Natural Language Processing

Derek Erwin, Jay Kaiser,
Kenny Kha, Connor Russell

Chung-chieh (Ken) Shan - NLP Expert

Ken is an assistant professor in the School of Informatics and Computing at Indiana University.

He wrote his Harvard PhD dissertation on the relationship between the semantics of natural languages and programming languages.

Ken Shan's Favorite Achievement

Years ago, Ken created an online course catalog for his college.

It was able to search through course records using keywords.

The program was actually used by students.

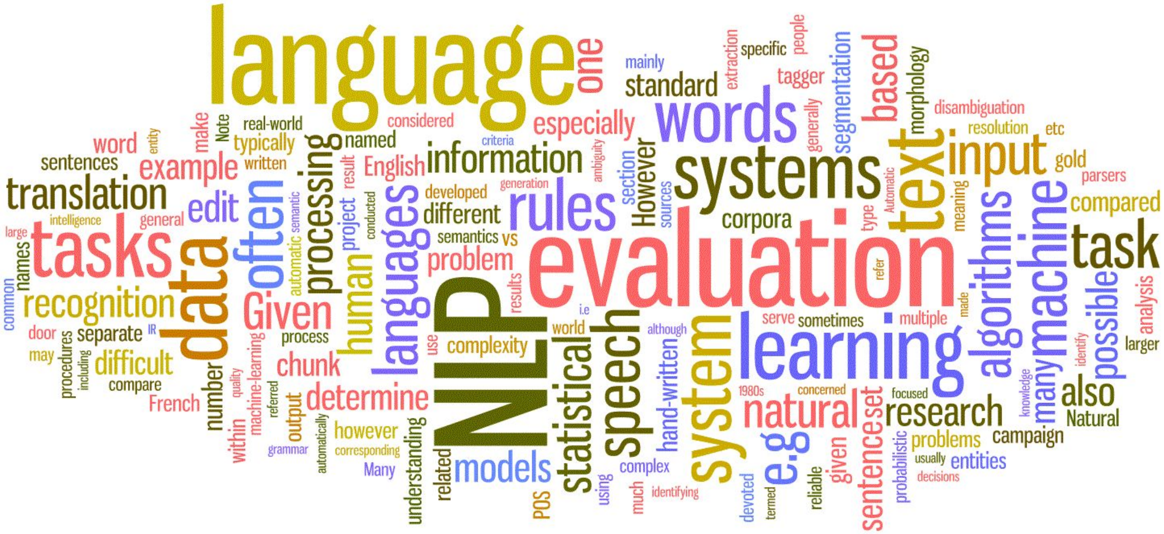
Today, such a program is common and uninteresting.

"Every time something is implemented, it loses its uniqueness and starts losing its definition as NLP." ~ Ken Shan

CR

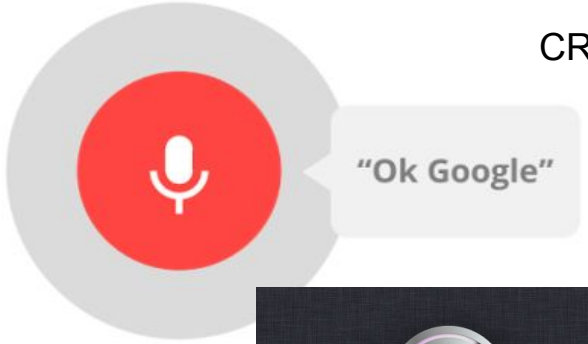
What is NLP?

“Natural language processing is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages.” - Wikipedia



Where is NLP?

- Virtual assistants (Siri/Cortana/"Ok Google")
- Text-to-speech/Speech-to-text software
- Grammar and spelling checkers
- Chatterbots (Cleverbot)
- Spam filters
- Machine translation (Google Translate)
- Automatic summarization
- Automatic article creation
- Information retrieval (Google.com)



NLP is the interaction between humans and technology through language.

"Computational linguistics is science. NLP is engineering." ~Ken Shan

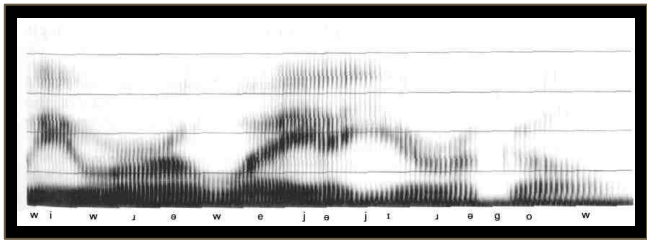
Common Issues in NLP

- Part-of-speech tagging
- Word-sense disambiguation
- Syntactic parsing
- Named entity recognition
- Coreference resolution
- Sentiment analysis
- Speech segmentation
- Relationship extraction
- Information extraction

All this together makes Siri!

“NLP is a little like garbage collectors. You don’t really notice they’re there until they mess up.” ~Ken Shan

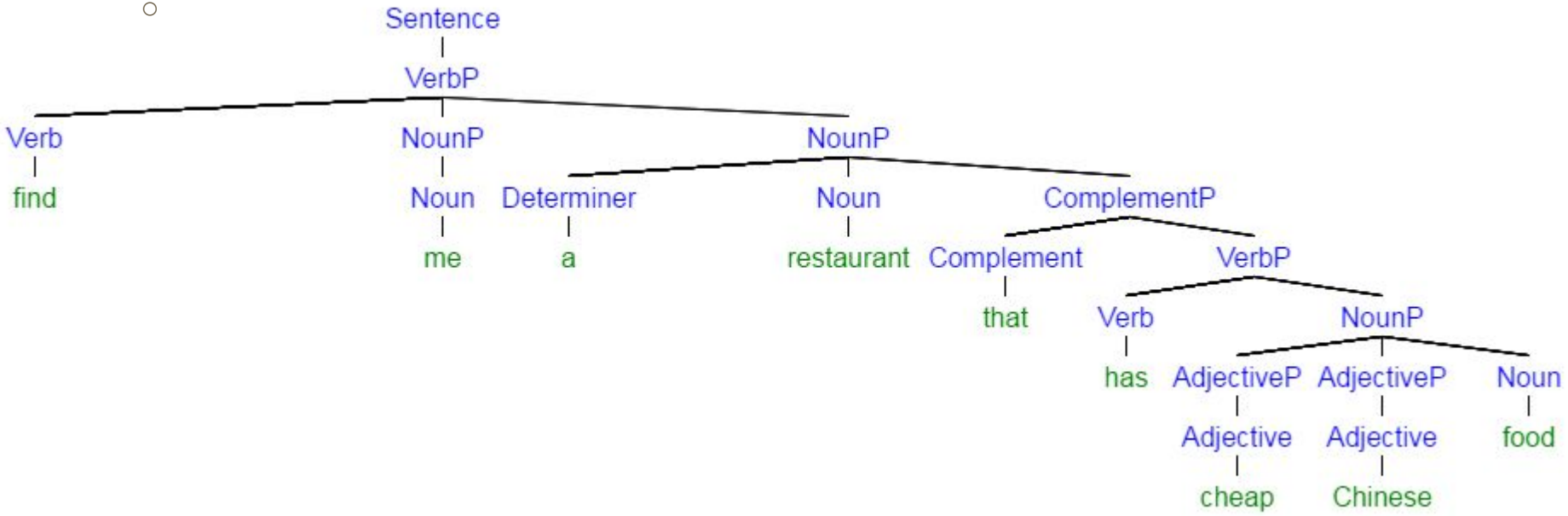
“Hey Siri, find me a nearby restaurant that has cheap Chinese food.”



- Speech-to-text and token segmentation
  - /faɪndmiəˈnɪrˈbaɪˈrɛstəˌrɑntðæthæzʧipʧaɪˈnizfud/
  - /faɪnd mi ə ˈnɪrˈbaɪ ˈrɛstəˌrɑnt ðæt hæz ʧip ʧaɪˈniz fud/
  - find me a nearby restaurant that has cheap chinese food
- Named entity recognition
  - find me a nearby restaurant that has cheap **Chinese** food
- Part-of-speech tagging
  - $\text{find}_{\text{verb}}$ me_{noun} a_{det} $\text{nearby}_{\text{adj}}$ $\text{restaurant}_{\text{noun}}$ $\text{that}_{\text{coord}}$ has_{verb} $\text{cheap}_{\text{adj}}$ $\text{Chinese}_{\text{adj}}$ $\text{food}_{\text{noun}}$
- Word-sense disambiguation
  - *find* **(show$_{\text{verb}}$)** me a nearby restaurant that has *cheap* **(in price)** *Chinese* **(culture)** food
- Relationship extraction
  - find **me$_{\text{the user}}$** a nearby restaurant that has cheap Chinese food

K

“Hey Siri, find me a nearby restaurant that has cheap Chinese food.”

- Syntactic parsing
  - 

K

“Hey Siri, find me a nearby restaurant that has cheap Chinese food.”

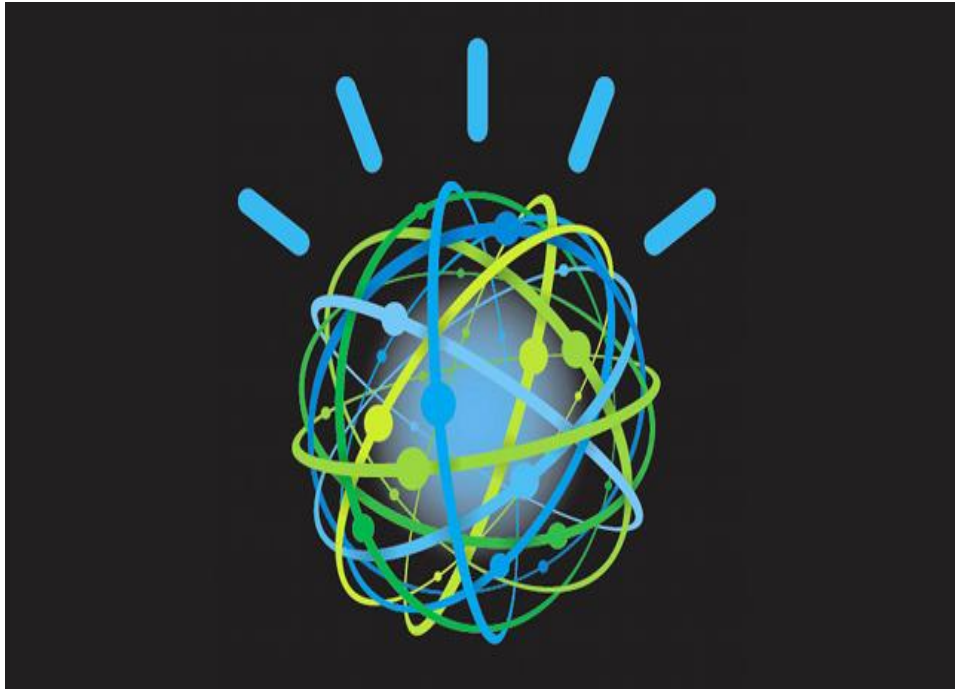
- Information retrieval
  - find me a nearby restaurant that has cheap Chinese food

actually means:

  - search online for:
    - open restaurants (that will still be open for at least another hour)
    - within a mile radius (or more if the user has shown evidence of owning a car)
    - of the user’s current location (taking into account their direction of movement)
    - that serve Chinese food (or food of similar ethnic background)
    - for under 15 dollars (per person not including drinks)
    - and that does not require a reservation

What is Watson?

- Cognitive System
- Combines information retrieval and natural language processing
- Builds its domain knowledge from structured and unstructured data
- A core set of technologies that can be customized and targeted to specific industries



How Watson Operates

1. Question Analysis - Parsing and finding keywords
2. Primary Search - Search is performed on the keywords to find relevance
3. Candidate Hypothesis Generation - Possible answers (candidate answers) are found in documents
4. Answer Scoring - Answers are given scores based on a large number of answer scoring analytics
5. Supporting Evidence - Candidate answers are used to do another search and further scored
6. Merging Candidate Scores and Scoring the Confidence - detects variants of the same answer and merges their feature scores together; use machine learning techniques to determine final confidence scores

Issues with this process: Language is ambiguous

- The horse raced past the barn fell.
  - The horse (that was) raced past the barn, fell.
- Time flies like an arrow. Fruit flies like a banana.
- The cat the man hated died.

And language can be stupid.

- “Colorless green ideas sleep furiously.” - Noam Chomsky
- The cat the man the boy loved hated died.
  - The cat, [that the man [(who) the boy loved,] hated,] died.
- Buffalo buffalo buffalo Buffalo buffalo buffalo Buffalo buffalo.
  - New York bison bully Minnesota bison (that) bully New York bison.
  - New York bison (that) bully Minnesota bison, bully New York bison.

Issues with this process: Lexical Ambiguity

- Named entity ambiguity
  - The **North Pole** is cold.
  - The **north pole** is where the flag is tied.
- Part-of-speech ambiguity
  - **Book** the flight.
  - Read the **book**.
- Homonyms
  - The animals were in the **pen**.
  - The **pen** was on the table
- Word-sense ambiguity
  - **Kill** somebody.
  - **Kill** a process.

Issues with this process: Speech is messy

Excerpts taken from Donald Trump's victory speech:

- She congratulated us — ***it's about us*** — on our victory, and I congratulated her and her family on a very, ***very*** hard-fought campaign. I mean, she — ***she*** fought very hard.
- ***Tremendous potential.*** I've gotten to know our country so well — ***tremendous potential***. It's going to be a beautiful thing.
- I want to tell the world community that while we will always put America's interests first, we will deal fairly with everyone, ***with everyone — all people and all other nations***.

Issues with this process: Text is messy

Taken from *tweet.onerandom.com*:

- Yasmeen is desperate to find her children in Syria LISTEN: http://bit.ly/1NYFvZC (@abcMatt @ Syria/Turkey border http://ab.co/1oiSssK)
- Good Morning Everyone , Happy Nice Day :D
- @Louis_Tomlinson You should do a concert in Greece at the next tour just saying
- Урожай собран - 20 еды! Ты тоже проверь свои грядки! http://gigam.es/imtw_Tribez #android, #androidgames, #gameinsight

Issues with this process: Other

- Semantic Confuzzlement
  - I ate spaghetti with meatballs. (ingredient of spaghetti)
  - I ate spaghetti with salad. (side dish of spaghetti)
  - I ate spaghetti with abandon. (manner of eating)
  - I ate spaghetti with a fork. (instrument of eating)
  - I ate spaghetti with a friend. (accompanier of eating)
- Idioms
  - Mrs. O'Leary's cow *kicked the bucket* over, starting the Great Chicago fire.*
  - Mrs O'Leary's cow finally *kicked the bucket* this morning. She's super dead. Good riddance.
  - Crap. My shoes are *broken*.
  - Hooray! My shoes are finally *broken in*.
- Sarcasm
  - Oh yeah, I *really wanted* you to spill your drink all over me.
  - Right... That polo *totally doesn't* make you look like a total douche-bro.

How can we possibly handle all this?

STATISTICS

and a lot of machine learning...

For many of these tasks:

1. Build or find a large body of text in a target language (corpus).
2. Find counts/frequencies/correlations/n-grams.
3. Apply the magic of statistics!
4. Run aforementioned magic through **machine learning** algorithms.
5. ????
6. PROFIT!!!

For example:

On Swype, the next word is suggested for each word you type.

"I will have a great day and I will be a good idea to *have a great day and I will...*"

I will have... (finished class by 7)
have a great day and... (I'll see you later!)
and I will be... (all ready to leave by 12)
(it would) ...be a good idea to... (grab some cash before you go)
(try) ...to have... (fun despite the rain)

"**Why** is it possible to get the best regards to the office and I will be in the office *and I will be in...*"

"**You** can get a chance to get the keys to the office *and I will be a good idea to...*"

Machine Learning

- Supervised learning:
  - Use already-labelled or classified data to automatically classify new data
  - Pro: easy to do
  - Con: requires lots of pre-processed data (4-10 hours for every hour of unprocessed data)
- Unsupervised learning:
  - Let the computer find patterns in the data which you can then classify
  - Pro: data doesn't need to be pre-processed
  - Con: hard to do well

- Classification tasks
  - Given two or more classes of data, determine which one new data should be put into
- Clustering
  - Cluster data into appropriate regions that are initially unknown

Ken Shan's Current Work

In 2014, Ken received a $1.4 million grant from DARPA to work on probabilistic programming.

With this, he hopes to create a means to allow for easier integration and application of machine learning algorithms, not only in NLP.

Researchers will be able to spend more time analyzing and optimizing results and less time writing equations for the algorithms.

Future of NLP

“Humans define what is ‘correct’. Because humans are always right, computers will always play catch-up.” ~ Ken Shan

Immediately, here are the most difficult problems in NLP in *English*:

- analyzing discourse structure (how sentences are connected and related)
- building relations between sentences (finding cause and effect)
- implementing and updating databases automatically with world knowledge

*Virtual assistants with whom you can converse and ask to solve any task.

Thank You!