What is NLP?
CMSC 473/673

T-REX IN: "COMPUTATIONAL LINGUISTICS"
I LOVE LEARNING
LEARNING IS MY FAVORITE
Today’s Learning Goals

• NLP vs. CL
• Terminology:
  – NLP: vocabulary, token, type, one-hot encoding, dense embedding, parameter/weight, corpus/corpora
  – Linguistics: lexeme, morphology, syntax, semantics, “discourse”
• NLP Tasks (high-level):
  – Part of speech tagging
  – Syntactic parsing
  – Entity id/coreference
• Universal Dependencies
T-REX IN: "COMPUTATIONAL LINGUISTICS"

Computational linguistics is the study of computer-based language processing!
Natural Language Processing
≈
Computational Linguistics
Natural Language Processing

≈

Computational Linguistics

science focus

computational bio
computational chemistry
computational X
Natural Language Processing

≈

Computational Linguistics

build a system to translate
create a QA system

engineering focus

science focus

computational bio
computational chemistry
computational X
Natural Language Processing ≈ Computational Linguistics

Both have impact in/contribute to/draw from:

- Machine learning
- Information Theory
- Data Science
- Systems Engineering
- Logic
- Theory of Computation
- Linguistics
- Cognitive Science
- Psychology
- Political Science
- Digital Humanities
- Education
build a system to translate
create a QA system

*engineering focus*

**Natural Language Processing**

≈

**Computational Linguistics**

*science focus*

computational bio
computational chemistry
computational X

*these views can co-exist peacefully*
What Are Words?

Linguists don’t agree

(Human) Language-dependent

White-space separation is a sometimes okay (for written English longform)

Social media? Spoken vs. written? Other languages?
What Are Words? Tokens vs. Types

The film got a great opening and the film went on to become a hit.

**Type**: an element of the vocabulary.

**Token**: an instance of that type in running text.

**Vocabulary**: the words (items) you know

How many of each?
Terminology: Tokens vs. Types

The film got a great opening and the film went on to become a hit.

Types
- The
- film
- got
- a
- great
- opening
- and
- the
- went
- on
- to
- become
- hit
- .

Tokens
- The
- film
- got
- a
- great
- opening
- and
- the
- film
- went
- on
- to
- become
- a
- hit
- .
Terminology: Tokens vs. Types

The film got a great opening and the film went on to become a hit.

**Types**
- The
- film
- got
- a
- great
- opening
- and
- the
- went
- on
- to
- become
- hit
- .

**Tokens**
- The
- film
- got
- a
- great
- opening
- and
- the
- film
- went
- on
- to
- become
- a
- hit
- .
Representing a Linguistic “Blob”

1. An array of sub-blobs
   word $\rightarrow$ array of characters
   sentence $\rightarrow$ array of words

   How do you represent these?
Representing a Linguistic “Blob”

1. An array of sub-blobs
   word $\rightarrow$ array of characters
   sentence $\rightarrow$ array of words
   How do you represent these?

2. Integer representation/one-hot encoding

3. Dense embedding
Representing a Linguistic “Blob”

1. An array of sub-blobs
   word → array of characters
   sentence → array of words

2. Integer
   representation/one-hot
   encoding

Let $V = \text{vocab size (\# types)}$

1. Represent each word type
   with a unique integer $i$,
   where $0 \leq i < V$

3. Dense embedding
Representing a Linguistic “Blob”

1. An array of sub-blobs
   word → array of characters
   sentence → array of words

2. Integer representation/one-hot encoding

3. Dense embedding

Let $V = \text{vocab size ( \# types) }$

1. Represent each word type with a unique integer $i$, where $0 \leq i < V$

2. Or equivalently, ...
   - Assign each word to some index $i$, where $0 \leq i < V$
   - Represent each word $w$ with a $V$-dimensional binary vector $e_w$, where $e_{w,i} = 1$ and 0 otherwise
One-Hot Encoding Example

- Let our vocab be \{a, cat, saw, mouse, happy\}

Q: What is V (# types)?
One-Hot Encoding Example

• Let our vocab be \{a, cat, saw, mouse, happy\}
• $V = \# \text{ types} = 5$
• Assign:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4</td>
</tr>
<tr>
<td>cat</td>
<td>2</td>
</tr>
<tr>
<td>saw</td>
<td>3</td>
</tr>
<tr>
<td>mouse</td>
<td>0</td>
</tr>
<tr>
<td>mouse</td>
<td>1</td>
</tr>
</tbody>
</table>

How do we represent “cat?”
One-Hot Encoding Example

- Let our vocab be \{a, cat, saw, mouse, happy\}
- \( V = \# \text{ types} = 5 \)
- Assign:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4</td>
</tr>
<tr>
<td>cat</td>
<td>2</td>
</tr>
<tr>
<td>saw</td>
<td>3</td>
</tr>
<tr>
<td>mouse</td>
<td>0</td>
</tr>
<tr>
<td>happy</td>
<td>1</td>
</tr>
</tbody>
</table>

How do we represent “cat?”

\[
e_{\text{cat}} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}
\]

How do we represent “happy?”
One-Hot Encoding Example

• Let our vocab be \{a, cat, saw, mouse, happy\}
• \(V = \# \text{ types} = 5\)
• Assign:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4</td>
</tr>
<tr>
<td>cat</td>
<td>2</td>
</tr>
<tr>
<td>saw</td>
<td>3</td>
</tr>
<tr>
<td>mouse</td>
<td>0</td>
</tr>
<tr>
<td>happy</td>
<td>1</td>
</tr>
</tbody>
</table>

How do we represent “cat?”

\[ e_{\text{cat}} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \]

How do we represent “happy?”

\[ e_{\text{happy}} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \]
Representing a Linguistic “Blob”

1. An array of sub-blobs
   word → array of characters
   sentence → array of words

2. Integer representation/one-hot encoding

3. Dense embedding

   Let E be some embedding size (often 100, 200, 300, etc.)

   Represent each word w with an E-dimensional real-valued vector $e_w$
A Dense Representation (E=2)
Where Do We Observe Language?

- All around us

- NLP/CL: from a corpus (pl: corpora)
  - Literally a “body” of text

- In real life:
  - Through curators (the LDC)
  - From the web (scrape Wikipedia, Reddit, etc.)
  - Via careful human elicitation (lab studies, crowdsourcing)
  - From previous efforts

- In this class: the Universal Dependencies
<table>
<thead>
<tr>
<th>Language</th>
<th>UD Version</th>
<th>Part of Speech</th>
<th>Syntax</th>
<th>Arcade</th>
<th>Treebank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afriken</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ancient Greek</td>
<td>202K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ancient Greek-PROIEL</td>
<td>211K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabic</td>
<td>242K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabic-HYUAD</td>
<td>622K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabic-PUD</td>
<td>20K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basque</td>
<td>121K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belarusian</td>
<td>8K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bulgarian</td>
<td>158K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buryat</td>
<td>16K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catalan</td>
<td>593K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>123K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese-CFL</td>
<td>7K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese-PUD</td>
<td>21K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coptic</td>
<td>4K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Croatian</td>
<td>107K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td>1,593K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech-CAC</td>
<td>483K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech-CLLT</td>
<td>37K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech-PUD</td>
<td>18K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Danish</td>
<td>109K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>209K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch-LassySmall</td>
<td>101K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>254K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English-ESL</td>
<td>58K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English-LiDE</td>
<td>82K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English-PUD</td>
<td>21K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English-ParTUT</td>
<td>49K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estonian</td>
<td>47K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish</td>
<td>203K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish-FTB</td>
<td>159K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish-PUD</td>
<td>15K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>359K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French-FTB</td>
<td>566K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French-PUD</td>
<td>24K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French-ParTUT</td>
<td>27K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French-Sequlia</td>
<td>68K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Galician</td>
<td>126K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Galician-TreeGal</td>
<td>23K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>287K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German-PUD</td>
<td>29K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gothic</td>
<td>55K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greek</td>
<td>61K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hebrew</td>
<td>115K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td>351K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindi-PUD</td>
<td>23K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungarian</td>
<td>42K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indonesian</td>
<td>121K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indonesian-PUD</td>
<td>28K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Irish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian-PUD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese-KTC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese-PUD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kazakh</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean-PUD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kurdish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latin-ITTB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latin-PROIEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latvian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lithuanian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maltese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malyalam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marathi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old Church Slavonic</td>
<td>57K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persia</td>
<td>181K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persian</td>
<td>181K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polish</td>
<td>52K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td>210K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese-BR</td>
<td>297K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese-PUD</td>
<td>21K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romanian</td>
<td>216K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>99K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian-PUD</td>
<td>19K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian-SynTagRus</td>
<td>1,107K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanskrit</td>
<td>1K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sorbian</td>
<td>86K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovak</td>
<td>100K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovenian</td>
<td>140K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovenian-SST</td>
<td>29K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>423K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish-AnCora</td>
<td>547K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish-PUD</td>
<td>22K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td>96K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish-ES</td>
<td>79K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish-PUD</td>
<td>19K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish Sign Language</td>
<td>&lt;1K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tamil</td>
<td>8K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thai</td>
<td>23K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td>56K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkish-PUD</td>
<td>16K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ukrainian</td>
<td>25K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper Sorbian</td>
<td>11K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urdu</td>
<td>139K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uyghur</td>
<td>11K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vietnamese</td>
<td>43K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Upcoming UD Treebanks

- Amharic
- Armenian
- Bangla
- Bengali-ODS
- Cantonese
- Chinese-HK

http://universaldependencies.org/
part-of-speech & syntax for > 120 languages
Computational linguistics is the study of computer-based language processing!

A major area of computational linguistics is that of "ambiguity resolution". It turns out that many things people say in a language – English, for example – can have more than one meaning!

Consider the phrase "fruit flies like a banana". Is it describing the taste of fruit flies, or rather flying fruit? How can a computer hope to figure this out?

"Language is Productive"
orthography
orthography

morphology: study of how words change

Adapted from Jason Eisner, Noah Smith
Watergate
Troopergate
Watergate ➔ Bridgegate
Deflategate
orthography

morphology

lexemes: a basic “unit” of language
Ambiguity

Kids Make Nutritious Snacks
Ambiguity

Kids Make Nutritious Snacks
Kids *Prepare* Nutritious Snacks
Kids *Are* Nutritious Snacks

sense ambiguity
orthography

morphology

lexemes

syntax: study of structure in language

Adapted from Jason Eisner, Noah Smith
Ambiguity

British Left Waffles on Falkland Islands
Lexical Ambiguity...

British Left Waffles on Falkland Islands
British Left Waffles on Falkland Islands
British Left Waffles on Falkland Islands
... yields the “Part of Speech Tagging” task

British Left Waffles on Falkland Islands

- Adjective
- Noun
- Verb

British Left Waffles on Falkland Islands

- Noun
- Verb
- Noun
Parts of Speech

Classes of words that behave like one another in “similar” contexts

Pronunciation (stress) can differ: object (noun: OB-ject) vs. object (verb: ob-JECT)

It can help improve the inputs to other systems (text-to-speech, syntactic parsing)
Syntactic Ambiguity...

Pat saw Chris with the telescope on the hill.

I ate the meal with friends.
Pat saw Chris with the telescope on the hill.

I ate the meal with friends.
Syntactic parsing: perform a “meaningful” structural analysis according to grammatical rules.

I ate the meal with friends.
Syntactic Parsing Can Help Disambiguate

I ate the meal with friends.

Diagram:

- **S**
  - **NP**
    - **VP**
      - **ate**
        - **NP**
        - **the meal**
      - **with friends**
      - **PP**
Syntactic Parsing Can Help Disambiguate

I ate the meal with friends.
Clearly Show Ambiguity...
But Not Necessarily All Ambiguity

I ate the meal with a fork
I ate the meal with gusto
I ate the meal with friends
orthography

morphology

lexemes

syntax

semantics: study of (literal?) meaning
orthography
morphology
lexemes
syntax
semantics

pragmatics: study of (implied?) meaning

Adapted from Jason Eisner, Noah Smith
orthography
morphology
lexemes
syntax
semantics
pragmatics
discourse: study of how we communicate

Adapted from Jason Eisner, Noah Smith
Semantics $\rightarrow$ Discourse Processing

John stopped at the donut store.
Semantics → Discourse Processing

John stopped at the donut store.
Semantics $\rightarrow$ Discourse Processing

John stopped at the donut store before work.
John stopped at the donut store on his way home.
Semantics $\rightarrow$ Discourse Processing

John stopped at the donut shop.

John stopped at the trucker shop.

John stopped at the mom & pop shop.

John stopped at the red shop.
Discourse Processing through Coreference

I spread the cloth on the table to protect it.

I spread the cloth on the table to display it.
Discourse Processing through Coreference

I spread the cloth on the table to protect it.

I spread the cloth on the table to display it.
Discourse Processing through Coreference

I spread the cloth on the table to protect it.

I spread the cloth on the table to display it.
NLP + Latent Modeling

explain what you see/annotate

with things “of importance” you don’t

observed text
orthography
morphology
lexemes
syntax
semantics
pragmatics
discourse

Adapted from Jason Eisner, Noah Smith
T-REX IN: "COMPUTATIONAL LINGUISTICS"

Computational linguistics is the study of computer-based language processing!

A major area of computational linguistics is that of "ambiguity resolution". It turns out that many things people say in a language - English, for example - can have more than one meaning!

Consider the phrase "fruit flies like a banana". Is it describing the taste of fruit flies, or rather flying fruit? How can a computer hope to figure this out?

Many have focused on statistical modelling of language, but this approach is approximate. I agree!
NLP <-> Machine Learning

Goal: Learn parameters (weights) $\theta$ to develop a scoring function that says how “good” some provided text is
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.

Goal: Learn parameters (weights) $\theta$ to develop a scoring function that says how “good” some provided text is
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.

Goal: Learn parameters (weights) \( \theta \) to develop a scoring function that says how “good” some provided text is
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.

Goal: Learn parameters (weights) $\theta$ to develop a scoring function that says how “good” some provided text is
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.

Goal: Learn parameters (weights) $\theta$ to develop a scoring function that says how “good” some provided text is.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.

Q: If we make $p$ to be a probability distribution, what are the minimum and maximum values of $s$?

Goal: Learn parameters (weights) $\theta$ to develop a scoring function that says how “good” some provided text is.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.

Goal: Learn parameters (weights) $\theta$ to develop a scoring function that says how “good” some provided text is.
Use ML Techniques to Learn the Weights

probabilistic model

\[ p_\theta(X) \]

objective

\[ F(\theta) \]
Gradient Ascent

$$\underset{\theta}{\text{arg max}} \; F(\theta)$$
Gradient Ascent

\[ \arg \max_{\theta} F(\theta) \]
Gradient Ascent

\[ \arg \max_{\theta} F(\theta) \]
Gradient Ascent

arg max \( F(\theta) \)

\[ \nabla_{\theta} F(\theta) \]

“gradient of F with respect to \( \theta \)”
Gradient Ascent

$$\arg \max_{\theta} F(\theta)$$

“gradient of $F$ with respect to $\theta_k$ while holding all other variables constant”
A major area of computational linguistics is the study of "ambiguity resolution". It turns out that many things people say in a language – English, for example – can have more than one meaning! Consider the phrase "fruit flies like a banana". Is it describing the taste of fruit flies, or rather flying fruit? How can a computer hope to figure this out?

Many have focused on statistical modelling of language, but this approach is approximate. I agree!

What do YOU know about computational linguistics?

Ever read a little paper called "Non-Statistical Models for Unsupervised Prepositional Phrase Attachment?"

That was me!

It was some of my earliest work on head word tuples!
Today’s Learning Goals

• NLP vs. CL
• Terminology:
  – NLP: vocabulary, token, type, one-hot encoding, dense embedding, parameter/weight, corpus/corpora
  – Linguistics: lexeme, morphology, syntax, semantics, “discourse”
• NLP Tasks (high-level):
  – Part of speech tagging
  – Syntactic parsing
  – Entity id/coreference
• Universal Dependencies