HG2052
Language, Technology and the Internet

The Web as Corpus

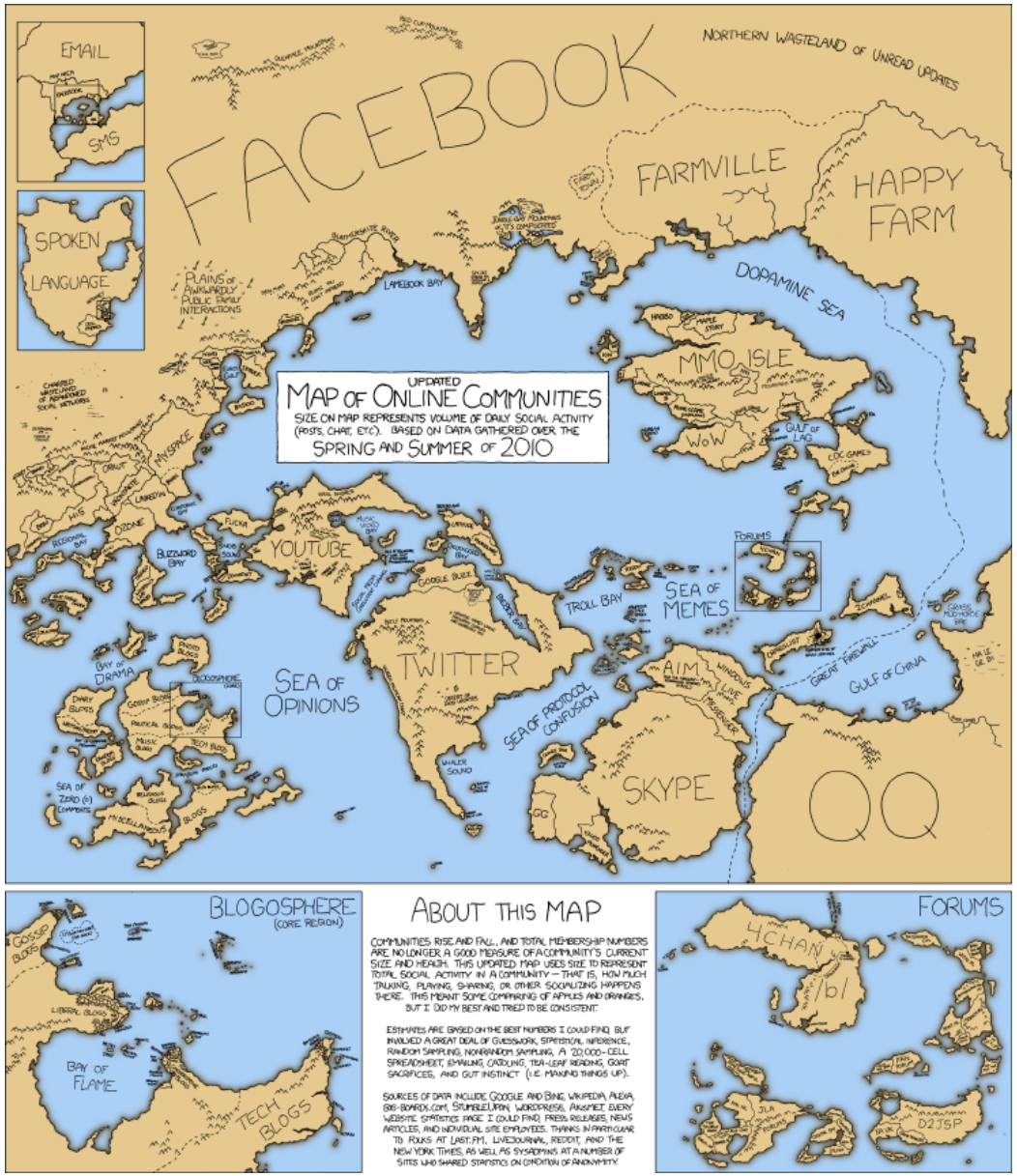
Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 7

Revision of the World Wide Web and HTML

- The Internet
  more than just the web (email, VoIP, FTP, Streaming, Messaging, ...)
- The structure of Markup: Visual vs Logical
  WSISWYG; WYSIAYG; WYSIWYM
- The structure of the Web — hypertext
  pages linked to pages
- The future of the Web
- Linguistic features of the web
  un-edited; large volume; editable; multi-media

Map of online communities



The Web as Corpus

The Web as Corpus

- the web is a collection of text, thus it is a corpus
- the largest available corpus: more than 7.2×10^{11} words
  (10 times bigger than the English Gigaword Corpus)
- nearly all kinds of text and lots of languages present
- not preprocessed, lots of ungrammatical (and linguistically useless) text
- how to access it?

Direct Query or Sample?

- Direct Query: Search Engine as Query tool and WWW as corpus
  (Objection: Results are not reliable)
  - Population and exact hit counts are unknown ⇒ no statistics possible.
  - Indexing does not allow us to draw conclusions on the data.
  - ⊗ Search engines miss functionalities that linguists / lexicographers would like to have (POS, lemmatization, ...).

- Web Sample: Use search engine to download data from the net and build a corpus from it.
  - known size and exact hit counts ⇒ statistics possible.
  - people can draw conclusions over the included text types.
  - (limited) control over the content.
  - ⊗ sparser data

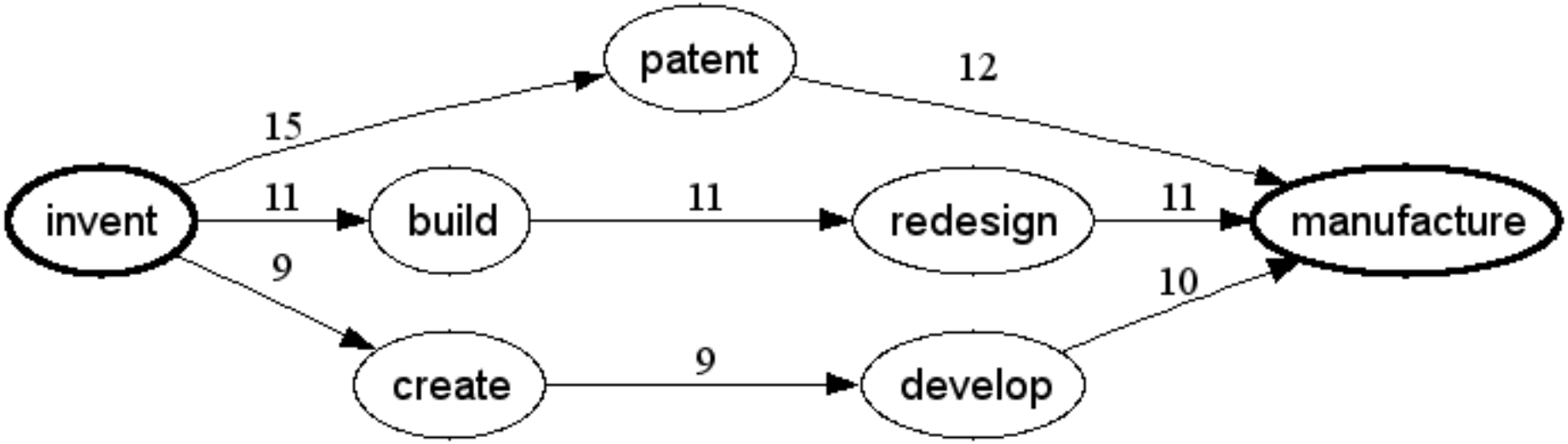
Direct Query

- Document counts are shown to correlate directly with "real" frequencies (Keller and Lapata, 2003), so search engines can help - but...
- lots of repetitions of the same text (not representative)
- very limited query precision (no upper/lower case, no punctuation...)
- only estimated counts, often hard to reproduce exactly
- how to access (Google API, Yahoo API, Scripts)
- Alexa: "buy" (parts of) web, and process it on their machines

Direct Query Example

- Directly using web counts (instead of corpus counts),
  VerbOcean (Chklovski & Pantel 2004)
  - gather verb pairs which are semantically related but the relation is unknown (DIRT: Lin and Pantel 2001)
    example pair: *love* – *marry*
  - pick a semantic relation (e.g. `happens-before`) and design typical patterns for this relation (e.g. *to X and then Y*)
  - instantiate the patterns (*to love and then marry*) and count Google hits (here: 6)
  - estimate whether or not the number of hits indicates a significant correlation, then assign the relation (or not)

VerbOcean


patent
12
15
invent
11
build
11
redesign
11
manufacture
9
10
create
9
develop

Limitations of web search interfaces

- Search engines only provide limited context
- Search engines do not allow for linguistically complex queries
- Results are organized according to relevance to the topic, not to left/right context
- Search engine counts cannot generally be trusted
- Access may be limited to n hits/second|day

Some Issues with Google Web Counts

Search for:

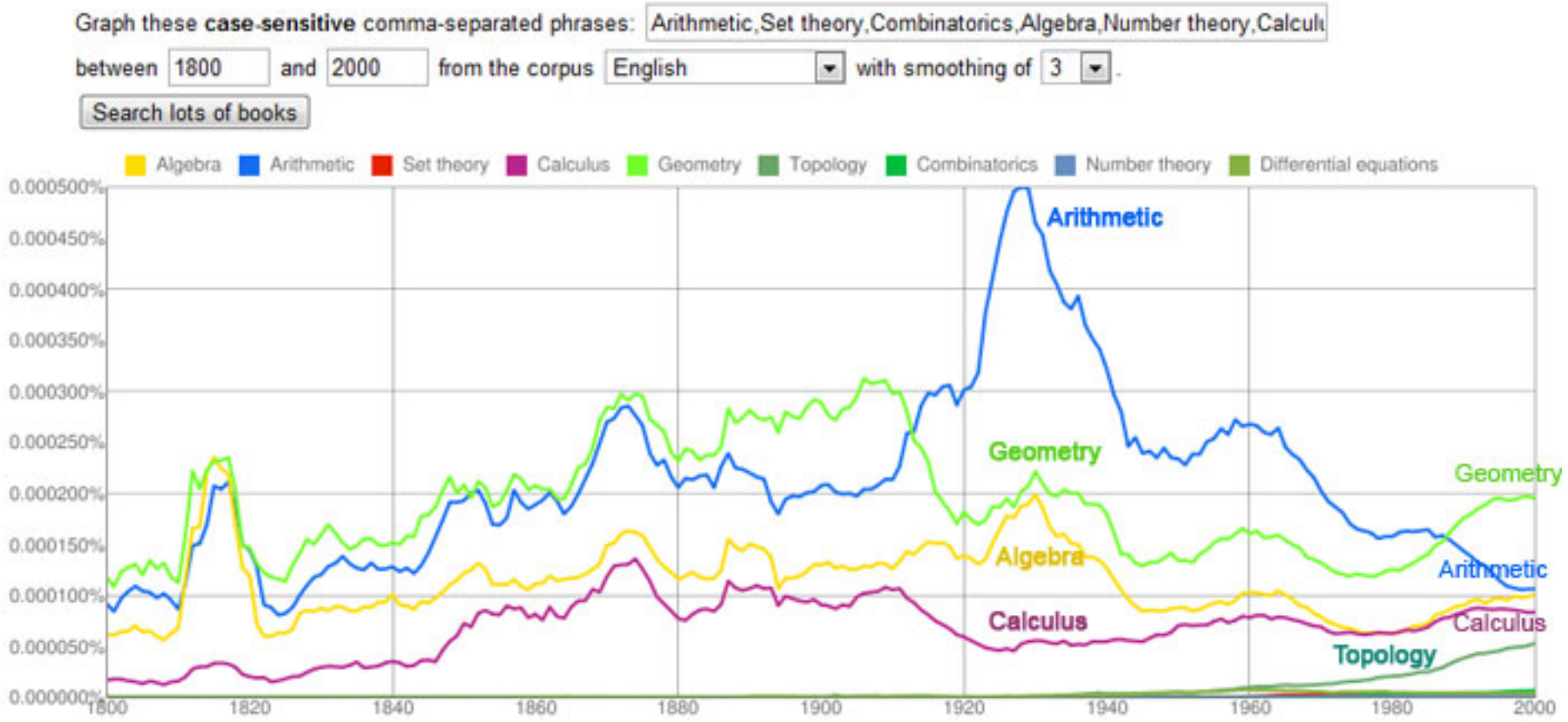
- machine "machine"
- machine machine
- the machine
- machine translation
- "machine translation"

The bottom line, said Mr. Norvig, is that getting an accurate estimate isn't that important for most of Google's users, so the company hasn't invested much time and computing power. "*It's only reporters and computational linguists who care if it's really precise*," he said.

Web Count Units

- ghit or "google hit" is the most common unit used to count web snippets (in the early 2000s)
  - it is document frequency not term frequency
- whit or "web hit" is the more general term
- Normally you compare two phenomena to get a unitless ratio
  (e.g. *different from* vs *different than*)
  251,000,000 ghits vs 71,500,000 or 3.5:1 (accessed 2012-04-04)
- GPB, for "Ghits per billion documents" is good if want a more stable number (suggested by Mark Liberman)
  but Google no longer releases the index size
- So always say when you counted, and try to use ratios

Google Books Ngram Viewer



Google Books Ngram Viewer

- Display a graph showing how phrases have occurred in a corpus of books
  - A variety of corpora (different languages, different genres, different times)
- Can do various searches
  - Wildcards: *I like to **
  - Inflection: *book_INF a hotel*
  - Simple POS
  - You can do combinations: **_NOUN rocks*; *green_**
  - Dependencies: *has=>*_NOUN*

Let's try it

- Compare similar words (e.g. *Miss* vs *Mrs* vs *Mdm* vs *Ms*)
  try some synonyms (*poop, shit, ...*)
- Find common objects
- Compare translation equivalents (*eat* vs 吃 *chi*)

Web Sample

Sample the Web

- Extracting and filtering web documents to create linguistically annotated corpora (Kilgarriff and Dhonnchadha, 2006)
  - gather documents for different topics (balance!)
  - exclude documents which cannot be preprocessed with available tools (here taggers and lemmatizers)
  - exclude documents which seem irrelevant for a corpus (too short or too long, word lists,...)
  - do this for several languages and make the corpora available

Internet Corpora: Outline

1. Select Seed Words (500)

2. Combine to form multiple queries (6,000)

3. Query a search engine and retrieve the URLs (50,000)

4. Download the files from the URLS (100,000,000 words)

5. Postprocess the data (encoding; cleanup; tagging and parsing)

Sharoff, S (2006) Creating general-purpose corpora using automated search engine queries. In M. Baroni, S. Bernardini (eds.) *WaCky! Working papers on the Web as Corpus*, Bologna, 2006.

Internet Corpora

Select about 500 words from a list of the most frequent word forms in your language. It is important that selected words are sufficiently general, i.e. they do not belong to a specific domain, but they are not function words. For instance, *picture, extent, raised, events* are good query words for English.

Produce a list of 5000-6000 queries each of which consists of 4 words (you may need more to get more links). If the language for which you want to collect a corpus is not listed by Google yet, add a couple of very frequent function words that are not used in cognate languages. If you collect a corpus for a language with relatively few Internet pages, you may decrease the number of words in a query (however, this will also decrease the amount of connected text in the pages returned, so you'll get more price lists, forms, catalogues, etc). Collect the top 10 URLs produced by Google for each query.

Download URLs . The list of successfully downloaded URLs constitutes an open-source corpus.

Postprocess the set of downloaded files — correction of encodings, conversion of all texts to Unicode, filtering out duplicate pages, removing navigation frames, etc, followed by lemmatisation and part-of-speech tagging.

Composition assessment : take a sample of about 200 texts from the corpus and describe them according to a text typology: Authorship; Mode; Audience; Domain; Aim.

Compare : If you have another corpus for the same language, you can compare their frequency lists using the log-likelihood score (see Paul Rayson's Log-likelihood calculator)

Domain Specific Corpora

- You can tune a web corpus in various ways
  - adjusting the seed query terms (different content)
  - restricting the URLs (e.g. only `.edu` or `.sg`)
  - restricting the page type (only blogs, ...)
  - restricting the license of the web page
    The English CC corpus has been compiled from webpages with the Creative Commons permissive licences. The corpus is less balanced than the main Internet-English (less professional news, more blogs and fanzines), but it can be redistributed without limitations.

Disposable Corpora

- The web provides unprecedented opportunities for gathering data
- Viable source of "disposable corpora" , built ad hoc for specific purposes
- Essential for working with specialized languages
- Need to automatically extract web corpus
  extraction is time-consuming

Post-processing

- Filter documents by size
  - Small documents ($< 5KB$) contain very little real text
  - Large documents ($> 200KB$) tend to be indices, catalogues, lists, etc.
- Remove perfect duplicates
  - Actually, removed both the original & the duplicate:

  ... tend to be warning messages

Boilerplate stripping

- Boilerplate: HTML markup, javascript, other non-linguistic material
- Removing boilerplate information is crucial to obtaining linguistic data only
  - Content-rich sections of a document will have a low html tag density
  - Boilerplate sections have a wealth of html
  - This heuristic is "relatively independent of language and crawling strategy"
- If a text does not have enough function words, it is likely non-linguistic material (e.g., a list)
  - Require at least 10 function word types and 30 tokens on a page
  - ... which must make up at least 25% of the total words

Near-duplicate detection

- Take fingerprints of a fixed number of randomly-selected n-grams (ignoring function words)
  - e.g., extract 25 5-grams from each document
- Near-duplicates have a high overlap
  - e.g., at least 2 5-grams in common

Linguistic Post-processing

- Prepare the data for searching:
  - Run a POS tagger over it
  - Clean the documents further, using POS tags
    - Where the POS tag distribution is unusual,
    - ... perform another round of anomalous document finding
    - Look for problematic (erroneous) POS tags and remove those documents
    - Use cues such as number of unrecognized words, proporition of words with upper-case initial letters, ...
- Index the document by word, POS and lemma

Benefits of Web Corpora

- Help address data sparseness issues
- Provide more interpersonal material
- Check the claims made with other corpora
- Expensive to build corpora otherwise, yet they are needed for under-resourced languages
- Current corpora are often restricted in size and/or variety
- News corpora do not represent general language
- Need a variety of text types

Web Corpus Statistics

Corpus	tokens	words	lemmas	URLs	words/doc
I-EN	126,643,151	2,003,056	1,608,425	42,133	3,006
I-DE	126,117,984	3,384,491	3,081,197	31,195	4,043
I-RU	156,534,391	2,036,503	791,311	33,811	4,630

British National Corpus (BNC)

- 100-million-word text corpus
  - 90% written
  - 10% speech (transcribed)
- balanced across a wide variety of genres
- a representative sample of British English of the late 20th century
- POS tagging and domain/genre tagging

The corpus gold standard!

Text assessment

To determine whether the corpora is balanced like the BNC, Sharoff assesses a variety of factors

- Authorship:
  - Single
  - Multiple
  - Corporate: 44% for I-EN, 18% for BNC
  - Unknown
- Gender: Female writers are underrepresented: 23%/3% male/female split in I-EN vs. 28%/13% for BNC

- Mode
  - Written
  - Spoken: 0-1% for web corpora, 10% for BNC
  - Electronic: 16% for Russian, 13% for English, 9% for German; 0% for BNC
  - Important type of data, as is similar to speech in some ways, similar to writing in some

- Audience
  - General: 33% in I-EN
  - Informed: 45% in I-EN
  - Professional: 22% in I-EN

  Overall, I-EN seems somewhat balanced w.r.t. this classification (similar to BNC)

- Aims of text production
  - Discussion: 45% in I-EN
  - Recommendation (Hard to tell apart from discussion)

- Recreation: fiction=17% in BNC vs. recreation=4% in I-EN
- Instruction
- Information

- Domain
  - natsci (natural sciences)
  - appsci (applied sciences): 7% in BNC vs. 29% in I-EN
  - socsci (social sciences): 16% in RRC vs. 5% in I-RU
  - politics
  - business
  - life; arts; leisure

Corpus of web URLs

- One strategy for releasing a corpus is to organize a list of appropriate URLs
- If a corpus consists of a list of URLs and associated software for extracting them, how stable is such a corpus?
- We can measure a corpus's half-life by seeing how many pages are left after a certain amount of time
- Initial experiments show that some links are gone after a few months
  - Feb 2005 $\rightarrow$ Aug 2005: 934/1000 remaining
  - Jun 2005 $\rightarrow$ Aug 2005: 982/1000 remaining
- Need longer term studies and studies testing different parameters

Some other corpora

- http://corpus.byu.edu/glowbe/
- http://commoncrawl.org/

Summary

- The web can be used as a corpus
  - Direct access
    - Fast and convenient
    - Huge amounts of data
    - ⊗ unreliable counts
  - Web sample
    - Control over the sample
    - Some setup costs (semi-automated)
    - ⊗ Less data
- Richer data than a compiled corpus
- ⊗ Less balanced

References

- Keller, Frank and Mirella Lapata. 2003. Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics* 29:3, 459-484.

- Timothy Chklovski and Patrick Pantel. 2004. VERBOCEAN: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*. pp. 33-40. Barcelona, Spain.

- Kilgarriff, Michael Rundell and Elaine Uí Dhonnchadha. 2006. Efficient corpus development for lexicography: building the New Corpus for Ireland. *Language Resources and Evaluation Journal* 40 (2): 127-152.

➢ Lin, Dekang and Patrick Pantel. 2001. DIRT - Discovery of Inference Rules from Text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*. pp. 323-328. San Francisco, CA.

➢ Sharoff, S (2006) Creating general-purpose corpora using automated search engine queries. In M. Baroni, S. Bernardini (eds.) *WaCky! Working papers on the Web as Corpus*, Bologna, 2006.