COR: Corpus Linquistics

Lecture 11
Representativeness and Balance

Francis Bond
Department of Asian Studies
Palacký University
https://fcbond.github.io/
bond@ieee.org

https://github.com/bond-lab/Corpus-Linguistics

Overview

- Revision of Corpora and Language Engineering
  - Language as a Statistical Model
  - NLP and the Empirical Revolution
  - The Empirical Revolution and Linguistics
- Representativeness and balance
- Copyright and Ethics

Review

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

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

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

Conclusions

- A surprising amount of variation is possible in MT

- 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

- The closer the training is to the test, the better the result

- The general conclusion: more data is better data

Representativeness and Balance

This section is based on: “Corpus and Text: Basic Principles” by John Sinclair (2004) in *Developing Linguistic Corpora: a Guide to Good Practice* Martin Wynne ed, University of Oxford

Representativeness

To define representativeness we need to consider the following questions about the users of the language we will represent:

- What sort of documents do they write and read, and what sort of spoken encounters do they have?
- How can we allow for the relative popularity of some publications over others, and the difference in attention given to different publications?
- How do we allow for the unavoidable influence of practicalities such as the relative ease of acquiring public printed language, e-mails and web pages as compared with the labour and expense of recording and transcribing private conversations or acquiring and keying personal handwritten correspondence?
- How do we identify the instances of language that are influential as models for the population, and therefore might be weighted more heavily than the rest?

There are no definite answers to these questions.

How to be representative

1. decide on the structural criteria that you will use to build the corpus, and apply them to create a framework for the principal corpus components;

2. for each component draw up a comprehensive inventory of text types that are found there, using external criteria only;

3. put the text types in a priority order, taking into account all the factors that you think might increase or decrease the importance of a text type — the kind of factors discussed above;

4. estimate a target size for each text type, relating together (i) the overall target size for the component (ii) the number of text types (iii) the importance of each (iv) the practicality of gathering quantities of it;

5. as the corpus takes shape, maintain comparison between the actual dimensions of the material and the original plan;

6. (most important of all) document these steps so that users can have a reference point if they get unexpected results, and that improvements can be made on the basis of experience.

> *The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.*

Balance

- for a corpus to be balanced, the proportions of different kinds of text it contains should correspond with informed and intuitive judgments.

- Most general corpora of today are biased towards text: estimates of the optimal proportion of spoken language range from 50% to 90%
  because speech is so expensive to collect this imbalance is likely to remain.

- Balance can conflict with representativeness
  - Consider popular magazines in English
  - there are a large number of them and most use a highly specialised language
  - It is an important text type, but it is almost impossible to select a few texts which can claim to be representative
  - Can magazines for fly fishermen, personal computers and popular music really represent the whole variety of popular magazines (as is the case in The Bank of English)?

- Specialised corpora are constructed after some initial selectional criteria have been applied, for example a blog corpus or a patent corpus. More delicate criteria are used to partition them, but the issues of balance and representativeness remain cogent and central in the design.

> *The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.*

Summary

- The extent to which conclusions from a test can be generalized depend on both the nature of the evaluation function and the size and representativeness of the test set

- The extent to which conclusions drawn from a corpus study can be generalized to all language depend crucially on the design of the corpus

- In general more **representative** data is better data

- But we need to be aware of limitations:
  - Data sparsity
  - Out of vocabulary items
  - Over fitting
  - Domain adaption

Copyright and Licensing

Copyright

- Governments grant certain rights to authors of creative works, typically called copyrights in order encourage them to produce more
  - The most basic right is the right to forbid people from copying it without permission
  - Any work produced is by default copyright of the author
- Some or all of these rights can be waived or transferred
  - An author may sell the rights for a manuscript to a publisher
  - A blogger may place their postings in the public domain
  - A publisher may give permission to an author to post their paper on their website
  - A work may be distributed under a license that allows copying only under some conditions

- Copyright laws are national laws, although they may be harmonized by treaties
  - A text may be illegal to copy in one country, but legal in another
- Copyright laws change over time
  - E.g. in the U.S. originally 14 years for books only
  - Now 70 years after the death of the author for almost everything (but not recipes)
- New technology complicates things
  - Sending email involves making multiple copies on different servers
  - Recording speech can happen without the creator's knowledge

Copyright issues are very complicated

Some Rough Guidelines

- Copying something which is under copyright is illegal unless specific permission is granted or it falls under **fair dealing**, such as for the purpose of research or education

- How can you get permission?
  - You can buy it (for some works)
  - You can get signed permission from the copyright holder (or recorded permission for preliterate speakers)
  - You can get implicit permission (e.g. for email or web pages)
  - It can be permitted by a license
    - **CC-by** allows you to copy and redistribute if you acknowledge
    - **CC-by-nconc** allows you to copy and redistribute if you acknowledge and it is for non-commercial use

- The following factors will be considered to decide if it is fair dealing (in Singapore)
  - purpose and character of the dealing, including whether such dealing is of a commercial nature or is for non-profit educational purposes
  - nature of the work or adaptation
  - amount copied, relative to the whole work
  - effect of the dealing upon the potential market for the work, and effect upon its value
  - the possibility of obtaining the work or adaptation within a reasonable time at an ordinary commercial price
  - whether the copy is for the purpose of criticism or review; for the purpose reporting of news; for the purpose of judicial proceedings or professional advice (a sufficient acknowledgment of the work is required)
  - it is not an infringement if a person makes a copy from an original copy of a computer program as a back-up

http://en.wikipedia.org/wiki/Fair_dealing

Copyright for Corpora

- Arguments for restrictive licensing
  - Competitive advantage (common for speech corpora)
  - Compensation for the effort of creation
  - Minimize effect on the value of the original work

- Arguments for open licensing
  - Annotation is expensive, making the data open gets the best return on this investment
  - Annotation is typically ongoing, opening the data gets you more feedback
  - Researchers are evaluated by the impact that their work has. Open data generally has more impact.
  - Language data is part of our shared heritage

Choice of License

- Should be considered early on (before you start compiling your corpus)
- May depend on the funding body
- Depends on the source data
- General trend is to open licensing
  - Open Science Project
  - Open Access Journals
  - Open Source Software
- Try to chose a standard license (such as Creative Commons)

NTU's policy on Data Sharing

5.6.1 "The final research data from projects carried out at NTU shall be made available for sharing (via the NTU Data Repository) unless there are prior formal agreements with external collaborators and parties on non-disclosure or proprietary use of the data." http://research.ntu.edu.sg/rieo/RI/Pages/Research-Data-Policies.aspx

5.6.2 "The sharing and use of research data shall be based on Creative Commons license CC:BY:NC, where others may use data for non-commercial applications only and must correctly attribute the data source in NTU."

Creative Commons Licenses

License	Derivative Works	Same License	Commercial Use
CC-BY	+	-	+
CC-BY-SA	+	+	+
CC-BY-NC	+	-	-
CC-BY-ND	-	-	+
CC-BY-NC-SA	+	+	-
CC-BY-NC-ND	-	-	-

BY Attribution (all licenses)

SA Share Alike (requires copies to have the same license)

NC Non-Commercial (Not Open)

ND No Derivatives (allows only exact copies) (Not Open)

Many, many other license also exist (GPL, MIT, BSD, Apache, ...)

http://creativecommons.org/licenses/

The Open Definition

- The Open Definition sets out principles that define **openness** in relation to data and content.

- It makes precise the meaning of **open** in the terms **open data** and **open content** and thereby ensures quality and encourages compatibility between different pools of open material.

- It can be summed up in the statement that:

  > “Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness).”

- Put most succinctly:

  > “Open data and content can be freely used, modified, and shared by anyone for any purpose”

https://opendefinition.org/

Conclusions

Be careful of copyright