a given domain
  - What kinds of entities exist (in that domain)
  - What kinds of relationships hold among them

- Ontologies usually assume a particular level of granularity
  - doesn’t capture all details
How to build Resources?

➢ Boot strap ontologies from MRDs

1. Parse definitions to find the genus
2. Take it as hyponym or parse further if it is relational
   abbreviation, nickname, kind, polite form, . . .

➢ Boot strap bilingual dictionaries from other bilingual dictionaries

➢ Link through a pivot language (≈ 65\% precision)
➢ Add in semantic links (≈ 80\% precision)
➢ Link through two pivot languages (≈ 97\% precision)

➢ Text mining . . .
How to build Resources?

➢ Take advantage of the fact that syntax is motivated by semantics

➢ Bounded individual things are countable
➢ Divisible substances are uncountable

1. Predict countability from semantic classes

➢ <animal> is countable
➢ <meat> is uncountable

2. Predict verbal structure from semantic classes

➢ Learn from corpora

➢ Find defining patterns:
  * many $N$ implies $N$ is countable
  * much $N$ implies $N$ is uncountable
Text Mining and Knowledge Acquisition
Review of Text Mining and Knowledge Acquisition

➢ Too much information for people to handle: Information Overload

➢ Text mining is:

The discovery by computer of new, previously unknown information, by automatically extracting information from a usually large amount of different unstructured textual resources.
LARGE amounts of data

➢ You can tolerate some noise
  ➢ conversion errors, spelling errors, etc.

➢ Shallow robust techniques are needed

➢ Typically only consider more things with more than \( n \) instances
  ➢ Hope that errors are infrequent
Template Filling

➢ Looking for known relations in text
  ➢ fill slots in a template

➢ Restricted search space gives high accuracy
Named Entity Recognition

➢ Identify interesting things
   (People, Organizations, Places, Dates, Times, . . . )

➢ Typically done as a sequence labeling task
  ➢ Tag as Inside, Outside, Beginning, (IOB)

➢ Train a classifier with annotated text
  ➢ Features include: Words, Stems, Shape, POS, Chunks, Gazetteers
Relation Detection

- We can use patterns to find relation tuples

- \(< S, \text{ hypernym, } A >\)
  - \(S (\text{such as}|\text{like}|\text{e.g.}) \ A; A \text{ and other } S; S (\text{including}|\text{especially}) \ A\)

- \(< A, \text{ synonym, } B >\)
  - \(\text{both } A \text{ and } B; \text{either } A \text{ or } B; \text{neither } A \text{ nor } B\)

- Simple patterns are easy to find in vast data sources

- High frequency patterns can be quite reliable
  - multiple patterns increase confidence
Sample Question

➢ Outline a method of deriving an ontology from a text corpus using patterns

➢ Give two patterns, and examples of text they would match, for English

➢ Give two patterns, and examples of text they would match, for a non-English language (don’t forget to gloss)
## Evaluation Measures

<table>
<thead>
<tr>
<th>System</th>
<th>Actual</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>target</td>
<td>not target</td>
</tr>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>

Precision = \( \frac{tp}{tp + fp} \); Recall = \( \frac{tp}{tp + fn} \); \( F_1 = \frac{2PR}{P+R} \)

| tp     | True positives: system says Yes, target was Yes |
| fp     | False positives: system says Yes, target was No |
| tn     | True negatives: system says No, target was No |
| fn     | False negatives: system says No, target was Yes |
Text Mining Summary

➢ There is a lot of information out there

➢ Much of it is unstructured text

➢ Using NLP techniques we can extract this information
  ➢ But we can’t trust it all

➢ Well defined tasks on restricted domains work best
Structured Text
Why Structured Text?

➢ Reduce Ambiguity
  ➢ Need to make meaning explicit

➢ Traditionally this is done by annotating text in some way

➢ The best way is using Logical Markup
Visual Markup vs Logical Markup

➤ Visual Markup (Presentational)
  ➤ What you see is what you get (WYSIWYG)
  ➤ Equivalent of printers’ markup
  ➤ Shows what things look like

➤ Logical Markup (Structural)
  ➤ Show the structure and meaning
  ➤ Can be mapped to visual markup
  ➤ Less flexible than visual markup
  ➤ More adaptable (and reusable)
XML: eXtensible Markup Language

- XML is a set of rules for encoding documents electronically.
- Based on a simplified SGML
- XML’s design goals emphasize simplicity, generality, and usability.
- It is a textual data format
- It supports many encodings, with Unicode preferred
- It can represent arbitrary data structures, for example in web services.
- XML syntax can be specified and validated
Validation

➢ Validation is very important

➢ Ill-formed data makes parsing complex

➢ Early detection of errors is cost-effective

➢ Validated data is easy to maintain
Semantic Web
Goals of the Semantic Web

➢ Web of data
  ➢ provides common data representation framework
  ➢ makes possible integrating multiple sources
  ➢ so you can draw new conclusions

➢ Increase the utility of information by connecting it to definitions and context

➢ More efficient information access and analysis

E.G. not just "color" but a concept denoted by a Web identifier:
<http://pantone.tm.example.com/2002/std6#color>
Semantic Web Architecture (details)

➤ Identify things with Uniform Resource Identifiers

➤ Universal Resource Name: urn:isbn:1575864606

➤ Identify relations with Resource Description Framework

➤ Triples of <subject, predicate, object>
➤ Each element is a URI
➤ RDFs are written in well defined XML
➤ You can say anything about anything

➤ You can build relations in ontologies (OWL)
➤ Then reason over them, search them, . . .
Citation, Reputation and PageRank
Citation Networks

How can we tell what is a good scientific paper?

- **Content-based**
  - Read it and see if it is interesting (hard for a computer)
  - Compare it to other things you have read and liked
- **Context based: Citation Analysis**
  - See who else read and thought it interesting enough to cite
Reputation and Citation Analysis

➤ One major use of citation networks is in measuring productivity and impact of the published work of a scientist, scholar or research group.

➤ Some scores are

➤ Total Number of Citations (Pretty Useful)
➤ Total Number of Citations minus Self-citations
➤ Total Number of (Citations / Number of Authors)

➤ Problems

➤ Not all citations are equal: citations by ‘good’ papers are better
➤ Newer publications suffer in relation to older ones

➤ Weight Citations by Quality of the paper
Gaming Citations

➤ Least/Minimum Publishable Unit
  ➤ Break research into small chunks to increase the number of citations
  ➤ Sometimes there is very little new information

➤ Self citation, in-group citation

➤ Write only proceedings (some journals are not often read)

➤ Submitting only to High Impact factor journals
  
  You improve what gets measured
  not necessarily what you want to improve
Recall how hyperlinks are written:

\[ \text{<a href="http://path.to.there/page/HG803/"}>HG803: Language, Technology and the Internet.</a> \]

For more information about Language, Technology and the Internet, see the <a href="http://..">HG803 Course Page.</a>

Link analysis builds on two intuitions:

1. The hyperlink from A to B represents an endorsement of page B, by the creator of page A.
2. The (extended) anchor text pointing to page B is a good description of page B.

This is not always the case; for instance, most corporate websites have a pointer from every page to a page containing a copyright notice.
PageRank as Citation analysis

➤ Citation frequency can be used to measure the impact of an article.
   ➤ Simplest measure: Each article gets one vote – not very accurate.

➤ On the web: citation frequency = inlink count
   ➤ A high inlink count does not necessarily mean high quality . . .
   ➤ . . . mainly because of link spam.

➤ Better measure: weighted citation frequency or citation rank
   ➤ An article’s vote is weighted according to its citation impact.
   ➤ This can be formalized in a well-defined way and calculated.
PageRank as Random walk

Imagine a web surfer doing a random walk on the web

- Start at a random page
- At each step, go out of the current page along one of the links on that page, equiprobably

In the steady state, each page has a long-term visit rate. What proportion of the time someone will be there

This long-term visit rate is the page's PageRank.

PageRank = long-term visit rate = steady state probability
Teleporting – to get us out of dead ends

➢ At a **dead end**, jump to a random web page with prob. $1/N$.

➢ At a **non-dead end**, with probability 10%, jump to a random web page (to each with a probability of $0.1/N$).

➢ With remaining probability (90%), go out on a random hyperlink.

➢ For example, if the page has 4 outgoing links: randomly choose one with probability $(1-0.10)/4=0.225$

➢ 10% is a parameter, the **teleportation rate**.

➢ Note: “jumping” from dead end is independent of teleportation rate.
Each inbound link is a positive vote.
Example Graph: Weighted

Pages with higher PageRanks are lighter.
Gaming PageRank

➤ Link Spam adding links between pages for reasons other than merit. Link spam takes advantage of link-based ranking algorithms, which gives websites higher rankings the more other highly ranked websites link to it. Examples include adding links within blogs.

➤ Link Farms creating tightly-knit communities of pages referencing each other, also known humorously as mutual admiration societies.

➤ Scraper Sites “scrape” search-engine results pages or other sources of content and create "content" for a website. The specific presentation of content on these sites is unique, but is merely an amalgamation of content taken from other sources, often without permission.
Comment spam is a form of link spam in web pages that allow dynamic user editing such as wikis, blogs, and guestbooks. Agents can be written that automatically randomly select a user edited web page, such as a Wikipedia article, and add spamming links.

The nofollow link: a value that can be assigned to the rel attribute of an HTML hyperlink to instruct some search engines that a hyperlink should not influence the link target’s ranking in the search engine’s index.

- Google does not index the target of a link marked nofollow.
- Yahoo! does not include the link in its ranking
- ...

There is a continuous battle between

- Search companies, who want to get the most useful page to the user
- Page writers, who want to get their page read

All metrics get gamed
Digital object identifier

➢ DOI: a string used to uniquely identify an electronic document or object

➢ Metadata about the object is stored with the DOI name
➢ The metadata includes a location, such as a URL
➢ The DOI for a document is permanent, the metadata may change
➢ Gives a Persistent Identifier (like ISBN)

➢ The DOI system is implemented through a federation of registration agencies coordinated by the International DOI Foundation

➢ By late 2013 approximately 85 million DOI names had been assigned by some 9,500 organizations

➢ DOI: 10.1007/s10579-008-9062-z
   http://www.springerlink.com/content/v7q114033401th5u/
Machine Translation Revisited
Review of Machine Translation

MT is difficult

- Inherent ambiguity in language
  * Lexical: sense mismatches, lexical gaps
  * Structural: head switching, reference resolution

A full solution requires world knowledge

But we can approximate the solution

- contextual rules (RBMT)
- learned examples (EBMT)
- frequencies and ‘language models’ (SMT)
- hybrid combinations (SMT+syntax, RBMT+models)
  * combing output of several systems (system combination)
Internationalization and Localization

➢ Internationalization (i18n)
  ➢ designing a software application so that it can be adapted to various languages and regions without engineering changes

➢ Localization (L12n)
  ➢ adapting internationalized software for a specific region or language by adding locale-specific components and translating text
    * Text and menus
    * Government assigned numbers (such as the Social Security number in the US, National Insurance number in the UK, PIN in Singapore)
    * Telephone numbers, addresses and international postal codes
    * Currency (symbols, positions of currency markers)
    * Culturally sensitive examples
Empirical NLP
Review of Empirical NLP

Empirical denotes information gained by means of observation, experience, or experiment.

Emphasises testing systems by comparing their results on held-out gold standard data.

1. Create a gold standard or reference (the right answer)
2. Compare your result to the reference
3. Measure the error
4. Attempt to minimize it globally (over a large test set)
Error Measures

➢ Word Error Rate
   * Error is the minimum edit distance between system and reference

➢ BLEU
   * compares word \(n\)-gram overlap with reference translations

\[
W E R = \frac{S + D + I}{N}
\]

\[
B L E U \approx \sum_{i=1}^{n} \frac{n\text{-grams in sentence and reference}}{|n\text{-grams}|}
\]

* The dog bark ⇔ The dog barks

  the ⇔ the
dog ⇔ dog
bark ⇔ barks
the dog ⇔ the dog
dog bark ⇔ dog barks
the dog barks ⇔ the dog barks
Error Measures

➤ Manual Evaluation

➤ Fluency: How natural does the translation sound?
➤ Adequacy: How much of the meaning is translated?
BLEU pros and cons

➤ Good
   ➢ Easy to calculate (if you have reference translations)
   ➢ Correlates with human judgement to some extent

➤ Bad
   ➢ Doesn’t deal well with variation
   ➢ Biased toward $n$-gram models

➤ How to improve the reliability?
   ➢ Use more reference sentences
   ➢ Use more translations per sentence
   ➢ Improve the metric: METEOR
     * add stemmed words; add WordNet matches (partial score)
Problems with MT Testing

➢ You get better at what you test

➢ You may over-fit your model to the data

➢ If the metric is not the actual goal things go wrong
  ➢ BLEU score originally correlated with human judgement
  ➢ As systems optimized for BLEU
  ➢ . . . they lost the correlation
  ➢ You can improve the metric, not the goal

➢ The solution is better metrics, but that is hard for MT

➢ We need to test for similar meaning: a very hard problem
Why do we test in general?

Testing is important for the following reasons

1. Confirm Coverage of the System

2. Discover Problems

3. Stop Backsliding
   ➢ Regression testing — test that changes don’t make things worse

4. Algorithm Comparison
   ➢ Discover the best way to do something

5. System comparison
   ➢ Discover the best system for a task
How do we test?

- **Functional Tests** (Unit tests)
  - Test system on test suites

- **Regression Tests**
  - Test different versions of the system

- **Performance Tests**
  - Test on normal input data

- **Stress Tests** (Fuzz tests)
  - Test on abnormal input data
Morphological Analysis and Tagging
Review of Morphological Analysis and Tagging

➢ Morphological analysis is the analysis of units within the word
  ➢ Segmentation: splitting text into words
  ➢ Lemmatization: finding the base form
  ➢ Tokenization: splitting text into tokens (for further processing)

➢ Part of Speech tagging assigns POS tags to words or tokens
  ➢ Often combined with morphological analysis
Segmentation

➢ Separate a stream into units
   ➢ non-spaced languages (Chinese, Thai, ...)
   ➢ speech input

➢ Need both good lexicons and unknown word handling

➢ Typically learn rules from a tagged corpus
   ➢ treat rare words as unknown words

➢ Can pass ambiguity to the next stage
Lemmatization

- Lemmatization is the process of finding the stem or canonical form
- You must store all irregular forms
- You need rules for the rest (inflectional morphology)
- Rare words tend to be regular
  - For languages without much morphology, you can expand everything offline
- Most rules depend on the part-of-speech
  - So lemmatization is done with (or after) part-of-speech tagging
Tokenization

- Splitting **words** into **tokens** — the units needed for further parsing
  - Separating punctuation
  - Adding BOS/EOS (Beginning/Eng of sentence) markers
  - Splitting into stem+morph: *went* → *go*+*ed*
  - Normalization
    * *data base*
    * *data-base*
    * *database*
  - Possibly also chunking
    * *in order to* → *in_order_to*

- This process is very task dependent
Parts of Speech (POS)

➢ Four main open-class categories

  **Noun**  heads a noun phrase, refers to things
  **Verb**  heads a verb phrase, refers to actions
  **Adjective**  modifies Nouns, refers to states or properties
  **Adverb**  modifies Verbs, refers to manner or degree

➢ Closed class categories vary more

  **Preposition**  *in, of*: links noun to verb (postposition)
  **Conjunction**  *and, because*: links like things
  **Determiner**  *the, this, a*: delimits noun’s reference
  **Interjection**  *Wow, um*
  **Number**  *three, 125*: counts things
  **Classifier**  *頭 “animal”*: classifies things
Part of Speech Tagging

➢ Exploit knowledge about distribution
  ➢ Create tagged corpora

➢ With them, it suddenly looks easier
  ➢ Just choose the most frequent tag for known words (I pronoun, saw verb, a article, . . .)
  ➢ Make all unknown words proper nouns
  ➢ This gives a baseline of 90% (for English)

➢ The upper bound is 97-99% (human agreement)
  ➢ The last few percent are very hard
Representing ambiguities

- Two opposite needs:
  - Disambiguate early
    → Improve speed and efficiency
  - Disambiguate late
    → Can resolve ambiguities with more information

- Several Strategies:
  - Prune: Discard low-ranking alternatives
  - Use under-specification (keep ambiguity efficiently)
  - Pack information in a lattice (keep ambiguity efficiently)

- Combine tasks instead of pipe-lining
Parsing and Generation
Review of Parsing and Generation

➤ Parsing
➤ Words to representation

➤ Generation
➤ Representation to words

➤ Two main syntactic representations:
➤ Dependencies (word-to-word)
➤ Phrase Structure Trees (with phrasal nodes)


Efficiency is important

➢ Need to avoid exponential processing

➢ Least complex is best

  constant < linear < Quadratic dependency pars. < polynomial HPSG pars. < exponential

➢ May sacrifice some accuracy for speed

  ➢ Discard low ranked paths (known as pruning)
Dependency and PSG

➢ Dependency Grammars $O(n^2)$

People saw her duck

➢ Phrase Structure Grammars $O(n^3)$
Sample Question

➢ Show the ambiguity in *He gave her cat food* using:

- Brackets
- Paraphrases
- Dependencies
- Phrase structure trees

➢ Give an example of an ambiguous sentence in a language other than English, and show the ambiguity using:

- Different English glosses
- Dependencies
- Phrase structure trees
Dependencies, Brackets and Paraphrases

(N V:give D N N)
He gave her cat food

(He (gave (her cat) (food)))
He gave food to her cat

(N V:give Pr N N)
He gave her cat food

(He (gave (her) (cat food)))
He gave cat food to her
He gave her cat food

(He (gave (her cat food)))
The cat food which was hers he gave [to someone]

To show the ambiguity, you can minimally show:

- He gave her (cat food)
- He gave (her cat) food
- He gave (her cat food)
Generation: Process

- Generation is the process of producing language

- At the lowest level, it involves:
  - Taking a semantic realization and producing a string
  - Normally multiple strings are possible
    * Over-generate and rank with a language model
Example-based Machine Translation
Example-based Machine Translation

➢ When translating, reuse existing knowledge:

  0 Compile and align a database of examples
  1 Match input to a database of translation examples
  2 Identify corresponding translation fragments
  3 Recombine fragments into target text

➢ Example:

➢ Input: He buys a book on international politics
➢ Data:
  * He buys a notebook – Kare wa noto o kau
  * I read a book on international politics – Watashi wa kokusai seiji nitsuite kakareta hon o yomu
➢ Output: Kare wa kokusai seiji nitsuite kakareta hon o kau
Example-based Translation: Advantages/Disadvantages

➢ Advantages

➢ Correspondences can be found from raw data
➢ Examples give well structured output if the match is big enough

➢ Disadvantages

➢ Lack of well aligned bitexts
➢ Generated text tends to be incohesive
Translation Memories

➢ Translation Memories are aids for human translators

➢ Store and index entire existing translations
➢ Before translating new text
  * Check to see if you have translated it before
  * If so, reuse the original translation

➢ Checks tend to be very strict ⇒ translation is reliable

➢ Identical except for white-space differences
➢ The translator is in control
➢ Translation companies can pool memories, giving them an advantage
Statistical Machine Translation
Statistical Machine Translation (SMT)

➢ Find the translation with the highest probability of being the best.

➢ Probability based on existing translations (bitext)

➢ Balance two things:

➢ Adequacy (how faithful the translation to the source)
➢ Fluency (how natural is the translation)

➢ These are modeled by:

➢ Translation Model: $P(T|S)$
  how likely is it that this translation matches the source

➢ Language Model: $P(T)$
  how likely is it that this translation is good English

➢ Overall: $\hat{T} = \arg\max_T P(S|T) = \arg\max_T P(T|S)(T)$
Translation Model (IBM Model 4)

\[ P(J, A|E) \]

Fertility Model
\[ \prod n(\phi_i|E_i) \]

NULL Generation Model
\[ (m-\phi_0) p_0^{m-2\phi_0} p_1^{\phi_0} \]

Lexicon Model
\[ \prod t(J_j|E_{A_j}) \]

Distortion Model
\[ \prod d_1(j - k|\mathcal{A}(E_i)\mathcal{B}(J_j)) \]
\[ \prod d_1>(j - j'|\mathcal{B}(J_j)) \]

could you recommend another hotel

could could recommend another another another hotel

could could recommend NULL another another another hotel

ていただけます紹介しを他のホテルか

他のホテルを紹介していただけますか

Millions of candidates are produced and ranked.

Final Review and Conclusions
More data improves BLEU:  

Doubling the translation model data gives a 2.5% boost.  
Doubling the language model data gives a 0.5% boost.  
For linear improvement in translation quality the data must increase exponentially  
* BLEU +10% needs $2^4 = 16$ times as much bilingual data  
* BLEU +20% needs $2^8 = 256$ times as much bilingual data  
* BLEU +30% needs $2^{12} = 4096$ times as much bilingual dbilingual data
Transfer in Machine Translation
Review of Transfer

➤ Approaches to Transfer

➤ Particular Problems (and solutions)

➤ Ways to improve
Approaches to Transfer

✧ The place of transfer

✧ Parse source text to source representation (SR)
✧ Transfer this to some target representation (TR)
✧ Generate target text from the TR

✧ The depth of transfer

Direct Transfer  Source representation is words or chunks
Syntactic Transfer  Source representation is trees
Semantic Transfer  Source representation is meaning
Interlingua  Transfer to a universal meaning representation
The Vauquois Triangle

- Analysis
  - Semantic Transfer
  - Syntactic Transfer
  - Direct Translation

Interlingua

Generation

Source Language

Target Language

Final Review and Conclusions
a) Transfer: $n(n - 1)$ engines $L1 \rightarrow L2$, $L1 \rightarrow L3$, $L1 \rightarrow L4$, $L2 \rightarrow L1$, …

b) Interlingua: $2n$ engines $L1 \rightarrow LI$, $LI \rightarrow l1$, $L2 \rightarrow LI$, $LI \rightarrow L2$, … (but LI is hard)
Problems and Solutions

➢ Lexical Choice: single words don’t give enough context to chose
  ➢ Add context dependent rules
  ➢ Add multiword expressions to lexicons (typically 60-70%)
  ➢ Use document information (User dictionaries)
  ➢ Use the most frequent translation as a default

➢ Language Differences
  ➢ Use richer representations: syntax, semantics
  ➢ Use bigger chunks
  ➢ Over-generate and rank with a statistical model
Fully Automatic High Quality Machine Translation

➤ **METEO**
- Canadian English ↔ French system
- Translates meteorology text (weather reports)
- Short, repetitive sentences
- 30 million sentences a year
- MT with human revision (< 9% of sentences revised)

➤ **ALT-FLASH**
- Japanese → English system
- Translates Stock market flash reports
- Short, repetitive sentences, speed very important
- 10 thousand sentences a year
- MT with human revision (< 2% of sentences revised)
Some well studied problems

➢ Head-switching: head is dependent in the other language

I swam across the river
→ J’ai traversé le fleuve en nageant “I crossed the river by swimming”

I went to Orchard road by Taxi.
→ Saya naik taksi ke Orchard “I rode a taxi to Orchard”

➢ Relation-changing: e.g. verb → adjective

濡れている紙 nurete iru kami “paper which is wet” → wet paper

➢ Lexical Gaps: translation missing in the source or target language

herd, pack, mob, crowd, group → mure

➢ Possessive Pronoun Drop: possessive pronouns sometimes required

鼻がかゆい hana-ga kayui “nose itchy” → my nose itches
Number mismatch: number required in one language but not the other

鼻は感覚器官だ hana-w kankakukikan da “noses sensory organ is”
→ Noses are sensory organs

Argument mismatch: Verb structure is different

watashi-ni kodomo-ga iru “to me children are” → I have children
to→SUBJECT; SUBJECT→OBJECT

Idiom mismatch: Idiomatic in one language but not the other

I lost my head “I got angry”
→ atama-ni kita “it came to my head”
I racked my brains “I thought hard”
→ chie-wo shibotta “I squeezed knowledge”
The following phenomena are hard to translate:

- Long sentences
- Coordination
- Unknown words (either new words or spelling errors)
  * new genre
  * poorly edited text
- Different language families

We can identify these and give a translatability score

This is useful to identify problems for pre-editing
This is useful to identify output for post-editing
Ways to Improve Machine Translation Quality

➢ Pre-editing: fix the text before it is translated
   Controlled language restricts the syntax and vocabulary

➢ Post-editing: fix the text after it is translated

➢ Domain-Specific: narrow the domain to restrict ambiguity

➢ User Dictionary: tune the system by developing a dictionary for a specific task

➢ Training Data:
   ➢ get more training data
   ➢ get training data that better matches the task
Word Sense Disambiguation (WSD)
Word Sense Disambiguation (WSD)

- Many words have several meanings

- We need to determine which sense of a word is used in a specific text

- With respect to a dictionary (WordNet)
  - chair = a seat for one person, with a support for the back; "he put his coat over the back of the chair and sat down"
  - chair = the officer who presides at the meetings of an organization; "address your remarks to the chairperson"

- With respect to the translation in a second language
  - chair = chaise
  - chair = directeur
Attempt to disambiguate all open-class words in a text

He put his suit over the back of the chair

Knowledge-based approaches

- Use information from dictionaries
- Definitions / Examples for each meaning
- Find similarity between definitions and current context

Position in a semantic network

- Find that “table” is closer to “chair/furniture” than to “chair/person”

Use discourse properties

- A word exhibits the same (single) sense in a discourse / in a collocation
Lesk Algorithm

Identify senses of words in context using definition overlap (Michael Lesk 1986)

1. Retrieve from MRD all sense definitions of the words to be disambiguated

2. Determine the definition overlap for all possible sense combinations
   ➢ number of words overlapping in both definitions
   ➢ context can be a window larger or smaller than a sentence

3. Choose senses that lead to highest overlap
Simplified Lesk

➢ Original Lesk definition: measure overlap between sense definitions for all words in context
   ➢ Identify simultaneously the correct senses for all words in context
   ➢ Compare the definitions of words to the definitions of words

➢ Simplified Lesk: measure overlap between sense definitions of a word and current context
   ➢ Identify the correct sense for one word at a time
   ➢ Search space significantly reduced
Sample Question

➢ Outline how to disambiguate words using Lesk and simplified Lesk.

➢ Given the following definition sentences

➢ disambiguate pine and cone in pine cone using the Lesk algorithm

➢ disambiguate pine and cone in Pine cones hanging in a tree

➢ PINE
1. kinds of evergreen tree with needle-shaped leaves
2. waste away through sorrow or illness

➢ CONE
1. solid body which narrows to a point
2. something of this shape whether solid or hollow
3. fruit of certain evergreen trees
Discourse based Methods

➢ One Sense per Discourse

➢ A word preserves its meaning across all its occurrences in a discourse
  ➢ 98% of the two-word occurrences in the same discourse carry the same meaning

➢ One Sense per Collocation

➢ A word tends to preserve its meaning when used in the same collocation
  ➢ Strong for adjacent collocations
  ➢ Weaker as the distance between words increases

➢ 97% precision on words with two-way ambiguity
Conclusions
Technologically Speaking: Conclusions

➢ Natural language is ambiguous and has a lot of variation

➢ We can model this in many ways
  ➢ $n$-grams
  ➢ trees
  ➢ graphs & vectors

➢ Using these models we can do:
  ➢ Rich Indexing (the Semantic Web); Machine Translation; . . .

➢ I hope you enjoyed at least some of the class

➢ Please contact me if you have any further questions:
  bond@ieee.org
Traditionally linguistic analysis was done largely by hand, but computer-based methods and tools are becoming increasingly more widely used in contemporary research. This course provides an introduction to the key instruments and resources available on the personal computer that can assist the linguist in performing fast and accurate quantitative analyses. Frequency lists, tagging and parsing, concordancing, collocation analysis and applications of Natural Language Processing will be discussed.

This will be teach you to use Python and the NLTK toolkit.
Good luck with the exam!