Informatics 1B: Data and Analysis

Lecture 6: Semi-structured Data: Basic Concepts

Frank Keller, with material added by Helen Pain

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

February 15, 2006

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

From Structured to Semi-structured Data

Structured data:

- properties are know at the outset;
- fixed ways of describing, acquiring, and processing the data can be defined;
- formalized using entity/relationship model or relational model.

Basic Concepts
Corpus Data
Questions Corpora Can Answer
Obtaining Corpus Counts
Building Applications Using Corpora

From Structured to Semi-structured Data

Structured data:

- properties are know at the outset;
- fixed ways of describing, acquiring, and processing the data can be defined;
- formalized using entity/relationship model or relational model.

Unstructured data:

- properties not known at the outset;
- hard to define fixed set of entities and relationships;
- example: bacteria in a soil sample: species and their relationships and interactions not known in advance.

From Structured to Semi-structured Data

Semi-structured data:

- unstructured data enriched with *annotation;*
- certain parts of the data are assigned predefined *labels;*
- annotation heavily dependent on specific application.

From Structured to Semi-structured Data

Semi-structured data:

- unstructured data enriched with *annotation;*
- certain parts of the data are assigned predefined *labels;*
- annotation heavily dependent on specific application.

Example: Text

- no predefined structure (trivial: one word precedes the other);
- does not make sense to enter text into a database;
- becomes semi-structured by annotation: e.g., label all verbs; label all proper names.

Corpora as Semi-structured Data

Corpus data are a form of semi-structured data.

A corpus is a collection of textual or spoken data meeting the following criteria:

Corpora as Semi-structured Data

Corpus data are a form of semi-structured data.

A corpus is 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.

Corpora as Semi-structured Data

Corpus data are a form of semi-structured data.

A corpus is 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.

Example

Is a single email a corpus? Is a collection of 1,000 emails a corpus? Is the novel *Harry Potter* a corpus?

Automated humour

JAPE (Binsted and Ritchie 1994, 1997) is capable of producing punning riddles such as:
What's the difference between leaves and a car?
One you brush and rake, the other you rush and brake.
What do you get when you cross a monkey and a peach?
An ape-ricot.
What do you call a murderer with fibre?
A cereal killer.

It searches a general purpose dictionary to find words that fit pre-defined structures called schemas and templates.

Will all the words be understood? How can you tell?

What kind of vegetable can jump? *A spring onion.*
What do you get when you cross cars and sandwiches? *Traffic Jam*
How does a whale cry? *Blubber blubber.*
How is a car like an elephant? *They both have trunks.*
What do you get when you cross a vitellus and a saddlery? *A yolk yoke.*

Harry Potter and the order of the Phoenix by J.K.Rowling

"The Headmaster has sent me to tell you, Potter, that it is his wish for you to study Occlumency this term."
"Study what?" said Harry blankly.
Snape's sneer became more pronounced.
"Occlumency, Potter....

Revolting Rhymes by Roald Dahl

Ah, Piglet, you must never trust
Young ladies from the upper crust.
For now, Miss Riding Hood, one notes,
Not only has two wolfskin coats,
But when she goes from place to place,
She has a PIGSKIN TRAVELLING CASE.

Goal! by Colin McNaughton

He dribbles past the great Gigs, swerves past the magical Owen, sweeps past Shearer, puts the ball through the legs of Beckham and shoots....

Extract from a Corpus

> "And the ark rested in the seventh month, on the seventh day of the month upon the mountains of Ararat." Noah's ark did not come to rest on Mount Ararat (Massis in Armenian) but on the mountains of Ararat or Armenia. It is very strongly believed that the first five books (Pentateuch) of the Bible were written by Moses and another Moses of Khorene stated in his history that Ararat was the central portion of Armenia.

Where is this excerpt from?

Extract from a Another Corpus

> David Beckham is lending his voice to Vodafone's official voice-mail service, the first celebrity ever to allow such an extraordinary endorsement. Subscribers to the mobile phone company can now have a recording of Beckam's squeaky voice informing callers that "This is the voicemail service for X. Please leave a message after the tone". The move is part of a multi-million pound deal with the Manchester United star and opens up a whole new line of business for celebrities seeking extra cash while they are still in the limelight.

Where is this excerpt from?

Questions Corpora Can Answer

Empirical questions in linguistics and cognitive science:

- corpora can be analyzed using statistical tools;
- hypotheses about language processing and language acquisition can be tested;
- new facts about language structure can be discovered.

Questions Corpora Can Answer

Empirical questions in linguistics and cognitive science:

- corpora can be analyzed using statistical tools;
- hypotheses about language processing and language acquisition can be tested;
- new facts about language structure can be discovered.

Engineering questions in AI and computer science:

- corpora represent the data that language processing system have to handle;
- algorithms exist to extract regularities from corpus data;
- text-based or speech-based computer applications can learn automatically from corpus data.

Questions Corpora Can Answer

Example

Assume we have a corpus that consists of the Sherlock Holmes story *A Case of Identity*. Simple questions we could ask are:

1. Find all lines containing the word *Holmes*.

Questions Corpora Can Answer

Example

Assume we have a corpus that consists of the Sherlock Holmes story *A Case of Identity*. Simple questions we could ask are:

1. Find all lines containing the word *Holmes*.
2. Find all lines beginning with the word *Holmes*.

Questions Corpora Can Answer

Example

Assume we have a corpus that consists of the Sherlock Holmes story *A Case of Identity*. Simple questions we could ask are:

1. Find all lines containing the word *Holmes*.
2. Find all lines beginning with the word *Holmes*.
3. Find all lines starting with an upper case letter.

Extract from *A Case of Identity*

Find all lines containing the word *Holmes*.

> "My dear fellow." said Sherlock **Holmes** as we sat on either
> a realistic effect," remarked **Holmes**. "This is wanting in the
> said **Holmes**, taking the paper and glancing his eye down
> "I have seen those symptoms before," said **Holmes**, throwing
> merchant-man behind a tiny pilot boat. Sherlock **Holmes** welcomed
> "You've heard about me, Mr. **Holmes**," she cried, "else how

Extract from *A Case of Identity*

Find all lines beginning with the word *Holmes*.

Holmes, when she married again so soon after father's death,
Holmes alone, however, half asleep, with his long, thin form
Holmes. "He has written to me to say that he would be here at
Holmes had been talking, and he rose from his chair now with a

Extract from *A Case of Identity*

Find all lines starting with an upper case letter.

A Case of Identity
The husband was a teetotaler, there was no other woman
Take a pinch of snuff, Doctor, and acknowledge that I
The larger crimes are apt to be the simpler, for the
And yet even here we may discriminate.
When a woman has a secret
Etherege, whose husband you found so easy when the

Frequencies

Frequency information obtained from corpora is often useful for answering scientific or engineering questions.

Frequencies

Frequency information obtained from corpora is often useful for answering scientific or engineering questions.

Token count N: number of tokens (words, punctuation marks, etc.) in a corpus (i.e., size of the corpus).

Frequencies

Frequency information obtained from corpora is often useful for answering scientific or engineering questions.

Token count N: number of tokens (words, punctuation marks, etc.) in a corpus (i.e., size of the corpus).

Type count: number of different tokens in a corpus.

Frequencies

Frequency information obtained from corpora is often useful for answering scientific or engineering questions.

Token count N: number of tokens (words, punctuation marks, etc.) in a corpus (i.e., size of the corpus).

Type count: number of different tokens in a corpus.

Absolute frequency $f(t)$ of a type t: number of tokens of t in a corpus.

Frequencies

Frequency information obtained from corpora is often useful for answering scientific or engineering questions.

Token count N: number of tokens (words, punctuation marks, etc.) in a corpus (i.e., size of the corpus).

Type count: number of different tokens in a corpus.

Absolute frequency $f(t)$ of a type t: number of tokens of t in a corpus.

Relative frequency of a type t: absolute frequency of t normalized by the token count, i.e., $f(t)/N$.

Frequencies

Example

The British National Corpus (BNC) is an important reference. Let's compare some counts from the BNC with counts from our sample corpus *A Case of Identity*.

Frequencies

Example

The British National Corpus (BNC) is an important reference. Let's compare some counts from the BNC with counts from our sample corpus *A Case of Identity*.

	BNC	A Case of Identity
Token count N	100,000,000	7,006
Type count	636,397	1,621
$f(Holmes)$	890	46
$f(Sherlock)$	209	7
$f(Holmes)/N$	.0000089	.0066
$f(Sherlock)/N$	.00000209	.000999

Unigrams

We can now ask questions such as: what are the most frequent words in a corpus?

Unigrams

We can now ask questions such as: what are the most frequent words in a corpus?

- Count absolute frequencies of all word types in the corpus;
- tabulate them in an ordered list;
- results: list of *unigram* frequencies (frequencies of individual words).

Unigrams

We can now ask questions such as: what are the most frequent words in a corpus?

- Count absolute frequencies of all word types in the corpus;
- tabulate them in an ordered list;
- results: list of *unigram* frequencies (frequencies of individual words).

Example

Compare the unigram frequencies for BNC and *A Case of Identity*. Notice: article *the* most frequent word in both corpora; prepositions like *of* and *to* appear in both lists, etc.

Unigrams

Example

BNC		A Case of Identity	
6184914	the	350	the
3997762	be	212	and
2941372	of	189	to
2125397	a	167	of
1812161	in	163	a
1372253	have	158	I
1088577	it	132	that
917292	to	117	it

n-grams

The notion of unigram can be generalized:

- *bigram* frequencies (pairs of words);
- *trigram* frequencies (triples of words);
- *n-gram* frequencies (*n*-tuples of words).

n-grams

The notion of unigram can be generalized:

- *bigram* frequencies (pairs of words);
- *trigram* frequencies (triples of words);
- *n-gram* frequencies (n-tuples of words).

Example

Compute the most frequent n-grams in *A Case of Identity*, for $n = 2 \ldots 4$. Notice: The larger the n, the more linguistically meaningful the units.

n-grams

Example

Bigrams		Trigrams		4-grams	
40	of the	5	there was no	2	very morning of the
23	in the	5	Mr. Hosmer Angel	2	use of the money
21	to the	4	to say that	2	the very morning of
21	that I	4	that it was	2	the use of the
20	at the	4	that it is	2	the King of Bohemia

n-grams

Example

Bigrams		Trigrams		4-grams	
40	of the	5	there was no	2	very morning of the
23	in the	5	Mr. Hosmer Angel	2	use of the money
21	to the	4	to say that	2	the very morning of
21	that I	4	that it was	2	the use of the
20	at the	4	that it is	2	the King of Bohemia

Notice: *n*-gram frequencies get smaller with increasing *n*: more possible word combinations, increased *data sparseness.*

Natural Language Processing

Corpora used extensively in two areas of informatics:

- *natural language processing* (NLP) builds computer systems that understand or produce text;
- *speech processing* develops systems that understand or produce spoken language.

Natural Language Processing

Corpora used extensively in two areas of informatics:

- *natural language processing* (NLP) builds computer systems that understand or produce text;
- *speech processing* develops systems that understand or produce spoken language.

Rely on probability theory, information theory, machine learning to develop algorithms that extract statistical regularities from corpora.

Natural Language Processing

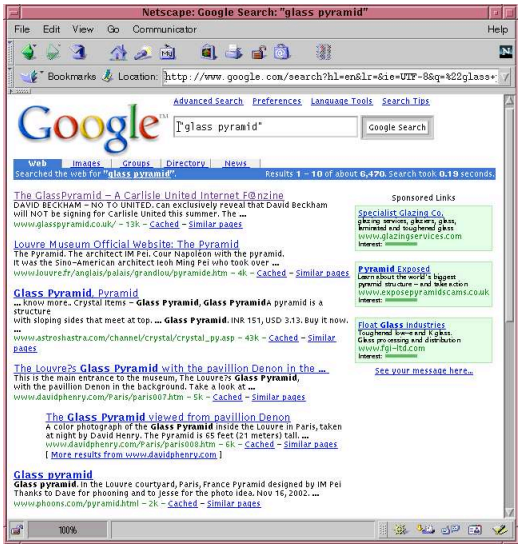
Corpora used extensively in two areas of informatics:

- *natural language processing* (NLP) builds computer systems that understand or produce text;
- *speech processing* develops systems that understand or produce spoken language.

Rely on probability theory, information theory, machine learning to develop algorithms that extract statistical regularities from corpora.

What are examples for NLP applications that use corpus data?

NLP Applications



NLP Applications

NLP applications often rely heavily on corpus data. Typically applications include:

- *Information retrieval (IR):* retrieve information from document collections. Examples: Google.

NLP Applications

NLP applications often rely heavily on corpus data. Typically applications include:

- *Information retrieval (IR):* retrieve information from document collections. Examples: Google.
- *Summarization:* take a text and compress it, i.e., produce an abstract or summary. Example: Newsblaster.

NLP Applications

NLP applications often rely heavily on corpus data. Typically applications include:

- *Information retrieval (IR):* retrieve information from document collections. Examples: Google.
- *Summarization:* take a text and compress it, i.e., produce an abstract or summary. Example: Newsblaster.
- *Machine Translation (MT):* take a text in a source language and turn it into a text in the target language. Example: Babelfish. http://world.altavista.com/

Machine Translation Example: Original

> Charlie Cook's favourite book has many tales to tell! Lose yourself in the wonderful adventures that unfold as Charlie discovers the hidden worlds in everyone's favourite book. From pirates to aliens to ghosts, there's something for everyone in Charlie's amazing book! A brand new picture book from the award-winning and hugely popular team behind THE GRUFFALO.

English Original

Machine Translation Example: English to French

> Le livre du favori du cuisinier de Charlie a beaucoup de contes à le dire ! Perdez-vous dans les aventures merveilleuses qui dévoilent pendant que Charlie découvre les mondes cachés dans chacun livre de favori. Des pirates aux étrangers aux fantômes, il y a quelque chose pour chacun en livre étonnant de Charlie ! Un livre nouveau d'image de l'équipe award-winning et énormement populaire derrière LE GRUFFALO.

Babelfish French translation

Machine Translation Example: French back to English

> The book of the favourite of the cook of Charlie has many tales to say it! You in the marvellous adventures lose which reveal while Charlie discovers the worlds hidden in each one delivers of favourite. Pirates the abroads with the phantoms, there is something for each one in astonishing book of Charlie! A book new of image of the award-winning team and énormement popular behind the GRUFFALO.

Translated back to English....

Summary

- Structured vs. semi-structured data;
- corpora as a form of semi-structured data;
- created using balancing and sampling;
- can be used answer scientific question or solve engineering problems;
- typical NLP/speech applications:
  - information retrieval;
  - summarization;
  - machine translation.
- type/token distinction; absolute/relative frequency;
- *n*-grams frequencies become sparser with increasing *n*.

References

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