HG8003 Technologically Speaking:
The intersection of language and technology.

Machine Translation Revisited: Empirical Natural Language Processing

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

Lecture 08
Location: LT8

HG8003 (2014)
## 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>Quiz</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recess</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
Overview

➢ Review of Citation, Reputation and PageRank

➢ Citation and Reputation
  * Identifying high quality research
➢ PageRank (Google’s algorithm for ranking web pages)
  * Identifying interesting web pages

➢ Machine Translation Revisited (from a different perspective)

➢ The Empirical Revolution in NLP
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)
➢ Average (Citation * Impact Factor / Number of Authors)

➤ Problems

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

```html
<a href="http://path.to.there/page/HG8003/">HG8003: 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.
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.
Example Graph

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 Farms** creating tightly-knit communities of pages referencing each other.

➢ **Scraper Sites** "scrape" search-engine results pages or other sources of content and create "content" for a website.

➢ **Comment spam** is a form of link spam in web pages that allow dynamic user editing such as wikis, blogs, and guestbooks.

! 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.
Current Status

➤ 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
There is a great demand for Machine Translation

But there are many problems

- Linguistic
- Technical
- Interface

Kinds of Machine Translation

- Rule-based (Knowledge-based):
  - **Transfer**: $n(n - 1)$ engines for $n$ languages
  - **Interlingua**: $2n$ engines for $n$ languages
- Data-driven: Example-based, Statistical

Successful and Unsuccessful Applications
Machine translation: problems and issues

➤ An overall look at the current state of machine translation based on a panel presentation by John Hutchins (13 December 2007)

➤ Abbreviations:

RBMT  Rule-based Machine Translation
EBMT  Example-based Machine Translation
SMT   Statistical Machine Translation
SL    Source Language
TL    Target Language
PE    Post-Editing
HT    Human Translation
MT    Machine Translation
Kinds of Machine Translation

➤ Knowledge-driven

➤ Rule-based (Knowledge-based): Transfer, Interlingual

Attempt to understand the text in some way and then translate it, normally with a hand built model.

➤ Data-driven

➤ Example-based MT, Statistical MT

Attempt to build a model based on existing translations, don’t attempt understanding, just approximation.
Components of Transfer based MT

- Parse source text to source representation (SR)
- Transfer this to some target representation (TR) (next week)
- Generate target text from the TR
- If the source = target, then we are paraphrasing
How Deep Should We Go?

The Vauquois Triangle

Source Language → Interlingua → Target Language

Direct Translation → Syntactic Transfer → Semantic Transfer → Generation

Analysis
Rule Based Machine Translation

The Ikehara Discontinuity

Also known as the Copestake Inverted Funnel.
Inherent (linguistic) problems:

Bilingual lexical differences

➢ Bilingual Lexical Ambiguity
(more than one equivalent, whether ambiguous in SL or not):

➢ Schraube: screw/bolt/propeller
➢ corner: coin/angle; Ecke/Winkel
➢ light: léger, clair, facile, allumer, lumière, lampe, feu
➢ look: regarder, chercher, sembler
➢ bank: 銀行 ginkou “financial institution”, 土手 dote “embankment”
Lexical Gap: the absence of a word in a particular language
Lexicalized in one language but not the other

- river: fleuve/rivière
- Taube: dove/pigeon
- 牛油 nūuyōu: butter, ghee
- shallow waters: ape puțin adânci “not so deep waters” (Romanian)
- wear: kiru “wear clothes on upper body”, haku “wear clothes on lower body”
- *ungood

Solved (?) by

- one-to-many contextual rules (RBMT)
- examples (EBMT)
- frequencies and ‘language models’ (SMT)
Inherent (linguistic) problems:

Structural ambiguity

(1) Peter mentioned the book I sent to Mary
    Peter mentioned the book which I sent to Mary
    Peter mentioned to Mary the book which I sent [to Peter/David]
(2) We will meet the man you told us about yesterday
    \(\ldots\) the man you told us about yesterday
(3) We will meet the man you told us about tomorrow
    we will meet tomorrow the man \(\ldots\)
(4) pregnant women and children [unambiguous for HT] des femmes et des enfants enceintes [produced by MT system]

(5) a. Smog and pollution control are important factors  
b. Smog and pollution control is under consideration  
c. The authorities encouraged smog and pollution control

Often, problems such as (1), (2), and (3) are problematic for RBMT, but they may be ‘solved’ by SMT ‘language model’ and by EBMT databases. But problem (4) requires ‘knowledge’ (i.e. rule-based KBMT)
Inherent (linguistic) problems:

**Bilingual structural differences**

(6) Young people like this music

*Cette musique plaît aux jeunes gens*
This music played by young people

(7) The boy likes to play tennis

*Der Junge spielt gern Tennis*
The boy plays happily tennis

(8) He happened to arrive in time

*Er ist zufällig zur rechten Zeit angekommen*
He is accidentally on right time came
Difficult to specify transfer rules (RBMT) to cover all circumstances and contexts; but example-based (EBMT) and statistics-based (SMT) approaches no better. You need some kind of parsing.
A multi word expression is one in which the meaning of the whole is not fully predictable from the meanings of the individual words.

- **Kick the bucket**
- **Look up**
- **New York Giants**

In MT a MWE is when a translation is not word to word, but rather a set of words translates to a set of words.

(9) 1993 年 の 頭
1993 nen-no atama
1993 year ’s head
“the beginning of 1993”
“early 1993”
Non-linguistic problems of ‘reality’

(10) The soldiers$_i$ shot at the women$_j$ and some of them$_j$ fell

(11) The soldiers$_i$ shot at the women$_j$ and some of them$_i$ missed

➢ must know what ‘them’ refers to e.g. if translating into French (ils or elles)

✗ No easy solutions with linguistic rule-based approaches

✗ No easy solutions with corpus-based approaches

➢ Need to add reference resolution (currently a hot topic!)
Perhaps only solution using Artificial Intelligence approaches Knowledge-based *machine translation*, e.g. Carnegie-Mellon University)

However, perhaps this problem is exaggerated: no need to understand what AIDS and HIV are in order to translate:

(12) *The AIDS epidemic is sweeping rapidly through Southern Africa. It is estimated that more than half the population is now HIV positive.*
Problems of stylistic difference

(13) The possibility of rectification of the fault by the insertion of a valve was discussed by the engineers.

(14) The engineers discussed whether it was possible to rectify the fault by inserting a valve.

➢ **English:** Advances in technology created new opportunities.

➢ **Japanese:** Because technology has advanced, opportunities have been created.

➢ **Japanese:** Technology has advanced. There are new opportunities.

All methods of MT tend to retain SL structural features; however, theoretically SMT *language model* approach should be more TL-oriented.
Hybrid systems

clearly, none of the current MT ‘models’ are capable of solving all problems

hence search for hybrid architectures

in theory, it would seem that (on average):

- RBMT better for SL analysis
- EBMT better for transfer
- SMT best for TL generation
Problem is that different approaches not easily compatible:

- there are however research prototypes combining:
  - EBMT with statistical methods
  - EBMT using rules similar to those in RBMT systems
  - perhaps a version of EBMT will be the answer

- Currently ‘hybrid’ systems are mainly parallel systems with a selection mechanism
  - i.e., translate with several systems and chose the best translation

- However, more RBMT systems use statistical models and more SMT systems use parsers, so there is gradual convergence
Translation demand

➢ Dissemination: production of ‘publishable quality’ texts
   ➢ but, since raw output inadequate:
     * post-editing
     * control of input (pre-editing, controlled language)
     * domain restriction (reducing ambiguities)

➢ Assimilation: for extracting essential information
   ➢ use of raw output, with or without light editing

➢ Interchange: for cross-language communication (correspondence, email, etc.)
   ➢ if important: with post-editing; otherwise: without editing

➢ Information access: to databases and document collections
Post-editing:

Types of errors for correction

- Misspelling in original not recognised, therefore not translated;

- missing punctuation

- e.g. *The Commission vice president* translated as *Le président du vice de la Commission* (because no hyphen between *vice* and *president*)
Post-Editing: Complex syntax

➤ prepositions:

➢ ... *el desarrollo de programas de educación nutricional* ...
  * MT: ... *the development of programs of nutritional education*
  * PE: ... *in nutritional education* ...

➤ verb phrases:

➢ ... *el procedimiento para registrar los hogares* ...
  * MT: *the procedure in order to register the households*
  * PE: *the procedure for registering households*
Inversions:

\[ \ldots \text{la inversión de la Argentina en las investigaciones de malaria} \]

- MT: \ldots \text{the investment of Argentina in the research of malaria}
- PE: \text{Argentina’s investment in malaria research}

Reflexive verbs with inversions:

\[ \text{Se estudiarán todos los pacientes diagnostocados como} \ldots \]

- MT: \text{There will be studied all the patients diagnosed as} \ldots
- PE: \text{Studies will be done on all patients diagnosed as} \ldots

\[ \text{En 1972 se formuló el Plan Decenal de Salud para las Américas.} \]

- MT: \text{In 1972 there was formulated the Ten-Year Health Plan for the Americas}
- PE: \text{The year 1972 saw the formulation of the Ten-Year Health Plan for the Americas.}
Translators and post-editors

- post-editing by translators:
  - not foreseen initially
  - skills acquired over time and practice in real working conditions
  - requires perseverance
    (initially post-editing takes longer than complete translation)

- High quality human translation is also normally edited
Post-editing pros and cons

➤ advantages:
  ➢ translators can maintain quality control
  ➢ consistency of terminology
  ➢ repetitive matter produced by MT, linguistic quality by HT

➤ disadvantages:
  ➢ correction of ‘trivial’ mistakes; too often correcting same type of error
  ➢ style too much SL oriented
  ➢ translators as ‘slaves’ to machine

➤ need for special post-editing tools (not always provided)

➤ specially trained post-editors [still rare]
Controlled language

- Controlled authoring of the source text in standard manner, suitable for unambiguous translation

- Typical rules:
  - use only approved terminology, e.g. *windscreen* rather than *windshield*
  - use only approved sense: *follow* only as ‘come after, not ‘obey’
  - avoid ambiguous words: *replace*, either (a) remove and put back, or (b) remove and put something else in place; not *appear* but: come into view, be possible, show, think
  - only one ‘topic’ per sentence, e.g. one instruction, command
  - use short sentences, e.g. maximum 20 words
  - ...
Typical rules (cont):

- do not omit articles
- do not use pronouns instead of nouns if possible
- do not use phrasal verbs, such as *pour out*
- do not omit implied nouns
- do not drop subjects
- avoid co-ordination of phrases and clauses
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)
MT for assimilation

- publication quality not necessary
- fast/immediate
- readable (intelligible), for information use
  - intelligence services (e.g. NAIC)
  - occasional translation (home use)
➤ as draft for translation

➤ aid for writing in foreign language
  ➤ as used by EC administrators

➤ emails, Web pages

➤ any system type can be used (including those originally for mainframes and PCs)
  ➤ online MT has all types of rule-based systems - and now also SMT
Empirical Natural Language Processing
The Empirical revolution in NLP

➢ As systems get bigger, behavior is harder to predict

➢ Looking at system output one sentence at a time is slow

➢ Can we automate testing?

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)
The Empirical approach

1. Develop an algorithm and gather examples/rules from training data

2. Optimize any parameters on development data

   - Normally about 10% of the training data

3. Test on held-out, unseen test data

   This gives a fair estimate of how good the algorithm is — if the test criteria are appropriate.
Word Error Rate in Speech Recognition

➢ The first successful wide spread testing:
  ➢ Compare your output to a reference
  ➢ Calculate the number of substitutions, deletions and insertions to make them match (Minimum edit distance)
  ➢ Normalize by dividing by the length of the reference

\[
WER = \frac{S+D+I}{N}
\]

➢ Reference: I want to recognize speech today
System: I want wreck a nice peach today
Eval: S S I S

➢ \[ WER = \frac{3+0+1}{6} = 0.667 \]
Some properties of WER

- Correlates well with the task

- Reducing WER is always a good thing

- A WER of 0 implies perfect results (assuming the reference is correct)

- $WER < 0.05$ generally considered the minimum to be useful

- Competitions were held to see who could get the lowest WER
  - Speech Recognition had 10 years of rapid improvement
  - It has slowed down now
### How good are the systems?

<table>
<thead>
<tr>
<th>Task</th>
<th>Vocab</th>
<th>WER (%)</th>
<th>WER (%) adapted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digits</td>
<td>11</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Dialogue (travel)</td>
<td>21,000</td>
<td>10.9</td>
<td>—</td>
</tr>
<tr>
<td>Dictation (WSJ)</td>
<td>5,000</td>
<td>3.9</td>
<td>3.0</td>
</tr>
<tr>
<td>Dictation (WSJ)</td>
<td>20,000</td>
<td>10.0</td>
<td>8.6</td>
</tr>
<tr>
<td>Dialogue (noisy, army)</td>
<td>3,000</td>
<td>42.2</td>
<td>31.0</td>
</tr>
<tr>
<td>Phone Conversations</td>
<td>4,000</td>
<td>41.9</td>
<td>31.0</td>
</tr>
</tbody>
</table>

Speaker adapted systems have a lower WER.
Speaker Adaptation as you use the system

- **Personal Recognition** launched by Google in 2010
  - associate the user's speech input with the user
  - build a small speech corpus
  - use these words to build a speech model specifically for the user
  - accuracy improvements begin fairly quickly and build over time
  - you must opt in to use the system

- Improve the data, not the algorithm

- Build specialized models

Empirical vs Rational NLP

➤ The 1990s went through an empirical revolution

➤ Funding agencies sponsored competitions
  ➤ TREC: Text REtrieval Conference
  ➤ MUC: Message Understanding Conference
  ➤ DARPA Machine Translation Competitions

➤ Data to test with became more available

➤ Reviewers demanded evaluation in papers

➤ A lot of research on evaluation methods
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
MT Evaluation

- Evaluating MT output is non-trivial
  - There may be multiple correct answers.
    * I like to swim
    * I like swimming
    * Swimming turns me on

- Hand evaluation requires a bilingual evaluator - expensive

- Automatic evaluation can be done by comparing results (in a held out test set) to a set of reference translations
  - The most common metric is BLEU
  - Other scores are: Word Error Rate; METEOR
MT Evaluation: Fluency and Adequacy

➤ Fluency: How do you judge the fluency of this translation?

➢ 5 = Flawless English
➢ 4 = Good English
➢ 3 = Non-native English
➢ 2 = Disfluent English
➢ 1 = Incomprehensible

➤ Adequacy: How much of the meaning expressed in the reference translation is also expressed in the hypothesis translation?

➢ 5 = All
➢ 4 = Most
➢ 3 = Much
➢ 2 = Little
➢ 1 = None
MT Evaluation: The BLEU score

- BLEU score compares n-grams (normally up to 4) with those in the reference translation(s) (with a brevity penalty)

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

- 0.3–0.5 typical; 0.6+ approaches human

- Only really meaningful summed over a test set
  - individual sentences are too short
An example of BLEU

| Cand 1: | It is a guide to action which ensures that the military always obeys the commands of the party |
| Cand 2: | It is to insure the troops forever hearing the activity guidebook that party direct |
| Ref 1:  | It is a guide to action that ensures that the military will forever heed Party commands |
| Ref 2:  | It is the guiding principle which guarantees the military forces always being under the command of the Party |
| Ref 3:  | It is the practical guide for the army always to heed the directions of the party |

Intuition for BLEU (from Jurafsky and Martin Fig 25.31)
BLEU pros and cons

➤ Good

➢ Easy to calculate (if you have reference translations)
➢ Correlates with human judgement to some extent
➢ Used in standard competitions

➤ Bad

➢ Doesn’t deal well with variation
   * Exact string match
   * Near misses score zero: $cat \neq cats$!
➢ Biased toward n-gram models
   * SMT systems optimize for BLEU
Misleading Bleu Scores

➢ 信号は赤でした。
  ➢ The light was red.
  ➢ The signal was red. (0.35)

➢ 大丈夫です。
  ➢ I'm all right.
  ➢ I am all right. (0.27)

➢ 空港から電話しています。
  ➢ I’m calling from the airport.
  ➢ I am telephoning from the airports. (0.22)
How to improve the reliability?

➢ Use more reference sentences

➢ Use more translations per sentence
  ➢ Can be automatically created by paraphrasing

➢ Improve the metric: METEOR
  ➢ add stemmed words (partial score): \( \text{cat} \approx \text{cats}! \)
  ➢ add WordNet matches (partial score): \( \text{cat} \approx \text{feline}! \)

➢ Unfortunately this adds noise
  ➢ Errors in stemming
  ➢ Uneven cover in WordNet

➢ Still better than BLEU (so far) — but harder to calculate
Problems with testing

➤ You get better at what you test

➤ 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
Conclusion

➢ A surprising amount of variation is possible in MT

➢ This makes evaluation difficult
  ➢ If we know the correct answer, the problem is solved

➢ But evaluation is very important in NLP
  ➢ Use automatic evaluation
  ➢ Recognize the risks
These slides were made by Francis Bond

Much of the Machine Translation section was taken from a panel discussion by John Hutchins (13 December 2007)
http://www.hutchinsweb.me.uk/