CS 545: Natural Language Processing

Benjamin Snyder

Administrivia

• Prereqs:
  • Comfort doing simple math (probability, stats, a tiny bit of calculus and linear algebra -- we will review as necessary)
  • Programming experience
  • Basic algorithms knowledge (e.g. dynamic programs)
  • Interest in language / linguistics

Administrivia

• Grading:
  • 6-8 Homework assignments (50% of grade)
  • Midterm quiz (15% of grade)
  • Project / Final (20% of grade)
  • Attendance / participation (15% of grade)
Administrivia

• Room change!
• Starting on Thursday, we will meet in Room 2255 EH.
• Will send an email reminder.

Administrivia

• Communication via classlist and Piazza
• Benjamin Snyder: bsnyder@cs.wisc.edu; Room 6395 in CS building; Office hours TBD
• TA: Nisha Kiran, nkiran@cs.wisc.edu; Office hours time and location TBD

Tentative Syllabus
Questions?

Survey

- How many people here?
- languages
- **Terms:** gaussian distribution, Maximum likelihood estimator, entropy, eigenvector, lagrange multiplier, morpheme, dynamic program
- know python, perl
All that stuff is important, but...

What can computers do with human language?

A Dream

• Make computers more useful by getting them to ...
• Answer questions using the Web
• Translate documents from one language to another
• Do library research (what papers to read? summarize!)
• Manage email intelligently (what's urgent? what's spam?)
• Help us make informed decisions (which phone to buy)
• Follow directions given by any user (your Grandma)
• Fix your spelling or grammar
• Write poems or novels
• Give advice, psychotherapy
• Predict world events (elections, financial markets)
• Major obstacle to this fantasy: language!

Different Levels of Linguistic Knowledge

- discourse
- pragmatics
- semantics
- syntax
- morphology
- phonology
- orthography
- phonetics
The sound of air moving through my larynx while my tongue is raised in the back of my throat is \textit{iy} (as in teach).

The sound of air gurgling up from my stomach is not speech.

How to distinguish/generate the waveforms for speech sounds?
“Let’s pee in the corner. Let’s pee in the spotlight.”

Sociological factors: “I left my brains down in Africa”
- merry/marry/Mary
- pin/pen/pan/paean
- caught/cot
- goin’/going
- something/suttin’/sumpin’/sumthin’

Words like because and about can be realized many different ways.

How can sounds combine?

Different Levels of Linguistic Knowledge

- discourse
- pragmatics
- semantics
- syntax
- morphology
- phonology
- orthography

What are the symbols?
Different Levels of Linguistic Knowledge

- phonetics
- phonology
- morphology
- syntax
- semantics
- pragmatics
- discourse
- orthography

What are the words?

fax, google, w00t, OMG, Man-fucking-hattan, lol, lolz, unfriend, tweet, Obamacare, coo af

After it sorts each sub-part, it merges them.
After they sort each sub-part, they merge them.

How many merges are needed?
One merge.
Merging is fast.
To split is human, to merge divine.

morphology What are the words?

One house among many houses
One mouse among many mouses

uygurlaştıramadıklarımızdan mı sizinince
"(behaving) as if you are among those whom we could not civilize"

Different Levels of Linguistic Knowledge

- discourse
- pragmatics
- semantics
- syntax
- morphology
- orthography
- phonetics
Noah gave Kevin the book.
= Noah gave the book to Kevin.
= The book was given to Kevin by Noah.
= The book was given by Noah to Kevin.
*Gave Noah Kevin the book.

syntax
What are the utterances?

I want a flight to Tokyo.
I want to fly to Tokyo.
I found a flight to Tokyo.
*I found to fly to Tokyo.

Different Levels of Linguistic Knowledge

- discourse
- pragmatics
- semantics
- syntax
- morphology
- phonology
- orthography
- phonetics

What does it mean?

Colorless green ideas sleep furiously.

"Jerusalem - there is no such city!"

In this country a woman gives birth every fifteen minutes. Our job is to find that woman and stop her.

What does it mean?
Different Levels of Linguistic Knowledge

- phonetics
- phonology
- morphology
- syntax
- semantics
- pragmatics
- discourse
- orthography

What are the intentions?

“Would you mind passing the salt?”

pragmatics

“I’m sorry Dave, I’m afraid I can’t do that.”
“You’re so funny.”
“I can’t believe I ate the whole thing.”

Different Levels of Linguistic Knowledge

- discourse
- pragmatics
- semantics
- syntax
- morphology
- phonology
- orthography

What’s going on in context?
The Tin Woodman went to the Emerald City to see the Wizard of Oz and ask for a heart. After he asked for it, the Woodman waited for the Wizard’s response.

Any time you got nothing to do - and lots of time to do it - come on up.

**Different Levels of Linguistic Knowledge**

- discourse
- pragmatics
- semantics
- syntax
- morphology
- orthography
- phonology
- phonetics

**Ambiguity**  Students hate annoying professors

From Groucho:
- Last night I shot an elephant in my pajamas. What he was doing in my pajamas I’ll never know.

Headlines:
- Kids Make Nutritious Snacks
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Iraqi Head Seeks Arms

From Facebook:
- I’d rather have Kissed a Girl stuck in my head than the Girl from Ipanema.
Research has focused on English
- Most languages beyond reach of NLP:
  - Lack of data
  - Variations in linguistic structure

4,000 living languages

Linguistic Typology: The study of language difference

Variations in Ambiguity

English: fish (noun) / fish (verb)
French: poisson (noun) / pêcher (verb)
Variations in Ambiguity

- Differences in morphology
  English: in my country separate words
  Hebrew: בארצי

- Differences in syntax
  Japanese: チーズのスパゲティを食べた genitive marker
  English: I ate pasta with cheese.

A Multilingual Probabilistic Model

Corpus

Orwell's Nineteen Eighty Four (~100k words)
- Slavic: Bulgarian, Czech, Serbian, Slovene
- Uralic: Hungarian, Estonian
- Romance: Romanian
- Germanic: English

Task: Part-of-speech Induction
As we add Languages...

Number of Languages

Tag accuracy

Archeological Decipherment

lost language

known languages

Linguistic Assumptions

- Related languages have cognates
  - Arabic: dheker
  - Syriac: dukra
  - Hebrew: zakhar

- Systematic mapping between alphabets
  - (dh) -> (d)
The Ugaritic Language

Family: Northwest Semitic
Tablets from: 14th – 12th century BCE
Discovered: 1928
Deciphered: 1932 (by WW1 code breakers)

Large portion of vocabulary covered by cognates with Semitic languages

Arabic: malik
Syriac: malîk
Hebrew: melek
Ugaritic: Malk

Corpus: 34,105 tokens, 7,386 unique types

The Epic Of Ba‘al
(English translation)

As soon as El sees Her,
He cracks a smile and laughs.
He has His feet on the footstool,
And twiddles His fingers.

He lifts His voice
And shouts:
"Why has Lady Asherah of the Sea come?
Why came the Creatress of Gods?
Art Thou hungry?
Then have a morsel!
Or art Thou thirsty?
Then have a drink!
Eat!

And Lady Asherah of the Sea replied:
"Thou art great, O El,
Thou art verily wise!
The gray of Thy beard hath verily instructed Thee!
Here are pectorals of gold for Thy breast.
Lo, also it is the time of His rain.
Baal sets the season,
And gives forth His voice from the clouds.
He flashes lightning to the earth.
As a house of cedars let Him complete it,
Or a house of bricks let Him erect it!
Let it be told to Aliyan Baal:
The mountains will bring Thee much silver.
The hills, the choicest of gold;
The mines will bring Thee precious stones,
And build a house of silver and gold.
A house of lapis gems!"

And Lady Asherah of the Sea replied:
"Thou art great, O El,
Thou art verily wise!
The gray of Thy beard hath verily instructed Thee!
Here are pectorals of gold for Thy breast.
Lo, also it is the time of His rain.
Baal sets the season,
And gives forth His voice from the clouds.
He flashes lightning to the earth.
As a house of cedars let Him complete it,
Or a house of bricks let Him erect it!
Let it be told to Aliyan Baal:
The mountains will bring Thee much silver.
The hills, the choicest of gold;
The mines will bring Thee precious stones,
And build a house of silver and gold.
A house of lapis gems!"
The Decipherment Task

- Given:
  - Corpus of undeciphered language
  - Lexicon of related language (non-parallel)

- Learn:
  - Alphabetic mapping
  - Word mappings

Decipherment Intuition I

- True alphabetic mapping
  \[ \Rightarrow \text{similar character-level distributions} \]

Decipherment Intuition II

Interplay between learning:
- Alphabetic mapping
- Higher level morpheme & word correspondences
Deciphering Wingdings

\[ p = d + 4 \]

\[ d + 4 \]

\[ d + 2 \]

Deciphering Wingdings

\[ p = d \]

\[ d + 4 \]

\[ d + 2 \]

Deciphering Wingdings

\[ p = d \]

\[ d + 4 \]

\[ d + 2 \]
Deciphering Wingdings

| l = d   | + = e |

\[ \text{Deciphering Wingdings} \]

| l = d   | + = e |

\[ \text{Deciphering Wingdings} \]

| l = d   | + = e |

\[ \text{Deciphering Wingdings} \]

| l = d   | + = e |
Deciphering Wingdings

\[ s \circ \star \sharp \text{ s} \quad s \circ \star \sharp \text{ ed} \]
\[ \circ \text{ s} \circ \text{ s} \quad \circ \text{ s} \circ \text{ ed} \]
\[ \text{d e s s} \quad \text{d e s s} \]

\[ l = d \]
\[ \# = e \]
\[ \& = s \]
\[ \sharp = k \]

Deciphering Wingdings

\[ s \circ \star k \text{ s} \quad s \circ \star k \text{ ed} \]
\[ \circ \text{ s k} \circ \text{ s} \quad \circ \text{ s k} \circ \text{ ed} \]
\[ \text{d e s k} \quad \text{d e s k} \]

\[ l = d \]
\[ \# = e \]
\[ \& = s \]
\[ \sharp = k \]

Deciphering Wingdings

\[ s a \star k \text{ s} \quad s a \star k \text{ ed} \]
\[ a s k \text{ s} \quad a s k \text{ ed} \]
\[ \text{d e s k} \quad \text{d e s k} \]

\[ l = d \]
\[ \# = a \]
\[ \& = e \]
\[ \& = s \]
\[ \sharp = k \]
Deciphering Wingdings

\[
\begin{align*}
\text{sa} & \rightarrow \text{sa} \\
\text{ck} & \rightarrow \text{ck} \\
\text{ed} & \rightarrow \text{ed} \\
\text{desk} & \rightarrow \text{desk}
\end{align*}
\]

\[
\begin{align*}
\text{a} & = \text{d} \\
\text{e} & = \text{a} \\
\text{s} & = \text{c} \\
\text{k} & = \text{k}
\end{align*}
\]

• Used knowledge of English lexicon & morphology (\text{ask}, \text{ed})

• Discovery of morpheme correspondences ↔ Discovery of character correspondences

A Probabilistic Decipherment Model
Results

- Alphabetic Mapping
- Word Mapping (type-level, words with cognates)
- Morpheme Mapping

Accuracy:
- HMM Baseline
- Model no structural sparsity
- Complete Model

- Accuracy: (29/30) 60% 75%

Play Jeopardy (IBM)

Infer Gibbon Grammar (M. Coen)
Read your Mind
(T. Mitchell)

Predictive model

stimulus word
“celery”

Intermediate
semantic features
extracted from
trillion-word text
corpus

Mapping learned
from fMRI
training data

predicted
activity for
“celery”

Trace Language History

Locate Origin of Human Language
Analyze Fiction

**Persuasion** by Jane Austen

- Charlotte
- Mr Elliot
- Sir Hugo
- Lady Russell
- Anne

- Captain Wentworth

---

Analyze Fiction

**David Copperfield** by Charles Dickens

- I

---

Decipher Messages from Serial Killers

- Zodiac 408 (solved, 1969, by F.B.I.)
- Zodiac 340 (still unlocked)
Model the evolution of topics over time

**Translate (?)**

We do not see ourselves as others see us.

We see us as others do not see myself.

Others do not see us as one, please refer to us.

We are not all happy people.