LING 120: Language and Computers

Semester: FALL 2017

Instructor: Sowmya Vajjala

Iowa State University, USA

25 Aug 2017

Class Outline

- Quick recap of last class
- More on Unicode
- Encoding speech on computer
- Assignment 1 description

Encoding text on computers

- Everything is seen as bits and bytes by computer. So, there should be someway to encode text in binary system.

Encoding text on computers

- Everything is seen as bits and bytes by computer. So, there should be someway to encode text in binary system.
- ASCII is one such 7 bit encoding that covers English, numbers, punctuation etc.
- It can be extended to other languages by creating 8 bit variations to ASCII.
- However, there are so many languages and writing systems. What can we do about that?

Unicode

- Aim: a single representation to represent all characters in all existing writing systems (`unicode.org`).

Unicode

- Aim: a single representation to represent all characters in all existing writing systems (`unicode.org`).
- How does it do this?: it uses a 32 bit representation instead of 8 bit!
- So, how many characters can it represent? - $2^{32} = 4,294,967,296$!

Unicode

- Aim: a single representation to represent all characters in all existing writing systems (`unicode.org`).
- How does it do this?: it uses a 32 bit representation instead of 8 bit!
- So, how many characters can it represent? - $2^{32} = 4,294,967,296$!
- As we discussed towards the end of last class, 32 bits for every letter is a problem - waste of space, makes it slow etc.
- What to do??

UTF-8, UTF-16, UTF-32

- Unicode has three representations (UTF- Unicode Transformation Format) - 8, 16, 32 represent the number of bits needed to represent a character in that representation.

UTF-8, UTF-16, UTF-32

- Unicode has three representations (UTF- Unicode Transformation Format) - 8, 16, 32 represent the number of bits needed to represent a character in that representation.
- How can 2^{32} combinations be represented with 2^{16} or 2^{8} combinations itself??

UTF-8, UTF-16, UTF-32

- Unicode has three representations (UTF- Unicode Transformation Format) - 8, 16, 32 represent the number of bits needed to represent a character in that representation.
- How can 2^{32} combinations be represented with 2^{16} or 2^{8} combinations itself??
- The idea is to use variable number of bytes to represent a character (instead of 1 byte all the time or 4 bytes all the time)
- How to do that?: Use the left most bits as "flags" to tell the computer about number of bytes used per character. i.e., if the starting bit is 1, it means there is only character. Starting two bits are 11 means - you should expect two bytes per character, and so on.

UTF-8, UTF-16, UTF-32

- Unicode has three representations (UTF- Unicode Transformation Format) - 8, 16, 32 represent the number of bits needed to represent a character in that representation.
- How can 2^{32} combinations be represented with 2^{16} or 2^{8} combinations itself??
- The idea is to use variable number of bytes to represent a character (instead of 1 byte all the time or 4 bytes all the time)
- How to do that?: Use the left most bits as "flags" to tell the computer about number of bytes used per character. i.e., if the starting bit is 1, it means there is only character. Starting two bits are 11 means - you should expect two bytes per character, and so on.
- Good thing about this is: ASCII is already UTF-8, you don't have to change anything.

UTF-8 Details

- First byte tells you how many bytes to expect. e.g., if you see something like 11110xxx, you know you should expect this character to be of four bytes.
- Second byte on, everything starts with 10 to indicate that it is not the first byte in that sequence.
- Let us take the example of the Greek character α. In Unicode, its value is 945, which in binary is 11 10110001. What is this with 32 bits?

UTF-8 Details

- First byte tells you how many bytes to expect. e.g., if you see something like 11110xxx, you know you should expect this character to be of four bytes.
- Second byte on, everything starts with 10 to indicate that it is not the first byte in that sequence.
- Let us take the example of the Greek character α. In Unicode, its value is 945, which in binary is 11 10110001. What is this with 32 bits?
- 00000000 00000000 00000011 10110001
- How can we represent this number with UTF-8?

UTF-8 Details

- First byte tells you how many bytes to expect. e.g., if you see something like 11110xxx, you know you should expect this character to be of four bytes.
- Second byte on, everything starts with 10 to indicate that it is not the first byte in that sequence.
- Let us take the example of the Greek character α. In Unicode, its value is 945, which in binary is 11 10110001. What is this with 32 bits?
- 00000000 00000000 00000011 10110001
- How can we represent this number with UTF-8?
- **11**001110 **10**110001

Question

- The question I did not manage to ask in last class: open `zh.wikipedia.org` in Firefox browser, and find out what the encoding of that page is. Usually, you will also see a host of other encodings - what other options do you see?. What happens if you choose a different encoding instead of the one shown?

Question

- The question I did not manage to ask in last class: open `zh.wikipedia.org` in Firefox browser, and find out what the encoding of that page is. Usually, you will also see a host of other encodings - what other options do you see?. What happens if you choose a different encoding instead of the one shown?
- You see unreadable text! :-)
- Any questions on so far?

If that made you curious about encodings...

- Follow the khan academy lectures on number systems if you are not familiar with them.
- Browse through unicode.org to know more about different language representations
- Browse through wikipedia.org, see all different languages in which it exists - try to notice differences between them (scripts, long words, very short words, no punctuation etc)
- Think in terms of what this means for a computer

Note: Make use of the office hours. Send me an email to schedule a time if the office hours does not work for you.

Encoding Speech on Computer

Why?

- Many languages have no written form! How can we work with that language? (why bother?)

Why?

- Many languages have no written form! How can we work with that language? (why bother?)
- Hands free interfaces can be made possible with spoken language encoding (Siri, Echo etc). Always convenient than typing, isn't it?

Why?

- Many languages have no written form! How can we work with that language? (why bother?)
- Hands free interfaces can be made possible with spoken language encoding (Siri, Echo etc). Always convenient than typing, isn't it?
- What if I cannot provide a interface to type that script on computer?

Why?

- Many languages have no written form! How can we work with that language? (why bother?)
- Hands free interfaces can be made possible with spoken language encoding (Siri, Echo etc). Always convenient than typing, isn't it?
- What if I cannot provide a interface to type that script on computer?
- How can we examine differences in accents, dialects etc when we see them just as text versions?

Why?

- Many languages have no written form! How can we work with that language? (why bother?)
- Hands free interfaces can be made possible with spoken language encoding (Siri, Echo etc). Always convenient than typing, isn't it?
- What if I cannot provide a interface to type that script on computer?
- How can we examine differences in accents, dialects etc when we see them just as text versions?
- Teaching pronunciation, Helping speech pathologists diagnose problems, etc.

Making sense of speech signals-1

- One way: Transcribe into a phonetic alphabet (IPA is one such alphabet) that is universal.
- Problem?

Making sense of speech signals-1

- One way: Transcribe into a phonetic alphabet (IPA is one such alphabet) that is universal.
- Problem? : how do we do that? Doing it manually is very expensive and time consuming!
- So how do we represent speech?

Making sense of speech signals-1

- One way: Transcribe into a phonetic alphabet (IPA is one such alphabet) that is universal.
- Problem? : how do we do that? Doing it manually is very expensive and time consuming!
- So how do we represent speech?
- by studying the acoustic properties of its sound waves.

Making Sense of Speech Signals-2

- When we record sound, it is a continuous audio wave. However, they are stored as discrete points, based on something called "sampling rate" (how many times per second do we extract a sound snippet). This tells us about the quality of the recording.
- High sampling rate indicates what quality of recording (better or worse?)

Making Sense of Speech Signals-2

- When we record sound, it is a continuous audio wave. However, they are stored as discrete points, based on something called "sampling rate" (how many times per second do we extract a sound snippet). This tells us about the quality of the recording.
- High sampling rate indicates what quality of recording (better or worse?) - better
- Why don't we just take a large sampling rate all the time, then?

Making Sense of Speech Signals-2

- When we record sound, it is a continuous audio wave. However, they are stored as discrete points, based on something called "sampling rate" (how many times per second do we extract a sound snippet). This tells us about the quality of the recording.
- High sampling rate indicates what quality of recording (better or worse?) - better
- Why don't we just take a large sampling rate all the time, then? - more space!
- Telephone conversations usually are recorded with 8000 samples per second, general speech recording is 16K or 32K.

What speech properties are interesting?

- speech rate (fluency, number of pauses etc)
- Loudness/amplitude
- What sound frequencies correspond to different characters in human speech?
- How can we tell sounds apart with this frequency information?
- Pitch - how high or low is a sound (useful especially for identifying vowels)
- Intonation - rise and fall of pitch

Assignment 1 description

Check the assignment file on Canvas.
10 marks, Deadline: 8th September 2017, upload a PDF.

Next Week

- Topics: Encoding text and speech - conclusion; Writing aids - introduction
- To Do: Go through Chapter 1 and Notes, Exercises after it. Ask questions - either on the forum on Canvas or in the class next week.
- To Do: Start thinking about Assignment 1
- Note: This class is not difficult at all - you just need to be curious about language-computer interaction!

Question for Today

The following phrases/sentences represent some mishearings of songs and possible errors that a speech recognition software can also make. Try to guess an alternate version and post your responses on Canvas forum for today. That is your attendance for today:

- Example: "How to wreck a nice beach" - "How to recognise speech"
- "Secret agent man"
- "when the rainbow shaves you clean, you'll know"
- "with my knee on my mind"
- "language interpreters"
- "synthetic meditation"

Note: This is not an exam. Just a fun activity to make you think about the topic!