Natural Language Processing (aka NLP, Computational Linguistics, Human Language Technology)

Benjamin Snyder
All that stuff is important, but...

What can computers do with human language?
A Dream
A Dream

• Make computers more useful by getting them to ...
A Dream

• Make computers more useful by getting them to...
  • Answer questions using the Web
A Dream

• Make computers more useful by getting them to ...
  • Answer questions using the Web
  • Translate documents from one language to another
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!)
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?)
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)
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)
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
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
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
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)
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
  |  phonetics
  |  orthography
Different Levels of Linguistic Knowledge

discourse
  \ |  
pragmatics
  \ |  
semantics
  \ |  
syntax
  \ |  
morphology
  \ |  
phonology
  \ |  
phonetics
  \ |  
orthography
What are the sounds?
phonetics  What are the sounds?
The sound of air moving through my larynx while my tongue is raised in the back of my throat is \( \text{ɪɨ} \) (as in teach).
The sound of air moving through my larynx while my tongue is raised in the back of my throat is $\text{i} \text{y}$ (as in teach).

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

phonetics  What are the sounds?
The sound of air moving through my larynx while my tongue is raised in the back of my throat is $\text{i}_\text{y}$ (as in teach).

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

How to distinguish/generate the waveforms for speech sounds?
Different Levels of Linguistic Knowledge

- discourse
  - pragmatics
    - semantics
      - syntax
        - morphology
          - phonology
            - phonetics
          - orthography
Different Levels of Linguistic Knowledge

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

- orthography

How can sounds combine?
phonology  How can sounds combine?
“Let’s pee in the corner. Let’s pee in the spotlight.”

phonology How can sounds combine?
“Let’s pee in the corner. Let’s pee in the spotlight.”

“I left my brains down in Africa”

phonology How can sounds combine?
“Let’s pee in the corner. Let’s pee in the spotlight.”

www.kissthisguy.com

“I left my brains down in Africa”

phonology How can sounds combine?
“Let’s pee in the corner. Let’s pee in the spotlight.”

Sociological factors:
- merry/marry/Mary
- pin/pen/pan/paean
- caught/cot
- goin’/going
- something/suttin’/sumpin’/sumthin’

phonology How can sounds combine?

“I left my brains down in Africa”

www.kissthisguy.com
“Let’s pee in the corner. Let’s pee in the spotlight.”

www.kissthisguy.com

Sociological factors:
- merry/marry/Mary
- pin/pen/pan/paean
- caught/cot
- goin’/going
- something/suttin’/sumpin’/sumthin’

“I left my brains down in Africa”

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

phonology

How can sounds combine?
Different Levels of Linguistic Knowledge

- discourse
  - pragmatics
  - semantics
    - syntax
    - morphology
    - phonology
    - phonetics
  - orthography
Different Levels of Linguistic Knowledge

- discourse
  - pragmatics
    - semantics
      - syntax
        - morphology
          - phonology
          - phonetics
          - What are the symbols?
          - orthography
What are the symbols?  orthography
第二阶段的奥运会体育比赛门票与残奥会开闭幕式门票的预订工作已经结束，现在进入门票分配阶段。在此期间，我们不再接受新的门票预订申请。

What are the symbols?   orthography
第二阶段的奥运会体育比赛门票与残奥会开闭幕式门票的预订工作已经结束，现在进入门票分配阶段。在此期间，我们不再接受新的门票预订申请。

What are the symbols? orthography
第二阶段的奥运会体育比赛门票与残奥会开闭幕式门票的预订工作已经结束，现在进入门票分配阶段。在此期间，我们不再接受新的门票预订申请。

NATRL LNGUAG PRCSSNG

“Let’s go see the *praade*!”
“They have a mouse in *they’re* house.”
“What do you want for *desert*?”

*What are the symbols?*  orthography
Different Levels of Linguistic Knowledge

- discourse
- pragmatics
- semantics
- syntax
- morphology
  - phonology
  - phonetics
  - orthography
Different Levels of Linguistic Knowledge

- discourse
  - pragmatics
  - semantics
  - syntax
  - morphology
    - What are the words?
- phonology
- phonetics
- orthography
morphology  What are the words?
fax, google, w00t, OMG, Man-fucking-hattan, lol, lolz, unfriend, tweet, Obamacare, coo af

morphology  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?
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**
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**

uygarlaştıramadıklarımızdanmışsınızcasına
“(behaving) as if you are among those whom we could not civilize”
Different Levels of Linguistic Knowledge

- discourse
  - pragmatics
    - semantics
      - syntax
        - morphology
          - phonology
            - phonetics
          - orthography
Different Levels of Linguistic Knowledge

- discourse
  - pragmatics
  - semantics
    - syntax
      - What are the utterances?
    - morphology
- phonology
  - phonetics
- orthography
syntax  What are the utterances?
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?
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
      - phonetics
  - orthography
Different Levels of Linguistic Knowledge

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

What does it mean?
semantics  What does it mean?
“Jerusalem - there is no such city!”

semantics  What does it mean?
“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.

semantics What does it mean?
“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.

semantics What does it mean?

Colorless green ideas sleep furiously.
Different Levels of Linguistic Knowledge

- Discourse
  - Pragmatics
    - Semantics
      - Syntax
        - Morphology
          - Phonology
            - Phonetics
          - Orthography
Different Levels of Linguistic Knowledge

What are the intentions?
pragmatics  What are the intentions?
“Would you mind passing the salt?”

pragmatics  What are the intentions?
“Would you mind passing the salt?”

pragmatics  What are the intentions?

“I’m sorry Dave, I’m afraid I can’t do that.”
“Would you mind passing the salt?”

pragmatics  What are the intentions?

“I’m sorry Dave, I’m afraid I can’t do that.”
“You’re so funny.”
“Would you mind passing the salt?”

pragmatics  What are the intentions?

“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
            - Phonetics
          - Orthography
Different Levels of Linguistic Knowledge

- discourse
  - pragmatics
  - semantics
  - syntax
  - morphology
- phonology
- orthography
- What’s going on in context?
discourse  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.
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.
Ambiguity Students hate annoying professors

Headlines:

From Facebook:

- I’d rather have Kissed a Girl stuck in my head than the Girl from Ipanema.
Ambiguity  Students hate annoying professors

From Groucho:

Headlines:

From Facebook:

• I’d rather have Kissed a Girl stuck in my head than the Girl from Ipanema.
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:

From Facebook:

• I’d rather have Kissed a Girl stuck in my head than the Girl from Ipanema.
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:

From Facebook:
• I’d rather have Kissed a Girl stuck in my head than the Girl from Ipanema.
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

From Facebook:

• I’d rather have Kissed a Girl stuck in my head than the Girl from Ipanema.
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

From Facebook:
• I’d rather have Kissed a Girl stuck in my head than the Girl from Ipanema.
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

From Facebook:

• I’d rather have Kissed a Girl stuck in my head than the Girl from Ipanema.
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
Linguistic Typology: The study of language difference

Subject Verb Object Positioning
Number of Genders
Definite Article

- None (145 languages)
- Two (50 languages)
- Three (26 languages)
- Four (12 languages)
- Five or more (24 languages)
Variations in Ambiguity

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

• Differences in morphology

  English:  in my country
  Hebrew:  בארצי

• Differences in syntax

  Japanese:  チーズのスパゲティを食べた
  English:  I ate pasta with cheese.

  genitive marker
A Multilingual Probabilistic Model

I love fish

J’ adore les poisson

ani ohev dagim

Mujhe machchli pasand hai
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...

Tag accuracy

Number of Languages
As we add Languages...

Tag accuracy

Number of Languages

Mono
Multi
Archeological Decipherment

lost language

? →

known languages
Linguistic Assumptions

- Related languages have *cognates*
  
  Arabic: dheker نَكَر
  Syriac: dukraً دَكْرَة
  Hebrew: zakhar זָכָר

- Systematic mapping between alphabets

  \[
  \begin{array}{c}
  
  ن \leftrightarrow \ز \leftrightarrow \ذ \\
  (dh) \leftrightarrow (z) \leftrightarrow (d)
  
  \end{array}
  \]
successful archaeological decipherment has turned out to require a synthesis of logic and intuition based, as already remarked, on wide linguistic, archaeological, historical and cultural knowledge that computers do not (and presumably cannot) possess.
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 مَلِك
Syriac: malkā مَلِك
Hebrew: melek מֶלֶך
Ugaritic: malku

Corpus: 34,105 tokens, 7,386 unique types
As soon as El sees Her,
He cracks a smile and laughs.
His feet He sets 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!
Or drink!
Eat bread from the tables!
Drink wine from the goblets!
From a cup of gold, the blood of vines!
If the love of El moves Thee,
Yea the affection of The Bull arouses Thee!"

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

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

\[
\begin{align*}
P(x|a) \quad & \quad P(x'|a') \\
b & c & d & e & f & g & h & \quad b' & c' & d' & e' & f' & g' & h'
\end{align*}
\]

Ugaritic letters \quad Hebrew letters
Interplay between learning:

- Alphabetic mapping
- Higher level morpheme & word correspondences
Deciphering Wingdings

✌❦❄⦨

Deciphering Wingdings

✌❦❄⦨

Deciphering Wingdings

✌❦❄⦨⧞⧘

Deciphering Wingdings

✌❦❄⦨⧞⧘

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

++){swift:"

Deciphering Wingdings

){}
Deciphering Wingdings

\[ \text{♥️ Keywords} \]

\[ \text{△ Keywords} \]

\[ d \]

\[ = d \]
Deciphering Wingdings

\[
\begin{align*}
\text{✌️❤️☆△✌️} & \quad \text{✌️❤️☆△❤️} \\
\text{❤️✌️△✌️} & \quad \text{❤️✌️△❤️} \\
\text{d✌️△} & \quad \text{d❤️△} \\
\downarrow & = \ d
\end{align*}
\]
Deciphering Wingdings

\[ \begin{align*}
\text{✌️❤️☆△✌️} & \rightarrow \text{ed} \\
\text{❤️✌️△❤️} & \rightarrow \text{ed} \\
d \text{e} & \rightarrow \text{ede}
\end{align*} \]

\[ \begin{align*}
\text{✈️}= d \\
\text{荪}= e
\end{align*} \]
Deciphering Wingdings

\[ \text{de} \quad \Delta \]

\[ \text{d} = \text{d} \]

\[ \phi = \text{e} \]
Deciphering Wingdings

= d
= e
Deciphering Wingdings

s ♡ ♦ ♤  s
s ♡ ♦  s
d e s ♤

↓ = d
♀ = e
✌ = s
Deciphering Wingdings

\[ s \heartsuit \star \kappa | s \]
\[ \heartsuit \ s \ k | s \]
\[ d e s k \]

\[ \heartsuit \ k e d \]
\[ \heartsuit \ s \ k | e d \]
Deciphering Wingdings

sakt|s
ask|s
desk

\(\uparrow\) = d \(\uparrow\wedge = a\)
\(\varphi\) = e
\(\ominus\) = s
\(\Delta\) = k
Deciphering Wingdings

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

\[\text{ัสค์} = \text{d} \quad \text{ฆ่า} = \text{a}\]
\[\text{รัศมี} = \text{e} \quad \text{หิมะ} = \text{c}\]
\[\text{เฉียง} = \text{s} \quad \text{ค้อ} = \text{k}\]
Deciphering Wingdings

- Used knowledge of English lexicon & morphology (ask, -ed)

- Discovery of morpheme correspondences

- Discovery of character correspondences
A Probabilistic Decipherment Model

\[ \bar{\lambda} \rightarrow \nu \]

string-edit base
distribution

\[ G_0 \]

concentration
parameter

\[ \phi_k, \pi_k \]

\[ \pi_k^\infty \]

\[ \lambda \rightarrow \nu \]

\[ \phi_k, \pi_k \]

\[ \pi_k^\infty \]

\[ \phi_k, \pi_k \]

\[ \pi_k^\infty \]

\[ G_{pre} \rightarrow G_{stem} \rightarrow G_{suf} \]

Hebrew lexicon

\[ \mathcal{H} \]

pre-pair\(_i\)

stem-pair\(_i\)

suf-pair\(_i\)

\[ c_i \]

has cognate?

\[ w_i \]

Ugaritic word

\[ N \]
Results

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

(type-level, words with cognates)
Results

**Alphabetic Mapping**

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

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Alphabetic Mapping</th>
<th>Word Mapping (type-level, words with cognates)</th>
<th>Morpheme Mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>75%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

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

Accuracy:

- Alphabetic Mapping: 75%
- Word Mapping: 60%
- Morpheme Mapping: 75% (type-level, words with cognates)

(29/30)
Play Jeopardy (IBM)
Infer Gibbon Grammar

(M. Coen)
Read your Mind

(T. Mitchell)
Trace Language History
Locate Origin of Human Language
Analyze Fiction

Persuasion by Jane Austen
Analyze Fiction

David Copperfield by Charles Dickens
Decipher Messages from Serial Killers

Zodiac 408 (solved, 1969, by couple at breakfast)

Zodiac 340 (still unsolved)
Model the evolution of topics over time
We do not see ourselves as others see us.

他の人が私たちを見るように私たちは自分が表示されません。

We see us as others do not see myself.

他の人は自分が表示されないように我々は我々を参照してください。

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

……

我々は、すべての人々が幸せではないです。

We are not all happy people.
Want to Learn More?

- CS 545 ("Natural Language and the Computer")
- Prereq: CS 367 or programming experience
- Spring 2012
- If you have any questions: bsnyder@cs.wisc.edu