HG2052
Language, Technology and the Internet

Review and Conclusions

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

Lecture 12

Introduction

What have we learned?

- How technology affects our use of language
- How language is used on the internet
  - Some interesting things we can now do, that we couldn't before
- Collaboration on the Web
- The web as a source of linguistic data: Direct Query and Sample
- The Semantic Web: meaning for non-humans
- Citation and reputation

Reflection

- What was the most surprising thing in this class?
- What do you think is most likely wrong?
- What do you think is the coolest result?
- What do you think you're most likely to remember?
- How do you think this course will influence you as a linguist?
- What (if anything) did you hope to learn that you didn't?

Goals

- Gain an understanding of how technology affects language use
- Develop familiarity with markup and meta information in texts
- Get a feel for what research is all about, especially relating to web mining and online frequency counting

Upon successful completion, students will:

- have an understanding of how technology shapes language use
- be able to test linguistic hypotheses against web data.
- know how to edit Wikipedia

Themes

- Language and Technology
  - Writing and Speech Technology
- Language and the Internet
  - Email; Chat; Virtual Worlds; WWW; IM; Blogs; Facebook; Wikis; Twitter
- The Web as Corpus
- The Web beyond Language
  - Semantic Web and Networks

Revision

Language and Technology

- The two things that separate humans from animals (Sproat, 2010, §1.1)
  - Language
    - large vocabulary (10,000+)
    - complicated syntax (no upper length; recursion; embedding)
  - Technology
    - Widespread tool use
    - Widespread tool manufacture

- Speech — the start of language

- Writing — the first great intersection

Language and the Internet

- New forms of communication
  - Neither speech nor text
  - Massively interactive
- Extremely rapid change
- A first hand narrative (I was online before the internet :-)
  - but I am probably behind you all now

New forms

- Email (from PC, phone, other)
- Chat; Usenet
- Virtual Worlds
- WWW
- Blogs (overlap)
- Facebook, LinkedIn
- Wikis
- Twitter

Representing Language

- Writing Systems
- Encodings
- Speech
- Bandwidth

What is represented?

- Phonemes: /ma d la ks æv@kadoz/ (45)
  - Not a simple correspondence between a writing system and the sounds
  - Some logographs (*3*, *@*, *$*)
  - Even when a new alphabet is designed, pronunciation changes.
- Syllables: (ma) (d) (la ks) (æ)(v@)(ka)(doz) (10,000+)
- Morphemes: ma /me + 's d la k + s æv@kado + z (100,000+)
- Words: *my dog likes avocados* (200,000+)
- Concepts: `speaker poss canine fond avacado+PL` (400,000+)

Three Major Writing Systems

- Alphabetic (Latin)
  - one symbol for consonant or vowel
  - Typically 20-30 base symbols (1 byte)
- Syllabic (Hiragana)
  - one symbol for each syllable (consonant+vowel)
  - Typically 50-100 base symbols (1-2 bytes)
- Logographic (Hanzi)
  - pictographs, ideographs, sounds-meaning combinations
  - Typically 10,0000+ symbols (2-3 bytes)

Computational Encoding

- Need to map characters to bits
- More characters require more space
  - English: ASCII (7 bits); Western Europe: ISO-8859-1 (8 bits); Japanese: EUC (16 bits); Everything: UTF-8 (8-32 bits)
- Moving towards unicode for everything
- If you get the encoding wrong, it is gibberish
  - Web pages should state their encoding
  - Sometimes they are wrong
  - Encoding Detection usually involves statistical analysis of byte patterns
    - But an encoding can be shared by many languages

Speech and Language Technology

- The need for speech representation
- Storing sound
- Transforming Speech
  - Automatic Speech Recognition (ASR): sounds to text
  - Text-to-Speech Synthesis (TTS): texts to sounds
- Speech technology — the Telephone!

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

Results of various DARPA competitions (from Richard Sproat's slides)

Why is it so difficult?

- Pronunciation depends on context
  - The same word will be pronounced differently in different sentences
- Speaker variability
  - Gender
  - Dialect/Foreign Accent
  - Individual Differences: Physical differences; Language differences (idiolect)
- Many, many rare events
  - 300 out of 2000 diphones in the core set for the AT&T NextGen system occur only once in a 2-hour speech database

Two steps in a TTS system

1. Linguistic Analysis
   - ➢ Sentence Segmentation
   - ➢ Abbreviations: *Dr Smith lives on Nanyang Dr. He is ...*
   - ➢ Word Segmentation:
     - ➢ 森山前日銀総裁 *Moriyama zen Nichigin Sousai*
     - ⊗ 森山前日銀総裁 *Moriyama zennichi gin Sousai*

2. Speech Synthesis
   - ➢ Find the pronunciation
   - ➢ Generate sounds
   - ➢ Add intonation

Speech Synthesis

- Articulatory Synthesis: Attempt to model human articulation.
- Formant Synthesis: Bypass modeling of human articulation, and model acoustics directly.
- Concatenative Synthesis: Synthesize from stored units of actual speech

Prosody of Emotion

- Excitement: Fast, very high pitch, loud
- Hot anger: Fast, high pitch, strong, falling accent, loud
- Fear: Jitter
- Sarcasm: Prolonged accent, late peak
- Sad: Slow, low pitch

The main determinant of "naturalness" in speech synthesis is not "voice quality", but natural-sounding prosody (intonation and duration)

Richard Sproat

Speed is different for different modalities

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

Reading	300	200 (proof reading)
Writing	31	21 (composing)
Speaking	150	
Hearing	150	210 (speeded up)
Typing	33	19 (composing)

➢ Reading >> Speaking/Hearing >> Typing

⇒ Speech for input
⇒ Text for output

The Telephone

Speech like	Text like
time bound	space bound
spontaneous	contrived
face-to-face	visually decontextualized
loosely structured	elaborately structured
socially interactive	factually communicative
immediately revisable	repeatedly revisable
prosodically rich	graphically rich

New Modalities

Email; Usenet; Chat and Blogs

- All share some characteristics of speech and text
- Usage norms not fixed
- Communication methods may disappear before the norms are fixed
  Usenet, Bulletin Boards, Gopher, Archie, ...
- Large scale discourse studies still to be done
- Some genuinely new things
  - time-lagged, multi-person conversation
  - raw un-edited text
  - extreme multi-authorship

Email

Speech like	Text like
time bound*	space bound (deletable)
spontaneous*	contrived*
face-to-face	visually decontextualized
loosely structured*	elaborately structured*
socially interactive*	factually communicative
immediately revisable	repeatedly revisable*
prosodically rich	graphically rich *

Usenet (asynchronous)

Speech like	Text like
time bound*	space bound
spontaneous*	contrived*
face-to-face	visually decontextualized
loosely structured*	elaborately structured
socially interactive*	factually communicative
immediately revisable	repeatedly revisable
prosodically rich	graphically rich

Chat (synchronous)

Speech like	Text like
time bound*	space bound
spontaneous*	contrived
face-to-face	visually decontextualized
loosely structured*	elaborately structured
socially interactive*	factually communicative
immediately revisable	repeatedly revisable
prosodically rich	graphically rich

Blogs

Speech like	Text like
time bound	space bound
spontaneous*	contrived*
face-to-face	visually decontextualized
loosely structured*	elaborately structured*
socially interactive^{comments}	factually communicative
immediately revisable	repeatedly revisable*
prosodically rich	graphically rich *

Netspeak

- Inspired by shared background
  - `<rant>`I can't stand this`</rant>`
  - I hate`^H^H^H^H`love that idea (`^H` is backspace on a vt100)
  - lusers (users as seen by Systems Administrators)
  - suits
- Need to be in-group:

  cow orker Coworker

  clue "You couldn't get a clue during the clue mating season in a field full of horny clues if you smeared your body with clue musk and did the clue mating dance." Edward Flaherty (`talk.bizaare`)

- Gricean Principals
  - Posting (top/middle/bottom, quoting and trimming)
- Inspired by medium limitations
  - `GREAT;*great*;great`
  - `:-)` `(^_^)` `(;o;)` ‿
  - `brb`, `RTFM`, `IMHO`
  - lower case
  - lack of punctuation (hard on phone keyboards)

Collaboration and Wikis

- Version Control Systems
- Wikipedia
- Licensing and Ownership

Version Control Systems

- Versioning file systems
  - every time a file is opened, a new copy is stored
- CVS, Subversion, Git
  - changes to a collection of files are tracked
  - simultaneous changes are merged
- Revision Tracking
  - Revisions are stored within a file
- Authorship in shared writing; Explicit responsibility for changes

Wikipedia

- The core aim of the Wikimedia Foundation, is to get a free encyclopedia to every single person on the planet. (Jimmy Wales)
- Wikipedia makes it easy to share your knowledge
  people like to do this
- Most edits are done by insiders
- Most content is added by outsiders
- Content comparable to Britannica

The five pillars of Wikipedia

1. Wikipedia is an online encyclopedia
2. Wikipedia has a neutral point of view
   - ➢ Content policies: NPOV, Verifiability, and No original research
3. Wikipedia is free content
4. Wikipedians should interact in a respectful and civil manner
5. Wikipedia does not have firm rules

Licenses and Ownership

- Copyright
- Copyleft
- Creative Commons

What is a good article?

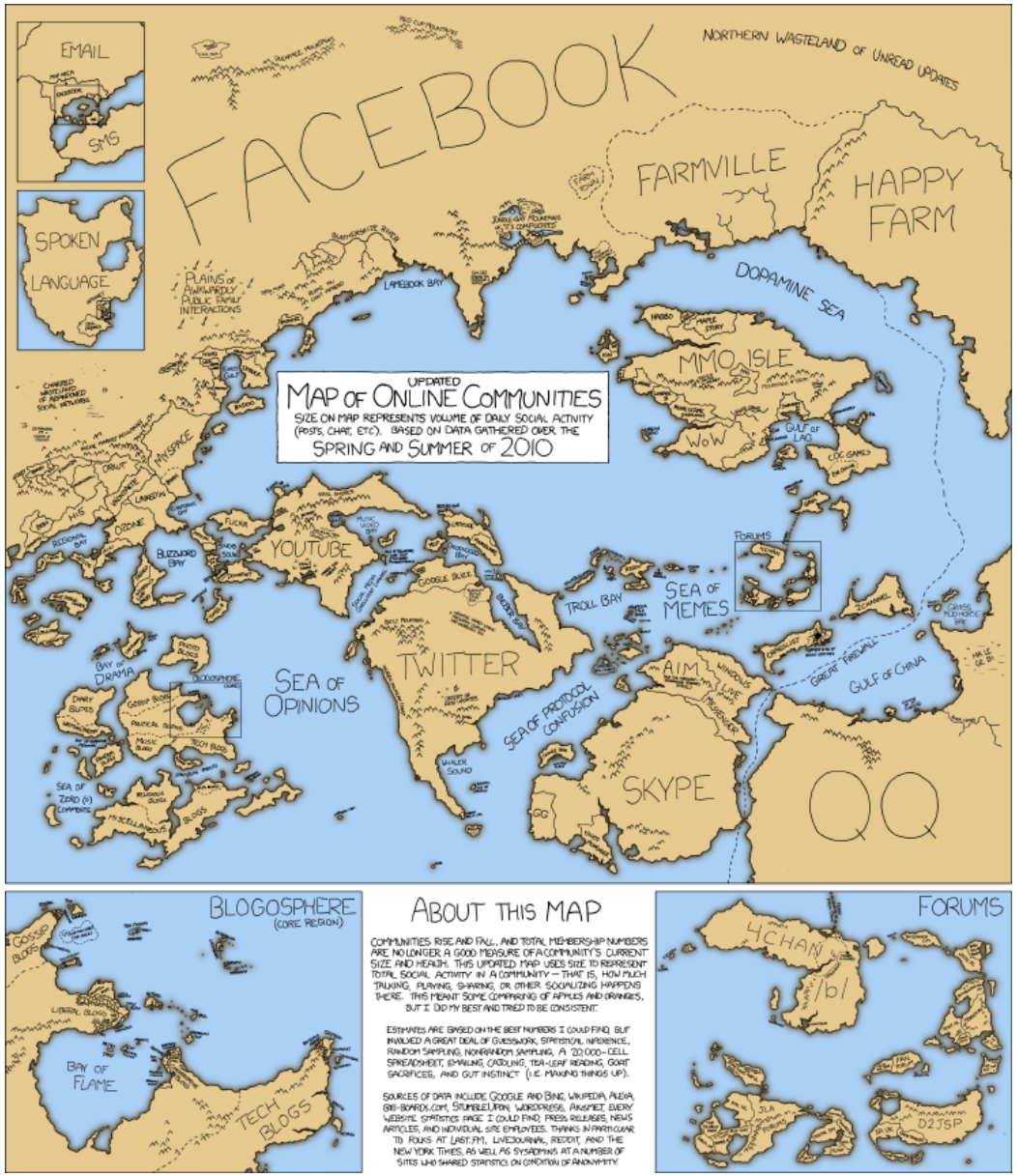
1. Well-written
2. Factually accurate and verifiable
3. Broad in its coverage
4. Neutral
5. Stable
6. Illustrated, if possible, by images

The World Wide Web and HTML

The InterWeb

- The Internet
  more than just the web (email, VoIP, FTP, Streaming, Messaging
- The structure of Markup: Visual vs Logical
  WSISWYG; WYSIAYG; WYSIWIM
- 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 Deep Web (the Invisible Web)

Dynamic content dynamic pages which are returned in response to a submitted query or accessed only through a form

Unlinked content pages which are not linked to by other pages (but clicking links them)

Private Web sites that require registration and login (Edventure)

Contextual Web pages with content varying for different access contexts (e.g., ranges of client IP addresses or previous navigation sequence).

Limited access content sites that limit access to their pages in a technical way (e.g., using the Robots Exclusion Standard)

Scripted content pages that are only accessible through links produced by JavaScript as well as content dynamically downloaded from Web servers via Flash or Ajax solutions.

Non-HTML/text content textual content encoded in multimedia (image or video) files or specific file formats not handled by search engines.

These pages all include data that search engines cannot find!

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)

The Web as Corpus

Two Approaches to using the Web as a Corpus

- 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 to draw conclusions on the data.
  - ⊗ Google is missing functionalities that linguists / lexicographers would like to have.

- 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

- Accessible through search engines (Google API, Yahoo API, Scripts)
- Document counts are shown to correlate directly with "real" frequencies (Keller 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
  - different queries give wildly different numbers

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

Web Sample

- Extracting and filtering web documents to create linguistically annotated corpora (Kilgarriff 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

Building 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 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, less markup

Text and Meta-text

- Explicit Meta-data
  - Keywords and Categories
  - Rankings
  - Structural Markup
- Implicit Meta-data
  - Links and Citations
  - Tags
  - Tables
  - File Names
  - Translations

Explicit Metadata

- You can get information from metadata within documents
  - When they are accurate they are very good
  - They are often deceitful
- HTML, PDF, Word, ...Metadata
- Keywords and Tags
- Rankings
- Links and Citations
- Structural Markup

Implicit Metadata

- You can get clues from metadata within documents
  - as they are non-intended, they tend to be noisy
  - but they are rarely deceitful
- HTML tags as constituent boundaries
- Tables as Semantic Relations
- File Names (content type and language)
- Translations — Bracketed Glosses; Cross-lingual Disambiguation
- Query Data
- Wikipedia Redirections and Cross-wiki Links

Language Identification and Normalization

Language Identification

- Need to identify the encoding and language of a page in order to process it (meta-data may be unreliable)
  - Linguistically-grounded methods
    - Diacritics
    - Character n-grams
    - Stop words
  - Similarity-based categorisation and classification
    - Character n-gram rankings
  - Machine Learning based methods
  - Context (under `.jp` or `.ko`?)
- Hard to do for short test snippets, similar languages, mixed text

Normalization

- Extracting text from various documents
- Segmenting continuous text
- Number Normalization: *$700K*, *$700,000*, *0.7 million dollars*, ...
- Date Normalization: *2000AD, 1421AH, Heisei 12, ...*
- Stripping stop words (***the***, ***a***, ***of***, ...)
- Lemmatization: *produces* $\rightarrow$ ***produce***
- Stemming: *producer* $\rightarrow$ *produc*; *produces* $\rightarrow$ *produc*
- Decompounding: *zonnecel* $\rightarrow$ *zon cel*

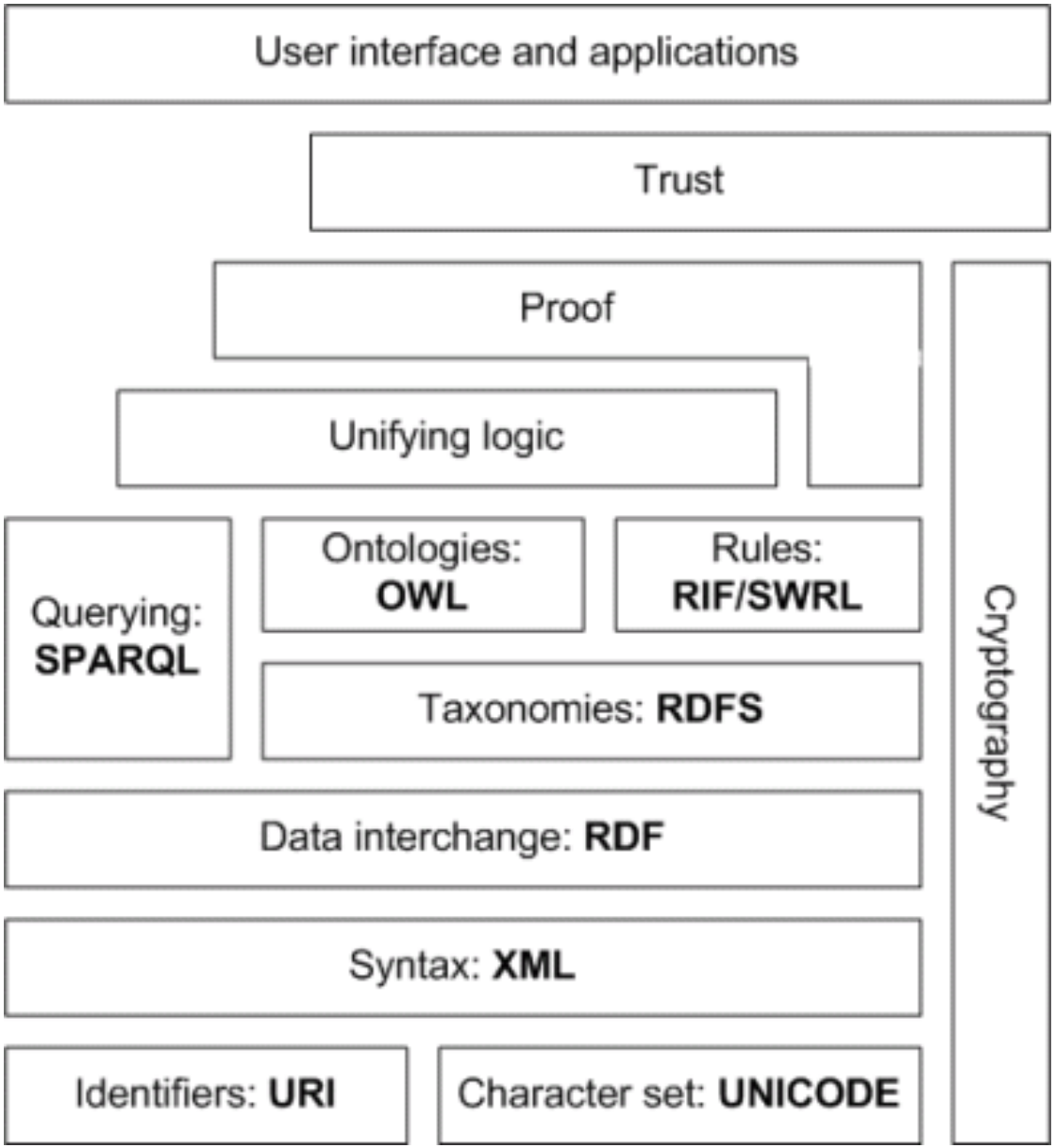
The Semantic Web

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



XML: eXtensible Markup Language

- XML is a set of rules for encoding documents electronically.
  - is a markup language much like HTML
  - was designed to carry data, not to display data
  - tags are not predefined. You must define your own tags
  - is designed to be self-descriptive
  - XML is a W3C Recommendation
- Can be validated for syntactic well-formedness

Semantic Web Architecture (details)

- Identify things with Uniform Resource Identifiers
  - Universal Resource Name: `urn:isbn:1575864606`
  - Universal Resource Locator: `http://www3.ntu.edu.sg/home/fcbond/`
- 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, ...

Criticism of the Semantic Web

Doctorow's seven insurmountable obstacles to reliable metadata are:

1. People lie
2. People are lazy
3. People are stupid
4. Mission Impossible: know thyself
5. Schemas aren't neutral
6. Metrics influence results
7. There's more than one way to describe something

Semantic Web and NLP

- The Semantic Web is about structuring data
- Text Mining is about unstructured data
- There is much more unstructured than structured data
  - NLP can infer structure
  - NLP makes the Semantic Web feasible
  - the Semantic Web can be a resource for NLP

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)
  - 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

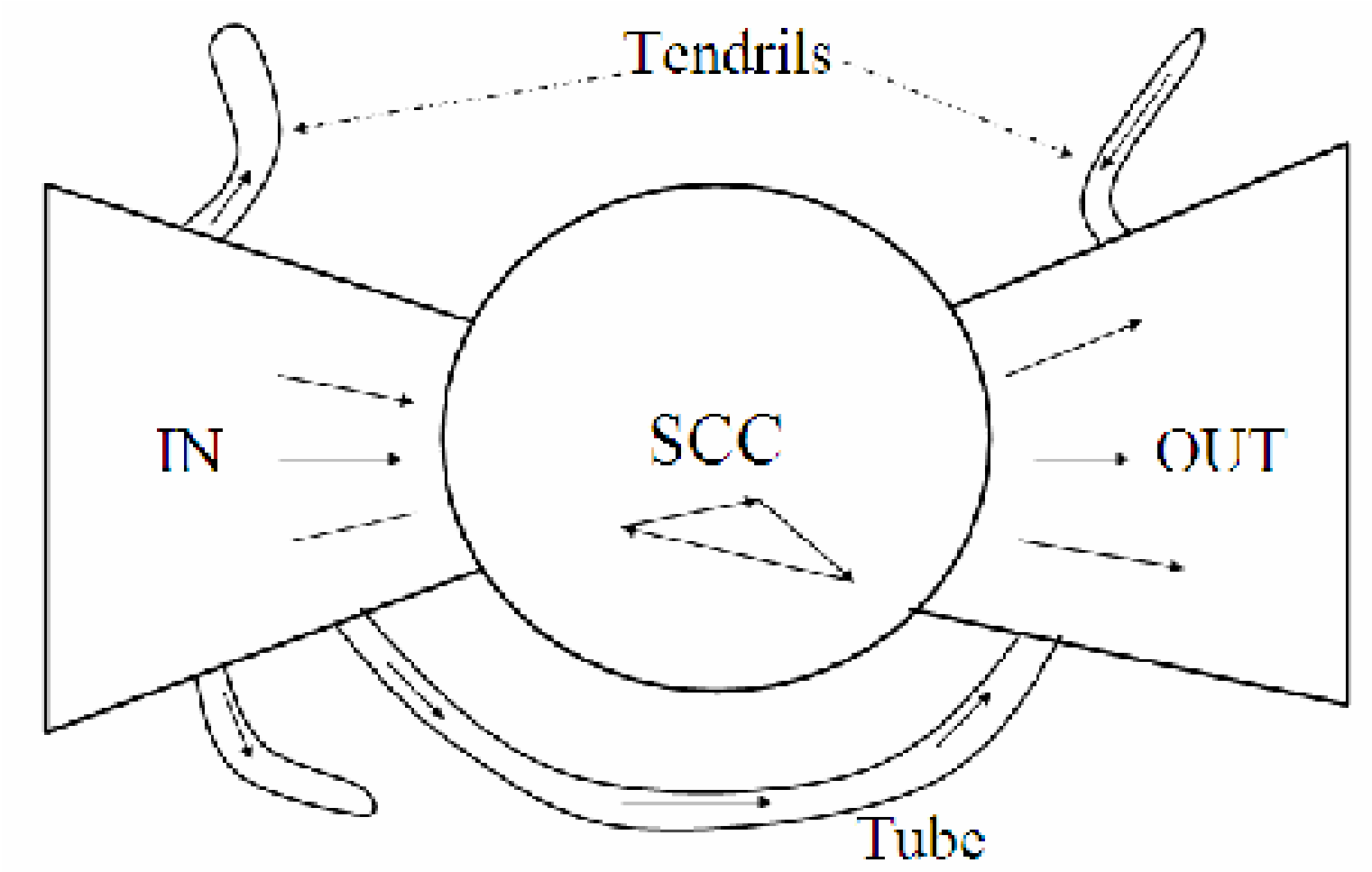
- Weight Citations by Quality of the paper

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

Characteristics of the Web: Bow Tie



SCC: Strongly Connected Core — can travel from any page to any page

Anchor Text

- Recall how hyperlinks are written:

```
<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

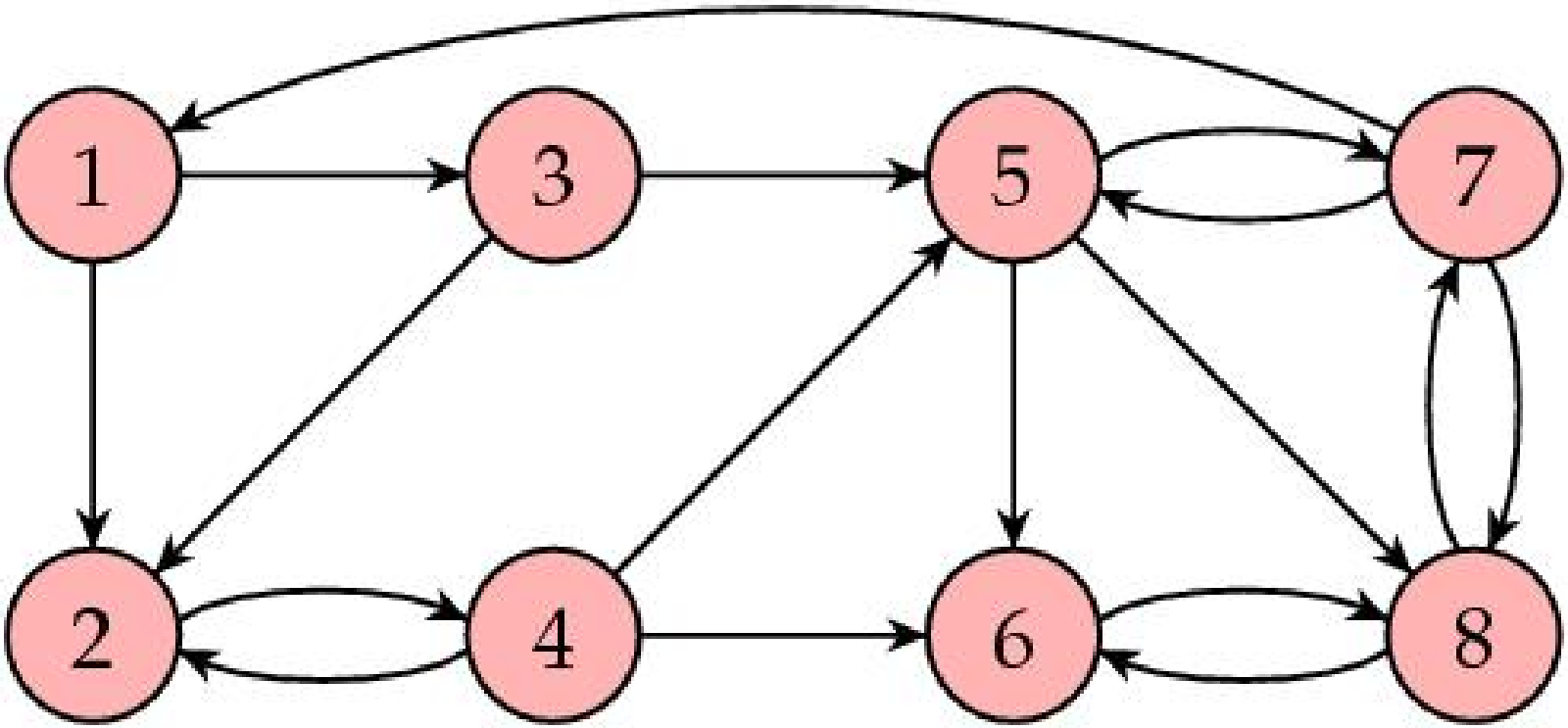
- This long-term visit rate is the page's PageRank.

- PageRank = long-term visit rate = steady state robability

Teleporting – to get us out of dead ends

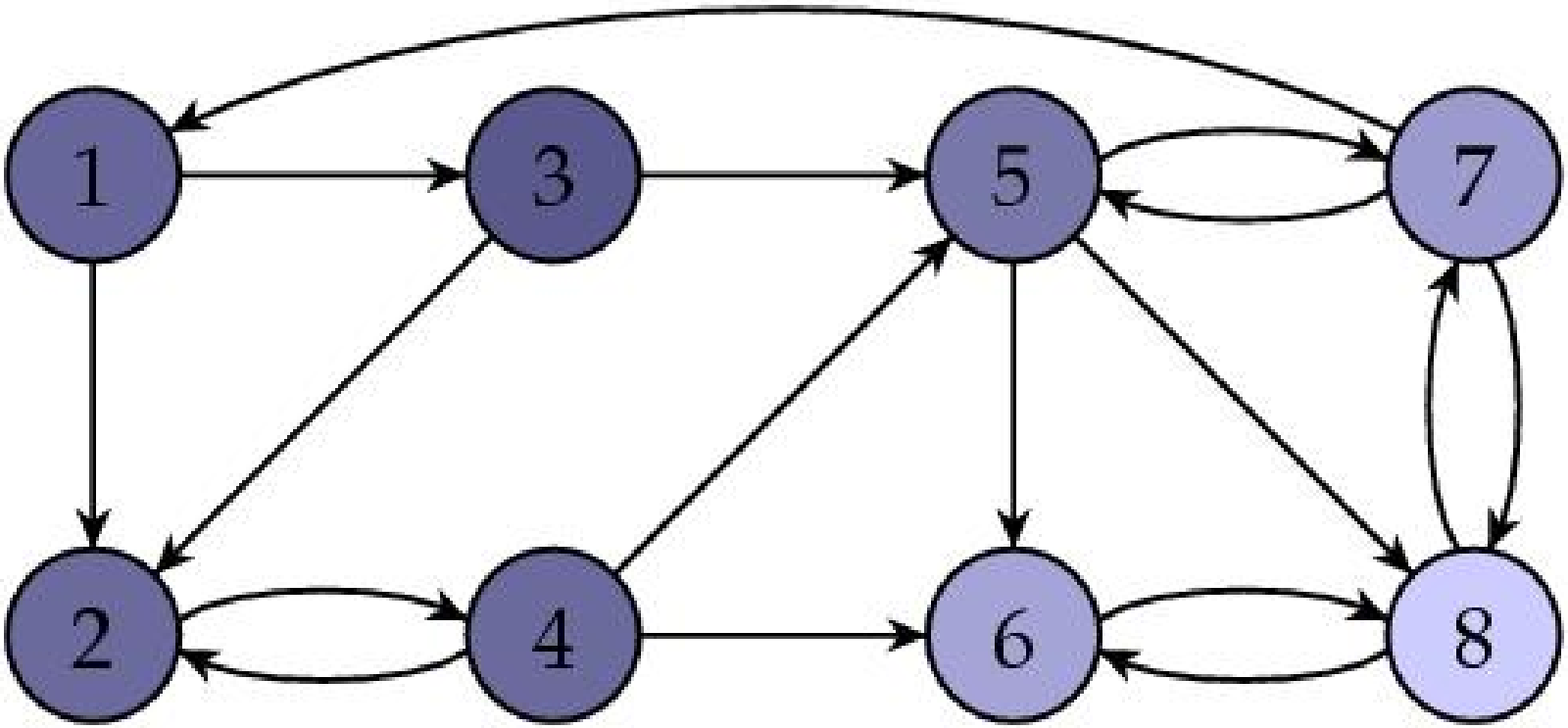
- At a dead end, jump to a random web page with probability $1/N$.
  (N is the total number of web pages)
- At a non-dead end, with probability 10%, jump to a random web page (to each with a probability of $0.1/N$).
- With remaining probability (90%), go out on a random hyperlink.
  - For example, if 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 a 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 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
  - ...

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 2009 approximately 43 million DOI names had been assigned by some 4,000 organizations
  - DOI: 10.1007/s10579-008-9062-z
    http://www.springerlink.com/content/v7q114033401th5u/

Conclusions

Revolutions in Language Technology

- Speech (Language itself)
- Writing (invented 3-5 times)
- Printing (made writing common)
- Digital Text (made writing transferable)
- Hyperlinking (taking writing beyond language)

The Internet

- The internet is useful as a tool
  - Passively for information access
  - Actively for collaboration
- The internet is interesting in and of itself
  - As a source of data about existing language
  - As a source of innovation in language

Recommended Texts

- *Wikipedia*
- Crystal, D. (2001). *Language and the Internet*. Cambridge University Press
- Sproat, R. (2010). *Language, Technology, and Society*. Oxford University Press
- Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* Prentice Hall, 2nd edition
- Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press

Complementary Courses

- **HG2051 Language and the Computer** — solving NLP problems with Python: introduces both programming and linguistics

- **HG3051 Corpus Linguistics** — (Pre Req-HG251 waived) This course is an introduction to the fast growing field of corpus linguistics.