How to Read *War and Peace* in 30 Seconds

Denis Griffis
PhD Student, Speech and Language Technologies lab
The Ohio State University
How to Read *War and Peace* in 30 Seconds

*Or: An introduction to Natural Language Processing*

Denis Griffis
PhD Student, Speech and Language Technologies Lab
The Ohio State University
Sorry, I missed that.
How the heck do we get a computer to understand text?
Natural language processing

From Wikipedia, the free encyclopedia

This article is about language processing by computers. For the processing of language by the human brain, see Language processing.

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many
OK Google, where are all the cats?
OK Google, where are all the cats?

```
SELECT CurrentLocation FROM AllAnimals WHERE AnimalType = 'Cat'
```
Speech Recognition

OK Google, where are all the cats?

SELECT CurrentLocation FROM AllAnimals WHERE AnimalType = 'Cat'
OK Google, where are all the cats?

SELECT CurrentLocation FROM AllAnimals WHERE AnimalType = 'Cat'
“Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for $250”
“Yess! Yess! Its official [Nintendo] announced today that they Will release the [Nintendo 3DS] in [north America] [march 27] for [$250]”
“Yess! Yess! Its official [Nintendo] announced today that they Will release the [Nintendo 3DS] in [north America] [march 27] for [$250]”

<table>
<thead>
<tr>
<th>Company</th>
<th>Product</th>
<th>Date</th>
<th>Price</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nintendo</td>
<td>Nintendo 3DS</td>
<td>March 27</td>
<td>$250</td>
<td>North America</td>
</tr>
</tbody>
</table>
“Yess! Yess! Its official [Nintendo] announced today that they Will release the [Nintendo 3DS] in [north America] [march 27] for [$250]”

<table>
<thead>
<tr>
<th>Company</th>
<th>Product</th>
<th>Date</th>
<th>Price</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nintendo</td>
<td>Nintendo 3DS</td>
<td>March 27</td>
<td>$250</td>
<td>North America</td>
</tr>
</tbody>
</table>

Natural language understanding
<table>
<thead>
<tr>
<th>Company</th>
<th>Product</th>
<th>Date</th>
<th>Price</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nintendo</td>
<td>Nintendo 3DS</td>
<td>March 27</td>
<td>$250</td>
<td>North America</td>
</tr>
</tbody>
</table>
Nintendo will release the Nintendo 3DS in North America on March 27 for $250.

<table>
<thead>
<tr>
<th>Company</th>
<th>Product</th>
<th>Date</th>
<th>Price</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nintendo</td>
<td>Nintendo 3DS</td>
<td>March 27</td>
<td>$250</td>
<td>North America</td>
</tr>
</tbody>
</table>
Nintendo will release the Nintendo 3DS in North America on March 27 for $250.

<table>
<thead>
<tr>
<th>Company</th>
<th>Product</th>
<th>Date</th>
<th>Price</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nintendo</td>
<td>Nintendo 3DS</td>
<td>March 27</td>
<td>$250</td>
<td>North America</td>
</tr>
</tbody>
</table>
A Brief History of NLP
A Brief History of NLP

< 1950
Codes for translation; message encodings
A Brief History of NLP

< 1950
Codes for translation; message encodings

1950s
Turing, Chomsky; Start of machine translation
A Brief History of NLP

< 1950
Codes for translation; message encodings

1950s
Turing, Chomsky; Start of machine translation

Late 1960s
SHRDLU, ELIZA
A Brief History of NLP

< 1950
Codes for translation; message encodings

1950s
Turing, Chomsky; Start of machine translation

1960s
Late 1960s
SHRDLU, ELIZA

1970s
1970s
Dependency theory, FSA parsing, concept ontologies

My dog eats adamantium

My dog eats adamantium

Nmod:Poss Nsubj Root Dobj
A Brief History of NLP

< 1950
Codes for translation; message encodings

1950s
Turing, Chomsky; Start of machine translation

1960s
Late 1960s
SHRDLU, ELIZA

1970s
1970s
Dependency theory, FSA parsing, concept ontologies

1980s
1980s
Machine learning; HMMs, Decision Trees
A Brief History of NLP

< 1950
Codes for translation; message encodings

1950s
Turing, Chomsky; Start of machine translation

Late 1960s
SHRDLU, ELIZA

1970s
Dependency theory, FSA parsing, concept ontologies

1980s
Machine learning; HMMs, Decision Trees

1990s
Neural networks, high-performance computing
A Brief History of NLP

< 1950
Codes for translation; message encodings

1950s
Turing, Chomsky; Start of machine translation

Late 1960s
SHRDLU, ELIZA

1970s
Dependency theory, FSA parsing, concept ontologies

1980s
Machine learning; HMMs, Decision Trees

1990s
Neural networks, high-performance computing

2000s on
Explosion in Web text data, ontologies
A Brief History of NLP

< 1950
Codes for translation; message encodings

1950s
Turing, Chomsky; Start of machine translation

Late 1960s
SHRDLU, ELIZA

1970s
Dependency theory, FSA parsing, concept ontologies

1980s
Machine learning; HMMs, Decision Trees

1990s
Neural networks, high-performance computing

2000s on
Explosion in Web text data, ontologies

Rule-based methods

Statistical methods
Rule-Based NLP

- Regular Expressions

```java
…………………………………………………………………………
#    Per cause_of_death
…………………………………………………………………………
{
    ruletype: "composite",
    pattern: (((ner:PERSON}+) /died/ /of|from/ /a/?) ([tag:NN}+) ),
    result: Format("per:cause_of_death(%s,%s)", $1.word, $2.word),
    action: (Annotate($1, kbp, "per"), Annotate($2, kbp, "per_cause_of_death"))
}
```

Example from Alan Ritter
Rule-Based NLP

- Keywords and arguments
Rule-Based NLP

- Keywords and arguments
Rule-Based NLP

- Finite State Automata

Example from “Speech and Language Processing”, Jurafsky and Martin, 2009
• Hidden Markov Models (HMMs)
Statistical NLP

- Hidden Markov Models (HMMs)
Statistical NLP

- Hidden Markov Models (HMMs)
Statistical NLP

- Hidden Markov Models (HMMs); MEMMs

I can haz cheezburger
Statistical NLP

- Hidden Markov Models (HMMs); MEMMs, CRFs
• Support Vector Machines (SVMs)
Statistical NLP

- Support Vector Machines (SVMs)
• Support Vector Machines (SVMs)
Statistical NLP

- Neural Networks (NNs)

Word count
Average token length
“Knave” occurred?
Ratio of /[ATCG]/

Input | Hidden layer(s) | Output

Shakespeare
Biomedical abstract
News article
Statistical NLP

- Other methods: matrix factorization, logistic regression, etc.

\[
\begin{bmatrix}
\ldots \\
\ldots \\
\ldots \\
\end{bmatrix}
= 
\begin{bmatrix}
\cdot \\
\cdot \\
\cdot \\
\end{bmatrix}
\begin{bmatrix}
\ldots \\
\end{bmatrix}
\]
Lots of current work uses both approaches in *joint systems*!
These are models...
These are models...

...but models are only tools to solve problems.
Kinds of Machine Learning

Unsupervised

Supervised
Kinds of Machine Learning

- Unsupervised
- Supervised
- Semi-Supervised
Kinds of Machine Learning

Unsupervised

Supervised

Semi-Supervised

aka Distantly-supervised, weakly-supervised
Unsupervised Learning

Goal: Discover hidden structure in data
Unsupervised Learning

Goal: Discover hidden structure in data
Supervised Learning

Goal: Use known information to categorize data
Supervised Learning

Goal: Use known information to categorize data
Supervised Learning

Goal: Use known information to categorize data
Supervised Learning

Goal: Use known information to categorize data
Semi-Supervised Learning

*Goal*: Use some known information, along with hidden structure, to categorize data
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data
Semi-Supervised Learning

**Goal:** Use some known information, along with hidden structure, to categorize data
Semi-Supervised Learning

*Goal*: Use some known information, along with hidden structure, to categorize data
Semi-Supervised Learning

**Goal:** Use some known information, along with hidden structure, to categorize data.
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data.
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data
Human effort required
Human effort required

Unsupervised
Human effort required

Unsupervised

Supervised
Human effort required

Unsupervised

Semi-supervised

Supervised
Human effort required

Unsupervised  Semi-supervised  Supervised

Classification power
At this point, you may be asking yourself...
At this point, you may be asking yourself...

So what do you **do** with all this stuff?
Lots of things!
This translation sucks

Этот перевод отстой
Parsing / Tagging

Picard ordered tea.
Picard ordered tea.

- Picard: NNP
- ordered: VBD
- tea: NN
Picard ordered tea.

Part of Speech:
- Picard: NNP
- ordered: VBD
- tea: NN

Dependency:
- NSubj: Picard
- Root: ordered
- DObj: tea
Abraham Lincoln was born February 12, 1809, in Hardin County, Kentucky...
“Abraham Lincoln was born February 12, 1809, in Hardin County, Kentucky...”

Birth Dates

<table>
<thead>
<tr>
<th>ID</th>
<th>Month</th>
<th>Day</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honest Abe</td>
<td>February</td>
<td>12</td>
<td>1809</td>
</tr>
</tbody>
</table>

Birth Locations

<table>
<thead>
<tr>
<th>ID</th>
<th>County</th>
<th>State</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Lincoln</td>
<td>Hardin</td>
<td>Kentucky</td>
<td>'Murica</td>
</tr>
</tbody>
</table>
Information Retrieval
Information Retrieval
Information Retrieval

Web Search

Google

Bing

Bioinformatics

ATTACCGCAGAT

1 | CATTACCGGAGATCCTA
2 | CCCATTACGGCCGCAGATAA
3 | ATTACCGAA
Information Retrieval

Question Answering

Who played Malcolm Reynolds?
Nathan Fillion

Who played Real Madrid last week?
Barcelona; final score 3-2
Etc., etc., etc.
Etc., etc., etc.

Automatic summarization
Automatic summarization


Etc., etc., etc.

Automatic summarization


Etc., etc., etc.

Automatic summarization

Sentiment analysis
Etc., etc., etc.

Automatic summarization

Life is meh, but donatos is awesummmmm

Sentiment analysis

Jay-Z is great, 'Ye sucks!

*Not actual tweets*
Etc., etc., etc.

Automatic summarization

Sentiment analysis

Life is meh, but donatos is awesummmmm

Jay-Z is great, 'Ye sucks!

*Not actual tweets*
Etc., etc., etc.

Automatic summarization

Sentiment analysis

*Not actual tweets*
Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis
Etc., etc., etc.

Automatic summarization

U: I want Chinese food.

S: Here are 473 Chinese places.

Sentiment analysis

U: How about cheap ones on the south side?

S: Here is 1 restaurant.

Discourse analysis

U: Eh, let's do Thai food instead.

S: I'm sorry, Dave, I can't let you do that.
Etc., etc., etc.

Automatic summarization


t: I want Chinese food.

s: Here are 473 Chinese places.

Sentiment analysis


Discourse analysis


User Goals

<table>
<thead>
<tr>
<th>Turn</th>
<th>Type</th>
<th>Location</th>
<th>Cheap?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chinese</td>
<td>???</td>
<td>???</td>
</tr>
<tr>
<td>2</td>
<td>Chinese</td>
<td>South</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>Thai</td>
<td>South</td>
<td>Yes</td>
</tr>
</tbody>
</table>

U: I want Chinese food.

S: Here are 473 Chinese places.

U: How about cheap ones on the south side?

S: Here is 1 restaurant.

U: Eh, let's do Thai food instead.

S: I'm sorry, Dave, I can't let you do that.
Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation
Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Phonemes
Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Phonemes

Morphemes
Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Phonemes

Morphemes

Words

maytheforcebewithyou

May the force be with you
Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Phonemes

Words

May the force be with you

Sentences

[ I spoke to Mr. Spock. ]
[ His response was illogical. ]

Morphemes

Unbreakable
<table>
<thead>
<tr>
<th>Topics</th>
<th>Sentences</th>
<th>Morphemes</th>
<th>Words</th>
<th>Phonemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>...who I met at a Trek convention.</td>
<td>[I spoke to Mr. Spock.] [His response was illogical.]</td>
<td>Unbreakable</td>
<td>maytheforcebewithyou</td>
<td>Unbreakable</td>
</tr>
<tr>
<td>As for Star Wars...</td>
<td></td>
<td></td>
<td></td>
<td>I spoke to Mr. Spock.</td>
</tr>
</tbody>
</table>

- **Automatic summarization**
- **Sentiment analysis**
- **Discourse analysis**
- **Segmentation**

**Etc., etc., etc.**
Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Disambiguation and reference
After I put him in [check]₁, he wrote me a [check]₂.
After I put him in [check]$^1$, he wrote me a [check]$^2$.

I spoke to [the customer]$^1$, then told [my boss]$^2$ that [she]$^2$ should fire [her]$^1$. 
Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Disambiguation and reference

Word sense disambiguation

After I put him in [check]₁, he wrote me a [check]₂.

Coreference resolution

I spoke to [the customer]₁, then told [my boss]₂ that [she]₂ should fire [her]₁.

Named entity recognition

[ Bugs Bunny ]ₚ e r s o n bought 50% of [ Acme Corp. ]ₐ n i e s in [ 2004 ]ₚ e r i d .
Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Disambiguation and reference

And many more!
How can I get in on this?
## NLP Toolkits

<table>
<thead>
<tr>
<th>Toolkit</th>
<th>Language</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apache OpenNLP</td>
<td>Java</td>
<td><a href="https://opennlp.apache.org">https://opennlp.apache.org</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>General-purpose NLP toolkit; tends to use older models, but under Apache license.</td>
</tr>
<tr>
<td>Natural Language Toolkit (NLTK)</td>
<td>Python</td>
<td><a href="http://www.nltk.org/">http://www.nltk.org/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard NLP option for Python; easy to pick up and play with, and includes several common corpora.</td>
</tr>
<tr>
<td>Mallet</td>
<td>Java</td>
<td><a href="http://mallet.cs.umass.edu/">http://mallet.cs.umass.edu/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>More technical toolkit, focused on current, high-complexity models.</td>
</tr>
<tr>
<td>LingPipe</td>
<td>Java</td>
<td><a href="http://alias-i.com/lingpipe/">http://alias-i.com/lingpipe/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Another general-purpose NLP toolkit; offers industry licensing option.</td>
</tr>
<tr>
<td>Stanford CoreNLP</td>
<td>Java</td>
<td><a href="http://nlp.stanford.edu/software/corenlp.shtml">http://nlp.stanford.edu/software/corenlp.shtml</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard tools in academia, tends towards cutting edge models. Low ease-of-use, and academic licensing restrictions.</td>
</tr>
<tr>
<td>Alchemy API</td>
<td>Cloud API</td>
<td><a href="http://www.alchemyapi.com/">http://www.alchemyapi.com/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fanciest industry option (owned by IBM). Offers NLP, vision, other ML resources.</td>
</tr>
</tbody>
</table>
Other Resources

Speech Recognition Toolkit - http://kaldi-asr.org/

Association for Computational Linguistics
http://aclweb.org/

http://www.signalprocessingsociety.org/
Questions?

My contact info:
Denis Griffis
griffis.30@osu.edu
http://web.cse.ohio-state.edu/slate/