# HG351 Corpus Linquistics

# Corpora and Language Engineering

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

Lecture 10
http://compling.hss.ntu.edu.sg/courses/hg3051/

# Overview

- Revision of Contrastive and Diachronic Studies
  - Contrastive Studies
  - Diachronic Studies
- Corpora and Language Engineering
  - Language as a Statistical Model
  - NLP and the Empirical Revolution
    - Speech Recognition and Segmentation
    - POS tagging, Word Sense Disambiguation and Parsing
    - Machine Translation
  - The Empirical Revolution and Linguistics

# Contrastive linguistics

- a linguistic approach that seeks to describe the differences and similarities between a pair of languages

- Some of the many applications:
  - to avoid interference errors in foreign-language learning
  - to assist interlingual transfer in the process of translating texts from one language into another
  - to find lexical equivalents in the process of compiling bilingual dictionaries, as illustrated

- generalized to the differential description of one or more varieties within a language, such as styles (contrastive rhetoric), dialects, registers or terminologies of technical genres

# Three Common Approaches

- Look at examples in a translation lexicon
  - ➢ well aligned sample
  - x no frequency information
- Look at examples in comparable corpora
  - ➢ good for getting general trends
  - x hard to measure **comparability**
- Look at examples in aligned corpora
  - ➢ can get numbers at a very detailed level
  - x translated text differs from monolingual text
  - x translations can be very free

# Three Studies

- A Corpus based Comparison of Satellites in Chinese and English
  - Comparable corpora used to compare verb+satellites
  - General trends clear, but few conclusions
  - Not clear how parallel the constructions are.

- Contrastive connectors in English and Chinese
  - Parallel Corpora used to study *however* and its Chinese counterparts
    - the HLM parallel corpus
    - the Babel English-Chinese Parallel Corpus
  - a deep analysis of a small number of examples
  - very different results for the different samples

- A Parallel Corpus-based Study of Translational Chinese
  - Compared English Text (EST), Original Chinese Text (OCT), Translated Chinese Text (TCT)
  - Translational Chinese has the following features
    - TCT uses fewer monosyllabic words than OCT does
    - compared with OCT, TCT uses more function words
    - TCT can change or expand the compositionality of some words or morphemes in Chinese.
  - TCT use more types and longer segments than OCT. This does not support the hypothesis of lexical and syntactic simplification in translation.

# Historical Linguistics

- All Historical linguistics is corpus linguistics: we can not ask Old English speakers for grammatical judgments
- The texts of a historical period or a dead language form a closed corpus
  - can only be extended by the (re-)discovery of previously unknown texts
- Comparable corpora sampled over different times have made it possible to quantitatively study language change
  - Helsinki Corpus of English Texts: Diachronic and Dialectal
    1.6 million words of British English from 750 to 1700 (Old/Middle/Modern)
  - Corpus of Historical American English has been created
    400 million words of text of American English from 1810 to 2009.
  - `http://ngrams.googlelabs.com/`

# The *by*-agent in English

- Peitsara (1993) used four subperiods from the Helsinki corpus to calculate the frequencies of different prepositions introducing agent phrases
  - In late Middle English (*c.* 1350) *of* and *by* were in roughly equal distribution (10.6:9)
  - By the fifteenth century *by* was three times more common than *of*
  - By 1640 *by* was eight times as common
- There was marked influence of text type: statutes and official documents were much more likely to use *by*
  - This is probably due to bilingual influence from French

Peitsara, K. (1993) ”On the development of the by-agent in English”, in *Early English in the Computer Age*. Rissanen, Kytö and Palander-Collin eds, 1993 pp 217-33, Berlin, Mouton de Gruyter.

# The case of it's and 'tis

Peitsara, K. (2004) "Variants of contraction: The case of it's and 'tis" *ICAME* **28** pp 77-94

- two contracted forms of *it is*
  - *'tis* (*tys, 't is, t is, t' is*)
  - *it's*
- Contractions are more often used in speech than text: hard to study historically
- Big shift to *it's* between 1800 and 1850
- Part of a general trend from proclisis (initial) to enclisis (embedded)

# Issues with Historical Linguistics

Rissanen (1989) identifies three problems with using historical corpora

- The philologist's dilemma —the danger that the use of a corpus and a computer may supplant the in-depth knowledge of language history which is to be gained from the study of original texts in their context

- The God's truth fallacy — the danger that a corpus may be used to provide representative conclusions about the entire language period, without understanding its limitations in the terms of which genres it covers.

- The mystery of vanishing reliability — the more variables which are used in sampling and coding the corpus (periods, genres, age, gender etc) the harder it is to represent each one fully and achieve statistical reliability. The most effective way of solving this problem is to build larger corpora of course

# Corpora and Language Engineering

# Testing and Training

- Language Engineering use Corpora in two important ways:
  - Testing
    - Create a reference set or gold standard
    - Test how well a system can produce these results
  - Training
    - Create a collection of labeled examples
    - Use these to learn features to classify new examples with one of the labels
  - Models are implicit in the corpus, rather than written by hand
  - Even which features are useful can be learned
  - Annotation becomes the bottleneck

# Classical NLP

- In classical NLP, researchers developed rules based on their own intuition and tested them on small test suites they themselves created

- Such systems proved hard to scale to large problems

- In particular, many systems required a large amount of knowledge:

  - words given their pronunciation
  - words classified by part of speech
  - words classified by semantic class
  - predicates classified by argument structure

  Creating and maintaining these is expensive and difficult:
  the resource bottleneck

# The breakthrough in speech recognition

- There was a great breakthrough in speech recognition
  - transcribe corpora, align the text and speech, and then learn pronunciations from the aligned data
    - Easier to do than writing IPA for each word
    - More natural data — includes performance changes
    - Easy to go back and look at more features
  - share data (prompted by funding agencies)
  - share test data (with an agreed on metric)
- System quality increased very quickly
- Shallow (non-cognitive) approaches began to dominate

# 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: | | | D | S | I | I | S | |

- $WER = \frac{2+1+2}{6} = 0.8$

# 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)
- $WER < 0.05$ considered the threshold 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?

| Task | Vocab | WER (%) | WER (%) adapted |
|---|---|---|---|
| Digits | 11 | 0.4 | 0.2 |
| Dialogue (travel) | 21,000 | 10.9 | — |
| Dictation (WSJ) | 5,000 | 3.9 | 3.0 |
| Dictation (WSJ) | 20,000 | 10.0 | 8.6 |
| Dialogue (noisy, army) | 3,000 | 42.2 | 31.0 |
| Phone Conversations | 4,000 | 41.9 | 31.0 |

Speaker adapted systems have a lower WER.

(2014) New Deep-learning based approaches may have reduced errors by as much as 30%.

Another nice feature of training on data — new algorithms arise and can be exploited

# WER Task

Calculate the WER for the following pairs:

(1) a. HYP: *What alright day*
b. REF: *What a bright day*

(2) a. HYP: *uno fantas grandes de limon*
b. REF: *dos fantas grandes de limon*

(3) a. HYP: *excuse me while I kiss this guy*
b. REF: *'scuse me while I kiss the sky*

(4) a. HYP: *Baby come back, you can play Monopoly*
b. REF: *Baby come back, you can blame it all on me*

# The need for automatic testing

- 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 plural of anecdote is not data*
     - intuition tends to miss many examples of use

# 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.

# 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

# Machine Translation Evaluation

- Evaluating MT output is non-trivial
  - There may be multiple correct answers
    泳ぐのが好きだ *oyogu-no-ga suki-da*
    - *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

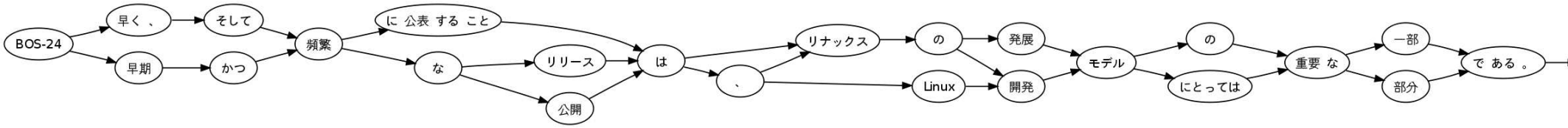
**Ref 3:** It is the practical guide for the army always to heed the directions of the party

Intuition for BLEU (from Jurafsky and Martin Fig 25.31)

# An Example of Variation

1. *Early and frequent releases are a critical part of the Linux development model*

   (a) 早期 かつ 頻繁 な 公開 は 、 リナックス の 開発 モデル の 重要 な 部分 である 。
   (b) 早く 、 そして 頻繁 に 公表 する こと は リナックス の 発展 モデル の 重要 な 一部 である 。
   (c) 早く 、 そして 頻繁 な リリース は 、 Linux 開発 モデル にとって は 重要 な 部分 である 。



# 1-grams (56)

な<>5
の<>4
は<>4
で<>3
モデル<>3
ある<>3
頻繁<>3
EOS<>3
BOS<>3
重要<>3
早く<>2
部分<>2
リナックス<>2
開発<>2
そして<>2
こと<>1

# 2-grams (55)

で<>ある<>3 3 3
重要<>な<>3 3 5
ある<>EOS<>3 3 3
な<>部分<>2 5 2
頻繁<>な<>2 3 5
は<>リナックス<>2 4 2
早く<>そして<>2 2 2
開発<>モデル<>2 2 3
EOS<>BOS<>2 2 2
BOS<>早く<>2 3 2
リナックス<>の<>2 2 4
モデル<>の<>2 3 4
の<>重要<>2 4 3
そして<>頻繁<>2 2 3
部分<>で<>2 2 3
は<>重要<>1 4 3
は<>Linux<>1 4 1

# 3-grams (54)

で<>ある<>EOS<>3 3 3 3 3 3 3
ある<>EOS<>BOS<>2 2 2 2 2 2 2
は<>リナックス<>の<>2 4 2 4 2 2 2
部分<>で<>ある<>2 2 3 3 2 2 3
重要<>な<>部分<>2 3 5 2 3 2 2
EOS<>BOS<>早く<>2 2 2 2 2 2 2
早く<>そして<>頻繁<>2 2 2 3 2 2 2
の<>重要<>な<>2 4 3 5 2 2 3
BOS<>早く<>そして<>2 3 2 2 2 2 2
な<>部分<>で<>2 5 2 3 2 3 2
モデル<>の<>重要<>2 3 4 3 2 2 2
こと<>は<>リナックス<>1 1 4 2 1 1 2
は<>重要<>な<>1 4 3 5 1 1 3
頻繁<>な<>公開<>1 3 5 1 2 1 1
そして<>頻繁<>に<>1 2 3 1 2 1 1

# 4-grams

- No four grams!

# 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): *cat* $\approx$ *cats*!
  - add WordNet matches (partial score): *cat* $\approx$ *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

# Variation in Translation (MRS Test set)

1. *The dog to chase is barking.*

   (a) 追う べき 犬 が 吠え て いる .
   (b) 追いかけよ う と する 犬 が 吠え て いる 。
   (c) 追いかけ られ て 、 その 犬 は 吠え て いる 。

2. *The dog was chased by Browne.*

   (a) 犬 が ブラウン に 追わ れ た .
   (b) その 犬 は ブラウン に 追いかけ られ た 。
   (c) その 犬 は 、 ブラウン さん に 追いかけ られ た 。

3. *The dog chased by Browne barked.*

(a) ブラウン に 追わ れ た 犬 が 吠え た .
(b) ブラウン に 追いかけ られ て いる 犬 が 吠え た 。
(c) ブラウン さん に 追いかけ られ た 犬 は 、 吠え た 。

4. *The dog is barking.*

(a) 犬 が 吠え て いる .
(b) 犬 が 吠え て いる 。
(c) その 犬 は 吠え て いる 。

5. *The dog has barked.*

(a) 犬 が 吠え た こと が ある .
(b) 犬 が 吠え た 。
(c) その 犬 は さっき から 吠え て いる 。

6. *The dog has been barking.*

(a) 犬 が 吠え て い た .
(b) 犬 が ずっと 吠え て いる 。
(c) その 犬 は さっき から ずっと 吠え て いる 。

7. *The dog had been barking.*

(a) 犬 が 吠え て い た .
(b) 犬 が ずっと 吠え て い た 。
(c) その 犬 は さっき まで ずっと 吠え て い た 。

8. *The dog will bark.*

(a) 犬 が 吠える だろ う .
(b) その 犬 は 吠える だろ う 。
(c) その 犬 は 吠え そう で ある 。

9. *The dog is going to bark.*

(a) 犬 が 吠える ところ だ .
(b) 犬 が 吠えよ う と し て いる 。
(c) その 犬 は 今 に も 吠え そう だ 。

10. *The dog could bark.*

(a) 犬 が 吠え ことが できる .
(b) 犬 が 吠え られる .
(c) その 犬 は 吠える ことが でき た 。
(d) その 犬 は 吠える 可能 性 が ある 。

11. *The dog couldn't bark.*

(a) 犬 が 吠える ことが でき ない .
(b) 犬 が 吠え られ ない .
(c) その 犬 は 吠える ことが でき なかっ た 。
(d) その 犬 が 吠える 可能 性 は ない 。

# Conclusion

- A surprising amount of variation is possible in MT
  so you need a lot of data for a reliable evaluation
- This makes evaluation difficult
  - If we know a correct answer, the problem is still not solved
- But evaluation is very important in NLP
  - Use automatic evaluation
  - Recognize the risks

# Coverage and OOV

- The resource bottleneck still exists in several forms
  - We don't have corpora for all tasks in all languages
  - In-domain training data much better than out-of-domain training data
  - More out-of-domain data only helps a little
- For those we do there are still Out of Vocabulary items (OOV)
- Distribution is domain dependent
- The bottleneck has shifted from lexicons to corpora

# Domain dependence

| Training Data | Test Set | Recall | Precision |
|---|---|---|---|
| WSJ | Brown | 80.3 | 81.0 |
| Brown | Brown | 83.6 | 84.6 |
| WSJ+Brown | Brown | 83.9 | 84.8 |
| WSJ | WSJ | 86.1 | 86.6 |
| WSJ+Brown | WSJ | 86.3 | 86.9 |

tested using similar test sets,
training data roughly twice as large for WSJ,
precision/recall measured with labeled parse constituents
Brown is non-homogeneous, WSJ is homogeneous

- Models over-fit to their training data
- Language use is slightly different in different genres

# Training from Corpora: POS tagging

- Learn rules automatically from tagged text
  - Many learning methods
  - Current popular learner is MIRA, before then CRF, before then SVM, . . .
  - Algorithms and CPU speeds are improving
- 96%+ accuracy using these features (probability based)
  - Previous $n$ words, (succeeding $n$ words)
  - Previous $n$ tags
  - Combinations of words and tags
  - Word Shape
- Learning methods relatively language independent
  but corpora and standards must exist

# Out of Vocabulary (OOV) words

- Unknown words are a big problem
  - Completely unknown words (not in lexicon)
  - Unknown uses of known words (derivation or lexicon gaps)
- Big, accurate lexicons are most useful!
- Otherwise guess from word shape (and context)
  - lowercase $\rightarrow$ common noun
  - uppercase $\rightarrow$ Proper noun
  - ends in *-ly* $\rightarrow$ adverb
  - ends in *-ing* and has vowel $\rightarrow$ verb
  - character type (Chinese character, alphabet, number, kana, . . . )
- You can learn these features (look at last $n$ letters, . . . )

# Some issues

- You get best results for things you have seen before
  - so if you are trying to extract knowledge, you reinforce what you already know
  - there are many underlying assumptions to text mining
    - you can tokenize
    - your sample is representative
  - in theory, data-driven is language agnostic
    - but tools are developed on major languages
    - our models work better for analytic languages
    - subversive text is likely to be harder to find

# Conclusion

- Testing on corpora objectively reveals system properties
- Training learns features humans don't predict
  - we are bad at simple frequency counts
- Training and Test have to be separate
- The closer the training is to the test, the better the result
- The general conclusion: more data is better data