Natural Language Processing

Chao Lan
Introduction
Main Reference

And many examples are from: http://people.ischool.berkeley.edu/~dbamman/nlp18.html
What is natural language processing?

“... giving computers the ability to process human language ...”

What are the NLP applications?
An Open Challenge
High-level and Low-level Applications
Common NLP Tasks

- Natural Language Understanding
- Natural Language Generation
- Speech Recognition
- Machine Translation
- Speech Synthesis
- Meaning

Text

Speech
Ambiguity

“One morning I shot an elephant in my pajamas.”

*Animal Crackers* (1930)
“One morning I shot an elephant in my pajamas.”

*Animal Crackers (1930)*
“One morning I shot an elephant in my pajamas.”

Is part-of-speech verb or noun?
Ambiguity

“One morning I shot an elephant in my pajamas.”

Animal Crackers (1930)
Ambiguity: Another Example

"I made her duck."
Approaches to Reduce Ambiguity

Part-of-Speech Tagging
- assign POS to each word

Word Sense Disambiguation
- assign a sense to a word in a given context

<table>
<thead>
<tr>
<th>Category</th>
<th>Target Word in Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>FISH/INSECT</td>
<td>...fish as Pacific salmon and striped bass and...</td>
</tr>
<tr>
<td>FISH/INSECT</td>
<td>...produce filets of smoked bass or sturgeon...</td>
</tr>
<tr>
<td>MUSIC</td>
<td>...exciting jazz bass player since Ray Brown...</td>
</tr>
<tr>
<td>MUSIC</td>
<td>...play bass because he doesn’t have to solo...</td>
</tr>
</tbody>
</table>

1. bass - (the lowest part of the musical range)
2. bass, bass part - (the lowest part in polyphonic music)
3. bass, basso - (an adult male singer with the lowest voice)
4. sea bass, bass - (flesh of lean-fleshed saltwater fish of the family Serranidae)
5. freshwater bass, bass - (any of various North American lean-fleshed freshwater fishes especially of the genus Micropterus)
6. bass, bass voice, basso - (the lowest adult male singing voice)
Where are AI and ML?

**AI Search**
- search in state space

**Machine Learning**
- learn to tag information
Ethical Issues in NLP

@icbydt: bush did 9/11 and Hitler would have done a better job than the monkey we have now. Donald Trump is the only hope we’ve got.
Ethical Issues in NLP
Semantics derived automatically from language corpora contain human-like biases

Aylet Galischan1, Joanna J. Bryson1,2, Arvind Narayanan1

1 See all authors and affiliations

Science 14 Apr 2017
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aai4133

Ethical Issues in NLP
One Source of NLP Bias

<table>
<thead>
<tr>
<th></th>
<th>annotator A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>puppy</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>fried chicken</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

observed agreement = 11/16 = 68.75%
Data Representation
## Term-Document Matrix

<table>
<thead>
<tr>
<th></th>
<th>Hamlet</th>
<th>Macbeth</th>
<th>Romeo &amp; Juliet</th>
<th>Richard III</th>
<th>Julius Caesar</th>
<th>Tempest</th>
<th>Othello</th>
<th>King Lear</th>
</tr>
</thead>
<tbody>
<tr>
<td>knife</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>dog</td>
<td>2</td>
<td></td>
<td>6</td>
<td>6</td>
<td></td>
<td>2</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>sword</td>
<td>17</td>
<td>2</td>
<td>7</td>
<td>12</td>
<td></td>
<td>2</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>love</td>
<td>64</td>
<td></td>
<td>135</td>
<td>63</td>
<td></td>
<td>12</td>
<td></td>
<td>48</td>
</tr>
<tr>
<td>like</td>
<td>75</td>
<td>38</td>
<td>34</td>
<td>36</td>
<td>34</td>
<td>41</td>
<td>27</td>
<td>44</td>
</tr>
</tbody>
</table>
Feature Vectors

Feature Vector of Document

size of vocabulary

<table>
<thead>
<tr>
<th>Term</th>
<th>1</th>
<th>2</th>
<th>17</th>
<th>64</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamlet</td>
<td>1</td>
<td>2</td>
<td>17</td>
<td>64</td>
<td>75</td>
</tr>
<tr>
<td>King Lear</td>
<td>2</td>
<td>12</td>
<td>17</td>
<td>48</td>
<td>44</td>
</tr>
</tbody>
</table>

Feature Vector of Term

number of documents

<table>
<thead>
<tr>
<th>Term</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>7</th>
<th>12</th>
<th>2</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>knife</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>sword</td>
<td>17</td>
<td>2</td>
<td>7</td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>17</td>
</tr>
</tbody>
</table>
TF-IDF Vector

Not all words are equally important!

$$tfidf(t, d) = tf_{t,d} \times \log \frac{N}{D_t}$$

TF(t, d): frequency of term t in document d

N: # of documents

Dt: # documents that contain term t

Example

<table>
<thead>
<tr>
<th></th>
<th>Hamlet</th>
<th>Macbeth</th>
<th>Romeo &amp; Juliet</th>
<th>Richard III</th>
<th>Julius Caesar</th>
<th>Tempest</th>
<th>Othello</th>
<th>King Lear</th>
</tr>
</thead>
<tbody>
<tr>
<td>knife</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>2</td>
<td></td>
<td>6</td>
<td>6</td>
<td>2</td>
<td></td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>sword</td>
<td>17</td>
<td>2</td>
<td>7</td>
<td>12</td>
<td>2</td>
<td></td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>love</td>
<td>64</td>
<td>135</td>
<td>63</td>
<td>12</td>
<td></td>
<td></td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>like</td>
<td>75</td>
<td>38</td>
<td>34</td>
<td>36</td>
<td>34</td>
<td>41</td>
<td>27</td>
<td>44</td>
</tr>
</tbody>
</table>

IDF

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>knife</td>
<td>dog</td>
<td>sword</td>
<td>love</td>
<td>like</td>
</tr>
<tr>
<td>0.12</td>
<td></td>
<td>0.20</td>
<td>0.12</td>
<td>0.20</td>
<td>0</td>
</tr>
</tbody>
</table>
# Term-Term Matrix

<table>
<thead>
<tr>
<th></th>
<th>Hamlet</th>
<th>Macbeth</th>
<th>Romeo &amp; Juliet</th>
<th>Richard III</th>
<th>Julius Caesar</th>
<th>Tempest</th>
<th>Othello</th>
<th>King Lear</th>
</tr>
</thead>
<tbody>
<tr>
<td>knife</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sword</td>
<td>17</td>
<td>2</td>
<td>7</td>
<td>12</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>love</td>
<td>64</td>
<td>135</td>
<td>63</td>
<td>12</td>
<td>12</td>
<td>48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>like</td>
<td>75</td>
<td>38</td>
<td>34</td>
<td>34</td>
<td>41</td>
<td>27</td>
<td>44</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>knife</th>
<th>dog</th>
<th>sword</th>
<th>love</th>
<th>like</th>
</tr>
</thead>
<tbody>
<tr>
<td>knife</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>dog</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>sword</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>love</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>like</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>
Cosine Similarity between Vectors

\[ \cos(x, y) = \frac{\sum_{i=1}^{F} x_i y_i}{\sqrt{\sum_{i=1}^{F} x_i^2} \sqrt{\sum_{i=1}^{F} y_i^2}} \]
Cosine Distance = 1 - Cosine Similarity

\[
\text{CosineDistance}(\text{Hamlet}, \text{King Lear}) = 0
\]

\[
\text{EuclideanDistance}(\text{Hamlet}, \text{King Lear}) = 900.67
\]