Introduction to NLP

Lecture #2
What will we cover today?

1. What is NLP?
2. Some application areas of NLP
3. A brief history of NLP
4. Famous NLP systems
5. Ambiguity and NLP
6. Overcoming ambiguity
   – Brief intro to machine learning/statistical methods
7. NLP and Social Computing – how are they related?
What is CL/NLP/NLU/HLT?

• Computational Linguistics
  – Language studies using computers
  – Understanding structure and usage of language

• Natural Language Processing
  – Computer analysis of structure and meaning
  – Speech transcription, translation, extraction

• Natural Language Understanding/Generation
  – Computer understanding of human language
  – Computer generation of human language

• Human Language Technology
  – Any of the above
NLU by computer

Natural language → COMPUTER → formal representation

NLU

NLG
Application Areas

• Information extraction
• Text summarization
• Story comprehension
• Question answering
• Speech recognition
• Machine translation
• Human-computer dialogue
• Others?
Information Extraction

• Goal: Pull out/highlight and classify references to various entities out of text
  – People, organizations, locations, artifacts
  – Relations between these

• Application: automated database population

EXAMPLE 1... *Henry Kaufman* is president of *Henry Kaufman & Co.*, a ...

EXAMPLE 2

*The Drum* @TheDrum · Dec 2
Trump turns to Periscope to boost (or save) his presidential hopes
goo.gl/26yiyB
Text Summarization

- Summarize content of text/media documents
- Informative vs. Indicative
- Generic vs. Topical
  - main content vs. topic-related
- Single text vs. Cross-document
  - topic & aspect detection
  - topical briefs, topic tracking
- Info Organization & Visualization
  - Multiple views of info space
  - Rapid comprehension
WAKEFIELD, Mass. (AP) -- The software engineer accused of gunning down seven co-workers stashed ammunition and a semiautomatic with a sniper scope in his office and bomb-making material in his home, authorities said Wednesday.

Michael M. McDermott, 42, pleaded innocent Wednesday to seven counts of murder and was ordered held without bail. [...] Wearing an orange jail jumpsuit and a bulletproof vest, the manacled and bearded McDermott stood impassively in court, his bushy dark hair draped over his shoulders. He looked around the courtroom frequently as prosecutor Tom O'Reilly described Tuesday's carnage: The shooter blasted through the offices of Edgewater Technology with 37 rounds from a semiautomatic rifle and several from a shotgun, striking workers in their heads and backs as they tried to flee. [...] Police found McDermott sitting silently in the reception area, a body nearby, his weapons within reach. He was arrested without gunfire. Officers who searched McDermott's work area Wednesday at the Internet consulting company found the ammunition in a cubby hole at his desk and shotgun shells in the trash basket, O'Reilly said. [...] Some of the four women and three men killed worked in the accounting department, which was recently ordered by the Internal Revenue Service to seize a portion of McDermott's wages. Last week, he had an

A software engineer who allegedly gunned down seven co-workers at their Internet consulting company may have been upset by a request to trim his wages to repay back taxes after the holidays, authorities said.
Story/Reading Comprehension

• Mapping a story on real world
  – More than one sentence
  – “Obvious” facts not mentioned
  – Must supply extra-linguistic knowledge

• A little story:
  – My shoe has a hole, Bob says. My foot is wet
  – Why did Bob’s foot get wet?

• Explanation – in a micro-world
  – Bob’s foot is wet because his shoe has a hole and he walked through water.
Question Answering

• Supply actual answers to user questions
  – How long does it take to fly from Paris to New York on a Concorde?
  – 3.5 hours

• Find relevant information:
  – In text, websites, etc.
  – In formatted databases, CIA World Factbook

• Extract information, convert into an answer

• Ranges from “trivia” to research problems

• IBM Watson Jeopardy Challenge
Question: What were the recent disasters that occurred in tunnels used for transportation?

Clarification Dialogue:
S: Are you interested in train accidents, automobile accidents or others?
U: Any that involved lost life or a major disruption in communication. Must identify losses.
Machine Translation

• Goal: translate text/speech in one language
  – E.g., Russian, Arabic, French, Chinese, Thai

• Into an equivalent in another language
  – E.g., English, Spanish, Japanese, Urdu, Polish

• Many applications
  – Such as?

• An extremely hard problem in general case
  – Why?
Machine-Lost in Translation?

• An infamous example (most likely bogus)
  
  the spirit indeed is willing, but the flesh is weak
  
  дух охотно готов но плоть будет неделей
  
  The vodka is good, but the meat is rotten

• Another infamous example (not bogus)
  
  Earth girls are easy
  
  Masse Mädchen sind einfach
  
  Mass girl are simple
Speech Recognition

• Transcribe spoken language into text
  – Equivalent of dictation, note taking, commands
  – Indispensable in hands-busy/eye-busy situations

• Varying conditions
  – Microphone, telephone, noise, meetings

• How accurate?
  – Depends upon application and what is said – why?

• Most interesting possibilities in H-C dialogues
Human-Computer Dialogue

• Communicating with a computer using NL commands
  – Typing, speaking

• Interacting with an artificial agent
  – Call center, help desk
  – Interactive programs that respond in NL

• Conversing like with a human
  – Artificial companions
A brief history of NLP: Part I

• 1947–54 some pioneering work in MT
  – Weaver, Booth – translating Russian scientific texts
• 1954–66 decade of optimism
  – Lots of simple approaches to MT
  – 1957 Chomsky publishes Syntactic Structures
• 1966 ALPAC report concludes MT not achievable
  – virtual end to MT research in the US
• 1970’s Post ALPAC trauma
  – Many theoretical frameworks emerge
  – Schank, Wilks, Fillmore, Quillian
A brief history of NLP: part II

- **1980’s emergence of new approaches:**
  - stochastic paradigm – large corpora
  - logic-based semantics, discourse modeling
  - domination of philosophical, symbolic school
    - e.g., Situation Semantics

- **1990 – two predominant approaches:**
  - stochastic NLP
  - unification-based grammars

- **2000 – practical applications**
  - Question Answering, Information Extraction
  - Human-Computer dialogue systems

- **2010 – Big Data, Deep learning**
  - Dramatic advances in speech recognition, MT
Some famous NLP systems

- Green’s BASEBALL - NLDB
- Woods et al.’s LUNAR – NLDB
- Winograd’s SHRDLU – Block World
- Grosz’ Task-oriented dialogues
- McKeown’s TEXT and McDonald’s MUMBLE – NLG
- Schank’s SAM – Story Understanding
- Lehnert’s QUALM
- Weizenbaum’s ELIZA
- Colby’s PARRY
ELIZA

• WHO: Weizenbaum (1966)
• WHAT: replicate the conversation between a psychotherapist and a patient
• HOW: by permuting or echoing the user input it created (momentarily) an impression of human-human conversation.
• FAMOUS FOR: nearly passing the Turing Test
ELIZA: example

U: men are all alike
E: what resemblance do you see
U: my boyfriend made me come here
E: Why do you say your boyfriend made you come here?
U: He thinks I am depressed most of the time
E: Do you think coming here will help you not to be depressed?
U: You sound like my father
E: tell me more about your family
PARRY

- WHO: Colby, 1972
- WHAT: simulated a paranoid schizophrenic
- HOW: implemented early conversational strategy, not just reflection of user input
- FAMOUS FOR: “talking” to ELIZA over ARPANET
PARRY: Example

parry: People get on my nerves sometimes.
eliza: I am not sure I understand you fully.
parry: You should pay more attention.
eliza: Suppose you should pay more attention?
parry: You're entitled to your opinion.
eliza: What makes you think I'm entitled to my opinion?
parry: It bothers me just to be around people.
eliza: Please go on.
State of the Art

- **Information extraction**
  - NE extraction – excellent if trained
  - Relationships/events – not as good as entities

- **Machine translation**
  - Narrow domains – workable
  - Open domain – ok, Google Translate (especially newest deep learning methods)

- **Speech recognition**
  - Dictation, trained to user – good
  - Open vocabulary, noise – getting there (Siri) – also here neural approaches

- **Dialogue systems**
  - Fixed domain – ok, but tedious
  - Data driven – getting there

- **QA** – good, ~70% average
Additional levels of language

- *Communicative Action* – using language to reach understanding or complete a task
  – This is what we would be concerned with

- *Social Action* – using language to influence other participants of a discourse
  – Overpower, be polite, oppose, etc.

- *Social Roles* – using language to establish a role in discourse
  – e.g., leader, opposition, spoiler, etc.
Why is NLP hard?

“At last, a computer that understands you like your mother”
Reason #1: Ambiguity

“At last, a computer that understands you like your mother”

1. It understands you as well as your mother understands you
2. It understands (that) you like your mother
3. It understands you as well as it understands your mother

• 1 and 3: Does this mean well, or poorly?
More ambiguity

• If the sentence is spoken it may be heard as:

1. “... a computer that understands you like your mother”
2. “... a computer that understands you lie cured mother”

• This is ambiguity at the acoustic level
  – speech recognition perplexity

• Why it did not occur to us?
Still more ambiguity

Different structures lead to different interpretations.

• This explains one reason for ambiguity
• But how real is it?
Semantic Ambiguity

• Two definitions of “mother”

1. a woman who has given birth to a child
2. a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar

• This is an instance of word sense ambiguity

• Ah, but why do we care?
Yet more ambiguity

Alice says they’ve built a computer that understands you like your mother but she doesn’t know any details.

Alice says they’ve built a computer that understands you like your mother but she doesn’t understand me at all.

- This is an instance of referential ambiguity
  - Involving resolution of an anaphor
  - At the discourse level
Where does it come from?

• Lack of linguistic context
  – Sentences occur as part of text, dialogue, book, ...
  – Yes, but we still can understand them correctly, even out of context... how come?

• Lack of commonsense knowledge
  – Sentences occur as part of situations, lectures, conversations... that supply extra-linguistic context
  – Yes, but we still can understand them correctly, even out of context... how come?
How to encode context into NLP?

Two possible solutions:

• **Symbolic approach**
  – Encode all the required information into computer
  – Knowledge about the world, sensors, learning...

• **Statistical approach**
  – Infer language properties from language samples
  – Learn to understand language by observing how people do it
A specific example

• Task: determine placement of articles (a, the, or none) in English text:

• Natural Language Processing (NLP) is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies. And, being a very active area of research and development, there is not a single agreed-upon definition that would satisfy everyone.
A Symbolic Approach

• Write explicit rules for article placement:
  1. Type of noun (countable, uncountable)
  2. Reference (specific, generic)
  3. Information value (given, new)
  4. Number (singular, plural)

• Then add more rules to handle exceptions
  – “The” is used with newspaper titles (*The Times*),

• And exceptions to exceptions...
  – No article used in names of magazines (*Time*)
A naïve statistical approach

• Collect a large collection of texts relevant to your domain (e.g., newspaper text)
• For each noun seen during training, compute its probability to take a certain determiner

\[ p(\text{determiner}|\text{noun}) = \frac{\text{freq}(\text{noun,determiner})}{\text{freq}(\text{noun})} \]

• Given a new noun, select a determiner with the highest likelihood as estimated on the training corpus
Does it work?

• Training:
  – A corpus of Wall Street Journal (WSJ) news
  – Approx. 25 million words

• Testing:
  – Set aside section of approx. 1 million words
  – Prediction accuracy: 71.5%

• Not great, but surprisingly high for such a simple method

• Advantages:
  – Many nouns always appear with the same determiner
    • “the FBI”, “the defendant”, . . . -- easily learned

• Disadvantages:
  – Would it hold up to other types of text?
  – What about unseen words?
A better statistical approach

• Learn generalized rules using features of nouns
  – $F = \{\text{noun, number, occurrence, countable, ...}\}$

<table>
<thead>
<tr>
<th>Noun</th>
<th>plural?</th>
<th>first appearance</th>
<th>determiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>defendant</td>
<td>no</td>
<td>yes</td>
<td>the</td>
</tr>
<tr>
<td>cars</td>
<td>yes</td>
<td>no</td>
<td>null</td>
</tr>
<tr>
<td>FBI</td>
<td>no</td>
<td>no</td>
<td>the</td>
</tr>
<tr>
<td>concert</td>
<td>no</td>
<td>yes</td>
<td>a</td>
</tr>
</tbody>
</table>

Learn classification function $D: F \rightarrow \{a, \text{the}, \text{null}\}$
Where else statistical methods?

• Speech recognition
  – Transforming signals to words

• Parsing
  – Transforming word strings to trees

• Machine Translation
  – strings in language X to strings in language Y

• Natural Language Generation
  – Internal representations to strings
Learning to parse

• Training: parallel corpus of sentences and parse trees
  – Penn Treebank = 50,000 sentences with associated trees
  – Usual set-up: 40,000 training sentences, 2400 to test
Learning to translate

• Training: parallel text in 2 languages
  – UN, Canadian Parliament, EU, Hong Kong
  – Human translations – literature, news, science

• Align: pretend target language an elaborate “code” – code breaking techniques

Он благополучно избегнул встречи с своей хозяйкой на лестнице.
He had successfully avoided meeting his landlady on the staircase.
Supervised vs. unsupervised

• Supervised learning = training data required
  – Significant burden to construct training corpus
  – Parsing, MT use supervised methods
• Unsupervised learning = no training data
  – Learn by discovering relationships in data
  – Word segmentation (Chinese), semantic classes
• Semi-supervised learning = bootstrapping
  – Learn initial rules from annotated data
  – Discover more rules by out-propagation
Vector Representation

- We can represent words by the contexts in which they occur
  - “You shall know a word by the company it keeps” J.R. Firth, 1957
  - In other words: a word meaning is decided by its context
- A context can be represented as a vector
  - A vector of documents a word occurs in (classic IR)
  - A vector of paragraphs
  - A vector of “context words”: k-word window around a word
    - So-called word embedding
- Common techniques
  - Latent semantic indexing – for reduced dimensionality space
    - Group similar meaning words together into one dimension
  - Neural networks
    - E.g., word2vec is 1-hidden layer NN that produces linear word semantic space
Word embedding

• The idea is to convert the space of words into a space of meanings
  – Assumption: there are far fewer meanings than words to describe them
• One-hot vectors represent word co-occurrences in language
  $x\rightarrow = [0 \ 0 \ ... \ 1 \ ... \ 0]$ – this has $N=10,000$ dimensions = vocabulary size
  the $i$-th bit is set if word $x$ occurs near $w$ (e.g., within $m$ words of $w$)
• We wish to learn an Embedding Matrix $W (N \times k)$ that represents all contexts where $w$ was found
  – $k$ is “small”, usually between 200 and 300
  – $W$ is seeded with random values that are updated as learning progresses
• Then $x\rightarrow * W = h\rightarrow$ is a vector representing $x$ in the low-dim embedding space
  – Since $x\rightarrow$ is one-hot vector, only one column from $W$ is extracted
• We need to learn $W$ so that for the word $w$ we have:
  – $h\rightarrow * w\rightarrow = P(x \text{ is near } w)$ – where $w\rightarrow$ we get from the data!
  – We update $W$ until this equation holds.
Word2vec: 1-hidden layer NN

Input Vector

A ‘1’ in the position corresponding to the word “ants”

Output Layer
Softmax Classifier

Hidden Layer
Linear Neurons

Probability that the word at a randomly chosen, nearby position is “abandon”

... “ability”

... “able”

... “zone”

10,000 positions

300 neurons

10,000 neurons
NLP and Social Computing/CSS

• Language provides great insight
  – Plentiful, relatively easy to handle
  – People like to talk, interact

• NLP used to extract plenty of information from language
  – Sociolinguistic structure of discourse
  – How this structure shapes opinions and beliefs
  – What is the role of cultural context
  – What are speakers’ attitudes, beliefs, and how they change
  – Social roles and relations between participants
What we covered today?

1. What is NLP?
2. Some application areas of NLP
3. A brief history of NLP
4. Famous NLP systems
5. Ambiguity and NLP
6. Overcoming ambiguity
   – Brief intro to machine learning/statistical methods
7. NLP and Social Computing – how are they related?
Assigned Reading: Paper 2


• Responses due 11:59 pm February 6th, 2018

• Discussion slides due 11:59 pm February 7th, 2018