Text is fun: Statistical exploration of large corpora

Siva Reddy
Lexical Computing Ltd, UK
http://sketchengine.co.uk

Centre for Exact Humanities (CEH)
IIIT Hyderabad
May 14 2012
Acknowledgments

Adam Kilgarriff

Michael Rundell
What is “meaning”?

- **Semantics**: Study of meaning in language.
- **Lexical semantics**: Study of meaning of words.

**Observation**

**Introspection**
What is “meaning”?  

- **Semantics**: Study of meaning in language.  
- **Lexical semantics**: Study of meaning of words.  

**Observation**  

**Introspection**
What is “meaning”?  

- **Semantics**: Study of meaning in language.  
- **Lexical semantics**: Study of meaning of words.
What is “meaning”?

- **Semantics**: Study of meaning in language.
- **Lexical semantics**: Study of meaning of words.
How are dictionaries built in pre-computer era?

James Murray and colleagues: Oxford English Dictionary
How are dictionaries built in pre-computer era?

Storage of Evidences
How are dictionaries built in pre-computer era?

Indexing
Revolution: Internet Era

corpus: a sample

dictionary: a distillation

language in use: speech, writing
Dictionary building: Requirements

- Corpus (Text) Collection
- Wordlist
- Evidence collection: Words in action.
- Word Profiles
Dictionary building: Requirements

- Corpus (Text) Collection
- Wordlist
- Evidence collection: Words in action.
- Word Profiles
Web as Corpus: Challenges

- Crawling
- Text extraction
- Spamming
- Duplication

Exercise 1: WebBootCaT

Collect corpus from web on a topic of interest.
(Baroni et al., 2006; Kilgarriff et al., 2010)
Web as Corpus: Challenges

- Crawling
- Text extraction
- Spamming
- Duplication

**Exercise 1: WebBootCaT**

Collect corpus from web on a topic of interest.
(Baroni et al., 2006; Kilgarriff et al., 2010)
Generalized dictionary

Domain-specific dictionary

Exercise 2: Keyword Extraction

Collect keywords from the corpus you collected above.
Wordlist

- Generalized dictionary
- Domain-specific dictionary

Exercise 2: Keyword Extraction

Collect keywords from the corpus you collected above.
Evidence collection

- Words in action
- Google like searching isn’t enough
- Get all the word forms of test?
- Words which are at a distance of three from test?
- Corpus Query Language: regular expressions
Regular expressions

Regular Expression Table:

Exercise 3: Write regular expressions for . . .
http://sketchengine.co.uk/exercises/regex/
Corpus Processing: Challenges

What are the noun forms of the word test?
Will "test.*" work?
Word Tokenization
Morphological analysis
Part-of-Speech Tagging
CQL: [lemma="treat" & tag="N.*"]
Corpus Processing: Challenges

- What are the noun forms of the word test?
- Will "test.*" work?
- Word Tokenization
- Morphological analysis
- Part-of-Speech Tagging
- CQL: [lemma="treat" & tag="N.*"]
Collocations (word associations)

- When do you say a word A is important to word B?
  - mouse: laser
  - mouse: food

Exercise 4: Collocations of the words girl and boy?

Download data from http://sivareddy.in/textisfun.tgz

Rank context words using mutual information:\(^a\frac{P(x,y)}{P(x)P(y)}\)

\(^a\)Removed log for simplicity
Collocations (word associations)

- When do you say a word A is important to word B?
- mouse: laser
- mouse: food

Exercise 4: Collocations of the words girl and boy?

Download data from http://sivareddy.in/textisfun.tgz

Rank context words using mutual information\(^a\): \(\frac{P(x, y)}{P(x)P(y)}\)

\(^a\)Removed log for simplicity
Word Sketch - a profile describing collocations

- The voice of the majority
- Sketch Grammar: describes the frequent constructions of words in language

Exercise 5: Objects of eat-v?
Write the Sketch Grammar capturing object relation?
My near-dream for Indian languages?

- Writing Sketch Grammar is not so time-taking.
- Exploit Sketch Grammar to build Syntactic Parser
- A parser for every language
- Cash the similarities between different languages
When do you say two words are similar?

Distributional Hypothesis (Harris, 1954)
- The words that occur in similar contexts tend to have similar meaning
- e.g: laptop, computer
- Backbone for *Vector Space Model of Semantics*.

Firth (Firth, 1957)
- You shall know a person from his friends - Chinese Proverb
- You shall know a word from its context - Firth’s Principle

Bag of words hypothesis
Two documents tend to be similar if they have similar distribution of similar words
When do you say two words are similar?

Distributional Hypothesis (Harris, 1954)
- The words that occur in similar contexts tend to have similar meaning
- e.g: laptop, computer
- Backbone for *Vector Space Model of Semantics*.

Firth (Firth, 1957)
- You shall know a person from his friends - Chinese Proverb
- You shall know a word from its context - Firth’s Principle

Bag of words hypothesis
Two documents tend to be similar if they have similar distribution of similar words
Vector Space Models (VSMs) of Semantics

- **Interpret semantics using VSM**
  - Backbone: Distributional Hypothesis
  - Text entity (we are interested in) as a Vector (point) in dimensional space.
  - Context of the entity as dimensions
  - Existing methods represent knowledge in VSMs mainly in three types (Turney and Pantel, 2010)
    - term-document
    - term-context
    - pair-pattern
Vector Space Models (VSMs) of Semantics

- **Interpret semantics using VSM**
  - Backbone: Distributional Hypothesis
- **Text entity (we are interested in) as a Vector (point) in dimensional space.**
- **Context of the entity as dimensions**
  - Existing methods represent knowledge in VSMs mainly in three types (Turney and Pantel, 2010)
    - term-document
    - term-context
    - pair-pattern
Vector Space Models (VSMs) of Semantics

- **Interpret semantics using VSM**
  - Backbone: Distributional Hypothesis
  - Text entity (we are interested in) as a Vector (point) in dimensional space.
  - Context of the entity as dimensions
  - Existing methods represent knowledge in VSMs mainly in three types (Turney and Pantel, 2010)
    - term-document
    - term-context
    - pair-pattern
Create a word-by-document matrix

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
<th>d5</th>
<th>d6</th>
<th>d7</th>
<th>d8</th>
<th>d9</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>interface</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>computer</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>system</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>response</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPS</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>survey</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>trees</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>graph</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>minors</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

d1: Human machine **interface** for Lab ABC **computer** applications

---

1 Image courtesy: (Landauer et al., 1998)
Document similarity can be found using Cosine similarity

\[ \text{sim}(D_1, D_2) = \frac{D_1 \cdot D_2}{\|D_1\| \|D_2\|} \]

\(^2\)Image courtesy: (Salton et al., 1975)
Document similarity can be found using Cosine similarity

\[ \text{sim}(D_1, D_2) = \frac{D_1 \cdot D_2}{\|D_1\| \|D_2\|} \]
Term-Context: Word Space Model

Meaning of a word as a vector (Schütze, 1998)

Meaning of a word is represented as a cooccurrence vector built from a corpus

<table>
<thead>
<tr>
<th></th>
<th>police-n</th>
<th>photon-n</th>
<th>speed-n</th>
<th>car-n</th>
<th>soul-n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic</td>
<td>142</td>
<td>0</td>
<td>293</td>
<td>347</td>
<td>1</td>
</tr>
<tr>
<td>Light</td>
<td>41</td>
<td>29</td>
<td>222</td>
<td>198</td>
<td>50</td>
</tr>
<tr>
<td>TrafficLight</td>
<td>5</td>
<td>0</td>
<td>13</td>
<td>48</td>
<td>0</td>
</tr>
</tbody>
</table>

Exercise 6: Compute similarity between car, bus, cycle

Hint: Represent words as vectors using above mutual information scores and compute Cosine similarity.
Term-Context: Word Space Model

Meaning of a word as a vector (Schütze, 1998)

Meaning of a word is represented as a cooccurrence vector built from a corpus

<table>
<thead>
<tr>
<th></th>
<th>police-n</th>
<th>photon-n</th>
<th>speed-n</th>
<th>car-n</th>
<th>soul-n</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Traffic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>142</td>
<td>0</td>
<td>293</td>
<td>347</td>
<td>1</td>
</tr>
<tr>
<td><strong>Light</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>41</td>
<td>29</td>
<td>222</td>
<td>198</td>
<td>50</td>
</tr>
<tr>
<td><strong>TrafficLight</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>13</td>
<td>48</td>
<td>0</td>
</tr>
</tbody>
</table>

Exercise 6: Compute similarity between car, bus, cycle

Hint: Represent words as vectors using above mutual information scores and compute Cosine similarity.
So far we represented a word with a single word sketch

- mouse vs mouse?
- Word Sense Disambiguation: collocations are the clue
- WordNet have been used extensively
- Can we guess the number of senses of a word?
Figure: Word Sense Induction in a Graph based setting
Semantic Word Sketches

Semantic Frames

Demo: http://corpdev.sketchengine.co.uk/run.cgi/first_form?corpname=5dcaa5fe

Exercise 7: abstract entities which modify boy and girl
Use word sense of context words as clue.
Beyond Words: Compositional Semantics

Given meanings of

- couch
- roast
- potato

Can we interpret the meanings of

- couch potato
- roast potato
Beyond Words: Compositional Semantics

Given meanings of
- couch
- roast
- potato

Can we interpret the meanings of
- couch potato
- roast potato
Roast Potato


