MANCHESTER 1824
The University of Manchester

COMP38210
Workshop 3: Preparing to index

John McNaught
&
Sandra Sampaio

MANCHESTER 1824
The University of Manchester

Overview

- Document conversion
- Language identification
- Tokenization
  - Writing systems
- Filtering: stop words
- Normalisation
  - Case folding
  - Synonyms and homonyms
  - Lemmatization
  - Stemming

(IIR book)

Recall basic indexing pipeline

Documents to be indexed.

Friends, Romans, countrymen.

Tokenizer

Token stream.

Friends | Romans | Countrymen

Linguistic modules

Modified tokens.

friend | roman | countryman

Indexer

Inverted index.

friend → 2 → 4

roman → 1 → 2

countryman → 13 → 16

MANCHESTER 1824
The University of Manchester

Earlier steps before tokenizer

- Document conversion
  - Map .pdf, .doc, ,xml, ... to internal representation format
  - “Converting PDF to XML is a bit like converting hamburgers into cows.” (Kay, Saxonica)
- What about scanned PDFs (images!)
- Results from OCR may be poor
  - Pre-processing to detect/correct errors?
  - OCR errors may appear as correct (but not intended) text
- Language detection
  - Determines choice of processors on a per language basis
- Character set(s) in use?
- Text type detection & domain detection
  - Punctuation varies with text type/domain. Select domain resources (controlled vocabularies, lexicons, grammars, trained ML models)

MANCHESTER 1824
The University of Manchester

(IIR book)

Complications

- Documents being indexed can include docs from many different languages
- A single index may have to contain terms of several languages
- A document or its components can contain multiple languages/formats
  - French email with a German pdf attachment
- What is a unit document ?
  - A file?
  - An email? (Perhaps one of many in an mbox)
  - An email with 5 attachments?
  - A group of files (PPT or LaTeX as HTML pages)

MANCHESTER 1824
The University of Manchester

Language identification

- http://tinyurl.com/nulxdvj
- Try the Google Language API from this site (ensure “Google” is ticked in left frame)
- See how good (or not) it is on text from various languages

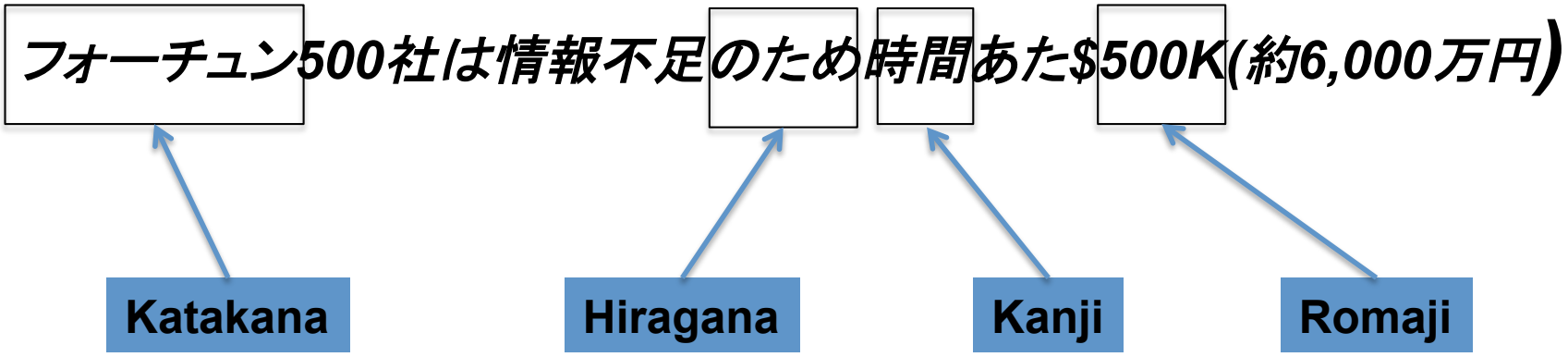
MANCHESTER 1824
The University of Manchester

Tokenization

- Goal: break input into tokens
- What is a word? Are all tokens words?
- 3 main classes of tokens often considered
  1. Morphosyntactic word
  2. Punctuation mark or special symbol
  3. A number
- Are these enough? Other types of token?
- Clitics? Compounds and multiwords?

Writing systems

- “token = whitespace-delimited char sequence”?
- Chinese, Japanese: no spaces between tokens
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - Unique tokenisation not guaranteed
- Japanese uses several alphabets…



Examples: IIR book

MANCHESTER 1824
The University of Manchester

A hard tokenization example

すももももももものうち。

Try Google Translate on this Japanese text.

(see worksheet 3 for file to download containing this example)

Then try inserting space(s) at various points.

Any idea what this means?

MANCHESTER 1824
The University of Manchester

Chinese example

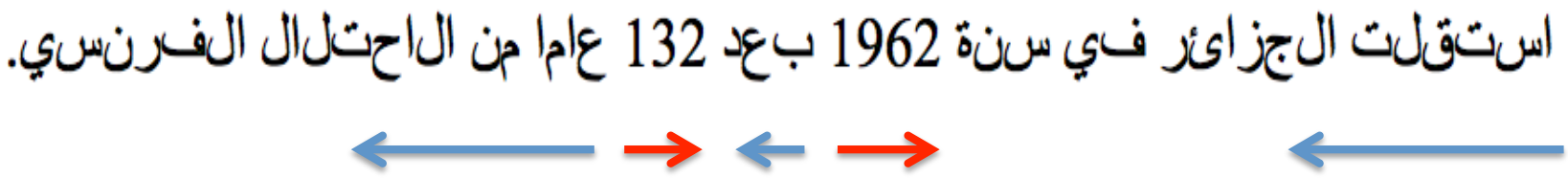
- 我喜欢新西兰花。

(Teahan et al., 2000, *CL* 26(3), p376)

- Again use an online translator and try inserting space at various points

Writing systems

- Arabic, Hebrew: generally written right to left, but not always
- Words separated, but complex ligatures used within words



- Right to left, but numbers written left to right

Example: IIR book

MANCHESTER 1824
The University of Manchester

Writing systems

- Alphabetic: symbols represent sounds
- Syllabic: symbols represent syllables
- Logographic: symbols represent words (or parts of words) – thousands of symbols
- A language's writing system often is complex
- English: mainly alphabetic, but what about characters such as %, &, £, 0, 1, 2, ...

MANCHESTER 1824
The University of Manchester

Writing systems

- Conventions denoting boundaries between units
- Amharic (Ethiopia) traditionally marks word (:) and sentence (::) boundaries
- Thai: neither marked
- English: whitespace plus punctuation marks, but still much ambiguity in tokenisation
- Chinese, Japanese: sentence boundaries but no word boundaries

MANCHESTER 1824
The University of Manchester

Tokenization issues

- How many tokens for forms such as *you're*?
  - 1? you're
  - 3? you + ' + re
  - 2? you + are
- president's speech vs. president + 's + speech
- The president's arriving now:
- president + 's vs. president + is
  - ' : genitive, contraction (is or has), quotation
- greengrocer's apostrophe (incorrect plural forms): GP's, 1980's, two orange's. Test: try to form plural possessive:
  - GP's' surgeries ??? a 1980's' fashion ???

MANCHESTER 1824
The University of Manchester

Tokenization issues

- Abbreviation
  - Problem at end of sentence
  - Previously unseen abbreviations problematic
- Hyphenation
  - Manchester-based
  - Sister-in-law
  - Fr.: il va vs. va-t-il - need to tokenise to recover grammatical information
  - EOLN word breaking (inserted vs. ‘real’ hyphen)

MANCHESTER 1824
The University of Manchester

Tokenization issues

- e-mail, co-operate: one token
- pro-Arab: two tokens
- a go-now-and-go-quickly ultimatum
- database, data-base, data base
- carbonaceous-chondrite-like
- Telephone numbers: many different formats (language conventions) with whitespace, dots, slashes, hyphens, parentheses, plus signs ...
- Dates
- Decimals: 0.05, 3.4, .6
- Language conventions for numbers:
  - English: 123,456.78
  - French: 123 456,78
  - German: 123.456,78

MANCHESTER 1824
The University of Manchester

Tokenization issues

- This dinner costs 5 pounds a dish
  - A £5-a-dish dinner
- pre- and post-conditions
  - Map to: pre-conditions and post-conditions ?
- {no-one, no one} is here
- no one person can tell how to tokenise

MANCHESTER 1824
The University of Manchester

Tokenization issues

- Clitics and contractions
- 's : already seen
- can’t vs. ca + n’t vs. can + not vs. cannot
- Also forms as in Fr. Je t’aime (Je te aime)
- And Fr. du (de + le)

MANCHESTER 1824
The University of Manchester

Tokenizing punctuation

- Punctuation usually treated as separate tokens
  - But often should be attached to adjacent token
  - Language dependent
- Abbreviation (already seen)
- Quotation marks
- Single quote and apostrophe often same char
- Abbreviated forms: fo’c’s’le (forecastle of a ship)

MANCHESTER 1824
The University of Manchester

Tokenization issues: domain specific

- Organism/species names, authorities, families, orders
  - E. coli
  - Saccharum spp. L. (authority L=Linnaeus)
- Coordinates: 41032'38"N
- 5'-TATGCTCGCCAGAGGATAATTA-3' (M/c-FI)
- Colony-stimulating factor 1 (CSF-1)-nullizygous mouse (Csf1(op)/Csf1(op)) phenotype
- p21-/- map to p21 -/- ?
- MEK1/2 map to Mek1 and Mek2 ?

MANCHESTER 1824
The University of Manchester

Tokenisation issues: chemistry

The importance of the phenolic quinolyl hydrazones arises from incorporating the quinoline ring with the phenolic compound; 2,4-dihydroxy benzaldehyde.

The present study is planned to check the effect of the counter anions on the type and geometry of the isolated copper(II)-complexes as well as the ligational behavior of the phenolic hydrazone; 4-[(2-(4,8-dimethylquinolin-2-yl)hydrazono)methyl] benzene-1,3-diol; (H2L).

MANCHESTER 1824
The University of Manchester

Scope of tokenisation

- Some tokenisers attempt to handle multiwords and map these to one token
  - San Francisco
  - in spite of
  - 12th January 2010
  - 12/01/2010
- Is this the job of tokenising? Or of e.g. named entity recognition?
- Tokenisation: knowing **when to split**, not when to combine

MANCHESTER 1824
The University of Manchester

Filtering: stop words

- In IR, a goal is to *distinguish* doc X from doc Y
- Hypothesis: highly frequently occurring words have low distinguishing power
  - Closed class function words (the, and, of, it, …)
    - Top 30 words account for ~30% of postings
  - 'content free' words (e.g. information)
- Therefore, should filter out such words
  - Use of stop list
- But trend is now away from doing this
  - Why?

MANCHESTER 1824
The University of Manchester

Normalization

- Map tokens to normalised form
  - {B.B.C., BBC} → BBC
  - {multi-national, multinational} → multinational
  - {résumé, resume} → resume ??
  - {Duesseldorf, Düsseldorf, Dusseldorf} → Dusseldorf
- Transliterations
  - Przhevalsky, Prjevalsky, Prejevalsky, Prezhevalsky, ...ski
- What do your users type? (even if normally use accents, may not type them in query)
- Cross-lingual considerations? E.g. Date forms
  - *7月30日 vs. 30/7*

MANCHESTER 1824
The University of Manchester

Normalization: case folding

- Map all upper case letters to lower case?
  - … Fed government shutdown …
  - … fed government shutdown …
  - General Electric vs. general electric
  - CAT vs. cat

MANCHESTER 1824
The University of Manchester

Synonyms, homonyms, etc.

- {car, automobile}
- {favor, favour}
- bank (river) vs. bank (finance)
- bow (ship), bow (tie), bow (salute), bow (weapon), bow (violin), bow (hair)
- **What we decide will affect what we index and what is retrieved**

MANCHESTER 1824
The University of Manchester

Lemmas and stems

- {am, are, is} → be
- {horse, horses, horse's, horses' } → horse
- the girls' horses are different heights → the girl horse be different height
- Lemmatization: reduction to "dictionary headword" form

MANCHESTER 1824
The University of Manchester

Stemming

- Chop “ends of words” (typically) before indexing
- Language dependent process, often heuristic, crude
- Goal: collapse similar forms to a canonical form (improves word counts, improves *recall*)
- May yield forms that are “not words”

MANCHESTER 1824
The University of Manchester

Aside: precision/recall

- 2 Measures (among others) used to evaluate the performance of search engines
- Precision: fraction of retrieved documents that are relevant
- Recall: fraction of relevant documents that are retrieved
- Precision increases as recall decreases and vice versa
- More later!

MANCHESTER 1824
The University of Manchester

Stemming errors

- Understemming fails to conflate related forms
  - divide → divid
  - division → divis
- Overstemming conflates unrelated forms
  - neutron, neutral → neutr

MANCHESTER 1824
The University of Manchester

Stemming algorithms

- Most used for English: Porter's stemmer
- Results?
  - Mixed for English
- Morphologically complex languages benefit more, e.g.:
  - Spanish, German, Finnish
    - 30% performance gain for Finnish

MANCHESTER 1824
The University of Manchester

Resources

- langid.py for 97 languages (paper and code):
  - http://aclweb.org/anthology//P/P12/P12-3005.pdf
  - https://github.com/saffsd/langid.py
- Downloadable C++/Python port:
  - https://code.google.com/p/chromium-compact-language-detector/
- Language identification for short texts (hard!) for those interested
  - http://www.lrec-conf.org/proceedings/lrec2010/pdf/279_Paper.pdf

MANCHESTER 1824
The University of Manchester

Resources

- Chapter 2 of Manning et al., Introduction to Information Retrieval (online, free)
- Palmer, D. (2000) Tokenisation and sentence segmentation. Ch 2 in Dale, R., Moisl, H. & Somers, H. (eds), *Handbook of Natural Language Processing*. Marcel Dekker. In UML
- Mikheev, A. (2002) Periods, Capitalized Words, etc. *Computational Linguistics* 28(3): 289-318. Via UML e-journals

MANCHESTER 1824
The University of Manchester

Resources

- Xerox tokenizer demo
  - http://open.xerox.com/Services/fst-nlp-tools/Consume/175
- NLTK tokenizer demo
  - http://text-processing.com/demo/tokenize/
- Illinois demos
  - http://cogcomp.cs.illinois.edu/page/demos
- FBK TextPro demo
  - http://hlt-services2.fbk.eu/textpro/?page_id=56
- The classic UPenn PTB tokenizer script (sed)
  - ftp://ftp.cis.upenn.edu/pub/treebank/public_html/tokenizer.sed

MANCHESTER 1824
The University of Manchester

Resources

- Van Rijsbergen's classic stoplist
  - http://www.dcs.gla.ac.uk/Keith/Chapter.2/Table_2.1.html

MANCHESTER 1824
The University of Manchester

Resources: stemming

- Porter stemmer (only for English)
  - http://snowballstem.org/demo.html
- Comparison Porter/Lancaster stemmers
  - http://smile-stemmer.appspot.com/
- Evaluation of stemming
  - https://www.cs.cmu.edu/~mbilotti/pubs/Bilotti04a.pdf
- Impact of stemming accuracy on IR
  - http://dx.doi.org/10.1016/j.ipm.2016.03.004
- Comparing effect of tokenization, stemming and stopword strategies
  - http://dx.doi.org/10.1007/s10791-007-9027-7