Informatics 1: Data & Analysis

Lecture 12: Corpora

Ian Stark

School of Informatics
The University of Edinburgh

Tuesday 5 March 2013
Semester 2 Week 7

Lecture Plan

XML

We start with technologies for modelling and querying *semistructured data*.

- Semistructured Data: Trees and XML
- Schemas for structuring XML
- Navigating and querying XML with XPath

Corpora

One particular kind of semistructured data is large bodies of written or spoken text: each one a *corpus*, plural *corpora*.

- Corpora: What they are and how to build them
- Applications: corpus analysis and data extraction

Homework

Tutorial Exercises

Tutorial sheet 6 is now online. This week you'll be using the *Corpus Query Processor* (CQP) tool and a library of all the works of Charles Dickens.

Before starting the exercise, you will need to read the opening sections of the CQP Tutorial, also available from the course web page.

Reading

T. McEnery and A. Wilson.
Corpus Linguistics. Second edition, Edinburgh University Press, 2001.
Chapter 2: What is a corpus and what is in it? (§2.2.2 optional)

Photocopied handout, also available from the ITO.

Natural Language as Data

Written or spoken natural language has plenty of internal structure: it consists of words, phrases and sentences, governed by spelling and grammatical rules, and so forth.

Nevertheless, on a computer, it standardly represented as a text file: a simple sequence of characters.

This is an example of unstructured data: the data format itself has no structure imposed on it. (Above the level of character encoding.)

Often, however, it is useful to annotate text by marking it up with additional information about its linguistic or semantic content.

Text with this kind of markup is a widespread and substantial example of semistructured data.

What is a Corpus?

The word *corpus* (plural *corpora* or *corpuses*) is Latin for "body".

In literature a corpus is a collection of written texts, in particular the complete works of a single author, or a body of writing on a single subject.

In *computational linguistics* and in *theoretical linguistics* a corpus is a body of written or spoken text used for study of a particular language or language variety.

This application domain depends on the following features in a corpus.

- Representative sampling
- Finite size
- Machine-readable form
- Use as a standard reference

The following slides expand on these: all are important for a corpus to be a useful linguistic resource.

Representative Sampling

Sampling

Corpora provide data for *empirical linguistics:* the scientific investigation of real-world language use, proposing and testing hypotheses.

However, any corpus can only contain a *sample* of language use — although it might be very large, it will usually be dwarfed by the actual language in the wild. (XKCD: What if? #34)

Representative

For meaningful linguistic analysis, the sample in a corpus should be chosen as *representative* of the language as it is used in practice.

For example, the complete works of Shakespeare is an appropriate corpus for analysing how Shakespeare used language; but would not give a representative sample for studying Elizabethan English.

Finiteness

It's natural that corpora should be finite. Most also have a fixed size. When building a corpus it is usually decided at the outset how the language is to be sampled and how much data to include. Once the samples have been taken, the corpus content is fixed.

There are exceptions to this: *monitor corpora* capture the continuing growth and change of a language. They remain finite, but may extend in size over time.

This finite size rule for corpora contrasts with the study of *grammars* in theoretical linguistics. These are sets of rules, such as *context-free grammars*, which generate potentially infinite collections of sentences.

Machine Readable

Historically, the word "corpus" referred to a body of printed (or even written) text.

Now, corpora are almost universally machine-readable — that is, stored on and transferred between computers.

Machine-readable corpora have several distinctive features in comparison with books of printed text.

- They can be huge in size, up to billions of words.
- They can be searched and analysed efficiently.
- They can be made available to many users simultaneously, at large distances.
- They can easily (and sometimes automatically) be annotated with additional useful information.

Standard Reference

A corpus is often a *standard reference* for the language variety it represents.

Having a corpus as a standard reference allows competing theories about the language variety to be compared against each other on the same sample data.

For this, the corpus has to be widely available to researchers, fitting their shared requirements and used by them in practice.

The likely usefulness of a corpus as a standard reference depends on all the preceding three features: representativeness, fixed finite size and machine readability.

Summary

A corpus is — in general — a widely available fixed-sized body of machine-readable text, appropriately sampled to properly represent a certain language variety.

Any particular corpus, however, may not have all of these characteristics.

Some Notable English Language Corpora

- The Brown Corpus of American English was compiled at Brown University and published in 1967. It contains around 1,000,000 words.
- The British National Corpus (BNC), published in the mid-1990's, is a 100,000,000-word text corpus intended to representative of written and spoken British English from the late 20th century.
- The American National Corpus (ANC) is an ongoing project to create a 100,000,000-word corpus of written and spoken American English since 1990.

  The ANC currently contains 22,000,000 million words and is published, with annotations, as XML.
- The Oxford English Corpus (OEC) is an English corpus used by the makers of the Oxford English Dictionary. It is the largest text corpus of its kind, containing over 2,000,000,000 words.

Some Notable English Language Corpora

- The Brown Corpus of American English was compiled at Brown University and published in 1967. It contains around 1 megaword.
- The British National Corpus (BNC), published in the mid-1990's, is a 100-megaword text corpus intended to representative of written and spoken British English from the late 20th century.
- The American National Corpus (ANC) is an ongoing project to create a 100-megaword corpus of written and spoken American English since 1990.

  The ANC currently contains 22 megaword and is published, with annotations, as XML.
- The Oxford English Corpus (OEC) is an English corpus used by the makers of the Oxford English Dictionary. It is the largest text corpus of its kind, containing over 2 gigaword.

Two Kinds of Corpus

Individual corpora may be *unannotated* — just consisting of bare text – or *annotated*, usually with some linguistic or semantic information.

- Unannotated corpora are examples of unstructured data.
- Annotated corpora are examples of semistructured data.

The four English language corpora on the previous slide are all annotated.

From here on we will be looking almost exclusively at annotated corpora.

Building a Corpus

Two tasks are central to building an annotated corpus:

- Collect data — this involves *balancing* and *sampling*;
- Add information — perform the *annotation*.

Balancing ensures that the linguistic content of a corpus represents the full variety of the language sources for which the corpus is intended to provide a reference. For example, a balanced *text* corpus (as opposed to spoken) includes materials from sources such as books, newspapers, magazines, letters, etc.

Sampling ensures that the material is representative of the types of source. For example, sampling from newspaper text involves selecting texts randomly from different newspapers, different issues, and different sections of each newspaper.

Balancing

Balancing may operate across several different dimensions of source material.

- Language type: Taking samples from some or all of:
  - edited text (e.g., articles, books, news wire);
  - spontaneous text (e.g., email, blog comments, letters);
  - spontaneous speech (e.g., conversations, dialogues);
  - scripted speech (e.g., formal speeches).
- Genre: Finer-grained resolution of material type (e.g., 18th century novels, scientific articles, movie reviews, parliamentary debates).
- Domain: What the material is about (e.g., crime, travel, biology, law).
- Media: The physical realization of a corpus (e.g., text, audio, transcribed speech, video).

Planning for a corpus involves fixing on which kinds of balancing are required, and how they will be realised.

Example Balanced Corpora

Brown Corpus

A balanced corpus of written American English:

- One of the earliest machine-readable corpora;
- Developed by Francis and Kučera at Brown in early 1960's;
- 1M words of American English texts printed in 1961;
- Sampled from 15 different genres.

British National Corpus:

A large, balanced corpus of British English:

- One of the main reference corpora for English today;
- 90M words text; 10M words speech;
- Text sampled from newspapers, magazines, books, letters, school and university essays;
- Speech recorded from volunteers balanced by age, region, and social class; also meetings, radio shows, phone-ins, etc.

Comparison of Some Standard Corpora

Corpus	Size	Genre	Mode	Language
Brown Corpus	1M	general	text	American English
British National Corpus	100M	general	mixed	British English
Penn Treebank	1M	news	text	American English
Broadcast News Corpus	300k	news	speech	7 languages
MapTask Corpus	147k	dialogue	speech	British English
CallHome Corpus	50k	dialogue	speech	6 languages

Pre-processing and Annotation

Going from raw linguistic data to an annotated corpus can be broken down into stages.

- Pre-processing identifies the basic units in the corpus, such as:
  - Tokenization;
  - Sentence boundary detection.
- Annotation adds information appropriate to the corpus, such as:
  - Parts of speech;
  - Syntactic structure;
  - Dialogue structure;
  - Prosody (rhythm, intonation and stress in speech).

Tokenization

Tokenization: Divide raw textual data into *tokens* such as words, numbers, punctuation marks.

Word: A continuous string of *alphanumeric* characters delineated by *whitespace* (space, tab, newline).

Unicode provides a lot of detailed information about individual characters to help classify them and support tokenization.

However, there remain many potentially difficult cases. For example:

- amazon.com, Micro$oft
- John's, isn't, rock'n'roll
- A final take-it-or-leave-it offer
- Led down a cul de sac

Sentence Boundary Detection

Sentence boundary detection: Identify the start and end of individual sentences.

Sentence: A string of words ending in a full stop, question mark or exclamation mark.

This is enough much of the time; and again, Unicode can help classify sentence-breaking punctuation elements.

There are still potential problem cases:

- Dr. Foster went to Gloucester.
- “Are you going?” John asked.
- He lost cash on lastminute.com.

Detection of word and sentence boundaries is particularly difficult for spoken data.

Corpus Annotation

Annotation adds information to the corpus that is not explicit in the data itself. This is often specific to a particular application; and a single corpus may be annotated in multiple ways.

Annotation scheme is a basis for annotation, made up of a *tag set* and *annotation guidelines*.

Tag set is an inventory of labels for markup.

Annotation guidelines tell annotators — domain experts — how a tag set should be applied. In particular, this is to ensure consistency across different annotators.

The next lecture will look at *part-of-speech* or *POS* annotation as an example of this.