Text Normalization

Julia Hirschberg

CS 4706

- TTS demos:
  - ScanSoft/Nuance
  - AT&T
  - IBM
  - Cepstral
- SNL Robot Repair
- An interesting application for TTS

Text Normalization (1)

A sworn deposition that Sen. John McCain gave in a lawsuit more than 5 years ago appears to contradict one part of a sweeping denial that his campaign issued this week to rebut a New York Times story about his ties to a Washington lobbyist. On Wednesday night the Times published a story suggesting that McCain might have done legislative favors for the clients of the lobbyist, Vicki Iseman, who worked for the firm of Alcalde & Fay. One example it cited were two letters McCain wrote in late 1999 demanding that the Federal Communications Commission act on a long-stalled bid by one of Iseman's clients, Florida-based Paxson Communications, to purchase a Pittsburgh TV station. Just hours after the Times's story was posted, the McCain campaign issued a point-by-point response that depicted the letters as routine correspondence handled by his staff—and insisted that McCain had never even spoken with anybody from Paxson or Alcalde & Fay about the matter. "No representative of Paxson or Alcalde & Fay personally asked Senator McCain to send a letter to the FCC," the campaign said in a statement e-mailed to reporters.

But that flat claim seems to be contradicted by an impeccable source: McCain himself. "I was contacted by Mr. [Lowell] Paxson on this issue," McCain said in the Sept. 25, 2002, deposition obtained by NEWSWEEK. "He wanted their approval very bad for purposes of his business. I believe that Mr. Paxson had a legitimate complaint." While McCain said "I don't recall" if he ever directly spoke to the firm's lobbyist about the issue—an apparent reference to Iseman, though she is not named—"I'm sure I spoke to [Paxson]." McCain agreed that his letters on behalf of Paxson, a campaign contributor, could "possibly be an appearance of corruption"—even though McCain denied doing anything improper. McCain's subsequent letters to the FCC—coming around the same time that Paxson's firm was flying the senator to campaign events aboard its corporate jet and contributing $20,000 to his campaign—first surfaced as an issue during his unsuccessful 2000 presidential bid. William Kennard, the FCC chair at the time, described the sharply worded letters from McCain, then chairman of the Senate Commerce Committee, as "highly unusual."

Text Normalization (2)

Dr. Julia Hirschberg
Dept. of Computer Science
450 CS Bldg, M/C 0401
1214 Amsterdam Ave.
New York NY 10027
julia@cs.columbia.edu
Tel: 212-939-7114
Fax: 212-666-0140
http://www.cs.columbia.edu/~julia/

Today

- Segmentation
- Tokenization
- Abbreviations
- Numbers
- TTS markup
- Concept to Speech

Segmentation: What is a sentence?

A sworn deposition that Sen. John McCain gave in a lawsuit more than 5 years ago appears to contradict one part of a sweeping denial that his campaign issued this week to rebut a New York Times story about his ties to a Washington lobbyist. On Wednesday night the Times published a story suggesting that McCain might have done legislative favors for the clients of the lobbyist, Vicki Iseman, who worked for the firm of Alcalde & Fay. One example it cited were two letters McCain wrote in late 1999 demanding that the Federal Communications Commission act on a long-stalled bid by one of Iseman's clients, Florida-based Paxson Communications, to purchase a Pittsburgh TV station. Just hours after the Times's story was posted, the McCain campaign issued a point-by-point response that depicted the letters as routine correspondence handled by his staff—and insisted that McCain had never even spoken with anybody from Paxson or Alcalde & Fay about the matter. "No representative of Paxson or Alcalde & Fay personally asked Senator McCain to send a letter to the FCC," the campaign said in a statement e-mailed to reporters.

But that flat claim seems to be contradicted by an impeccable source: McCain himself. "I was contacted by Mr. [Lowell] Paxson on this issue," McCain said in the Sept. 25, 2002, deposition obtained by NEWSWEEK. "He wanted their approval very bad for purposes of his business. I believe that Mr. Paxson had a legitimate complaint." While McCain said "I don't recall" if he ever directly spoke to the firm's lobbyist about the issue—an apparent reference to Iseman, though she is not named—"I'm sure I spoke to [Paxson]." McCain agreed that his letters on behalf of Paxson, a campaign contributor, could "possibly be an appearance of corruption"—even though McCain denied doing anything improper. McCain's subsequent letters to the FCC—coming around the same time that Paxson's firm was flying the senator to campaign events aboard its corporate jet and contributing $20,000 to his campaign—first surfaced as an issue during his unsuccessful 2000 presidential bid. William Kennard, the FCC chair at the time, described the sharply worded letters from McCain, then chairman of the Senate Commerce Committee, as "highly unusual."

Rule-based Approaches

- For a potential sentence-ending word *w* followed by a '.'
  - If *w* is an abbreviation (e.g. 'Mr' or 'Mrs' or 'Dr' or 'Sen' or ….) → *w* does ***not*** end the sentence
  - O.w. *w* ends the sentence
- How do we know whether w is an abbreviation?
- What if an abbreviation ends a sentence?

  He works for Cisco, Inc.

Machine Learning Approaches

- Labeled data
  - Mechanical Turk??
- What features best predict sentence boundaries?
    - Is preceding word a known abbreviation?
    - How long is preceding word?
    - Is preceding word capitalized?
    - Is succeeding word capitalized?
    - ….
- Create feature vectors for each potential boundary
- Apply ML algorithm to produce classifier
- Test on held-out data

Hybrid Approaches

- Combine rules (for ‘easy’ decisions) with ML
  - Use rules to label initial corpus and build classifier, or
  - Add rules directly to ML results

Tokenization: What is a word?

...On Wednesday night the Times published a story suggesting that McCain might have done legislative favors for the clients of the lobbyist, Vicki Iseman, who worked for the firm of Alcalde & Fay. One example it cited were two letters McCain wrote in late 1999 demanding that the Federal Communications Commission act on a long-stalled bid by one of Iseman's clients, Florida-based Paxson Communications, to purchase a Pittsburgh TV station. Just hours after the Times's story was posted, the McCain campaign issued a point-by-point response that depicted the letters as routine correspondence handled by his staff—and insisted that McCain had never even spoken with anybody from Paxson or Alcalde & Fay about the matter. "No representative of Paxson or Alcalde & Fay personally asked Senator McCain to send a letter to the FCC," the campaign said in a statement e-mailed to reporters.

But that flat claim seems to be contradicted by an impeccable source: McCain himself. "I was contacted by Mr. [Lowell] Paxson on this issue," McCain said in the Sept. 25, 2002, deposition obtained by NEWSWEEK. "He wanted their approval very bad for purposes of his business. I believe that Mr. Paxson had a legitimate complaint." While McCain said "I don't recall" if he ever directly spoke to the firm's lobbyist about the issue—an apparent reference to Iseman, though she is not named—"I'm sure I spoke to [Paxson]." McCain agreed that his letters on behalf of Paxson, a campaign contributor, could "possibly be an appearance of corruption"—even though McCain denied doing anything improper. McCain's subsequent letters to the FCC—coming around the same time that Paxson's firm was flying the senator to campaign events aboard its corporate jet and contributing $20,000 to his campaign—first surfaced as an issue during his unsuccessful 2000 presidential bid. William Kennard, the FCC chair at the time, described the sharply worded letters from McCain, then chairman of the Senate Commerce Committee, as "highly unusual."

Word Decisions are Arbitrary but must be Consistent

- Depend on dictionary
  - Typically, segment hyphenated words if components appear in dictionary
    - But…some words are *optionally* hyphenated
    - Multi-modal/ multimodal
  - Typically, rewrite numbers to words
    - E.g. 1 orthographic token → many
    - 1000 → one thousand
    - 212-555-1212 → two one two five five five…
    - So you have to figure out what kind of number it is to do the segmentation

Abbreviations and Acronyms

- Expanding abbreviations correctly
  - Dr. Smith lives on Elm St. but Ms. St. John lives on Oak Ave.
  - Dr. North lives on Maple Dr. South.
- Other abbreviations and acronyms
  - Tcl, DLX, SCSI
  - UFO, NAACL, NAACP
  - Citicorp, Marine Corp
- Conventions for symbols: &c, il8n, evalu8, f2f, cu, tsp, 5tet

- Online abbreviations
  - RTFM, IMHO, OTOH, ANFSCD
  - Emoticons: ☹, ☺
- Ambiguous acronyms/abbreviations
  - AFAIK
  - PNG
  - How do we disambiguate?
- Multiple possible abbreviations for the same thing – abbreviations are arbitrary
  - Fplc, frpl, fpl
  - Ornges, orangs, orngs
  - But can we find “rules”?

Abbreviation Identification/Resolution (Sproat et al ’99)

One example it cited were two letters McCain wrote in late 1999 demanding that the Federal Communications Commission act on a long-stalled bid by one of Iseman's clients, Florida-based Paxson Communications, to purchase a Pittsburgh TV station…"No representative of Paxson or Alcalde & Fay personally asked Senator McCain to send a letter to the FCC," the campaign said in a statement e-mailed to reporters.... McCain's subsequent letters to the FCC—coming around the same time that Paxson's firm was flying the senator to campaign events aboard its corporate jet and contributing $20,000 to his campaign—first surfaced as an issue during his unsuccessful 2000 presidential bid. William Kennard, the FCC chair at the time, described the sharply worded letters from McCain, then chairman of the Senate Commerce Committee, as "highly unusual."

Abbreviations and their Expansions

- Devise rules to ***create*** abbreviations
  - How does living room → lvgrm? lvrm?
- Find possible abbreviations occurring in *same context* as full phrase
  - What does ‘same’ mean?

    FCC Chair William Kennard

    William Kennard, chair of the Federal Communications Commission

Ambiguous Abbreviations

- Hypothesis: If you know the domain/topic, the abbreviation will be unambiguous
  - *MO* in names/addresses vs. crime logs
  - *RNP* in political news vs. medical texts
  - *SEC* in financial news vs. clock
- How do we know the domain/topic area?
  - Topic spotting
  - How do you know when the topic changes?

Numbers in Context

- Normalizing numbers
  - In 1996 she sold 1995 shares and deposited $42 in her 401(k).
  - The number is 212-555-1210.
  - That cc # is Visa 4444-3607-5959, expiration 2/07.
- Conventions:
  - Dates
  - Money
  - Phone numbers
  - ID numbers, CC numbers,…

- Again, how do we infer the context?

Markup Languages

- Allow domain to be specified for particular applications
  - Reverse telephone directory
  - License plate look-up
  - Banking
  - Airline or train reservations

Mark-up Languages

- Let the user specify domains and other information using inline markup
- SABLE
  - Sproat et al ‘98
  - Implementation in Festival

An Example 

<SABLE>
<SPEAKER NAME="male1">
The boy saw the girl in the park <BREAK/> with the telescope.
The boy saw the girl <BREAK/> in the park with the telescope.
Some English first and then some Spanish.
<LANGUAGE ID="SPANISH">Hola amigos.</LANGUAGE>
<LANGUAGE ID="NEPALI">Namaste</LANGUAGE>
Good morning <BREAK /> My name is Stuart, which is spelled <RATE SPEED="-40%"> <SAYAS MODE="literal">stuart</SAYAS> </RATE> though some people pronounce it <PRON SUB="stoo art">stuart</PRON>.
My telephone number is <SAYAS MODE="literal">2787</SAYAS>.
I used to work in <PRON SUB="Buckloo">Buccleuch</PRON> Place, but no one can pronounce that.
By the way, my telephone number is actually
<AUDIO SRC="http://www.cstr.ed.ac.uk/~awb/sounds/touchtone.2.au"/>
<AUDIO SRC="http://www.cstr.ed.ac.uk/~awb/sounds/touchtone.7.au"/>
<AUDIO SRC="http://www.cstr.ed.ac.uk/~awb/sounds/touchtone.8.au"/>
<AUDIO SRC="http://www.cstr.ed.ac.uk/~awb/sounds/touchtone.7.au"/>.
</SPEAKER>
</SABLE>

Concept-to-Speech

- Provide a semantic representation instead of text
  - An NLG system specifies what to say and how, e.g. in markup language
- Application controls text and speech parameters
  - Utterance status is known
    - Question vs. response to a question?
    - Name vs. street address?
  - Discourse context is known
    - What's already been generated?
  - Domain is known
    - Names/addresses vs. weather reports

  - Syntax and semantics are known
- Problems:
  - Application must specify all information needed for text normalization and downstream processing
    - …all the problems that text input has must still be solved, altho with more information
  - Application must also decide how to produce the desired effects, within the limits of the TTS system
    - E.g. emotion, personality, old vs. new information

Cultural Dependence

- Russia:
  - Article 3 of the rules attached to the Moscow Telephone Network Subscribers Directory, 1916:
    - “Numbers over a hundred are to be pronounced as follows: 1.23—one twenty three, 9.72—nine seventy two, 70.09—seventy zero nine. In numbers over 10,000 every figure of a hundred should be pronounced separately, for example, 1.20.48—one twenty forty eight, 2.08.35—two zero eight thirty five, 3.35.29—three thirty five twenty nine, 4.49.52—four forty nine fifty two, 5.15.86—five fifteen eighty six etc., not one hundred and twenty forty eight, two hundred and eight thirty five etc.”

- In France
  - A French phone number is 10 digits given in series of two:
    - 01-43-48-12-85
    - "Zéro un, quarante-trois, quarante-huit, douze, quatre-vingt-cinq".
  - Numbers in addresses are always pronounced as a full number:
    - Chambre 823, 240 rie Rivoli
    - Chambre huit-cent-vingt-trois. Deux-cent-quarante, rue de Rivoli
- TTS systems need lots of knowledge!

Next Class

- Pronunciation modeling: Read Ghoshal et al 2009