CS544: Natural Language Processing

January 11, 2011

Zornitsa Kozareva
USC/ISI
Marina del Rey, CA
kozareva@isi.edu
www.isi.edu/~kozareva

The Dream

• It would be great if machines could
  – Process our emails
  – Translate languages accurately
  – Help us manage, summarize, and aggregate information
  – Understand phone conversation
  – Talk to us / listen to us

• But they cannot:
  – Language is complex, ambiguous, flexible, and subtle
  – Good solutions need linguistics and machine learning knowledge

What is NLP?

• Goal: intelligent processing of human language
  – Not just string and keyword matching

• End systems we want to build:
  – Less ambitious: spelling correction, name entity extractors
  – Ambitious: machine translation, information extraction, question answering, summarization …
Information Extraction

- Goal: build database entries from unstructured text
- Simple Task: Named Entity Extraction

A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb—allegedly detonated by urban guerrilla commandos—blew up a power tower in the north western part of San Salvador at 0650.

Incident type: bombing
Date: March 11, 2010
Location: San Salvador (city)
Perpetrator: urban guerrilla commandos
Physical target: power tower
Effect on physical target: destroyed
Effect on human target: no injury or death
Instrument: bomb

Information Retrieval

- Given a huge collection of text and a query
- Goal: find documents that are relevant to the query
Question Answering

- Find answers to general comprehension questions in a document collection

Text Summarization

Machine Translation

Japón comprará el 20% de los bonos japoneses a fines de mes

Japan buys 20% of the eurozone bond

used Google Translate
Speech Processing

• Automatic Speech Recognition

  “will you move the clinic there?”

• Performance: 5% for dictation, 50%+TV

Linguistics Levels of Analysis

• Phonology: sounds / letters / pronunciation
• Morphology: construction of words
• Syntax: structural relationships between words
• Semantics: meaning of strings (words, phrases)
• Discourse: relationships across different sentences
• Pragmatics: how we use language to communicate
• World Knowledge: facts about the world, common sense

MORPHOLOGY
Morphological Analysis

- *Morphology* studies the internal structure of words
- A *morpheme* is the smallest linguistic unit that has semantic meaning (Wikipedia)
- *Morphological Analysis* is the task of segmenting a word into its morphemes
  - *carried* => carry + ed (past tense)
  - *disconnect* => dis (not) + connect
- Challenging for morphologically rich languages like Finish and Turkish

SYNTACTIC TASKS

Part-of-Speech Tagging (POS)

- Annotate each word in a sentence with a part-of-speech tag
  
  I ate the spaghetti with meatballs.
  
  Pro V Det N Prep N

- Useful for syntactic parsing and word sense disambiguation
- English POS tagging 95% accurate
Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.

[NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].

Syntactic Parsing

- Produce syntactic parse tree of a sentence

- Help figuring out questions like: Who did what and when?

More issues in Syntax

- Prepositional Attachment
  "I saw the man with the telescope"

Syntax does not tell us much about meaning
SEMANTIC TASKS

Word Sense Disambiguation

• Understand language! How?

I walked to the bank ... of the river.
  to get money.

• Useful for machine translation, information retrieval

How to learn the meaning of words?

• From dictionaries, lexical repository like WordNet
  bank -- sloping land, especially the slope beside a body of water
  ex. "they pulled the canoe up on the bank"
  bank – a financial institution that accepts deposits and channels
  the money into lending activities
  ex. "he cashed a check at the bank"

• Automatically from the Web
Semantic Role Labeling

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb

  agent  patient  source  destination
  John  drove  Mary  from  LA  to  San Diego.

Textual Entailment

- Determine whether one natural language sentence entails another

  The glass is half empty.
  The glass is half full.

  Google bought Youtube.
  Google acquired Youtube.
Anaphora Resolution

• Determine which phrases in a document refer to the same entity

“George woke up. He went to the kitchen.”
“Peter put the carrot on the plate and ate it.”

Pragmatics

• Studies how language is used to accomplish goals

What can we conclude from the following sentences?
“Could you please pass me the salt?”
“I am afraid I cannot do this”

World Knowledge

“George woke up. He went to the bathroom and started shaving. He took the car key and left.”
WHERE WE STAND TODAY

What cannot NLP do today?

• Do general-purpose text generation
• Deliver semantics—either in theory or in practice
• Deliver long/complex answers by extracting, merging, and summarizing web info
• Handle extended dialogues
• Read and learn (extend own knowledge)
• Use pragmatics (style, emotion, user profile...)
• Provide significant contributions to a theory of Language (in Linguistics or Neurolinguistics) or of Information (in Signal Processing)

What can NLP do (robustly) today?

• Surface-level preprocessing (POS tagging, word segmentation, named entity extraction): 94%+
• Shallow syntactic parsing: 92%+ for English
• IE: ~40% for well-behaved topics (MUC, ACE)
• Speech: ~80% large vocab; 20%+ open vocab, noisy input
• IR: 40% (TREC)
• MT: ~70% depending on what you measure
• Summarization: ? (~60% for extracts; DUC)
• QA: ? (~60% for factoids; TREC)
CLASS DETAILS

What is in this Class?

• Some linguistic basics
  – structure of English
• Syntactic parsing
• Semantics
  – Word sense disambiguation
  – Semantic relations
• Applications:
  – Information Extraction
  – Machine Translation
  – Question Answering
  – Speech Recognition
  – Text Summarization

Class Requirements and Goals

• Class requirements:
  – Basic linguistics background
  – Basic probability and statistics
  – Decent coding skills

• Class goals:
  – Learn issues and techniques in NLP
  – Learn about applications that can benefit from NLP
  – Understand issues involved in processing natural language
  – Develop skills necessary to build NLP tools
Course Work

• Recommended Readings:

• Assignments:
  – 3 coding assignments
    • late submissions will not be accepted
    • brief 1-2 paged description
    • power point presentation
  – 1 final project
Ph.D. Researchers and Topics

At ISI:
- David Chiang — parsing, statistical processing
- Ulf Hermjakob — parsing, QA, language learning
- Jerry Hobbs — semantics, ontologies, discourse
- Edward Hovy — summarization, ontologies, NLU, MT
- Liang Huang — parsing, MT
- Kevin Knight — MT, NLU, encryption
- Zornitsa Kozareva — IE, text mining, lexical semantics
- Daniel Marcu — MT, QA, summarization, discourse
- Donald Metzler — IR
- (Patrick Pantel — clustering, ontologies, learning by reading)

At ICT:
- David DeVault — NL generation
- Andrew Gordon — cognitive science and language
- Anton Usavl — IR
- Kenji Sagae — parsing
- Bill Lockett — NLG
- David Traum — dialogue

At USC/EE:
- Shri Narayanan — speech recognition