Informatics 1B, 2008

School of Informatics, University of Edinburgh

Data and Analysis

Note 9

Data Acquisition and Annotation

Alex Simpson

Part II — Semistructured Data

XML

Note 6 Semistructured data and XML

Note 7 Querying XML documents with XQuery

Corpora

Note 8 Introduction to corpora

Note 9 Data acquisition and annotation

Note 10 Querying a corpus

Last lecture

Defined a corpus as a collection of textual or spoken data:

- sampled in a certain way;
- finite in size;
- available in machine-readable form;
- often serving as a standard reference.

This lecture

- How to collect corpus data (*balancing* and *sampling*)
- How to add information to a corpus (*annotation*).

Balancing and sampling

Balancing ensures that a corpus representative of the language, reflects the linguistic material that speakers are exposed to.

Example A balanced text corpus includes texts from many diffeerent types of source (depending on the language variety); e.g., books, newspapers, magazines, letters, etc.

Sampling ensures that the material is representative of the types of source.

Example Sampling from newspaper text: select texts randomly from different newspapers, different issues, different sections of each newspaper.

Balancing

Things to take into account when balancing:

- *language type*: may wish to include samples from some or all of:
  - edited text (e.g., articles, books, newswire);
  - spontaneous text (e.g., email, Usenet news, letters);
  - spontaneous speech (e.g., conversations, dialogs);
  - scripted speech (e.g., formal speeches).
- *genre:* fine-grained type of material (e.g., 18th century novels, scientific articles, movie reviews, parliamentary debates)
- *domain*: what the material is about (e.g., crime, travel, biology, law);

Examples of balanced corpora

Brown Corpus: a balanced corpus of written American English:

- one of the earliest machine-readable corpora;
- developed by Francis and Kucera at Brown in early 1960's;
- 1M words of American English texts printed in 1961;
- sampled from 15 different genres.

British National Corpus: large, balanced corpus of British English.

- one of the main reference corpora for English today;
- 90M words text; 10M words speech;
- text part 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.

Genres and domains in the Brown Corpus

The 15 genres are labelled A to R (letters I, O and Q are omitted); e.g.:

Genre A: PRESS (Reportage) — 44 texts
Domains: Political; Sports; Society; Spot News; Financial; Cultural

Genre B: PRESS (Editorial) — 27 texts
Domains: Institutional Daily; Personal; Letters to the Editor

Genre C: PRESS (Reviews) — 17 texts
Domains: theatre; books; music; dance

Genre J: LEARNED — 80 texts
Domains: Natural Sciences; Medicine; Mathematics; Social and Behavioral Sciences; Political Science, Law, Education; Humanities; Technology and Engineering

Comparison of some standard corpora

Corpus	Size	Genre	Modality	Language
Brown Corpus	1M	balanced	text	American English
British National Corpus	100M	balanced	text/speech	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

Raw data from a linguistic source can't be exploited directly. We first have to perform:

- *pre-processing:* identify the basic units in the corpus:
  - tokenization;
  - sentence boundary detection;
- *annotation:* add task-specific information:
  - parts of speech;
  - syntactic structure;
  - dialogue structure, prosody, etc.

Tokenization

Tokenization: divide the raw textual data into tokens (words, numbers, punctuation marks).

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

Example: potentially difficult cases:

- amazon.com, Micro$oft
- John's, isn't, rock'n'roll
- child-as-required-yuppie-possession

  (As in: "The idea of a child-as-required-yuppie-possession must be motivating them.")
- cul de sac

Sentence Boundary Detection

Sentence boundary detection: identify the start and end of sentences.

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

This is correct 90% of the time.

Example: potentially difficult cases:

- Dr. Foster went to Gloucester.
- He said “rubbish!”.
- He lost cash on lastminute.com.

The detection of word and sentence boundaries is particularly difficult for *spoken data*.

Corpus Annotation

Annotation: adds information that is not explicit in the corpus, increases its usefulness (often application-specific).

Annotation scheme: basis for annotation, consists of a tag set and annotation guidelines.

Tag set: is an inventory of labels for labels for markup.

Annotation guidelines: tell annotators (domain experts) how tag set is to be applied; ensure consistency across different annotators.

Part-of-speech (POS) annotation

Part-of-speech (POS) tagging is the most basic kind of linguistic annotation.

Each linguistic token is assigned a code indicating its *part of speech*, i.e., basic grammatical status.

Examples of POS information:

- singular common noun;
- comparative adjective;
- past participle.

POS tagging forms a basic first step in the disambiguation of homographs.

E.g., it distinguishes between the verb “boot” and the noun “boot”.

But it does not distiguish between “boot” meaning “kick” and “boot” as in “boot a computer”, both of which are transitive verbs.

Example POS tag sets

- CLAWS tag set (used for BNC): 62 tags;
- Brown tag set (used for Brown corpus): 87 tags:
- Penn tag set (used for the Penn Treebank): 45 tags.

Category	Examples	CLAWS	Brown	Penn
Adjective	happy, bad	AJ0	JJ	JJ
Adverb	often, badly	PNI	CD	CD
Determiner	this, each	DT0	DT	DT
Noun	aircraft, data	NN0	NN	NN
Noun singular	woman, book	NN1	NN	NN
Noun plural	women, books	NN2	NN	NN
Noun proper singular	London, Michael	NP0	NP	NNP
Noun proper plural	Australians, Methodists	NP0	NPS	NNPS

POS Tagging

Idea: Automate POS tagging: look up the POS of a word in a dictionary.

Problem: POS ambiguity: words can have several possible POS's; e.g.:

Time flies like an arrow. (1)

time: singular noun or a verb;
flies: plural noun or a verb;
like: singular noun, verb, preposition.

Combinatorial explosion: (1) can be assigned $\mathbf{2 \times 2 \times 3 = 12}$ different POS sequences.

Need to take sentential context into account to get POS right!

Probabilistic POS tagging

Observation: words can have more than one POS, but one of them is more frequent than the others.

Idea: assign each word its most frequent POS (get frequencies from manually annotated training data). Accuracy: around 90%.

State-of-the-art POS taggers take the context into account; often use Hidden Markov Models. Accuracy: 96–98%.

Example output from a POS tagger (not XML format!):

```
Our/PRP$ enemies/NNS are/VBP innovative/JJ and/CC
resourceful/JJ ,/, and/CC so/RB are/VB we/PRP ./. They/PRP
never/RB stop/VB thinking/VBG about/IN new/JJ ways/NNS
to/TO harm/VB our/PRP$ country/NN and/CC our/PRP$
people/NN, and/CC neither/DT do/VB we/PRP ./.
```

Use of markup languages

An important general application of markup languages, such as XML, is to separate *data* from *metadata*.

In a corpus, this serves to keep different types of information apart;

- *Data* is just the raw data.

  In a corpus this is the text itself.

- *Metadata* is data about the data.

  In a corpus this is the various annotations.

Nowadays, XML is the most widely used markup language for corpora.

The example on the next slide is taken from the BNC XML Edition, which was released only in 2007.
(The previous BNC World Edition was formatted in SGML.)

```
<wtext type="FICTION">
  <div level="1">
   <head> <s n="1">
     <w c5="NN1" hw="chapter" pos="SUBST">CHAPTER </w>
     <w c5="CRD" hw="1" pos="ADJ">1</w>
   </s> </head>
   <p> <s n="2">
     <c c5="PUQ">˜</c>
     <w c5="CJC" hw="but" pos="CONJ">But</w>
     <c c5="PUN">,</c> <c c5="PUQ">˜ </c>
     <w c5="VVD" hw="say" pos="VERB">said </w>
     <w c5="NP0" hw="owen" pos="SUBST">Owen</w>
     <c c5="PUN">,</c> <c c5="PUQ">˜</c>
     <w c5="AVQ" hw="where" pos="ADV">where </w>
     <w c5="VBZ" hw="be" pos="VERB">is </w>
     <w c5="AT0" hw="the" pos="ART">the </w>
     <w c5="NN1" hw="body" pos="SUBST">body</w>
     <c c5="PUN">?</c> <c c5="PUQ">˜</c>
   </s> </p>
   ....
  </div>
</wtext>
```

Aspects of this example

The example is the opening text of J10, a novel by Michael Pearce.

Some aspects of the tagging:

- The **`wtext`** element stands for *written text*. The attribute **`type`** indicates the genre.
- The **`head`** element tags a portion of header text (in this case a chapter heading).
- The **`s`** element tags sentences. (N.B., a chapter heading counts as a sentence.) Sentences are numbered via the attribute **`n`**.
- The **`w`** element tags words. The attribute **`pos`** is a POS tag, with more detailed POS information given by the **`c5`** attribute, which contains the CLAWS code. The attribute **`hw`** represents the *root form* of the word (e.g., the root form of “said” is “say”).
- The **`c`** element tags punctuation.

Syntactic annotation (parsing)

Syntactic annotation: information about the structure of sentences. Prerequisite for computing meaning.

Linguists use phrase markers to indicates which parts of a sentence belong together:

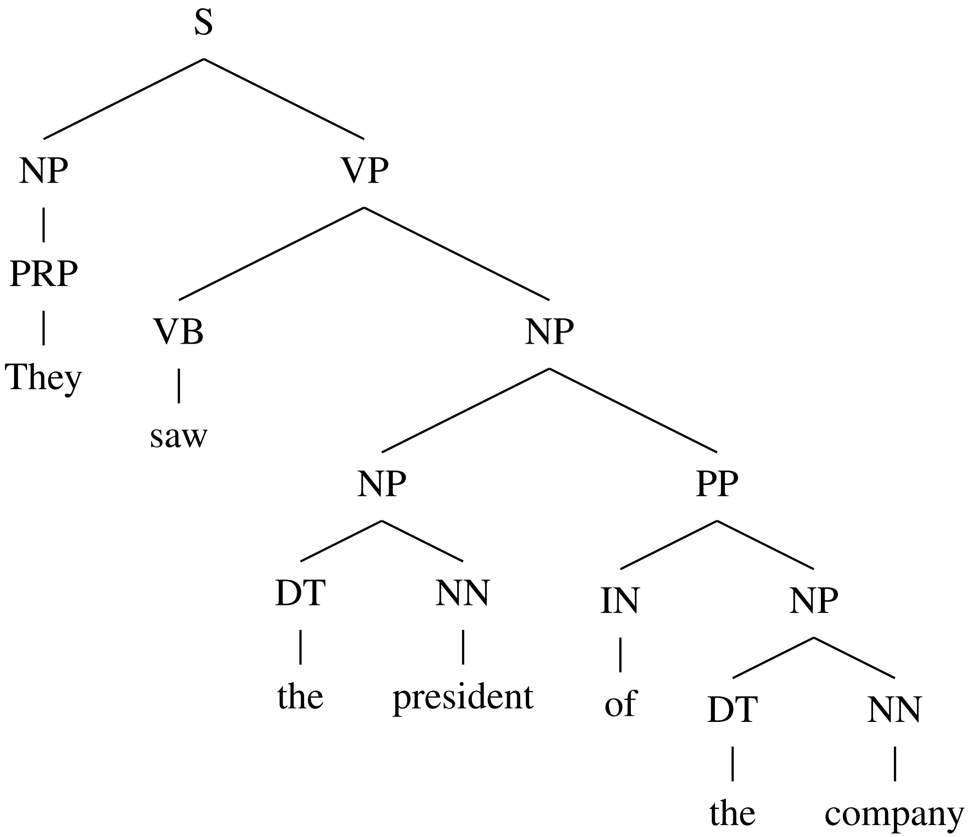
- noun phrase (NP): noun and its adjectives, determiners, etc.
- verb phrase (VP): verb and its objects;
- prepositional phrase (PP): preposition and its NP;
- sentence (S): VP and its subject.

Phrase markers group hierarchically in a *syntax tree*.

Syntactic annotation can be automated. Accuracy: around 90%.

Example syntax tree

Sentence from the Penn Treebank corpus:



The same syntax tree in XML:

```
<s>
  <np><w pos="PRP">They</w></np>
  <vp><w pos="VB">saw</w>
    <np>
      <np><w pos="DT">the</w> <w pos="NN">president</w></np>
      <pp><w pos="NN">of</w>
        <np><w pos="DT">the</w> <w pos="NN">company</w></np>
      </pp>
    </np>
  </vp>
</s>
```

Note the conventions used in the above document: phrase markers are represented as elements; whereas POS tags are given as attribute values.

N.B. The tree on the previous slide is *not* the XML element tree generated by this document.

Other Types of Annotation

- Edited text is comparatively easy to annotate;
- unscripted dialog is much harder (hesitations, false starts, slips of the tongue, cross talk);
- example for a corpus of unscripted dialog: HCRC MapTask corpus;
- rich annotation: dialog moves, disfluencies, gaze, parts of speech, syntax;
- we could also annotate prosodic structure, named entities, co-references, etc.

Background reading

Corpus Linguistics
Tony McEnery & Andrew Wilson
Edinburgh University Press,
2nd Edition, 2001

Chapter 2: What is a Corpus and What is in It?
Section 2.2.2.

Copies of this chapter are still available from the shelves outside room 5.03 Appleton Tower.