Intro to Text Processing
Lecture 2

Behrang Mohit

Today: **Text**

• Different types of text
• Corpus
• Levels of text processing
  – Corpus
  – Document
  – Sentence
  – Tokens
  – Characters and Encoding
So Much Text!

- Goal: Organize and process

Different forms of text

- Structured text
  - Databases
  - Knowledge-base
- Semi-structured
  - Wikipedia info-boxes
  - Excel sheets, tables in documents
  - XML documents
- Unstructured
  - Plain text: News, Webpages, Emails, SMS, tweets, ...
From text to knowledge

Organizing the knowledge

• Text: non-structured information

Charlie was born in London in 1889 and ....
Born in 1889, Chaplin spent his youth in ....
Chaplin (1889-1977) was a true genius of cinema ....
He was born 6 years before the birth of cinema!
Organizing the knowledge

• **Text: non-structured information**

Charlie was born in London in 1889 and ....
Born in 1889, Chaplin spent his youth in ....
Chaplin (1889-1977) was a true genius of cinema ....
He was born 6 years before the birth of cinema!

• **Knowledge-base: structured information**

<table>
<thead>
<tr>
<th>Person</th>
<th>Birth date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charlie Chaplin</td>
<td>1889</td>
</tr>
</tbody>
</table>
Need for textual data

- All text processing systems need data:
  - Translation
  - Question Answering
  - Summarization
  - ...
  - ...

Corpus/Corpora

- No access to all texts
  - We use a representative subset

- Corpus: A collection of documents
  - Used for linguistic analysis
  - Used for building a text processing system
    - Human labeled data for statistical systems
  - Used for benchmarking/testing a system
    - Performance of a search engine
Characteristics of a corpus

- Written, speech
- General (balanced), domain specific (e.g. news)
- Monolingual, Bilingual, Comparable
- Size of corpus
  - Web as corpora > 2,000 billion words
- Copyright issues and license

Annotated Corpus

- Linguistics annotation: adding some linguistic knowledge to the plain-text
- Need for annotated data
  - System that separates sentences.
  - System that finds names
  - System that answers questions
  - System that translates
  - Search Engine
  - ....
Examples of corpora

• British National Corpus
  – 90% written, 10% spoken
  – 100M words, British English
  – Plain text with some linguistic annotations

• Brown Corpus
  – English, balanced, written American, 1M words
  – Annotated with syntax and semantics information

Brown corpus: Text + syntax

```
[ The/DT Fulton/NNP County/NNP Grand/NNP Jury/NNP ]
said/VBD
[ Friday/NNP ]
[ an/DT investigation/NN ]
of/IN
[ Atlanta/NNP 's/POS recent/JJ primary/JJ election/NN ]
produced/VBD `\'/
[ no/DT evidence/NN ]
```
```
'/ that/IN
[ any/DT irregularities/NNS ]
took/VBD
[ place/NN ]

/.
```
Named Entity Corpus

[LOC Shanghai ] ( [LOC China ] ) 10–4 ( [ORG AFP ] ) – [MISC Croatian ] [PER Goran Ivanisevic ] [ 27 years of age ] was seeded first at [LOC Shanghai ]'s [ORG Chinese International Tennis Tournament ] , which begins tomorrow, Monday, with prizes totaling 340 thousand dollars.

Parallel Corpus

Katara
Katara is a cultural center in Doha.
It aims to promote the cultural activities in Qatar and support the creativity through its various facilities.
A forum for all intellectuals and artists, and a center to promote cultural awareness through festivals, exhibitions, seminars, concerts, and all forms of artistic expression.
Linguistic Annotation

• Creating gold-standard data
  – Used for building and training a system
  – Used for evaluating a system
• Requirements:
  – Robust annotation software
  – Clear guidelines
  – Quality control

Desktop and web-based annotation

• Text and GUI-based desktop annotation
• New web-based annotations
  – Work any-time, anywhere
  – Direct storage
  – Browser and bandwidth limitations
Web-based annotation

CMUQ’s QALB Project

• QALB: Qatar Arabic Language Bank
  – Automatic correction of Arabic text
    • Spelling, grammar, word-choice
    • Text: Natives (news comments), Learners (essays), MT
    • Target: A corpus of 2 million words
    • 10 annotators
  – Demo: The annotation interface
Quality Control

- Corpus needs to be an accurate language sample
- Quality of annotated corpus:
  - Human annotator’s agreement → high quality corpus

- Framework:
  - Randomly assign parts of the task to multiple annotators
  - Measure their agreement
    - Agreement measures

New Models: Crowd Sourcing

- Large scale, simple and cheap annotations
  - Quick and mass annotation
- Amazon Mechanical Turk
  - Cost: 1-10 cents/tasks
  - Frameworks to assess the annotators
    - Before and during the work
  - Limits: Quality and the range of expertise
Document structure

- Plain text
- Markups
  - SGML: Standard Generalized Markup language
  - HTML: Hyper text Mark-up Language
  - XML: Extensible Mark-up Language
  - Wikipedia
- Prints

Plain Text

CHAPTER 38

READER, I married him. A quiet wedding we had: he and I, the parson and clerk, were alone present. When we got back from church, I went into the kitchen of the manor-house, where Mary was cooking the dinner, and John cleaning the knives, and I said –

'Mary, I have been married to Mr Rochester this morning.' The housekeeper and her husband were of that decent, phlegmatic order of people, to whom one may at any time safely communicate a remarkable piece of news without incurring the danger of having one's ears pierced by some shrill ejaculation and subsequently stunned by a torrent of wordy wonderment. Mary did look up, and she did stare at me; the ladle with which she was basting a pair of chickens roasting at the fire, did for some three minutes hang suspended in air, and for the same space of time John's knives also had rest from the polishing process; but Mary, bending again over the roast, said only --
Reader, I married him. A quiet wedding we had: he and I, the parson and clerk, were alone present. When we got back from church, I went into the kitchen of the manor-house, where Mary was cooking the dinner, and John cleaning the knives, and I said &mdash; <p>

Mary, I have been married to Mr Rochester this morning.&lt;/q&gt; The housekeeper and her husband were of that decent, phlegmatic order of people, to whom one may at any time safely communicate a remarkable piece of news without incurring the danger of having one’s ears pierced by some shrill ejaculation and subsequently stunned by a torrent of wordy wonderment. Mary did look up, … &lt;/p&gt;
Today: **Text**

- Different types of text
- Corpus

Levels of text processing

- Corpus
- Document
- Sentence
- Tokens
- Characters and Encoding

**Sentences**

- Group of words which ...
  - Usually split by a full-stop, exclamation/question mark.
  - Many sentences have a verb

- Some languages have a word order:
  - English: Subject-Verb-Object (SVO)
  - Arabic: Verb Subject Object, SVO
Sentence splitting

• Sentence level processing ➔ Sentence boundary detection
• Baseline: split at ./!/?
• But:
  – The class was thought by Dr. Ali Eqbal and Prof. J. Patrick Jr.
  – P.O.Box, I.R. Pakistan, U.S.A.
• Use rules with list of acronyms

Statistical Sentence Splitting

• Collect different information (features) from the surrounding words:
  – Lexical category of the preceding/following words
  – Case of the preceding/following words
  – Is the preceding word an abbreviation
Sentence alignment

- Parallel corpus is ideal
  - Reality: Comparable corpus
    - Translators reorganize the text, creating new sentences.
  - Extremely non-aligned:
    - News articles in different languages
    - Wikipedia pages in different languages

Today: Text

 ✓ Different types of text
 ✓ Corpus
 Levels of text processing
   ✓ Corpus
   ✓ Document
   ✓ Sentence
   - Tokens
   - Characters and Encoding
Words vs. tokens

- Difference mainly in languages where white space does not necessarily mean

- Graphical words: token
  - What we see on the screen or in the book
    Her email is mariem@yahoo.com
    This is a half-cooked idea.

- Underlying word
  - Deeper linguistic
    - Half, cooked

Types

- How many tokens in a corpus?
  - The “wc” command

- How many token types in a corpus?
  - Number of unique tokens
  - Capitalization is not relevant (The = the)
Counting words and types

• *The chair of the session starts to chair the session*
  – Tokens?
  – Types?

Tokenization/Segmentation

“In Baghdad, Kirkuk (Kurdistan) and, Basra you hear different views.”

“ In Baghdad , Kirkuk ( Kurdistan ) and , Basra you hear different views . ”

• An important step for accurate text processing
  – Search Engine: looking for “Baghdad”
    • Not “Baghdad,”
  – Translation
  – ...

10/26/13
Tokenization

- Finding the word boundaries
  - Simple for Roman languages
    - White space and some rules
- Statistical/probabilistic approaches
- Not always easy for Asian languages

Word vs. Lemma

- Lemma is the collapsed form of a word
  - Goes, went, gone ➔ go
  - Relies, relied, relying ➔ rely
  - Boxes ➔ Box

- Lemmatization helps improving some text processing application
  - Search
  - Machine Translation
Counting words and types

- *The chair of the session starts to chair the session*
  - Tokens?
  - Types?

Stop words

- Tokens with little stand-alone information
  - The, a, an, there, here, is, am, ...
  - Punctuations

- ~ functions words
  - Content words
Character Encoding

- ASCII: American Standard Code for Information Interchange
  - $2^7 = 128$ characters (1 byte, with one control bit)
  - 99 numbers, punctuations, letters
    - Mainly English

ASCII characters

<table>
<thead>
<tr>
<th>Decimal</th>
<th>Binary (7-bit)</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>100 0001</td>
<td>A</td>
</tr>
<tr>
<td>66</td>
<td>100 0010</td>
<td>B</td>
</tr>
<tr>
<td>67</td>
<td>100 0011</td>
<td>C</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>97</td>
<td>110 0001</td>
<td>a</td>
</tr>
<tr>
<td>98</td>
<td>110 0010</td>
<td>b</td>
</tr>
<tr>
<td>99</td>
<td>110 0011</td>
<td>c</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>127</td>
<td>111 1111</td>
<td>(DEL)</td>
</tr>
</tbody>
</table>
Unicode

• A character encoding standard developed by the Unicode Consortium

• Provides a single representation for all world's writing systems

• "Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language.”
  – http://www.unicode.org

Unicode encodings

• Version 6.2 (2012) has codes for 110,182 characters
  – Full Unicode standard uses 32 bits (4 bytes): it can represent $2^{32} = 4,294,967,296$ characters!
  • In reality, only 21 bits are needed

• Unicode has three encoding versions
  – UTF-32 (32 bits/4 bytes): direct representation
  – UTF-16 (16 bits/2 bytes): 2^16 = 65,536 possibilities
  – UTF-8 (8 bits/1 byte): 2^8 = 256 possibilities

• Why UTF-16 and UTF-8?
  – They are more compact (for certain languages, i.e., English)
UTF-8

• Unicode has three encoding versions
  – UTF-32 (32 bits/4 bytes): direct representation
  – UTF-16 (16 bits/2 bytes): $2^{16}=65,536$ possibilities
  – UTF-8 (8 bits/1 byte): $2^8=256$ possibilities

• But how do you represent all of $2^{32}$ (=4 billion) code points with only one byte (UTF-8: $2^8=256$ slots)?
  – You don’t.
  – In reality, only $2^{21}$ bits are ever utilized for 110K characters.
  – UTF-8 and UTF-16 use a variable-width encoding.

Today: Text

✓ Different types of text
✓ Corpus

Levels of text processing
  ✓ Corpus
  ✓ Document
  ✓ Sentence
  ✓ Tokens
  ✓ Characters and Encoding
Python Lab

• Text processing warm up with standard Python and NLTK

• Lab-1 is available on course webpage:
  – [http://www.qatar.cmu.edu/~behrang/15383](http://www.qatar.cmu.edu/~behrang/15383)
  – Task-1: Reading a text corpus with Python and basic processing to find frequent word types.
  – Task-2: Repeating task-1, using some of NLTK utilities.

  – Start now and submit before 6 pm on Thursday