Building and Exploring (Web) Corpora

EMLS 2008, Stuttgart

23-25 July 2008

Pavel Rychlý

pary@fi.muni.cz

NLPlab, Masaryk University, Brno

Outline

(1)Introduction to text/web corpora

(2)Exercises 1

(3)Regular expressions, query language, CorpusBuilder

(4)Exercises 2

(5)Google as Web corpus, Unix tools

(6)Exercises 3

(7)Summary, Students' reports

Outline (1)

- Introduction to text corpora
  - Sketch Engine
  - WebBootCaT (building corpora from web)
- Exercises 1
  - Create own corpora of different languages using WebBootCaT
  - Try basic corpus searching
  - Find differences in corpora (languages, general/specialized)

Outline (2)

- Regular expressions, CQL, CorpusBuilder
  - Regular expressions tutorial
  - Corpus Query Language
    - defining word sketches
  - CorpusBuilder
    - building corpora from your own texts
- Exercises 2
  - RE examples
  - Create own corpus using CB
  - Define simple sketch grammar

Outline (3)

- Google as Web corpus, Unix tools
  - Why not to use a search engine a corpus
  - Simple but powerful tools for text processing
- Exercises 3
  - Googleology
  - Unix tools exercises

Who is who?

- Pavel Rychlý
  - Sketch Engine, Manatee, Bonito
- Jan Pomikálek
  - CorpusBuilder, WebBootCaT
- ???
  - computing/linguistics
  - languages
  - RE, SkE, programming languages

What is a text corpus?

- Wikipedia: In linguistics, a corpus (plural corpora) or **text corpus** is a large and structured set of texts (now usually electronically stored and processed).
- usually POS-tagged and lemmatized
  - demo (dream)
- a source of information about a natural language
- gives examples of how natural language is used

What can we do with corpora?

- search for occurrences (contexts) of
  - single words (or lemmas)
  - phrases
  - structures
    - e.g. the verb *look* followed by a preposition, an adjective and a noun
    - [lemma="look" & tag="V.*"] [tag="PR.*"] [tag="AJ.*"] [tag="N.*"]

What can we do with corpora? (2)

- compute statistics, e.g.
  - find collocates
    - demo: *feel*
  - word sketches
    - demo: *test*
  - frequency distributions
    - demo*: damn*, *feel*

Common usage of text corpora

- lexicography (writing dictionary entries)
  - recognize different senses of a given word
  - find strong collocations
- language learning/teaching
- building models for
  - machine translation
  - speech recognition

Size of text corpora

- Brown (English), DESAM (Czech)
  - 1 million words
- BNC (English), * National Corpus
  - 100 million words
- UkWaC (English), ItWaC (Italian)
  - 2 billion words
- BiWeC (English)
  - 5-10 billion words

Corpus manager

- software for working with corpora
- fast searching (corpora are large)
- powerful query language
- statistics

Sketch Engine

- http://corpora.sketchengine.co.uk/auth/
  - user name: emlsXX (e.g. emls05)
  - password: emls
- open BNC

Restricting searches to text types

- use the Text Types form

Working with concordances

- navigating through pages
- getting information about the source (document, structures)
- seeing wider context
- switching to/from sentence view
- changing view options
- sorting
- random samples
- frequencies

Word sketches and thesaurus

- Adam Kilgarriff: A **Word Sketch** is a corpus-based summary of a word's grammatical and collocational behaviour.
- word sketches divide collocations into grammatical relations
- thesaurus computed from word sketches

Web as corpus

- WWW is a very rich source of textual data (August 2005: 19.2 billion web pages)
- the data is available to everyone
- errors in texts – problem?
  - Google: acommodation/accomodation/accommodation

Advantages of web corpora

- common corpora
  - expensive
  - limited electronic resources
  - printed resources have to be used
  - building is time consuming
  - copyright issues
- web corpora
  - cheap
  - almost unlimited resources
  - building is fast (can be automated)

Using web as corpus (local)

- pre-create
  - crawl web
  - download web pages
  - clean data
  - annotate
  - output = large ballanced web corpus (itWaC, deWac)
- advantages
  - huge corpora can be build
- disadvantages
  - time consuming
  - computer experts required

Using web as corpus (on-line)

- on-the-fly
  - input = query
  - search engine
  - download web pages/snippets
  - (annotate)
  - output = concordance lines
- disadvantages
  - limited query language
  - slow

WebBootCaT

- BootCaT = Simple Utilities to Bootstrap Corpora and Terms from the Web
  - Marco Baroni et al (University of Bologna)
- medium size domain specific corpora
  - ca 1 million words
- input = seed words + options
- output = annotated domain specific corpus loaded into Sketch Engine

Domain specific corpora

- lexicography, speech recognition, machine translation
- less data is sufficient than for general corpora

SEED WORDS

climbing
rock
bouldering
ascent
route
on sight
el capitan
dolomiti

random n-grams generating

N-GRAMS

bouldering	climbing	climbing
rock	route	route
on sight	dolomiti	on sight

Google searching

URLS

```
http://en.wikipedia.org/wiki/Climbing
http://en.wikipedia.org/wiki/Rock-climbing
http://www.indoorclimbing.com/comp_types.html
http://www.cocc.edu/alish/intermclimb.htm
http://www.czechclimbing.com/
...
```

documents collecting

HTML DOCUMENTS

```
<!DOCTYPE html PUBLIC "-//W3C//DTD
<html xmlns="http://www.w3.org/199
    <html>
        <meta http-equiv="Content-
```

boilerplate stripping

TEXT DOCUMENTS

```
Climbing is going up, or depending
on context, also down or sidewards
(traversing). It may refer to
aircraft, a land vehicle, and ...
```

duplicates removal

POS tagging, lemmatisation

UNIQUE TEXT DOCUMENTS
WITH POS TAGS AND LEMMAS

Climbing	NN	climbing
is	VBZ	be
going	NN	going
up	RB	up

indexing

TEXT CORPUS IN
WORD SKETCH ENGINE

keywords extraction

EXTRACTED KEYWORDS

gear
climb
difficulty
bolts
wall
mountain
rope

WebBootCaT

- n-grams generating
- Yahoo! search (Yahoo! API)
- download web pages
- boilerplate stripping
  - strip tag heavy parts
- duplicates removal
  - Text::DeDuper (CPAN)
  - n-gram based

WebBootCaT

- POS-tagging, lemmatisation
  - TreeTagger
    - English, German, French, Italian, Spanish, Bulgarian
  - Czech tagging coming soon
- Indexing
  - manatee, Sketch Engine

Keywords extraction

- reference corpora
  - large web corpora (ca 500 million words)
- compare relative frequencies of words

word	WBC corpus	reference corpus
rope	$1.5 * 10^{-1}$ %	$8.3 * 10^{-4}$ %
wall	$1.1 * 10^{-1}$ %	$67.1 * 10^{-4}$ %
Yosemite	$1.2 * 10^{-1}$ %	$0.7 * 10^{-4}$ %

- multi-word expressions

KW extraction – problems

Kittyhawk: USS Kittyhawk calling. Request you alter course. Over and out.

Radio: Message received. Mission such we cannot alter cours. We request you alter course.

Kittyhawk: We are an aircraft carrier of the US Navy. We demand you alter course soonest to avoid collision.

Radio: We are unable to implement your request. We recommend you take avoiding action immediately.

Kittyhawk: If you continue to ignore our order we will open fire.

Radio: We are a lighthouse – your call!

Average reduced frequency

- look at the word distribution in the corpus
- the less uniform distribution the higher frequency reduction

Summary: What have you learned?

- Using corpus manager Sketch Engine
  - simple searching
  - working with concordances
    - sorting, random sampling, computing frequencies
  - viewing word sketches and thesaurus
- WebBootCaT
  - build a domain specific corpus from the web

Exercises 1: Concordance searches

- simple searches
  - find all occurrences of the word *play*
  - all words which have *play* as lemma
  - lemma *play* as noun
- using contexts
  - verb *fight* followed by
    - any preposition
    - preposition *for*
  - verb *fight* preceded by a noun
  - verb *fight* followed by the noun *independence* (window 5)

Exercises 1: WebBootCaT

- Create own corpora
  - different languages, same domain
  - different domains
- Compare corpora
  - find differences in collocations/word sketches of a word in different languages/domains
  - describe differences