Informatics 1B: Data and Analysis

Lecture 7: Semi-structured Data: Acquisition and Annotation

Frank Keller

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

February 3, 2005

1 Data Acquisition and Annotation
- Balancing and Sampling
- Pre-processing
- Markup Languages
- Corpus Annotation

Reading: lecture notes; McEnery and Wilson (2001: ch. 2).

Corpora

Last lecture: we defined a corpus as a collection of textual or spoken data meeting the following criteria:

- sampled in a certain way;
- finite in size;
- available in machine-readable form;
- often serves 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 books, newspapers, magazines, letters, etc.

Sampling ensures that the material is representative of the source.

Example

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

Language Types, Genres, Domains, Media

Things to take into account when balancing:

- *language type:*
  - edited text (e.g., articles, books, newswire);
  - spontaneous text (e.g., email, Usenet news);
  - 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);
- *media:* physical realization of a corpus (e.g., text, audio, transcribed speech, video).

Examples for Balanced Corpora

Brown Corpus: balanced corpus of written American English:

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

Genres and Domains in the Brown Corpus

Genre	Domain	Size
Press	Reportage	44 texts
Press	Editorial	27 texts
Press	Reviews	17 texts
–	Religion	17 texts
–	Skill and hobbies	36 texts
–	Popular lore	48 texts
–	Belles-lettres	75 texts
Miscellaneous	Government and house organs	30 texts
–	Learned	80 texts
Fiction	General	29 texts
Fiction	Mystery	24 texts
Fiction	Science	6 texts
Fiction	Adventure	29 texts
Fiction	Romance	29 texts
–	Humor	9 texts

Examples for Balanced Corpora

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.

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	dialog	speech	British English
CallHome Corpus	50k	dialog	speech	6 languages

Pre-processing

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;
  - dialog 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
- cul de sac, Zeitgeist

Sentence Boundary Detection

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

Sentence: string of words ending in a period, question mark or exclamation mark. This is correct 90% of the time.

Example: potentially difficult cases

- Dr. Foster went to Glasgow.
- He said "rubbish!".
- He lost cash on lastminute.com.

Detection of word and sentence boundary particularly difficult for *spoken data.*

Markup Languages

Markup languages are used to:

- keep different types of information in a corpus apart;
- separate data from *metadata* (i.e., data about the data);
- for corpora: separate words (data) and annotation (metadata).

Most widely used markup language: *XML* (Extensible Markup Language), see lecture 5.

Related standards: SGML and HTML.

Example from the BNC

```
<head type=MAIN>
<s n="233"><w NN2>Inspectors <w PRF>of <w NN2>schools <c PUQ>&bquo;
<w AV0>poorly <w VVN>equipped <w PRP>for <w NN1>curriculum
<c PUQ>&equo;
</head>
<head type=BYLINE>
<s n="234"><w PRP>By <w NP0>PETER <w NP0>WILBY
</head>
<p>
<s n="235"><w NN1>SCHOOL <w NN2>INSPECTORS <w VVN>employed
<w PRP>by <w AJ0>local <w NN2>authorities <w VBB>are <w AV0>poorly
<w VVN>equipped <w PRP>for <w DPS>their <w AJ0-NN1>future
<w NN1>role <w PRF>of <w VVG>monitoring <w AT0>the <w AJ0>national
<w NN1>curriculum<c PUN>, <w AT0>a <w NN1>report <w PRP>from
<w AT0> the <w NN1>Audit <w NN1>Commission <w PRP>for
<w NN1-AJ0>Local  <w NN2>Authorities <w VVZ>says <w AV0>today
<c PUN>.
</p>
```


XML

Entity tags:

- denote elements of the text, such as `&bquo;` (opening quotation mark);
- keep the content of the entity independent of its rendering (e.g., `&bquo;` rendered as " or ‘‘);
- encode non-standard characters (e.g., `&uuml;` represents ü)

Markup tags: encode the metadata proper; examples:

- `<head> ... </head>`: header, with argument `type`;
- `<p> ... </p>`: paragraph;
- `<s> ... </s>`: sentence, with argument `n`;
- `<w>`: word, with argument part of speech.

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.

Example: part of speech tag sets

1. CLAWS tag (used for BNC); 62 tags;
2. Brown tag (used for Brown corpus); 87 tags;
3. Penn tag set (used for the Penn Treebank); 45 tags.

Example: POS Tag Sets for English

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:* many words can have several POSs:

(1) Time flies like an arrow.

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

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

Need to take *sentential context* into account to get POS right.

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 of POS tagger

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 ./.

Syntactic Annotation

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:

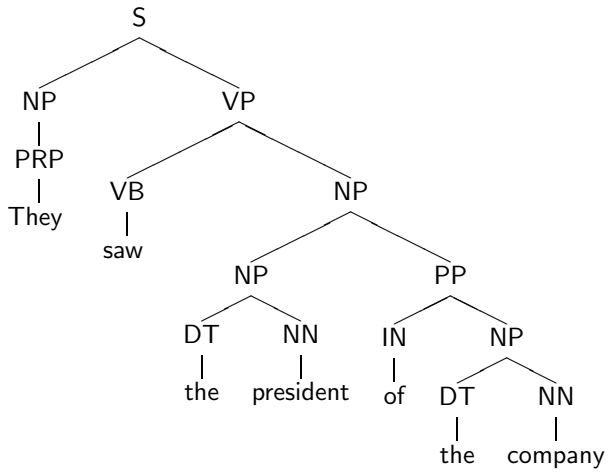
- verb phrase (VP): verb and its objects;
- noun phrase (NP): noun and its adjectives, determiners, etc.
- prepositional phrase (PP): preposition and its NP;
- sentence (S): VP and its subject.

Phrase markers group hierarchically in a *syntax trees.*

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

Example: syntax tree

Sentence from the Penn Treebank corpus:



Example: syntax tree

Same Tree in XML:

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


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 (see lecture notes);
- rich annotation: dialog moves, disfluencies, gaze, parts of speech, syntax;
- we could also annotate prosodic structure, named entities, co-references, etc.

Summary

- Balancing and sampling ensure that corpora are representative;
- corpora can be classified according to language types, genre, domain, and media;
- raw data needs to be pre-processed using tokenization and sentence-boundary detection;
- markup languages such as XML keep data and metadata (annotation) apart;
- annotation adds task-specific information to a corpus:
  - parts of speech;
  - syntactic structure;
  - dialog structure, prosody, named entities, etc.

References

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction*. Edinburgh University Press, Edinburgh, 2 edition.