

CMPT 825

Natural Language Processing




Anoop Sarkar

<http://www.cs.sfu.ca/~anoop>

Natural Language Processing (NLP)

- NLP is the application of a computational theory of human language
- Language is the predominant repository of human interaction and knowledge
- Goal of NLP: programs that “listen in”
- The AI Challenge: the Turing test
- Lots of speech and text data available

Information Retrieval

 [Advanced Search](#) [Preferences](#) [Language Tools](#) [Search Tips](#)
natural language processing
Search:  the web  pages from Canada
[Web](#) [Images](#) [Groups](#) [Directory](#) [News](#)
Searched the web for **natural language processing**. Results 1 - 10 of about 1,830,000. Search took 0.11 seconds.

[The Natural Language Group at ISI](#)

The **Natural Language Processing** group at the Information Sciences Institute of the University of Southern California (USC/ISI) is currently involved in various ...

Description: The **Natural Language Processing** group at the Information Sciences Institute of the University of Southern...

Category: Computers > Artificial Intelligence > ... > Research Groups

www.isi.edu/natural-language/nlp-at-isi.html - 5k - 3 Sep 2003 - Cached - Similar pages

Sponsored Links

Natural Language

Find Solutions for Your Business
Free Reports, Info. & Registration

www.KnowledgeStorm.com

Interest: 

See your message here...

[Natural Language Processing](#)

Natural Language Processing should make it possible for people to use computers in much the same way that they would use a human assistant to get their work ...

Description: Information on their projects, people, publications, and employment opportunities.

Category: Computers > Artificial Intelligence > ... > Research Groups

research.microsoft.com/research/nlp/ - 32k - Cached - Similar pages

[Natural Language Processing](#)

AI-related FAQs. **Natural Language Processing**. ...

www.cs.cmu.edu/afs/cs.cmu.edu/project/ai-repository/ai/html/faqs/ai/nlp/top.html - 1k - 3 Sep 2003 - Cached - Similar pages

[Natural Language Processing on the Web](#)

Ralf Brown's collection of links to **natural language processing** resources, including parsing, understanding, generation, machine translation, linguistics ...

www.cs.cmu.edu/~ralf/nlp.html - 9k - 3 Sep 2003 - Cached - Similar pages

[More results from www.cs.cmu.edu]

[Foundations of Statistical Natural Language Processing](#)

Foundations of Statistical **Natural Language Processing**. ... Chris Manning and Hinrich Schütze, Foundations of Statistical **Natural Language Processing**, MIT Press. ...

nlp.stanford.edu/fsnlp/ - 7k - Cached - Similar pages

Information Extraction

<SUCCESSION-1>

ORGANIZATION : <ORGANIZATION-2>

POST : "president"

WHO_IS_IN : <PERSON-1>

WHO_IS_OUT : <PERSON-2>

<IN> MEDIA (MED), PUBLISHING (PUB) </IN>

<TXT> <p>

<PERSON-1>

New York Times Co. named Russell T. Lewis, 45, president and
general manager of its flagship New York Times newspaper,

residing in New York City. Lewis will
replace

<ORGANIZATION-1>

vice president and deputy general manager

<ORGANIZATION-2>

Lance R. Primis, who in September was named president and chief
operating officer of the company.

</p> </>

<PERSON-2>

Language Summarization

Columbia Newsblaster

Summarizing all the news on the Web

Sunday, August 31, 2003

Articles from 08/28/2003 to 08/31/2003

Last update: 8:09 AM EST

Search for:

Go

in summaries

U.S.
World
Finance
Sci/Tech
Entertainment
Sports

View Today's
Images

View Archives

About Newsblaster

About today's run



Al-Qaida link to Iraq blast reported (World, 46 articles)

The Iraqi Governing Council said today that it would submit a security blueprint to the U.S.-led coalition demanding more control over the day-to-day safety of its citizens, after a car bombing Friday in Najaf that killed at least 95 people. Iraqi police have arrested four al-Qaida-linked suspects in the bombing of Iraq's holiest Shiite Muslim shrine, a senior police official told The Associated Press on Saturday. The official, who said the explosion death toll had risen to 107, said the men two Iraqis and two Saudis were caught shortly after Friday's car bombing. A car bomb ripped through a crowd of worshippers leaving Iraq's holiest Shiite shrine after Friday prayers, killing at least 85 people - including a top cleric - in the deadliest attack since the fall of President Saddam Hussein. Two Iraqis and two Saudis grabbed shortly after Friday's attack gave information leading to the arrest of the others, said the An Najaf police official, speaking on condition of anonymity. Here are extracts from the sermon delivered by the leading Shia Muslim politician, Ayatollah Mohammed Baqr al-Hakim, prior to his death in a car bomb attack in the holy city of Najaf in Iraq

Other stories about Iraq, Al and Iraqi:

- [The U.S. cannot cope with both Iraq and Palestine](#) (9 articles)
- [Poll: U.S. Losing Grip In Iraq](#) (10 articles)
- [A Suspected Operative of Al Qaeda Is Held in Iraq](#) (4 articles)

NLP: Lots of Applications

- Doc classification
- Doc clustering
- Spam detection
- Information extraction
- Summarization
- Machine translation
- Cross Language IR
- Multiple language summarization
- Language generation
- Plagiarism or author detection
- Error correction, language restoration
- Language teaching
- Question answering
- Knowledge acquisition (dictionaries, thesaurus, semantic lexicons)
- Speech recognition
- Text to Speech
- Speaker Identification
- (multi-modal) Dialog systems
- Deciphering ancient scripts

Natural Language: What is it?

- Answers from linguistics: the scientific study of human language

Natural Language (NL) vs. Artificial Language

- Genetic basis of human language
- Mysterially distinct from other species (human language is unique to humans)
- NL is complex, displays recursive structure

Natural Language: What is it?

- Learning of language is an inherent part of NL
- Language has idiosyncratic rules and a complex mapping to thought

For more read [The Great Eskimo Vocabulary Hoax](#) by Geoffrey Pullum

Language has structure

- What he did was climb a tree
- What he ran was to the store
- Drink your beer and go home!
- What are drinking and go home?
- Linus lost his security blanket
- Lost Linus blanket security his

Language is recursive

- This is the house
- This is the house that Jack built
- This is the grain that lay in the house that Jack built
- This is the rat that ate the grain that lay in the house that Jack built
- This is the cat that killed the rat that ate the grain that lay in the house that Jack built
- This is the dog that chased the cat that killed the rat that ate the grain that lay in the house that Jack built

Language is recursive

- Finite resources
- Infinite set of utterances
- Recursion

Facets of Language Structure

- **Phonetics** acoustic and perceptual elements
- **Phonology** inventory of basic sounds (phonemes) and basic rules for combination, e.g. vowel harmony
- **Morphology** how phonemes combine to form words, relationship of phonemes to meaning, e.g. delight-ed vs. de-light-ed
- **Syntax** sentence (utterance) formation, word order and the formation of constituents from word groupings
- **Semantics** how do word meanings recursively compose to form sentence meanings (from syntax to logical formulas)
- **Pragmatics** meaning that is not part of compositional meaning, e.g. *This professor dresses even worse than Anoop!*

Terminology: Grammar

- Grammar can be prescriptive or descriptive
- *Descriptive grammar* is a **model** of the form and meaning of a speaker of a language
- Grammar books for learning a language are *prescriptive grammars*, usually style manuals or rules for how to write clearly
- Except for some NLP apps like grammar checking or teaching, we are usually interested in creating models of language

Terminology: Parts of Speech

- Nouns: John, cow, **can**, tomorrow
- Pronouns: he, she, it, who
- Verbs: run, chase, teach
- Auxiliary verbs: be, **can**, will, might
- Modal verbs: **can**, might
- Determiners: the, a, each, two or more
- Prepositions: in, at, under

More parts of speech

- Adjectives: blue, former
- Adverbs: quickly, certainly
- Coordinating conjunctions: and, but, or
- Complementizers: that, whether, if
- Possessives: 's (Kim 's), whose
- Interjections: Hey!

Grammatical Relations

- Subject-Verb-Object

Kim likes olives

- Subject-Object-Verb

Kim-ka olivu-lul cohanta

Kim-Nom olives-Acc like-Present-Decl

- Modifiers: Kim likes olives on Tuesdays
- Optional arguments: Kim donated olives vs. Kim went to the store

Inflections

- Prefix: un-happy
- Suffix: olive-s
- Different types of prefix or suffix information:
 - Plurals: olive-s
 - Past tense: smash-ed
 - ...

Formal Languages and NLP

| | |
|------------------------------|-------------------------------------------------------------|
| Formal Language Theory | NLP |
| Language (possibly infinite) | Text Data, Corpus (finite) |
| Grammar | Grammar (usually inferred from data, produces infinite set) |
| Automata | Recognition/Generation algorithms |

Some more definitions

- **Classification:** assigning to the input one out of a finite number of classes, e.g.: Document -> spam, **formalization** -> Noun
- **Sequence learning/Tagging:** assigning a sequence of classes, e.g.: **I**/Pron **can**/Modal **open**/Verb **a**/Det **can**/Noun
- **Parsing:** assigning a complex structure, e.g.: formalization -> (Noun (Verb (Adj **formal**) **-ize**) **-ation**)
- **Grammar development:** human driven creation of a model for some linguistic data
- **Transduction:** transforming one linguistic form to another, e.g. summarization, translation, tokenization

Ambiguity: a key problem

- Lung cancer in women mushrooms
 - Mushrooms is noun or a verb?
- Teacher Strikes Idle Kids
 - Strikes is a verb or a noun?
- Two sisters reunited after 18 years in checkout counter
 - Is it reunited in checkout counter or 18 years in checkout counter?
- British Left Waffles on Falkland Islands
 - Is it British/Noun Left/Verb or British Left/Noun Phrase Waffles/Verb?

Ambiguity (cont'd)

- Kids make nutritious snacks
 - **make** can mean different things, which is it?
- Iraqi Head Seeks Arms
 - **Arms** can mean different things, which is it?
- Two Soviet Ships Collide, One Dies
 - What does **one** refer to in this case?
- Chef throws his heart into feeding needy
 - **Throws his heart** is not decomposed normally in this case: idiom finding