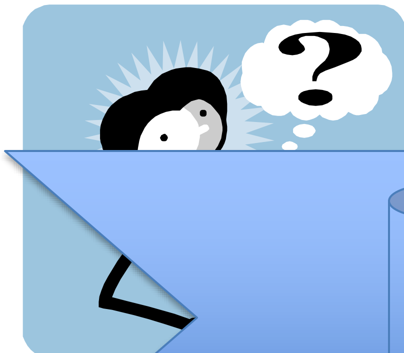
Text Processing/NLP for IR

Vasudeva Varma

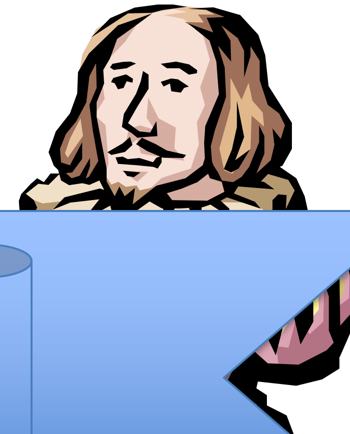
IIIT Hyderabad

The central problem in search

Searcher



Author



IR is hard because natural language is so rich (among other reasons)

Query Terms

“tragic love story”

Document Terms

“fateful star-crossed romance”

Do these represent the same concepts?

how do we represent text?

Remember

- Remember: computers don’t “understand” anything!
- “Bag of words”
  - Treat all the words in a document as index terms
  - Assign a “weight” to each term based on “importance” (or, in simplest case, presence/absence of word)
  - Disregard order, structure, meaning, etc. of the words
  - Simple, yet effective!
- Assumptions
  - Term occurrence is independent
  - Document relevance is independent
  - “Words” are well-defined

what’s a word?

天主教教宗若望保祿二世因感冒再度住進醫院。
這是他今年第二度因同樣的病因住院。

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

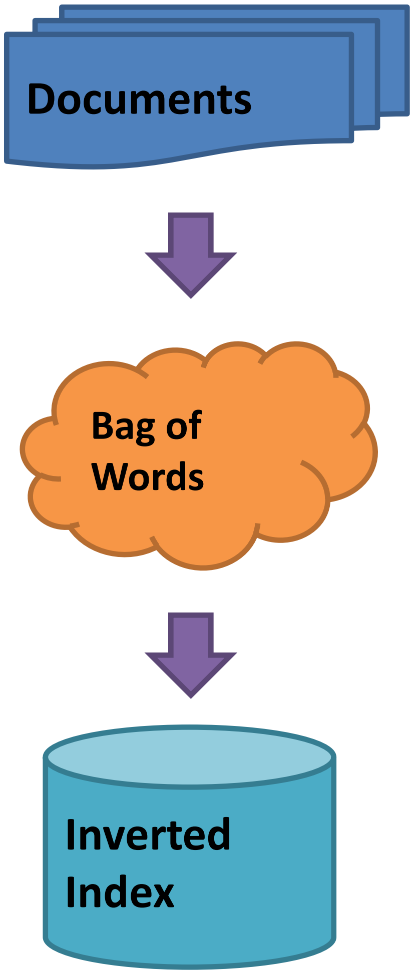
Выступая в Мещанском суде Москвы экс-глава ЮКОСа заявил не совершал ничего противозаконного, в чем обвиняет его генпрокуратура России.

भारत सरकार ने आर्थिक सर्वेक्षण में वित्तीय वर्ष 2005-06 में सात फ़ीसदी विकास दर हासिल करने का आकलन किया है और कर सुधार पर ज़ोर दिया है

日米連合で台頭中国に対処…アーミテージ前副長官提言

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

counting words…



case folding, tokenization, stop word removal, stemming

~~syntax~~, ~~semantics~~, ~~word knowledge~~, etc.

word as an indexing unit

Words = wrong indexing unit!

- 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?

- 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

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

nlp techniques

- Basic (used in most IR/IE systems)
  - *Linguistically motivated, but basic implementations*
  - Tokenizing
  - Stop words
  - Word stemming
- Advanced (sometimes used in IR/IE)
  - *Linguistically motivated, more complex implementations*
  - Phrase/name identification
  - Word sense disambiguation
  - Lexical acquisition
  - Parts of speech
  - Sentence parsing
  - Synonym expansion
  - Anaphoric resolution

natural language understanding

- NLU is a much larger field
  - Semantic interpretation
  - Knowledge representation
    - Logic, frames, …
  - Inference
  - Discourse structure
  - Natural language generation

General Framework of NLP

Slides from Prof. J. Tsujii, Univ of Tokyo and Univ of Manchester

General Framework of NLP

John runs.

Morphological and Lexical Processing

⇓

Syntactic Analysis

⇓

Semantic Analysis

⇓

Context processing Interpretation

General Framework of NLP

John runs.

John run+s.

P-N V 3-pre
N plu

Morphological and Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing Interpretation

General Framework of NLP

John runs.

John run+s.

P-N	V	3-pre
	N	plu

Morphological and Lexical Processing

⇩

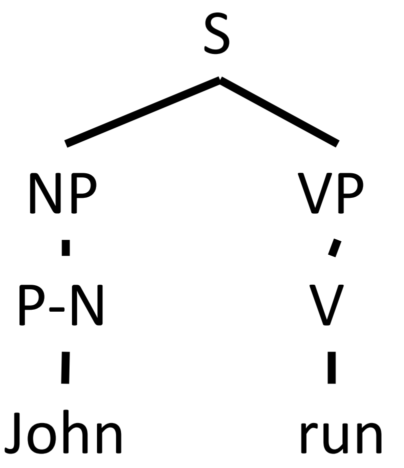
Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing Interpretation



General Framework of NLP

John runs.

Morphological and Lexical Processing

⇩

John run+s.

P-N V 3-pre

N plu

Syntactic Analysis

⇩

S

NP VP

P-N V

John run

Semantic Analysis

⇩

[Pred: RUN
Agent:John]

Context processing Interpretation

General Framework of NLP

John runs.

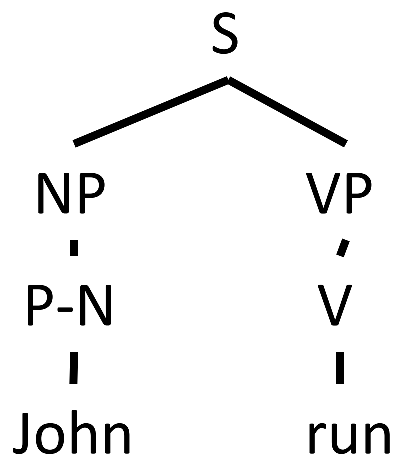
Morphological and Lexical Processing

John run+s.

P-N V 3-pre
N plu

⇩

Syntactic Analysis



⇩

Semantic Analysis

[Pred: RUN
Agent:John]

⇩

Context processing Interpretation

John is a student.
He runs.

General Framework of NLP

Morphological and Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing Interpretation

Tokenization

Part of Speech Tagging

Inflection/Derivation

Compounding

Term recognition

Domain Analysis

Difficulties of NLP

(1) Robustness:
Incomplete Knowledge

General Framework of NLP

Morphological and Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing Interpretation

Difficulties of NLP

(1) Robustness:
Incomplete Knowledge

General Framework of NLP

Morphological and Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing Interpretation

Incomplete Lexicons
- Open class words
- Terms

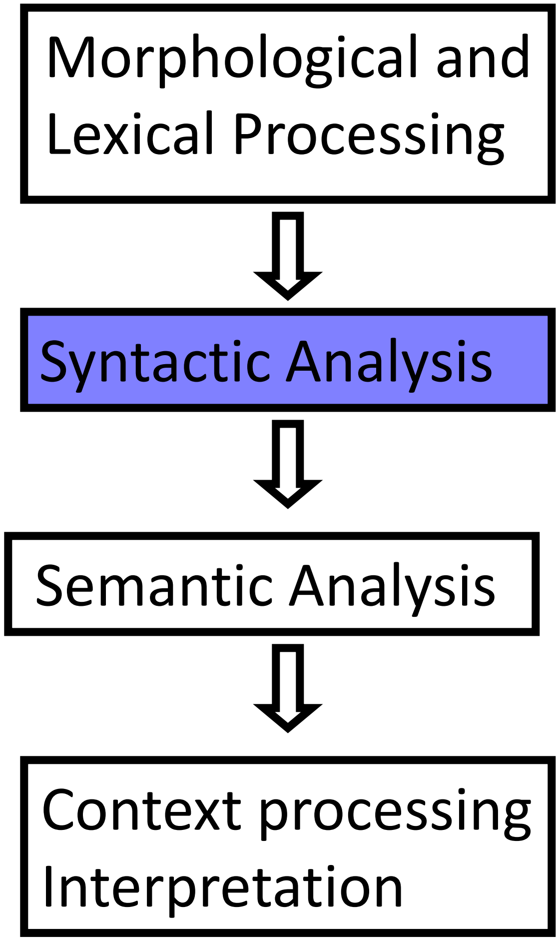
Term recognition

Named Entities
- Company names
- Locations
- Numerical expressions

Difficulties of NLP

(1) Robustness:
Incomplete Knowledge

General Framework of NLP



Incomplete Grammar

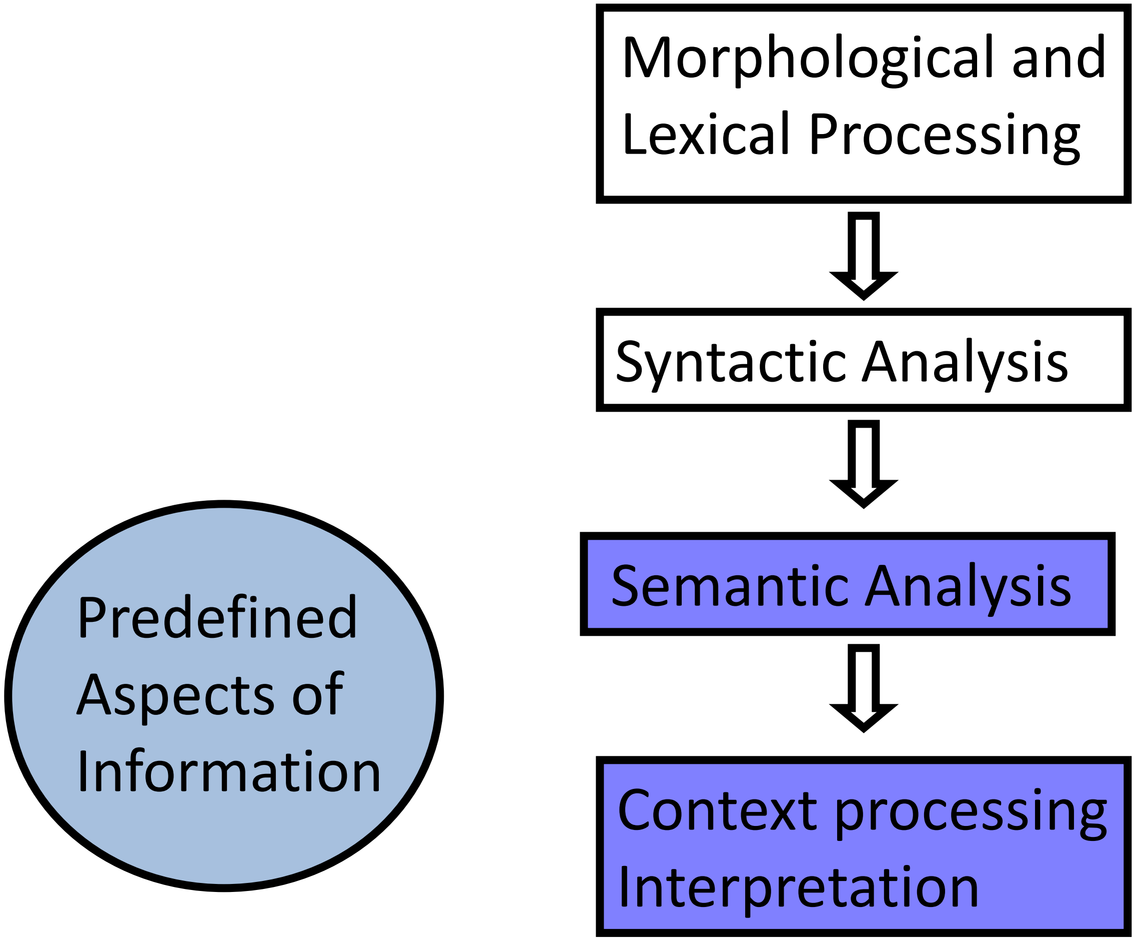
- Syntactic Coverage
- Domain Specific Constructions
- Ungrammatical Constructions

Difficulties of NLP

(1) Robustness:

Incomplete Knowledge

General Framework of NLP



Incomplete

Domain Knowledge

Interpretation Rules

Difficulties of NLP

(1) Robustness:
Incomplete Knowledge

(2) Ambiguities:
Combinatorial
Explosion

General Framework of NLP

Morphological and
Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing
Interpretation

Difficulties of NLP

(1) Robustness:
Incomplete Knowledge

(2) Ambiguities:
Combinatorial
Explosion

General Framework of NLP

Morphological and Lexical Processing

⇓

Syntactic Analysis

⇓

Semantic Analysis

⇓

Context processing
Interpretation

Most words in English are ambiguous in terms of their parts of speech.

runs: v/3pre, n/plu
clubs: v/3pre, n/plu
and two meanings

Difficulties of NLP

(1) Robustness:
Incomplete Knowledge

(2) Ambiguities:
Combinatorial
Explosion

General Framework of NLP

Morphological and Lexical Processing

⇩

Syntactic Analysis — Structural Ambiguities

⇩

Semantic Analysis — Predicate-argument Ambiguities

⇩

Context processing Interpretation

Structural Ambiguities

(1)Attachment Ambiguities

John bought a car with large seats.
John bought a car with $3000.

The manager of Yaxing Benz, a Sino-German joint venture
The manager of Yaxing Benz, Mr. John Smith

(2) Scope Ambiguities

young women and men in the room

(3)Analytical Ambiguities

Visiting relatives can be boring.

Semantic Ambiguities(1)

John bought a car with Mary.
$3000 can buy a nice car.

Semantic Ambiguities(2)

Every man loves a woman.

Co-reference Ambiguities

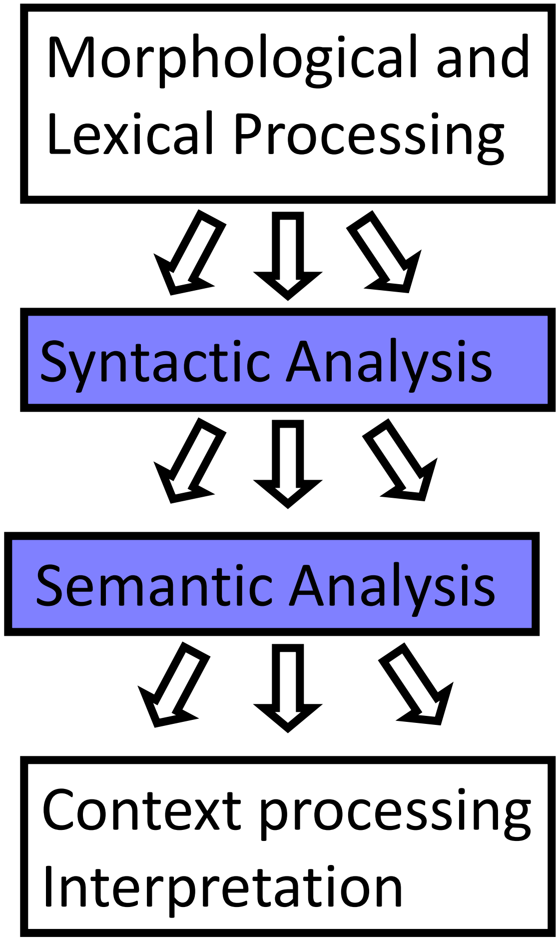
Difficulties of NLP

(1) Robustness:
Incomplete Knowledge

(2) Ambiguities:
Combinatorial
Explosion

Combinatorial Explosion

General Framework of NLP



Structural Ambiguities

Predicate-argument
Ambiguities

stemming, phrase identification, wsd

stemming (morphological roots)

- Stemming is commonly used in IR to conflate morphological variants
- Typical stemmer consists of collection of rules and/or dictionaries
  - Simplest stemmer is “suffix s”
  - Porter stemmer is a collection of rules
  - KSTEM uses lists of words plus rules for inflectional and derivational morphology
  - Similar approach can be used in many languages
  - Some languages are difficult – Indian Languages, Finnish, Arabic etc
- Small improvements in effectiveness and significant usability benefits

stemming (morphological roots)

- Stemming is commonly used in IR to conflate morphological variant~~s~~
- Typical stemmer consist~~s~~ of collection of rule~~s~~ and/or dictionarie~~s~~
  - Simplest stemmer is “suffix s”
  - Porter stemmer is a collection of rule~~s~~
  - KSTEM use~~s~~ list~~s~~ of word~~s~~ plu~~s~~ rule~~s~~ for inflecti derivational morphology
  - Similar approach can be used in many language~~s~~
  - Some language~~s~~ are difficult--e.g., Indian Language~~s~~
- Small improvement~~s~~ in effectivenes~~s~~ and significant usability benefit~~s~~

13/16 =
81.25%

rule-based stemming: porter

- Based on a measure of vowel-consonant sequences
  - measure *m* for a stem is $[C](VC)^m[V]$ where C is a sequence of consonants and V is a sequence of vowels (including y), [] indicates optional
  - m=0 (tree, by), m=1 (trouble,oats, trees, ivy), m=2 (troubl***es***, priv***at***e)
- Algorithm is based on a set of condition action rules
  - old suffix new suffix
  - rules are divided into steps and are examined in sequence
  - e.g., Step 1a: sses ss (caresses caress)
ies i (ponies poni)
s NULL (cats cat)
  - e.g., Step 1b: if m>0 eed ee (agreed agree)
if *V*ed NULL (plastered plaster *but* bled bled)
at ate (conflat(ed) conflate)
- Many implementations available
- Good average recall and precision

dictionary-based stemming

- KSTEM is an example (Krovetz,1993)
- Stems are dictionary headings
  - Consider the entries for word *stocking*
    - V: to put in stock or supplies
      - stocking stock
    - N: a usually knit close-fitting covering for the foot and leg
      - stocking stocking (no change)
  - So in KSTEM, *stocking* would not be stemmed
- For words not in dictionary, fall back on rules like those used by the Porter stemmer
- Most of the time, stems are real words

stemming examples

- Original text:
  Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales

- Porter Stemmer (plus some stopping):
  market strateg carr compan agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale stimul demand price cut volum sale

- KSTEM (plus stopping):
  marketing strategy carry company agriculture chemical report prediction market share chemical report market statistic agrochemic pesticide herbicide fungicide insecticide fertilizer predict sale stimulate demand price cut volume sale

problems with stemming

- Lack of domain-specificity and context can lead to occasional serious retrieval failures (which “stocking” is meant)
- Stemmers are often difficult to understand and modify
- Sometimes too aggressive in conflation
  - e.g., “policy”/“police”, “execute”/“executive”, “university”/“universe”, “organization”/“organ” are conflated by Porter
- Miss good conflations
  - e.g., “European”/“Europe”, “matrices”/“matrix”, “machine”/“machinery” are not conflated by Porter
- Produce stems that are not words and are often difficult for a user to interpret
  - e.g., with Porter, “iteration” produces “iter” and “general” produces “gener”
- Corpus analysis can be used to improve a stemmer or replace it

Discussion Point

stopping

Application: Spelling Correction
Reading: IR book Chapter 3.3

phrase identification

- Goal is to use phrases as indexing units
  - Makes general words more specific
  - blood   blood hound, blood test, blood brother, …
- Statistical approach
  - Index all pairs of adjacent words (“bigrams”)
  - Explosion in index elements makes this non-feasible
  - Also, it adds lots of “nonsense” phrases
    - “also it”, “it adds”, “adds lots”, “lots of”, “of nonsense”, “nonsense phrases”
- NLP approaches
  - Runs of words
  - Sentence parsing
  - Statistical models

phrases as runs of words

- Consider all runs of words between stop words
  - Can easily be extended to allow some stopwords
    - e.g., Library of Congress, cats and dogs
- Scan a large body of text for occurrences of phrases
- Any that occur more than n times are valid
  - Small n (e.g., 4) works impressively well

phrase identification

- Goal is to use phrases as indexing units
  - Makes general words more specific
  - blood   blood hound, blood test, blood brother, …
- Statistical approach
  - Index all pairs of adjacent words (“bigrams”)
  - Explosion in index elements makes this non-feasible
  - Also, it adds lots of “nonsense” phrases
    - “also it”, “it adds”, “adds lots”, “lots of”, “of nonsense”, “nonsense phrases”
- NLP approaches
  - Runs of words
  - Sentence parsing
  - Statistical models

“phrase identification”

- “Goal” is to “use phrases” as “indexing units”
  - Makes “general words” more “specific”
  - “blood” “blood hound”, “blood test”, “blood brother”, …
- “Statistical approach”
  - “Index” all “pairs” of “adjacent words” (“bigrams”)
  - “Explosion” in “index elements” makes this “non-feasible”
- “NLP approaches”
  - “Runs” of “words”
  - “Sentence parsing”
  - “Statistical models”

phrases and counts from trec

65824 United States
61327 Article Type
33864 Los Angeles
18062 Hong Kong
17788 North Korea
17308 New York
15513 San Diego
15009 Orange County
12869 prime minister
12799 first time
12067 Soviet Union
10811 Russian Federation
9912 United Nations
8127 Southern California
7640 South Korea
7620 end recording
7524 European Union
7436 South Africa
7362 San Francisco
7086 news conference
6792 City Council
6348 Middle East
6157 peace process
5955 human rights
5837 White House

5778 long time
5776 Armed Forces
5636 Santa Ana
5619 Foreign Ministry
5527 Bosnia-Herzegovina
5458 words indistinct
5452 international community
5443 vice president
5247 Security Council
5098 North Korean
5023 Long Beach
4981 Central Committee
4872 economic development
4808 President Bush
4652 press conference
4602 first half
4565 second half
4495 nuclear weapons
4448 UN Security Council
4426 South Korean
4219 first quarter
4166 Los Angeles County
4107 State Duma
4085 State Council
3969 market economy
3941 World War II

phrases and counts from u.s. patents

975362 present invention
191625 U.S. Pat
147352 preferred embodiment
95097 carbon atoms
87903 group consisting
81809 room temperature
78458 SEQ ID
75850 BRIEF DESCRIPTION
66407 prior art
59828 perspective view
58724 first embodiment
56715 reaction mixture
54619 DETAILED DESCRIPTION
54117 ethyl acetate
52195 Example 1
52003 block diagram
46299 second embodiment
41694 accompanying drawings
40554 output signal
37911 first end
35827 second end
34881 appended claims
33947 distal end
32338 cross-sectional view
30193 outer surface
29635 upper surface
29535 preferred embodiments
29252 present invention provides
29025 sectional view
28961 longitudinal axis
27703 title compound
27434 PREFERRED EMBODIMENTS
27184 side view
25903 inner surface
25802 Table 1
25047 lower end
25047 plan view
24513 third embodiment
24432 control signal
24296 upper end
24275 methylene chloride
24117 reduced pressure
23831 aqueous solution
23618 SEQUENCE DESCRIPTION
23616 SEQUENCE CHARACTERISTICS
22382 weight percent
22070 closed position
21356 light source
21329 image data
21026 flow chart
21003 PREFERRED EMBODIMENT

phrases from sentence parsing

- Run a shallow or deep parsing system
  - Simplest and common approach uses noun phrases
  - Can use other types, too, of course
    - Verb phrases, noun phrases with adjectives, prepositional phrases, noun+verb phrases, …

phrases from statistical models

- Build a dictionary of phrases using heuristic methods
  - Select High-frequency phrases (with 1-6 words)
  - POS tagging for (relatively?) lower-frequency phrases
    - e.g., throw away verbs or phrases ending with adjectives
- Estimate probabilities for Markov model
  - …that first word is the start of a phrase
  - …that next word is part of the same phrase
  - …that a phrase follows this phrase
  - Done on training data (WSJ 1987)
    - Smoothed for unknown words

named entities

- Perhaps identifying names can help
  - Proper names: Abdul Kalam
  - Place names: Hyderabad
  - Organizations: International Institute of Information Technology
- Various techniques for identifying named entities
  - Simple pattern matching: Mr.([A-Z][a-z]*)+
  - Hand-built or machine-learned rules
  - Hidden Markov models trained on tagged data

entity concept extraction

- More general version of named entity extraction
  - Chemical names
  - Countries, cities, states, provinces, …
  - Titles, dates, dollar amounts, percents, …
  - More general concepts--e.g., “information retrieval”
- Approaches are similar to named entities

anaphora and co-references

- Identifying references to the same object
  - Name resolution: “Ram Nath Kovind” Vs. “Honorable President of India”
  - Anaphora: “He denied all responsibility”, “He kicked it.”
- Techniques
  - Usually require deeper parsing of the text
  - Simple approaches: use closest name or noun phrase

word sense disambiguation

- Index by concept rather than words
- Does it help to disambiguate word senses?
  - Bank as a financial institution, bank as the edge of a river
  - Punch as in validate, punch as in hit, punch as a beverage
- Use NLP to identify the sense of a word
  - punch {punch-validate, punch-hit, punch-beverage}
- Obviously there are some queries it will help
  - Runs on a bank
  - Punch recipes
- But are they common enough that it helps?

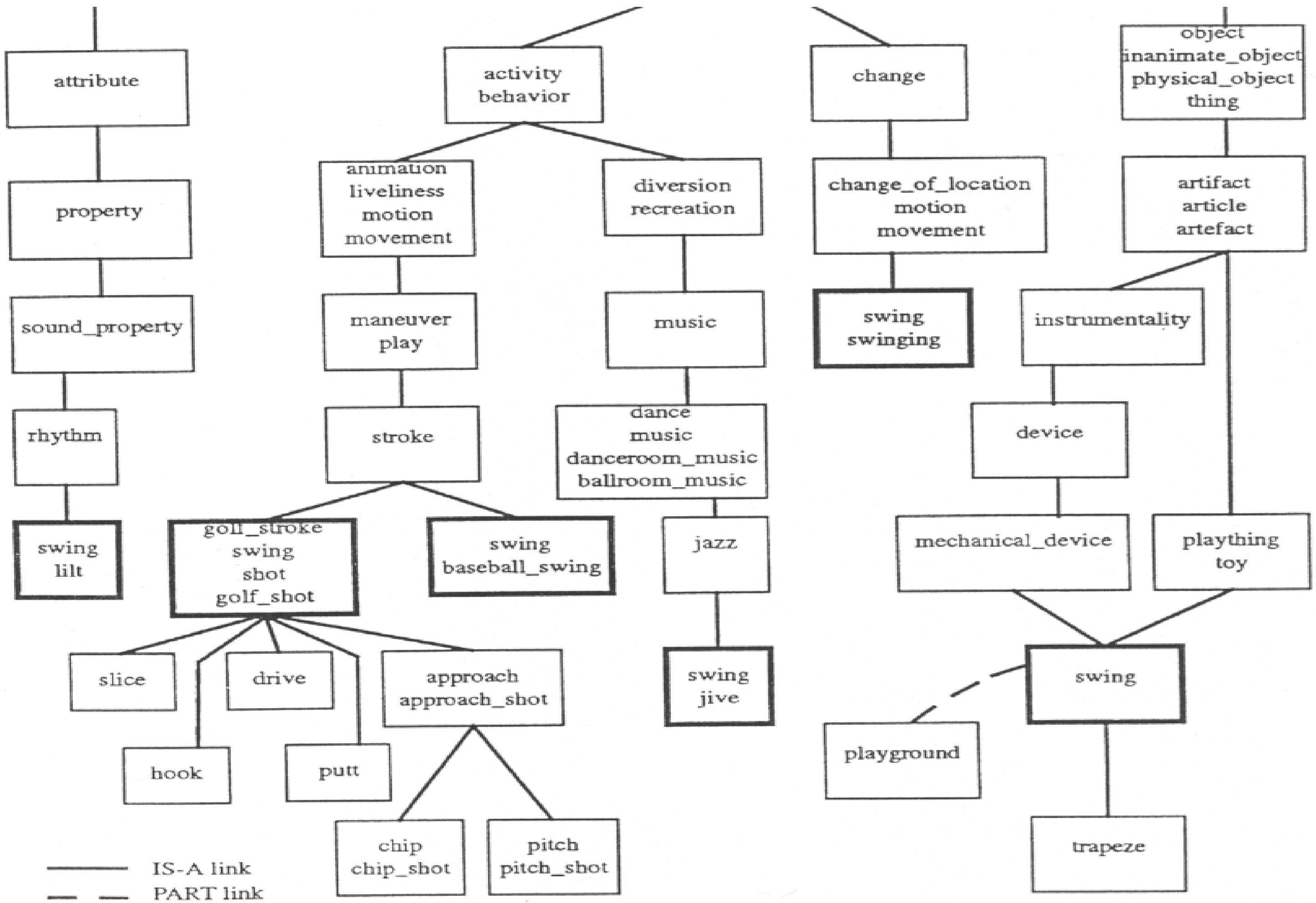
disambiguation experiment (voorhees, 1985)

- Idea: Use WordNet synsets for disambiguation
  - "WordNet® is an on-line ***lexical reference*** system whose design is inspired by current psycholinguistic theories of ***human lexical memory***. English nouns, verbs, adjectives and adverbs are organized into ***synonym sets***, each representing one underlying ***lexical concept***. Different relations link the synonym sets."
  - WordNet was developed by the Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller.
  - https://wordnet.princeton.edu/

synsets - examples

- Synsets are related in various ways
  - hypernym and hyponym (is-a relation) e.g.: (red, color)
  - meronym, holonym (part-of relation) e.g.: (wheel, car)
  - antonym
- Synset for “Calculate”
  - {calculate, cipher, cypher, compute, reckon, figure}
- 23 synsets for “stock”, including
  - broth, stock
  - livestock, stock, farm animal
  - stock certificate, stock
  - stock, gillyflower
  - stock, carry, stockpile (verb)
  - standard, stock (adjective)
- “Natural” has 17 senses
- “Language” has 6 senses
- “Processing” (process) has 8 senses

wordnet relationships for swing



use of synsets

- For each query word, find its synsets
  - Query “punch recipes”
  - punch (3 synsets), recipe (1 synset)
- Expand that synset into its “neighborhood”
  - Grow with WordNet hyponym relationships until any additional growth would include a different sense of any word in the core synset
- To disambiguate words in a document
  - Look at all synset neighborhoods for words in document
  - Compare to the way they overlap throughout collection
  - Choose the neighborhoods where local activity is greater than expected global activity

using synsets for retrieval

- Replace words with their sense-disambiguated form
- Do typical IR from there
- Results show a 6-40% drop in effectiveness
  - Depends on how disambiguated words are compared with non-disambiguated words
    - (Only nouns were disambiguated)
- What went wrong?
  - Different senses chosen when should have been same
  - Insufficient context in a query to select a sense
  - Fortuitous conflation of adjectives and nouns in original is suppressed

is ambiguity really a big problem?

- Consider the query “fly”
  - fly, the insect?
  - fly, the verb? In a plane? Running quickly?
  - fly, a zipper?
- But consider these queries
  - fly airplane, fly buzz, fly pants
- Even a single additional word can disambiguate
  - Note that NLP has no hope of disambiguating a single word
- Documents have many additional words
  - Ambiguity is essentially gone in a full document
  - Queries of moderate length have no ambiguity problem!

what does that suggest?

- Advanced NLP must be nearly perfect to help
- Queries are difficult to process
- Simple word-matching exploits linguistic knowledge
  - Extra words may disambiguate the meaning of words

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
- Words are really the wrong thing to index

thank you

Vasudeva Varma

@devvarma

vv@iiit.ac.in www.iiit.ac.in/~vasu