

LBSC 796/INFM 718R: Week 6
Representing the Meaning of Documents



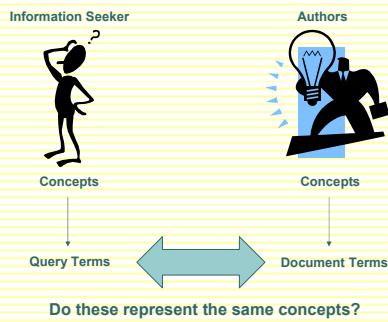
Jimmy Lin
College of Information Studies
University of Maryland

Monday, March 6, 2006

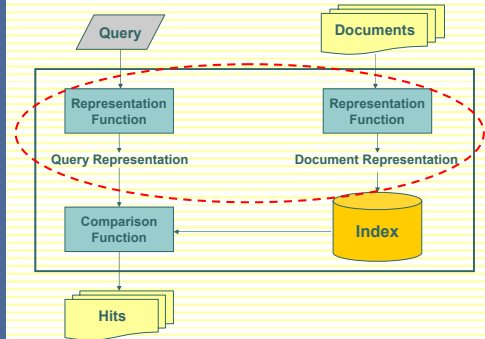
Muddy Points

- Binary trees vs. binary search
- Document presentation
- Algorithm running times
 - Logarithmic, linear, polynomial, exponential

The Central Problem in IR



Today's Class



Outline

- How do we represent the meaning of text?
- What are the problems?
- What are the possible solutions?
- How well do they work?

Why is IR hard?

- IR is hard because natural language is so rich (among other reasons)
- What are the issues?
 - Encoding
 - Tokenization
 - Morphological Variation
 - Synonymy
 - Polysemy
 - Paraphrase
 - Ambiguity
 - Anaphora

Possible Solutions

- Vary the unit of indexing
 - Strings and segments
 - Tokens and words
 - Phrases and entities
 - Senses and concepts
- Manipulate queries and results
 - Term expansion
 - Post-processing of results

Representing Electronic Texts

- A character set specifies the unit of composition
 - Characters are the smallest units of text
 - Abstract entities, separate from how they are stored
- A font specifies the printed representation
 - What each character will look like on the page
 - Different characters might be depicted identically
- An encoding is the electronic representation
 - What each character will look like in a file
 - One character may have several representations
- An input method is a keyboard representation

The Character 'A'

- ASCII = American Standard Code for Information Interchange


```
0 1 0 0 0 0 0 1 = 65 DEC = 'A'
0 1 0 0 0 0 1 0 = 66 DEC = 'B'
...
```
- 7 bits used per character
 - Number of representable characters = 128
 - Some character codes used for non-visible characters
- The visible characters:


```
!"#$%&'()*+,-./0123456789:;<=>?@
ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`
abcdefghijklmnopqrstuvwxyz{|}~
```

The Latin-1 Character Set

- ISO 8859-1: 8-bit characters for Western Europe
 - French, Spanish, Catalan, Galician, Basque, Portuguese, Italian, Albanian, Afrikaans, Dutch, German, Danish, Swedish, Norwegian, Finnish, Faroese, Icelandic, Irish, Scottish, and English

Printable Characters, 7-bit ASCII

	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
0	1	2	3	4	5	6	7	8	9	:	<	=	>	?	
@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	{		}	~	
	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î
	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î
	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Additional Defined Characters, ISO 8859-1

What about these languages?

天主教教宗若望保禄二世因感冒再度住进医院。
這是他今年第二次因同樣的病因住院。

وقال مارك ريجيف - الناطق باسم الخارجية الإسرائيلية - إن شارون قبل الدعوة وسيقوم للمرة الأولى بزيارة تونس، التي عكفت لفترة طويلة المقر الرسمي لمنظمة التحرير الفلسطينية بعد خروجها من لبنان عام 1982.

Выступая в Мещанском суде Москвы экс-глава ЮКОСа заявил не совершал ничего противозаконного, в чем обвиняет его генпрокуратура России.

भा रत सरका रने आर् थकिसर् वे क् षण मे वत् ती यवर् ष 2005-06 मे सा त फ्री सदी विका सदर हा सत्त करने का आकस्मन का या है और कर सु था र पर ज़ो र दया है

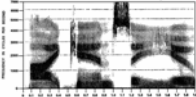
日米連合で中東中東に対処...ア-ミテージ前副長官通電

조제영 기자= 서울시는 25일 이명박 시장이 '행정중심복합도시' 건설안에 대해 '군대라도 등원해 막고싶은 심정'이라고 말했다는 일부 언론의 보도를 부인했다.

Tokenization

- What's a word?
 - First try: words are separated by spaces
The cat on the mat. → the, cat, on, the, mat
 - What about clitics?
I'm not saying that I don't want John's input on this.
- What about languages without spaces?

天主教教宗若望保禄二世因感冒再度住进医院。
→ 天主教 教宗 若望保禄二世 因 感冒 再度 住进 医院。
- Same problem with speech!



Where are the spaces?

Word-Level Issues

- Morphological variation
 - = different forms of the same concept
 - Inflectional morphology: same part of speech
 - break, broke, broken; sing, sang, sung; etc.
 - Derivational morphology: different parts of speech
 - destroy, destruction; invent, invention, reinvention; etc.
- Synonymy
 - = different words, same meaning
 - (dog, canine, doggy, puppy, etc.) → concept of *dog*
- Polysemy
 - = same word, different meanings
 - Bank:** financial institution or side of a river?
 - Crane:** bird or construction equipment?
 - Is:** depends on what the meaning of "is" is!

Paraphrase

- Language provides different ways of saying the same thing

Who killed Abraham Lincoln?

(1) John Wilkes Booth killed Abraham Lincoln.
 (2) John Wilkes Booth altered history with a bullet. He will forever be known as the man who ended Abraham Lincoln's life.

When did Wilt Chamberlain score 100 points?

(1) Wilt Chamberlain scored 100 points on March 2, 1962 against the New York Knicks.
 (2) On December 8, 1961, Wilt Chamberlain scored 78 points in a triple overtime game. It was a new NBA record, but Warriors coach Frank McGuire didn't expect it to last long, saying, "He'll get 100 points someday." McGuire's prediction came true just a few months later in a game against the New York Knicks on March 2.

Ambiguity

- What exactly do you mean?
 - I saw the man on the hill with the telescope.
Who has the telescope?
 - Time flies like an arrow.
Say what?
 - Visiting relatives can be annoying.
Who's visiting?
- Why don't we have problems (most of the time)?

Ambiguity in Action

- Different documents with the same keywords may have different meanings...

<p>What do frogs eat? keywords: frogs, eat</p> <p>✓(1) Adult frogs eat mainly insects and other small animals, including earthworms, minnows, and spiders.</p> <p>✗(2) Alligators eat many kinds of small animals that live in or near the water, including fish, snakes, frogs, turtles, small mammals, and birds.</p> <p>✗(3) Some bats catch fish with their claws, and a few species eat lizards, rodents, small birds, tree frogs, and other bats.</p>	<p>What is the largest volcano in the Solar System? keywords: largest, volcano, solar, system</p> <p>✓(1) Mars boasts many extreme geographic features; for example, Olympus Mons, is the largest volcano in the solar system.</p> <p>✗(2) The Galileo probe's mission to Jupiter, the largest planet in the Solar system, included amazing photographs of the volcanoes on Io, one of its four most famous moons.</p> <p>✗(3) Even the largest volcanoes found on Earth are puny in comparison to others found around our own cosmic backyard, the Solar System.</p>
---	---

Anaphora

- Language provides different ways of referring to the same entity

Who killed Abraham Lincoln?

(1) John Wilkes Booth killed Abraham Lincoln.
 (2) John Wilkes Booth altered history with a bullet. **He** will forever be known as the man **who** ended Abraham Lincoln's life.

When did Wilt Chamberlain score 100 points?

(1) Wilt Chamberlain scored 100 points on March 2, 1962 against the New York Knicks.
 (2) On December 8, 1961, **Wilt Chamberlain** scored 78 points in a triple overtime game. It was a new NBA record, but Warriors coach Frank McGuire didn't expect it to last long, saying, "**He**ll get 100 points someday." McGuire's prediction came true just a few months later in a game against the New York Knicks on March 2.

More Anaphora

- Terminology
 - Anaphor = an expression that refers to another
 - Anaphora = the phenomenon
- Other different types of referring expressions:
 - Fujitsu and NEC said they were still investigating, and that knowledge of more such bids could emerge... **Other major Japanese computer companies** contacted yesterday said they have never made such bids.
 - The hotel recently went through a \$200 million restoration... original artworks include an impressive collection of Greek statues in **the lobby**.
- Anaphora resolution can be hard!
 - The city council denied the demonstrators a permit because... **they** feared violence.
 - they** advocated violence.

What can we do?

- Here are the some of the problems:
 - Encoding, tokenization
 - Morphological variation, synonymy, polysemy
 - Paraphrase, ambiguity
 - Anaphora
- General approaches:
 - Vary the unit of indexing
 - Manipulate queries and results

The Encoding Problem

وقال مارك ريجيف - الناطق باسم الخارجية الإسرائيلية - إن شارون قبل الدعوة وسبقه للمرة الأولى بزيارة تونس، التي كانت لفترة طويلة المقر الرسمي لمنظمة التحرير الفلسطينية بعد خروجها من لبنان عام 1982.

Выступая в Мещанском суде Москвы экс-глава ЮКОСа заявил не совершал ничего противозаконного, в чем обвиняет его генпрокуратура России.

भा रत सरकार रने आर्थिक सर् वे क षण मे वित्ती य वर् ष 2005-06 मे सा त फी सदी वो का सदर हा सिल करने का आकलन किया है और कर सु धा र पर जो र दिया है

日米通商で自東中に対知...アーミテ-ジ新聞長官報告

조제영 기자= 서울시는 25일 이명박 시장이 행정중심복합도시 건설안에 대해 '군데라도 등원해 막고 싶은 심정'이라고 말했다는 일부 언론의 보도를 부인했다.

East Asian Character Sets

- More than 128 characters are needed!
 - Two-byte encoding schemes are used
- Several countries have unique character sets
 - GB in People's Republic of China
 - BIG5 in Taiwan
 - JIS in Japan
 - KS in Korea
 - TCVN in Vietnam
- Many characters appear in several languages

Unicode

- Goal is to unify the world's character sets
 - ISO Standard 10646
- Limitations:
 - Produces much larger files than Latin-1
 - Fonts are hard to obtain for many characters
 - Some characters have multiple representations, e.g., accents can be part of a character or separate
 - Some characters look identical when printed, but they come from unrelated languages
 - The sort order may not be appropriate

What do we index?

- In information retrieval, we are after the concepts represented in the documents
- ... but we can only index strings
- So what's the best unit of indexing?

The Tokenization Problem

- In many languages, words are not separated by spaces...
- Tokenization = separating a string into "words"
- Simple greedy approach:
 - Start with a list of every possible term (e.g., from a dictionary)
 - Look for the longest word in the unsegmented string
 - Take longest matching term as the next word and repeat

Probabilistic Segmentation

- For an input word: $c_1 c_2 c_3 \dots c_n$
- Try all possible partitions:
 $c_1 c_2 c_3 c_4 \dots c_n$
 $c_1 c_2 c_3 c_4 \dots c_n$
 $c_1 c_2 c_3 c_4 \dots c_n$
...
- Choose the highest probability partition
 - E.g., compute $P(c_1 c_2 c_3)$ using a language model
- Challenges: search, probability estimation

Indexing N-Grams

- Consider a Chinese document: $c_1 c_2 c_3 \dots c_n$
- Don't segment (you could be wrong!)
- Instead, treat every character bigram as a term
 $c_1 c_2 c_3 c_4 c_5 \dots c_n$
→ $c_1 c_2 c_2 c_3 c_3 c_4 c_4 c_5 \dots c_{n-1} c_n$
- Break up queries the same way
- Works at least as well as trying to segment correctly!

Morphological Variation

- Handling morphology: related concepts have different forms
 - Inflectional morphology: same part of speech
dogs = dog + PLURAL
broke = break + PAST
 - Derivational morphology: different parts of speech
destruction = destroy + ion
researcher = research + er
- Different morphological processes:
 - Prefixing
 - Suffixing
 - Infixing
 - Reduplication

Stemming

- Dealing with morphological variation: index stems instead of words
 - Stem: a word equivalence class that preserves the central concept
- How much to stem?
 - organization → organize → organ?
 - resubmission → resubmit/submission → submit?
 - reconstructionism?

Stemmers

- Porter stemmer is a commonly used stemmer
 - Strips off common affixes
 - Not perfect!
Errors of omission: Incorrectly lumps unrelated terms together
doe/doing
execute/executive
ignore/ignorant
Errors of omission: Fails to lump related terms together
create/creation
europe/european
cylinder/cylindrical
- Many other stemming algorithms available

Does Stemming Work?

- Generally, yes! (in English)
 - Helps more for longer queries
 - Lots of work done in this area
- Donna Harman (1991) How Effective is Suffixing? Journal of the American Society for Information Science, 42(1):7-15.
- Robert Krovetz. (1993) Viewing Morphology as an Inference Process. Proceedings of SIGIR 1993.
- David A. Hull. (1996) Stemming Algorithms: A Case Study for Detailed Evaluation. Journal of the American Society for Information Science, 47(1):70-84.
- And others...

Stemming in Other Languages

- o Arabic makes frequent use of infixes

the root *ktb* →
maktab (office),
kitab (book),
kutub (books),
kataba (he wrote),
naktubu (we write),
etc.

- o What's the most effective stemming strategy in Arabic? Open research question...

Words = wrong indexing unit!

- o Synonymy
 - = different words, same meaning
 - (dog, canine, doggy, puppy, etc.) → concept of *dog*
- o Polysemy
 - = same word, different meanings
 - Bank:** financial institution or side of a river?
 - Crane:** bird or construction equipment?
- o It'd be nice if we could index concepts!
 - Word sense: a coherent cluster in semantic space
 - Indexing word senses achieves the effect of conceptual indexing

Indexing Word Senses

- o How does indexing word senses solve the synonym/polysemy problem?

(dog, canine, doggy, puppy, etc.) → concept 112986
I deposited my check in the bank. bank → concept 76529
I saw the sailboat from the bank. bank → concept 53107

- o Okay, so where do we get the word senses?
 - WordNet: a lexical database for English
 - <http://wordnet.princeton.edu/>
 - Automatically find "clusters" of words that describe the same concepts
 - Other methods also have been tried...

Word Sense Disambiguation

- o Given a word in context, automatically determine the sense (concept)
 - This is the Word Sense Disambiguation (WSD) problem
- o Context is the key:
 - For each ambiguous word, note the surrounding words
 - bank (river, sailboat, water, etc.) → side of a river
 - bank (check, money, account, etc.) → financial institution
 - "Learn" a classifier from a collection of examples
 - Use the classifier to determine the senses of words in the documents

Does it work?

- o Nope!

Ellen M. Voorhees. (1993) Using WordNet to Disambiguate Word Senses for Text Retrieval. Proceedings of SIGIR 1993.
Mark Sanderson. (1994) Word-Sense Disambiguation and Information Retrieval. Proceedings of SIGIR 1994.
And others...

- o Examples of limited success....

Hinrich Schütze and Jan O. Pedersen. (1995) Information Retrieval Based on Word Senses. Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval.
Rada Mihalcea and Dan Moldovan. (2000) Semantic Indexing Using WordNet Senses. Proceedings of ACL 2000 Workshop on Recent Advances in NLP and IR.

Why Disambiguation Hurts

- o Bag-of-words techniques already disambiguate
 - Context for each term is established in the query
- o WSD is hard!
 - Many words are highly polysemous, e.g., *interest*
 - Granularity of senses is often domain/application specific
- o WSD tries to improve precision
 - But incorrect sense assignments would hurt recall
 - Slight gains in precision do not offset large drops in recall

An Alternate Approach

- Indexing word senses “freezes” concepts at index time
- What if we expanded query terms at query time instead?
 - dog AND cat →
(dog OR canine) AND (cat OR feline)
- Two approaches
 - Manual thesaurus, e.g., WordNet, UMLS, etc.
 - Automatically-derived thesaurus, e.g., co-occurrence statistics

Does it work?

- Yes... if done “carefully”
- User should be involved in the process
 - Otherwise, poor choice of terms can hurt performance

Handling Anaphora

- Anaphora resolution: finding what the anaphor refers to (called the antecedent)
 - John Wilkes Booth altered history with a bullet. **He** will forever be known as the man who ended Abraham Lincoln’s life.
 - He = John Wilkes Booth
- Most common example: pronominal anaphora resolution
 - Simplest method works pretty well: find previous noun phrase matching in gender and number

Expanding Anaphors

- When indexing, replace anaphors with their antecedents
- Does it work?
 - Somewhat
 - ... but can be computationally expensive
 - ... helps more if you want to retrieve sub-document segments

Beyond Word-Level Indexing

- Words are the wrong unit to index...
- Many multi-word combinations identify entities
 - Persons: George W. Bush, Dr. Jones
 - Organizations: Red Cross, United Way
 - Corporations: Hewlett Packard, Kraft Foods
 - Locations: Easter Island, New York City
- Entities often have finer-grained structures
 - Professor Stephen W. Hawking

```

graph TD
    A[Professor Stephen W. Hawking] --> B[title]
    A --> C[first name]
    A --> D[middle initial]
    A --> E[last name]
      
```
 - Cambridge, Massachusetts

```

graph TD
    A[Cambridge, Massachusetts] --> B[city]
    A --> C[state]
      
```

Indexing Named Entities

- Why would we want to index named entities?
- Index named entities as special tokens
 - In reality, at the time of Edison’s 1879 patent, the light bulb **PERSON DATE** had been in existence for some five decades
- And treat special tokens like query terms
 - Who patented the light bulb? → patent light bulb **PERSON**
 - When was the light bulb patented? → patent light bulb **DATE**
- Works pretty well for question answering

John Prager, Eric Brown, and Anni Coden. (2000) Question-Answering by Predictive Annotation. Proceedings of SIGIR 2000.

But First...

- We have to recognize named entities
- Before that, we have to first define a hierarchy
 - Influenced by text genres of interest... mostly news
- Decent algorithms based on pattern matching
- Best algorithms based on supervised learning
 - Annotate a corpus identifying entities and types
 - "Train" a probabilistic model
 - Apply the model to new text

Indexing Phrases

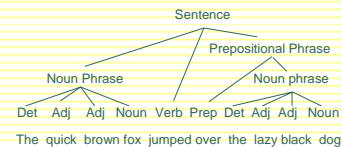
- Two types of phrases
 - Those that make sense, e.g., "school bus", "hot dog"
 - Those that don't, e.g., bigrams in Chinese
- Treat multi-word tokens as index terms
- Three sources of evidence:
 - Dictionary lookup
 - Linguistic analysis
 - Statistical analysis (e.g., co-occurrence)

Known Phrases

- Compile a term list that includes phrases
 - Technical terminology can be very helpful
- Index any phrase that occurs in the list
- Most effective in a limited domain
 - Otherwise hard to capture most useful phrases

Syntactic Phrases

- Parsing = automatically assign structure to a sentence



- "Walk" the tree and extract phrases
 - Index all noun phrases
 - Index subjects and verbs
 - Index verbs and objects
 - etc.

Syntactic Variations

- What does linguistic analysis buy?
 - Coordinations
 - lung and breast cancer → lung cancer, breast cancer
 - Substitutions
 - inflammatory sinonasal disease → inflammatory disease, sinonasal disease
 - Permutations
 - addition of calcium → calcium addition

Statistical Analysis

- Automatically discover phrases based on co-occurrence probabilities
 - $P(\text{"kick the bucket"}) = P(\text{"kick"}) \times P(\text{"the"}) \times P(\text{"bucket"})$?
- If terms are not independent, they may form a phrase
- Use this method to automatically learn a phrase dictionary

Does Phrasal Indexing Work?

- o Yes...
- o But the gains are so small they're not worth the cost
- o Primary drawback: too slow!

What about ambiguity?

- o Different documents with the same keywords may have different meanings...

<p>What do frogs eat? keywords: frogs, eat</p> <p>✓(1) Adult frogs eat mainly insects and other small animals, including earthworms, minnows, and spiders.</p> <p>✗(2) Alligators eat many kinds of small animals that live in or near the water, including fish, snakes, frogs, turtles, small mammals, and birds.</p> <p>✗(3) Some bats catch fish with their claws, and a few species eat lizards, rodents, small birds, tree frogs, and other bats.</p>	<p>What is the largest volcano in the Solar System? keywords: largest, volcano, solar, system</p> <p>✓(1) Mars boasts many extreme geographic features; for example, Olympus Mons, is the largest volcano in the solar system.</p> <p>✗(2) The Galileo probe's mission to Jupiter, the largest planet in the Solar system, included amazing photographs of the volcanoes on Io, one of its four most famous moons.</p> <p>✗(3) Even the largest volcanoes found on Earth are puny in comparison to others found around our own cosmic backyard, the Solar System.</p>
--	--

Indexing Relations

- o Instead of terms, index syntactic relations between entities in the text

Adult frogs eat mainly insects and other small animals, including earthworms, minnows, and spiders.

```

<-> frogs subject-of eat <->
  <-> insects object-of eat >->
  <-> animals object-of eat >->
  <-> adult modifies frogs >->
  <-> small modifies animals >->
  
```

Alligators eat many kinds of small animals that live in or near the water, including fish, snakes, frogs, turtles, small mammals, and birds.

```

<-> alligators subject-of eat <->
  <-> kinds object-of animals >->
  <-> small modifies animals >->
  
```

From the relations, it is clear who's eating whom!

Are syntactic relations enough?

- o Consider this example:

John broke the window. → < John subject-of break >
 The window broke. → < window subject-of break >

"John" and "window" are both subjects...
 But John is the person doing the breaking (or "agent"),
 and the window is the thing being broken (or "theme")

- o Syntax sometimes isn't enough... we need semantics (or meaning)!
- o Semantics, for example, allows us to relate the following two fragments:

The barbarians destroyed the city...
 The destruction of the city by the barbarians...

```

event: destroy
agent: barbarians
theme: city
  
```

Semantic Roles

- o Semantic roles are invariant with respect to syntactic expression

Mary loaded the truck with hay.
 Hay was loaded onto the truck by Mary.

```

event: load
agent: Mary
material: hay
destination: truck
  
```

- o The idea:
 - Identify semantic roles
 - Index "frame structures" with filled slots
 - Retrieve answers based on semantic-level matching

Does it work?

- o No, not really...
- o Why not?
 - Syntactic and semantic analysis is difficult: errors offset whatever gain is gotten
 - As with WSD, these techniques are precision-enhancers... recall usually takes a dive
 - It's slow!

Alternative Approach

- Sophisticated linguistic analysis is slow!
 - Unnecessary processing can be avoided by query time analysis
- Two-stage retrieval
 - Use standard document retrieval techniques to fetch a candidate set of documents
 - Use passage retrieval techniques to choose a few promising passages (e.g., paragraphs)
 - Apply sophisticated linguistic techniques to pinpoint the answer
- Passage retrieval
 - Find "good" passages within documents
 - Key Idea: locate areas where lots of query terms appear close together

Key Ideas

- IR is hard because language is rich and complex (among other reasons)
- Two general approaches to the problem
 - Attempt to find the best unit of indexing
 - Try to fix things at query time
- It is hard to predict *a priori* what techniques work
 - Questions must be answered experimentally
- Words are really the wrong thing to index
 - But there isn't really a better alternative...

One Minute Paper

- What was the muddiest point in today's class?