The Computer and Natural Language
(Ling 445/515)
Text and Speech Encoding

Markus Dickinson
Department of Linguistics, Indiana University
Autumn 2011

Language and Computers – where to start?

- If we want to do anything with language, we need a way
to represent language.
- We can interact with the computer in several ways:
  - write or read text
  - speak or listen to speech
- Computer has to have some way to represent
  - text
  - speech

Writing systems used for human languages

What is writing?

“a system of more or less permanent marks used
to represent an utterance in such a way that it can be
recovered more or less exactly without the
intervention of the utterer.”
(Peter T. Daniels, The World’s Writing Systems)

Different types of writing systems are used:

- Alphabetic
- Syllabic
- Logographic

Much of the information on writing systems and the graphics used are

Alphabetic systems

Alphabets (phonemic alphabets)

- represent all sounds, i.e., consonants and vowels
- Examples: Etruscan, Latin, Korean, Cyrillic, Runic,
  International Phonetic Alphabet

Abjads (consonant alphabets)

- represent consonants only (sometimes plus selected
  vowels; vowel diacritics generally available)
- Examples: Arabic, Aramaic, Hebrew

Alphabet example: Fraser

An alphabet used to write Lisu, a Tibeto-Burman language spoken by
about 657,000 people in Myanmar, India, Thailand and in the Chinese
provinces of Yunnan and Sichuan.

Consents:

\[
\begin{array}{c}
\{p\} \{b\} \{f\} \{v\} \{m\} \{n\} \{l\} \{r\} \{\theta\} \{\epsilon\} \\
\{\epsilon\} \{\theta\} \{\j\} \{\k\} \{\x\} \{\v\} \{\y\} \{\z\} \\
\end{array}
\]

Vowels:

\[
\begin{array}{c}
\{i\} \{e\} \{\epsilon\} \{\u\} \{\o\} \{\u\} \{\o\} \{\a\} \{\u\} \{\o\}
\end{array}
\]

Tones:

\[
\begin{array}{c}
\text{high tone} \quad \text{mid rising} \quad \text{mid} \quad \text{low} \quad \text{low} \quad \text{high} \quad \text{nasalization}
\end{array}
\]

(from: http://www.omniglot.com/writing/fraser.html)
**Abjad example: Phoenician**

An abjad used to write Phoenician, created between the 18th and 17th centuries BC; assumed to be the forerunner of the Greek and Hebrew alphabet.

<table>
<thead>
<tr>
<th>.Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>z</td>
<td>z</td>
</tr>
<tr>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>h</td>
<td>h</td>
</tr>
<tr>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>g</td>
<td>g</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>'sfer</td>
<td>'sfer</td>
</tr>
<tr>
<td>s</td>
<td>s</td>
</tr>
<tr>
<td>z</td>
<td>z</td>
</tr>
<tr>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>m</td>
<td>m</td>
</tr>
<tr>
<td>l</td>
<td>l</td>
</tr>
<tr>
<td>r</td>
<td>r</td>
</tr>
<tr>
<td>k</td>
<td>k</td>
</tr>
<tr>
<td>y</td>
<td>y</td>
</tr>
</tbody>
</table>

(from: http://www.omniglot.com/writing/phoenician.htm)

**More examples for non-transparent letter-sound correspondences**

**French**

1. **Versailles** → [versal]
2. **ete, etais, etait** → [ete]
3. **samhradh (summer)** → [sauruh]
4. **scri’obhaim (I write)** → [shgri:m]

What is the notation used within the []?

**Irish**

1. **Baile A’tha Cliath (Dublin)** → [bl’a: kli uh]
2. **samhradh (summer)** → [sauruh]
3. **scri’obhaim (I write)** → [shgri:m]

**The International Phonetic Alphabet (IPA)**

- Several special alphabets for representing sounds have been developed, the best known being the International Phonetic Alphabet (IPA).
- The phonetic symbols are unambiguous:
  - designed so that each speech sound gets its own symbol,
  - eliminating the need for multiple symbols used to represent simple sounds
  - one symbol being used for multiple sounds.
- Interactive example chart: http://web.uvic.ca/ling/resources/ipa/Charts/IPAlab/IPAlab.htm

**Syllabic systems**

**Syllabic alphabets (Alphasyllabaries)**

- writing systems with symbols that represent a consonant with a vowel, but the vowel can be changed by adding a **diacritic** (= a symbol added to the letter).
- Examples: Balinese, Javanese, Tibetan, Tamil, Thai, Tagalog

(cf. also: http://www.omniglot.com/writing/syllabaries.htm)

**Syllabaries**

- writing systems with separate symbols for each syllable of a language
- Examples: Cherokee, Ethiopic, Cypriot, Ojibwe, Hiragana (Japanese)

(cf. also: http://www.omniglot.com/writing/syllabaries.html#syll)

**A note on the letter-sound correspondence**

- Alphabets use letters to encode sounds (consonants, vowels).
- But the correspondence between spelling and pronunciation in many languages is quite complex, i.e., not a simple one-to-one correspondence.
- Example: English
  - same spelling — different sounds: ough: **ought**, **cough**
  - silent letters: **knee**, **knight**, **knife**, **debt**, **psychology**, **mortgage**
  - one letter — multiple sounds: **exit**, **use**
  - multiple letters — one sound: **the** (revolution)
  - alternate spellings: **jail** or **gaol**; but not possible **seagh** for **chef** (despite sure, dead, laugh)

**Syllabary example: Cypriot**

The Cypriot syllabary or Cypro-Minoan writing is thought to have developed from the Linear A, or possibly the Linear B script of Crete, though its exact origins are not known. It was used from about 800 to 200 BC.
Syllabic alphabet example: Lao

Script developed in the 14th century to write the Lao language, based on an early version of the Thai script, which was developed from the Old Khmer script, which was itself based on Mon scripts.

Example for vowel diacritics around the letter k:

<table>
<thead>
<tr>
<th>k</th>
<th>ᵏ</th>
<th>ʔ</th>
<th>kʰ</th>
<th>kʷ</th>
<th>kʼ</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>[k]</td>
<td>[k̥]</td>
<td>[k̂]</td>
<td>[kʰ]</td>
<td>[kʷ]</td>
<td>[kʼ]</td>
<td>[k̚]</td>
</tr>
</tbody>
</table>

(from: http://www.omniglot.com/writing/lao.htm)

Logographic writing systems

- Logographs (also called Logograms):
  - Pictographs (Pictograms): originally pictures of things, now stylized and simplified.
  - Ideographs (Ideograms): representations of abstract ideas
  - Compounds: combinations of two or more logographs.
  - Semantic-phonetic compounds: symbols with a meaning element (hints at meaning) and a phonetic element (hints at pronunciation).

Examples: Chinese (Zhongwen), Japanese (Nihongo), Mayan, Vietnamese, Ancient Egyptian

(from: http://www.omniglot.com/writing/chinese types.htm)

Logograph writing system example: Chinese

Pictographs

女子 口 日 月 山 川 鸟 木 麻雀
woman child mouth sun moon mountain river pig eye heart pain dead tree bird

Ideographs

一二三四上下中间力凹四
one two three above below middle strength (ough) convex concave

Compounds of Pictographs/Ideographs

好 安 明 家 思 神 武 男
peaceful safe bright home family thought spirit war man

An example from Ancient Egyptian

mīh (crocodile) - m i h (crocodile) - m i w - cat (fish)

(from: http://www.omniglot.com/writing/egyptian.htm)

Two writing systems with unusual realization

Tactile

- Braille is a writing system that makes it possible to read and write through touch; primarily used by the (partially) blind.
- It uses patterns of raised dots arranged in cells of up to six dots in a 3 x 2 configuration.
- Each pattern represents a character, but some frequent words and letter combinations have their own pattern.

Chromatographic

- The Benin and Edo people in southern Nigeria have supposedly developed a system of writing based on different color combinations and symbols.

(See: http://www.library.cornell.edu/africana/Writing_Systems/Chroma.html)

Braille alphabet
Japanese: logographic system kanji, syllabary katakana, syllabary hiragana
- kanji: 5,000-10,000 borrowed Chinese characters
- katakana
  - used mainly for non-Chinese loan words, onomatopoeic words, foreign names, and for emphasis
- hiragana
  - originally used only by women (10th century), but codified in 1946 with 48 syllables
  - used mainly for word endings, kids’ books, and for words with obscure kanji symbols
- romaji: Roman characters

Korean
“Korean writing is an alphabet, a syllabary and logographs all at once.”
(http://home.vicnet.net.au/~ozideas/writkor.htm)
- The hangul system was developed in 1444 during King Sejong’s reign.
  - There are 24 letters: 14 consonants and 10 vowels
  - But the letters are grouped into syllables, i.e. the letters in a syllable are not written separately as in the English system, but together form a single character.
  - E.g., “Hangeul” (from: http://www.omniglot.com/writing/korean.htm)
    - (han’gʌl) (hʌn) + (æ) + (ʌ) + (ɛ) + (ɛ) + (ə)
  - In South Korea, hanja (logographic Chinese characters) are also used.

Azeri
A Turkish language with speakers in Azerbaijan, northwest Iran, and (former Soviet) Georgia
- 7th century until 1920s: Arabic scripts. Three different Arabic scripts used
- 1929: Latin alphabet enforced by Soviets to reduce Islamic influence.
- 1939: Cyrillic alphabet enforced by Stalin
- 1991: Back to Latin alphabet, but slightly different than before.
  - Latin typewriters and computer fonts were in great demand in 1991
Converting decimal numbers to binary

**Tabular Method**

Using the first 4 bits, we want to know how to write 10 in bit (or binary) notation.

<table>
<thead>
<tr>
<th>Decimal</th>
<th>Remainder?</th>
<th>Binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/2</td>
<td>5</td>
<td>0101</td>
</tr>
<tr>
<td>2/2</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>1/2</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

**Division Method**

<table>
<thead>
<tr>
<th>Decimal</th>
<th>Remainder?</th>
<th>Binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/2</td>
<td>5</td>
<td>0101</td>
</tr>
<tr>
<td>2/2</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>1/2</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

**The ASCII chart**

Codes 1–31 are used for control characters (backspace, line feed, tab, ...).

<table>
<thead>
<tr>
<th>Decimal</th>
<th>Remainder?</th>
<th>Binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>0</td>
<td>0100000</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>0010001</td>
</tr>
<tr>
<td>34</td>
<td>0</td>
<td>0010000</td>
</tr>
<tr>
<td>35</td>
<td>1</td>
<td>0010001</td>
</tr>
<tr>
<td>36</td>
<td>0</td>
<td>0010000</td>
</tr>
<tr>
<td>37</td>
<td>1</td>
<td>0010001</td>
</tr>
</tbody>
</table>

**E-mail issues**

- Mail sent on the internet used to only be able to transfer the 7-bit ASCII messages. But now we can detect the incoming character set and adjust the input.
- Note that this is an example of meta-information = information which is printed as part of the regular message, but tells us something about that message.
- Multipurpose Internet Mail Extensions (MIME) provides meta-information on the text, which tells us:
  - which version of MIME is being used
  - what the character set is
  - if that character set was altered, how it was altered

- Extended the ASCII system with various other systems, for example:
  - ISO 8859-1: includes extra letters needed for French, German, Spanish, etc.
  - ISO 8859-7: Greek alphabet
  - ISO 8859-8: Hebrew alphabet
  - JIS X 0208: Japanese characters
  - Have one system for everything → Unicode
Unicode

Problems with having multiple encoding systems:

- Conflicts: two encodings can use:
  - same number for two different characters
  - different numbers for the same character
- Hassel: have to install many, many systems if you want to be able to deal with various languages

Unicode tries to fix that by having a single representation for every possible character.

"Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language." (www.unicode.org)

Compact encoding of Unicode characters

- Unicode has three versions
  - UTF-32 (32 bits): direct representation
  - UTF-16 (16 bits): $2^{16} = 65536$
  - UTF-8 (8 bits): $2^8 = 256$
- How is it possible to encode $2^{32}$ possibilities in 8 bits (UTF-8)?
  - Several bytes are used to represent one character.
  - Use the highest bit as flag:
    - highest bit 0: single character
    - highest bit 1: part of a multi byte character
  - Nice consequence: ASCII text is in a valid UTF-8 encoding.

UTF-8 details

- First byte unambiguously tells you how many bytes to expect after it
  - e.g., first byte of 11110xxx has a four total bytes
  - all non-starting bytes start with 10 = not the initial byte

<table>
<thead>
<tr>
<th>Byte 1</th>
<th>Byte 2</th>
<th>Byte 3</th>
<th>Byte 4</th>
<th>Byte 5</th>
<th>Byte 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0xxxxxx</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>110xxxx</td>
<td>10xxxx</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1110xxx</td>
<td>10xxxx</td>
<td>10xxxx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11110xx</td>
<td>10xxxx</td>
<td>10xxxx</td>
<td>10xxxx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>111110x</td>
<td>10xxxx</td>
<td>10xxxx</td>
<td>10xxxx</td>
<td>10xxxx</td>
<td>10xxxx</td>
</tr>
</tbody>
</table>

**Example:** Greek α (‘alpha’) has a code value of 945

- Binary: 11 10110001
- 11 10110001 = 011 10110001 = 01110 110001
- Insert these numbers into x’s in the second row:
  - 11011100 10110001

The need for speech

We want to be able to encode any spoken language

- What if we want to work with an unwritten language?
- What if we want to examine the way someone talks and don’t have time to write it down?

Many applications for encoding speech:

- Building spoken dialogue systems, i.e. speak with a computer (and have it speak back).
- Helping people sound like native speakers of a foreign language.
- Helping speech pathologists diagnose problems
What does speech look like?

We can transcribe (write down) the speech into a phonetic alphabet.

- It is very expensive and time-consuming to have humans do all the transcription.
- To automatically transcribe, we need to know how to relate the audio file to the individual sounds that we hear.

We need to know:
- some properties of speech
- how to measure these speech properties
- how these measurements correspond to sounds we hear

Articulatory properties: How it’s produced

We could talk about how sounds are produced in the vocal tract, i.e. articulatory phonetics

- place of articulation (where): [t] vs. [k]
- manner of articulation (how): [t] vs. [s]
- voicing (vocal cord vibration): [t] vs. [d]

But we need to know acoustic properties of speech which we can quantify.

Sampling rate

The higher the sampling rate, the better quality the recording … but the more space it takes.

- Speech needs at least 8000 samples/second, but most likely 16,000 or 22,050 Hz will be used
- The rate for CDs is 44,100 samples/second (or Hertz (Hz))

Now, we can talk about what we need to measure

Measuring sound

sampling rate = how many times in a given second we extract a moment of sound; measured in samples per second

- Sound is continuous, but we have to store data in a discrete manner.

![CONTINUOUS DISCRETE]

- We store data at each discrete point, in order to capture the general pattern of the sound

Acoustic properties: What it sounds like

Sound waves = “small variations in air pressure that occur very rapidly one after another” (Ladefoged, A Course in Phonetics), akin to ripples in a pond

The main properties we measure:

- speech flow = rate of speaking, number and length of pauses (seconds)
- loudness (amplitude) = amount of energy (decibels)
- frequencies = how fast the sound waves are repeating (cycles per second, i.e. Hertz)
  - pitch = how high or low a sound is
    - In speech, there is a fundamental frequency, or pitch, along with higher-frequency overtones.

Researchers also look at things like intonation, i.e., the rise and fall in pitch
Automatic speech recognition = process by which the computer maps a speech signal to text.

Uses/Applications:

- Dictation
- Dialogue systems
- Telephone conversations
- People with disabilities – e.g. a person hard of hearing could use an ASR system to get the text (closed captioning)
Steps in an ASR system

1. Digital sampling of speech
2. Acoustic signal processing = converting the speech samples into particular measurable units
3. Recognition of sounds, groups of sounds, and words

May or may not use more sophisticated analysis of the utterance to help.

- e.g., a [t] might sound like a [d], and so word information might be needed (more on this later)

Kinds of ASR systems

- Differing sizes and types of vocabularies
  - from tens of words to tens of thousands of words
  - might be very domain-specific, e.g., flight vocabulary
- continuous speech vs. isolated-word systems:
  - continuous speech systems = words connected together and not separated by pauses
  - isolated-word systems = single words recognized at a time, requiring pauses to be inserted between words

Text-to-Speech Synthesis (TTS)

Could just record a voice saying phrases or words and then play back those words in the appropriate order.

- This won’t work for, e.g., dialogue systems where speech is generated on the fly.

Or can break the text down into smaller units

1. Convert input text into phonetic alphabet (unambiguous)
2. Synthesize phonetic characters into speech

To synthesize characters into speech, people have tried:

- using formulas which adjust the values of the frequencies, the loudness, etc.
- using a model of the vocal tract and trying to produce sounds based on how a human would speak

Synthesizing Speech

In some sense, TTS really is the reverse process of ASR

- Since we know what frequencies correspond to which vowels, we can play those frequencies to make it sound like the right vowel.
- However, as mentioned before, sounds are always different (across time, across speakers)

One way to generate speech is to have a database of speech and to use the diphones, i.e., two-sound segments, to generate sounds.

- Diphones help with the context-dependence of sounds

Speech to Text

If we convert speech to text and then back to speech, it should sound the same, right?

- But at the conversion stages, there is information loss.
  To avoid this loss would require a lot of memory and knowledge about what exact information to store.
- The process is thus irreversible.
Demos

Text-to-Speech

- AT&T multilingual TTS system: http://www2.research.att.com/~ttsweb/tts.php
- various systems and languages: http://www.ims.uni-stuttgart.de/~moehler/synthspeech/

N-gram definition

An n-gram is a stretch of text n words long

- Approximation of language: information in n-grams tells us something about language, but doesn’t capture the structure
- Efficient: finding and using every, e.g., two-word collocation in a text is quick and easy to do

N-grams help a variety of NLP applications, including word prediction

- N-grams can be used to aid in predicting the next word of an utterance, based on the previous \( n - 1 \) words

Simple n-grams

Let’s assume we want to predict the next word, based on the previous context of “I dreamed I saw the knights in

- What we want to find is the likelihood of the next word, given that we’ve seen \( w_1, \ldots, w_7 \)
- So, we’ll have to examine \( P(w_1, \ldots, w_8) \)

In general, for \( w_n \), we are looking for:

\[
(3) \quad P(w_1, \ldots, w_n) = P(w_1)P(w_2|w_1) \cdots P(w_n|w_{n-1})
\]

But these probabilities are impractical to calculate: they hardly ever occur in a corpus, if at all.

- And it would be a lot of data to store, if we could calculate them.

Unigrams

So, we can approximate these probabilities to a particular n-gram, for a given \( n \). What should \( n \) be?

- Unigrams (\( n = 1 \)):

\[
(4) \quad P(w_n|w_1, \ldots, w_{n-1}) \approx P(w_n)
\]

- Easy to calculate, but we have no contextual information

\[
(5) \quad \text{The quick brown fox jumped}
\]

- We would like to say that over has a higher probability in this context than lazy does.

Bigrams

\( n = 2 \) are a better choice and still easy to calculate:

\[
(6) \quad P(w_n|w_1, \ldots, w_{n-1}) \approx P(w_n|w_{n-1})
\]

\[
(7) \quad P(\text{over}|\text{The, quick, brown, fox, jumped}) \approx P(\text{over}|\text{jumped})
\]

And thus, we obtain for the probability of a sentence:

\[
(8) \quad P(w_1, \ldots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2) \cdots P(w_n|w_{n-1})
\]
Bigram example

What is the probability of seeing the sentence *The quick brown fox jumped over the lazy dog*?

(9) \[ P(\text{The quick brown fox jumped over the lazy dog}) = P(\text{The}|\text{START})P(\text{quick}|\text{The})P(\text{brown}|\text{quick})...P(\text{dog}|\text{lazy}) \]

Or, for our ASR example, we can compare:

(10) \[ P(\text{need}|I) >> P(\text{neat}|I) \]