Natural Language Processing

(adapted from Jim Martin)

1/6/2009

What’s this story about?

17 the
13 and
10 of
10 a
8 to
7 a
6 in
6 Romney
6 Mr
5 that
5 state
5 for
4 industry
4 automotiv e
4 Michigan
3 on
3 his
3 have
3 are
2 would
2 with
2 up
1 think
2 speech
2 primary
2 neck
2 is
2 further
2 fuel
2 from
2 former
2 energy
2 campaigning
2 billion
2 bill
2 at
2 They
2 Senator
2 Republican
2 Monday
2 McCain
2 He
2 Gov
1 wrong
1 who
1 upon
1 unions
1 unfunded
1 ultimately
1 trade
1 top
1 took
1 together
1 throughout
1 they
1 there
1 task
1 it
1 support
1 successful
1 standards
1 one
1 signed
1 shake
1 set
1 science
1 said
1 rise
1 research
1 requires
1 representative
1 remarkably
1 recent
1 rebuild
1 raising
1 including
1 pushed
1 imposing
1 president
1 him
1 polls
1 heavily
1 policy
1 has
1 pledged
1 greenhouse
1 plan
1 gas
1 people
1 future
1 or
1 forever
1 off
1 focused
1 measure
1 flurry
1 materials
1 fluid
1 mandates
1 first
1 losses
1 final
1 litany
1 field
1 leading
1 federal
1 leadership
1 lawmakers
1 essentially
1 emphasizing
1 back
1 automobile
1 economics
1 jobs
1 efficiency
1 job
1 economic
1 its
1 don
1 indicated
1 do
1 independent
1 disinterested
1 again
1 development
1 advisers
1 acknowledged
1 With
1 Washington
1 There
1 Recent
1 President
1 New
1 Mitt
1 Michigan
1 Lieberman
1 Joseph
1 John
1 Iowa
1 In
1 I
1 Huckabee
1 Hampshire
1 Economic
1 Detroit
1 Connect
1 Congress
1 Club
1 Bush
1 Arkansas
1 Arizona
1 America
Detroit — With economic issues at the top of the agenda, the leading Republican presidential candidates set off Monday on a final flurry of campaigning in Michigan ahead of the state’s primary that could again shake up a remarkably fluid Republican field.

Recent polls have indicated the contest is neck-and-neck between former Gov. Mitt Romney of Massachusetts and Senator John McCain of Arizona, with former Gov. Mike Huckabee of Arkansas further back.

Mr. Romney’s advisers have acknowledged that the state’s primary is essentially do-or-die for him after successive losses in Iowa and New Hampshire. He has been campaigning heavily throughout the state, emphasizing his childhood in Michigan and delivering a policy speech on Monday focused on aiding the auto industry.

In his speech at the Detroit Economic Club, Mr. Romney took Washington lawmakers to task for being a “disinterested” in Michigan’s plight and imposing upon the state’s automakers a litany of “unfunded mandates,” including a recent measure signed by President Bush that requires the raising of fuel efficiency standards.

He criticized Mr. McCain and Senator Joseph I. Lieberman, independent of Connecticut, for a bill that they have pushed to cap and trade greenhouse gas emissions. Mr. Romney asserted that the bill would cause energy costs to rise and would ultimately be a “job killer.”

Mr. Romney further pledged to bring together in his first 100 days representatives from the automotive industry, unions, Congress and the state of Michigan to come up with a plan to “rebuild America’s automotive leadership” and to increase to $20 billion, from $4 billion, the federal support for research and development in energy, fuel technology, materials science and automotive technology.

---

The story

Romney Battles McCain for Michigan Lead
By MICHAEL LUO

Vector Representations

- The first slide was a basic vector representation for the meaning of a text
  - Also known as a “bag of words” representation
- Discourse segments, sentence boundaries, syntax, word order are all ignored.
- Roughly, all that matters is the set of words that occur and how often they occur
Vector Representations

• These representations are the basis for many interesting and useful systems
• Hypothesis: there has to be something better.
• Much of NLP is directed at finding representations that do a better job at capturing the meaning and intent behind texts.

Natural Language Processing

• What is it?  
  ◦ Getting computers to perform useful and interesting tasks involving human languages.  
  ◦ Secondarily concerned with the insights that such computational work gives us into human processing of language.
Why Should You Care?

Two trends
1. An enormous amount of knowledge is now available in machine readable form as natural language text
2. Conversational agents are becoming an important form of human-computer communication

Major Topics
1. Words
2. Speech
3. Syntax
4. Meaning
5. Discourse
6. Applications exploiting each
Applications

• First, what makes an application a language processing application (as opposed to any other piece of software)?
  • An application that requires the use of knowledge about human languages
    ▪ Example: Is Unix wc (word count) an example of a language processing application?

Applications

• Word count?
  • When it counts words: Yes
    ▪ To count words you need to know what a word is. That’s knowledge of language.
  • When it counts lines and bytes: No
    ▪ Lines and bytes are computer artifacts, not linguistic entities
Big Applications

- Question answering
- Conversational agents
- Summarization
- Machine translation

Big Applications

- These kinds of applications require a tremendous amount of knowledge of language.
- Consider the following interaction with HAL the computer from 2001: A Space Odyssey
HAL from 2001

- Dave: *Open the pod bay doors, Hal.*
- HAL: *I’m sorry Dave, I’m afraid I can’t do that.*

What’s needed?

- Speech recognition and synthesis
- Knowledge of the English words involved
  - What they mean
- How groups of words clump
  - What the clumps mean
What’s needed?

- Dialog
  - It is polite to respond, even if you’re planning to kill someone.
  - It is polite to pretend to want to be cooperative (I’m afraid, I can’t...)

Real Example

What is the Fed’s current position on interest rates?

- What or who is the “Fed”?
- What does it mean for it to have a position?
- How does “current” modify that?
Caveat

NLP has an **AI** aspect to it.

- We’re often dealing with ill-defined problems
- We don’t often come up with perfect solutions/algorithms
- We can’t let either of those facts get in our way

Topics: Linguistics

- Word-level processing
- Syntactic processing
- Lexical and compositional semantics
- Discourse processing
**Topics: Techniques**

- Finite-state methods
- Context-free methods
- Augmented grammars
  - Unification
  - Lambda calculus
- First order logic

- Probability models
- Supervised machine learning methods

**Topics: Applications**

- Small
  - Spelling correction
  - Hyphenation
- Medium
  - Word-sense disambiguation
  - Named entity recognition
  - Information retrieval
- Large
  - Question answering
  - Conversational agents
  - Machine translation

- Stand-alone
- Enabling applications
- Funding/Business plans
Killing Palestinians and wounding nine in the Gaza Strip
Nine Palestinians were wounded among civilians in an Israeli air raid in the neighboring result in the Gaza Strip. This comes immediately after the killing of two Palestinians by an Israeli army mortar attack on the border. Israeli occupying forces carried out air and infantry forces in the Shab Hanna camp in the West Bank.

Bashir meets Fraser, the Security Council will not impose forces Darfur
Sudanese President Omar al-Bashir meets with Secretary-General Kofi Annan on Monday in London and the Security Council will not impose forces in the Darfur region. Sudanese Darfur deployment of the United Nations is to be preventive in Darfur and to prevent violence within rebel groups in the region.

Rumsfeld and Cheney insist on keeping the American forces in Iraq
US Defense Secretary Donald Rumsfeld and Vice President Dick Cheney have called for the withdrawal of American forces in Iraq, but there are indications that the US forces will continue to be deployed in Iraq in order to prevent the rise of violence in the region.

Killing civilians and wounding officer suicide attack in Afghanistan
An international force to help establish security (ISAF) killed civilians and a soldier for an officer in an attack against Afghan forces convoy south Atlantic Afghanistan. In the capital Kabul, a hand grenade exploded at the passage of a man and the officer was not reported injured or damage.
Multi-document summarization can be used to address more complex kinds of questions.

Circa 2002:

*What’s going on with the Hubble?*
The U.S. orbiter Columbia has touched down at the Kennedy Space Center after an 11-day mission to upgrade the Hubble observatory. The astronauts on Columbia gave the space telescope new solar wings, a better central power unit and the most advanced optical camera. The astronauts added an experimental refrigeration system that will revive a disabled infrared camera. "Unbelievable that we got everything we set out to do accomplished," shuttle commander Scott Altman said. Hubble is scheduled for one more servicing mission in 2004.

Weblog Analytics

- Textmining weblogs, discussion forums, message boards, user groups, and other forms of user generated media.
  - Product marketing information
  - Political opinion tracking
  - Social network analysis
  - Buzz analysis (what’s hot, what topics are people talking about right now).
Education

- Hypothesis of this course!

Categories of Knowledge

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

Each kind of knowledge has associated with it an encapsulated set of processes that make use of it.

Interfaces are defined that allow the various levels to communicate.

This usually leads to a pipeline architecture.
Ambiguity

- I made her duck

**Sources**
- Lexical (syntactic)
  - Part of speech
  - Subcat
- Lexical (semantic)
- Syntactic
  - Different parses
Dealing with Ambiguity

• Four possible approaches:
  1. **Tightly coupled interaction among processing levels;**
     knowledge from other levels can help decide among choices at ambiguous levels.
  2. **Pipeline processing that ignores ambiguity as it occurs and hopes that other levels can eliminate incorrect structures.**

Dealing with Ambiguity

3. **Probabilistic approaches based on making the most likely choices**
4. **Don’t do anything, maybe it won’t matter**
Models and Algorithms

- **Models** are the formalisms that are used to capture the various kinds of linguistics knowledge we need.
- **Algorithms** are then used to manipulate the knowledge representations needed to tackle the task at hand.

Models

- State machines
- Rule-based approaches
- Logical formalisms
- Probabilistic models
**Algorithms**

- Many of the algorithms are **transducers**; algorithms that take one kind of structure as input and output another.
- Unfortunately, ambiguity makes this process difficult. This leads us to employ algorithms that are designed to handle ambiguity of various kinds.

**Paradigms**

- In particular...
  - **State-space search**
    - To manage the problem of making choices during processing when we lack the information needed to make the right choice
  - **Dynamic programming**
    - To avoid having to redo work during the course of a state-space search
      - CKY, Earley, Minimum Edit Distance, Viterbi, Baum-Welch
  - **Classifiers**
    - Machine learning based classifiers that are trained to make decisions based on features extracted from the local context
**State Space Search**

- States represent pairings of partially processed inputs with partially constructed representations.
- Goals are inputs paired with completed representations that satisfy some criteria.
- As with most interesting problems the spaces are normally too large to exhaustively explore.
  * We need heuristics to guide the search
  * Criteria to trim the space

**Dynamic Programming**

- Don’t do the same work over and over.
- Avoid this by building and making use of solutions to sub-problems that must be invariant across all parts of the space.
**Key Points**

- States in the search space are pairings of tape positions and states in the machine.
- By keeping track of as yet unexplored states, a recognizer can systematically explore all the paths through the machine given an input.

---

**Advanced Topics of Relevance**

- Information Extraction
  - pp 577-583: J&M V1
  - Chapter 22: J&M V2
- Discourse/Dialogue
  - Chapter 18/19: J&M V1
  - Chapter 21/24: J&M V2
- Prosody
  - Section 4.7: J&M V1
  - Section 8.3: J&M V2
- Automatic Speech Recognition
  - Chapter 7: J&M V1
  - Chapter 9: J&M V2
- Machine Translation
  - Chapter 21: J&M V1
  - Chapter 25: J&M V2
- Generation
  - Chapter 20: J&M V1
  - ??: J&M V2