LING 120: Language and Computers

Semester: FALL 2017

Instructor: Sowmya Vajjala

Iowa State University, USA

23 Aug 2017

Class outline

1. What is it about language that makes it difficult for computers?
2. Encoding language on computers
   - Writing systems
   - Storing different writing systems on computer
3. Small group exercise (computer based)

Some language processing scenarios for computers

Computers and human language-1

Google Home demo

Girl: Okay Google, what's apples in Spanish?
Google: (answers)
******
Woman: Change my dinner reservation tonight, 7:30 to 8pm.
Google: Your reservation for XXX is confirmed for 8pm.
source: `https://www.youtube.com/watch?v=2KpLHdAURGo`

Computers and human language-2

from 2011: Watson beats humans in Jeopardy

https://www.youtube.com/watch?v=WFR3lOm_xhE

Where do language and computers interact in real-world?

1. Apple Siri and other such software that can understand and interpret human speech (okay, partially)
2. Google Translate and the likes
3. Search Engines
4. Question Answering (e.g., IBM Watson)
5. News recommendation - related articles features in News websites
6. Sentiment analysis of product reviews on Amazon, for example
7. Spam classification in Gmail, Yahoomail etc
8. Information extraction from text (e.g., identifying calendar entries automatically in gmail)
9. Dialog systems (having interactive conversations with users, to do flight bookings etc)
10. Spelling and grammar checkers

... and many more.

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune (`http://goo.gl/zvx9Uw`)

1. When she says "I" in the first sentence, does she mean herself literally?

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune (`http://goo.gl/zvx9Uw`)

1. When she says "I" in the first sentence, does she mean herself literally?
2. What is she referring to? When will we know what is she referring to?

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune (`http://goo.gl/zvx9Uw`)

1. When she says "I" in the first sentence, does she mean herself literally?
2. What is she referring to? When will we know what is she referring to?
3. Who is "She"?

Let us take a small text snippet

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune (`http://goo.gl/zvx9Uw`)

1. When she says "I" in the first sentence, does she mean herself literally?
2. What is she referring to? When will we know what is she referring to?
3. Who is "She"?
4. What is "home country" in the last sentence?

Let us take a small text snippet -2

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune `http://goo.gl/zvx9Uw`

1. What is the main event of this text?

Let us take a small text snippet -2

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune `http://goo.gl/zvx9Uw`

1. What is the main event of this text?
2. What is "Chinese Homestyle Cooking" referring to?

Let us take a small text snippet -2

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune `http://goo.gl/zvx9Uw`

1. What is the main event of this text?
2. What is "Chinese Homestyle Cooking" referring to?
3. What is the relationship between "Chinese Homestyle cooking" and Tina?

Let us take a small text snippet -2

"'I'm not big, I'm not fancy," she said as she sat in a booth, looking out the window to Lincoln Way. "But I don't mind."

Chinese Homestyle Cooking's owners, Tina and Chung Song, have run the restaurant for almost 20 years from the small building off the corner of Sheldon Avenue and Lincoln Way. But in late October, they'll close their business when the lease on their building runs out.

Tina's start in the restaurant business came in 1982, when she emigrated from Taiwan and began working as a waitress for her sister's restaurants in Des Moines and Ankeny. While in the U.S., she got a call from Chung, who she had grown up with as a child in their home country.

Source: Ames Tribune `http://goo.gl/zvx9Uw`

1. What is the main event of this text?
2. What is "Chinese Homestyle Cooking" referring to?
3. What is the relationship between "Chinese Homestyle cooking" and Tina?
4. Is Lincoln Way something related to President Lincoln?

So ...

- Each question I asked is a language processing problem for the computer which is not completey solved yet!

So ...

- Each question I asked is a language processing problem for the computer which is not completey solved yet!
- ... and I only gave example of just one language (different languages may have different issues interms of processing)

So ...

- Each question I asked is a language processing problem for the computer which is not completey solved yet!
- ... and I only gave example of just one language (different languages may have different issues interms of processing)
- But before even getting in to that, how do we even represent language on a computer? What does a computer see when I type English or Greek or Chinese?

So ...

- Each question I asked is a language processing problem for the computer which is not completey solved yet!
- ... and I only gave example of just one language (different languages may have different issues interms of processing)
- But before even getting in to that, how do we even represent language on a computer? What does a computer see when I type English or Greek or Chinese?
- How do I type non-English characters anyway??

Encoding language on computers - Background

- Computer stores any kind of information (including language) in bits and bytes (did you hear this before?)

Encoding language on computers - Background

- Computer stores any kind of information (including language) in bits and bytes (did you hear this before?)
- What is a bit?

Encoding language on computers - Background

- Computer stores any kind of information (including language) in bits and bytes (did you hear this before?)
- What is a bit? (it is the short form of binary digit - 0 or 1)
- What is a byte?

Encoding language on computers - Background

- Computer stores any kind of information (including language) in bits and bytes (did you hear this before?)
- What is a bit? (it is the short form of binary digit - 0 or 1)
- What is a byte? (a unit of information comprising of 8 bits)
- How many different ways can we put 0s and 1s into 8 bit sequences?

Encoding language on computers - Background

- Computer stores any kind of information (including language) in bits and bytes (did you hear this before?)
- What is a bit? (it is the short form of binary digit - 0 or 1)
- What is a byte? (a unit of information comprising of 8 bits)
- How many different ways can we put 0s and 1s into 8 bit sequences? 2^8
  $\Rightarrow$ We can represent 256 different characters with 8 bits on a computer!

ASCII - a 7 bit encoding

- American Standard Code for Information Interchange (ASCII) is one of the early encoding systems for computers for storing English text.
- It used only 7 bits to encode different characters.

ASCII TABLE

Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	[NULL]	48	30	110000	60	0	96	60	1100000	140	`
1	1	1	1	[START OF HEADING]	49	31	110001	61	1	97	61	1100001	141	a
2	2	10	2	[START OF TEXT]	50	32	110010	62	2	98	62	1100010	142	b
3	3	11	3	[END OF TEXT]	51	33	110011	63	3	99	63	1100011	143	c
4	4	100	4	[END OF TRANSMISSION]	52	34	110100	64	4	100	64	1100100	144	d
5	5	101	5	[ENQUIRY]	53	35	110101	65	5	101	65	1100101	145	e
6	6	110	6	[ACKNOWLEDGE]	54	36	110110	66	6	102	66	1100110	146	f
7	7	111	7	[BELL]	55	37	110111	67	7	103	67	1100111	147	g
8	8	1000	10	[BACKSPACE]	56	38	111000	70	8	104	68	1101000	150	h
9	9	1001	11	[HORIZONTAL TAB]	57	39	111001	71	9	105	69	1101001	151	i
10	A	1010	12	[LINE FEED]	58	3A	111010	72	:	106	6A	1101010	152	j
11	B	1011	13	[VERTICAL TAB]	59	3B	111011	73	;	107	6B	1101011	153	k
12	C	1100	14	[FORM FEED]	60	3C	111100	74	<	108	6C	1101100	154	l
13	D	1101	15	[CARRIAGE RETURN]	61	3D	111101	75	=	109	6D	1101101	155	m
14	E	1110	16	[SHIFT OUT]	62	3E	111110	76	>	110	6E	1101110	156	n
15	F	1111	17	[SHIFT IN]	63	3F	111111	77	?	111	6F	1101111	157	o
16	10	10000	20	[DATA LINK ESCAPE]	64	40	1000000	100	@	112	70	1110000	160	p
17	11	10001	21	[DEVICE CONTROL 1]	65	41	1000001	101	A	113	71	1110001	161	q
18	12	10010	22	[DEVICE CONTROL 2]	66	42	1000010	102	B	114	72	1110010	162	r
19	13	10011	23	[DEVICE CONTROL 3]	67	43	1000011	103	C	115	73	1110011	163	s
20	14	10100	24	[DEVICE CONTROL 4]	68	44	1000100	104	D	116	74	1110100	164	t
21	15	10101	25	[NEGATIVE ACKNOWLEDGE]	69	45	1000101	105	E	117	75	1110101	165	u
22	16	10110	26	[SYNCHRONOUS IDLE]	70	46	1000110	106	F	118	76	1110110	166	v
23	17	10111	27	[ENG OF TRANS. BLOCK]	71	47	1000111	107	G	119	77	1110111	167	w
24	18	11000	30	[CANCEL]	72	48	1001000	110	H	120	78	1111000	170	x
25	19	11001	31	[END OF MEDIUM]	73	49	1001001	111	I	121	79	1111001	171	y
26	1A	11010	32	[SUBSTITUTE]	74	4A	1001010	112	J	122	7A	1111010	172	z
27	1B	11011	33	[ESCAPE]	75	4B	1001011	113	K	123	7B	1111011	173	{
28	1C	11100	34	[FILE SEPARATOR]	76	4C	1001100	114	L	124	7C	1111100	174	|
29	1D	11101	35	[GROUP SEPARATOR]	77	4D	1001101	115	M	125	7D	1111101	175	}
30	1E	11110	36	[RECORD SEPARATOR]	78	4E	1001110	116	N	126	7E	1111110	176	~
31	1F	11111	37	[UNIT SEPARATOR]	79	4F	1001111	117	O	127	7F	1111111	177	[DEL]
32	20	100000	40	[SPACE]	80	50	1010000	120	P					
33	21	100001	41	!	81	51	1010001	121	Q					
34	22	100010	42	"	82	52	1010010	122	R					
35	23	100011	43	#	83	53	1010011	123	S					
36	24	100100	44	$	84	54	1010100	124	T					
37	25	100101	45	%	85	55	1010101	125	U					
38	26	100110	46	&	86	56	1010110	126	V					
39	27	100111	47	'	87	57	1010111	127	W					
40	28	101000	50	(	88	58	1011000	130	X					
41	29	101001	51	)	89	59	1011001	131	Y					
42	2A	101010	52	*	90	5A	1011010	132	Z					
43	2B	101011	53	+	91	5B	1011011	133	[					
44	2C	101100	54	,	92	5C	1011100	134	\					
45	2D	101101	55	-	93	5D	1011101	135	]					
46	2E	101110	56	.	94	5E	1011110	136	^					
47	2F	101111	57	/	95	5F	1011111	137	_					

image source: commons.wikimedia.org

Writing Systems

- So English is covered by ASCII
- What should we do about several other languages written with different scripts?
- My language (Telugu) has 56 basic characters in the alphabet, and some 20 other additional characters that attach to these.
- Russian alphabet has has around 40 characters.
- There are several Indian language scripts like Telugu, having so many characters.
- There are languages such as Chinese which have 100s of characters.

Writing Systems

- So English is covered by ASCII
- What should we do about several other languages written with different scripts?
- My language (Telugu) has 56 basic characters in the alphabet, and some 20 other additional characters that attach to these.
- Russian alphabet has has around 40 characters.
- There are several Indian language scripts like Telugu, having so many characters.
- There are languages such as Chinese which have 100s of characters.

... clearly ASCII cannot account for all these! What is the solution?

Different encodings for different languages

- Just extend the ASCII to 8 bits and use the remaining numbers (128-255) for adding new characters.
- Several such encodings exist: ISO-8859-1 -adds additional characters for French, German Spanish; ISO-8859-8 for Hebrew etc.

Different encodings for different languages

- Just extend the ASCII to 8 bits and use the remaining numbers (128-255) for adding new characters.
- Several such encodings exist: ISO-8859-1 -adds additional characters for French, German Spanish; ISO-8859-8 for Hebrew etc.
- Problems with such an approach?

Different encodings for different languages

- Just extend the ASCII to 8 bits and use the remaining numbers (128-255) for adding new characters.
- Several such encodings exist: ISO-8859-1 -adds additional characters for French, German Spanish; ISO-8859-8 for Hebrew etc.
- Problems with such an approach?
  1. Two different encodings can have same number for different characters
  2. One character can get different numbers in two different encodings
- So what?:

Different encodings for different languages

- Just extend the ASCII to 8 bits and use the remaining numbers (128-255) for adding new characters.
- Several such encodings exist: ISO-8859-1 -adds additional characters for French, German Spanish; ISO-8859-8 for Hebrew etc.
- Problems with such an approach?
  1. Two different encodings can have same number for different characters
  2. One character can get different numbers in two different encodings
- So what?:
  1. If the encoding information is not provided in the webpage, a browser needs to guess. Guessing is difficult with those two problems.
  2. Each time I want to see a new language, I need a new encoding, install and setup process to work with it!

Solution: Unicode

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

Solution: 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?

Solution: 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?
  $2^{32} = 4,294,967,296$!
- Do we really need so many?
- What are the advantages and disadvantages of this 32 bit representation?

UTF-8, UTF-16, UTF-32

- Unicode has three representations (UTF- Unicode Transformation Format) - the numbers 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) - the numbers 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) - the numbers 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) - the numbers 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

A small exercise

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? Post the answer in today's forum to get attendance for today.

Next Class

- Topic: Encoding spoken language
- Assignment 1 description
- ToDo: Read chapter 1