‘Motifs, Shapelets and Discords’ is all you need!!

Yogesh Haribhau Kulkarni

Introduction to Sequences/Time-Series

Overview

xkcd

(Parables: verse/short stories with morals)

NLP is AI

Can Machine understand the way humans do?

Can Machine Answer?

(ENGLAND, June, 1989) - Christopher Robin is alive and well. He lives in England. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book. He made up a fairy tale land where Chris lived. His friends were animals. There was a bear called Winnie the Pooh. There was also an owl and a young pig, called a piglet. All the animals were stuffed toys that Chris owned. Mr. Robin made them come to life with his words. The places in the story were all near Cotchfield Farm. Winnie the Pooh was written in 1925. Children still love to read about Christopher Robin and his animal friends. Most people don't know he is a real person who is grown now. He has written two books of his own. They tell what it is like to be famous.

1. Who is Christopher Robin? 2. When was Winnie the Pooh written?
3. What did Mr. Robin do when Chris was three years old?
4. Where did young Chris live? 5. Why did Chris write two books of his own?

(Ref: CS598 DNR FALL 2005 Machine Learning in Natural Language - Dan Roth University of Illinois, Urbana-Champaign)

Understanding Questions

- What is the question asking? (different from Googling)
- Beyond finding candidate passages; choose the right one.
- Say, Q: What is the fastest automobile in the world?
- A1: . . . will stretch Volkswagen's lead in the world's fastest growing vehicle market. Demand for cars is expected to soar . . .
- A2: . . . the Jaguar XJ220 is the dearest (415,000 pounds), fastest (217mph) and most sought after car in the world.
- And, what if the answers require aggregation

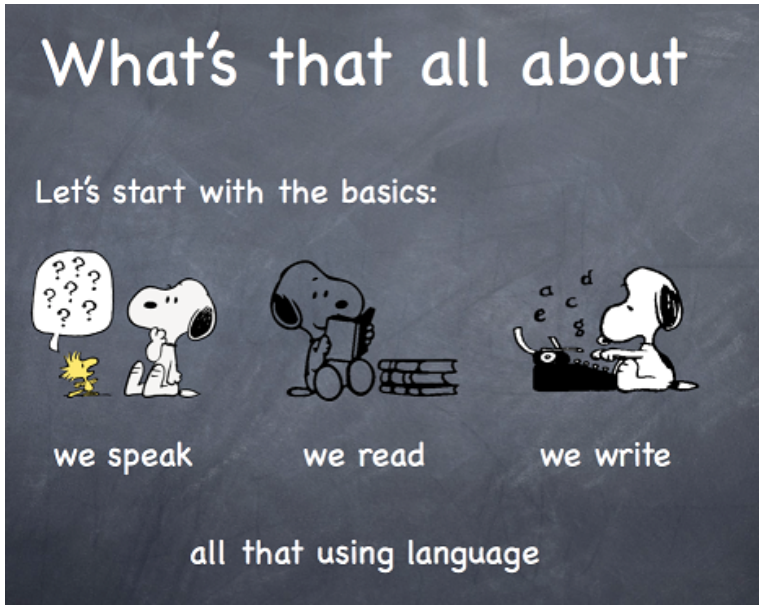
Not So Easy

Humans may not follow other humans, then what for the machines. But still we can attempt.

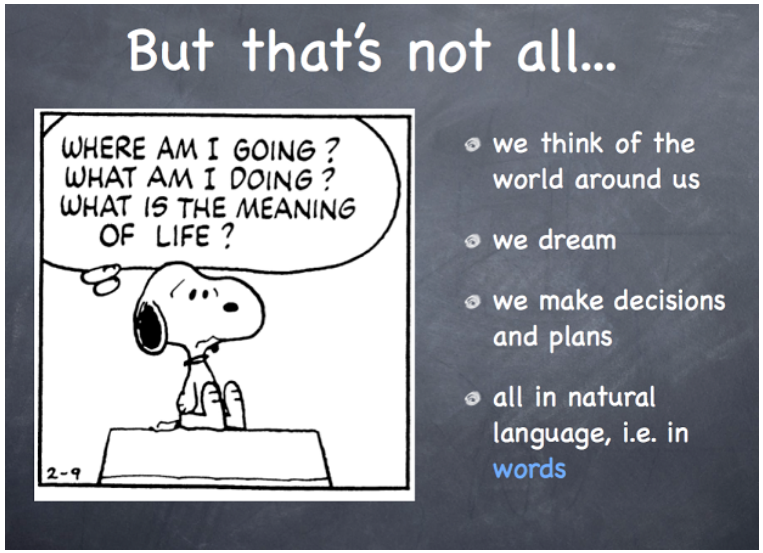
Need to study the language well!!!

(Ref: CS598 DNR FALL 2005 Machine Learning in Natural Language - Dan Roth University of Illinois, Urbana-Champaign)

What's a Language?



What's a Language?



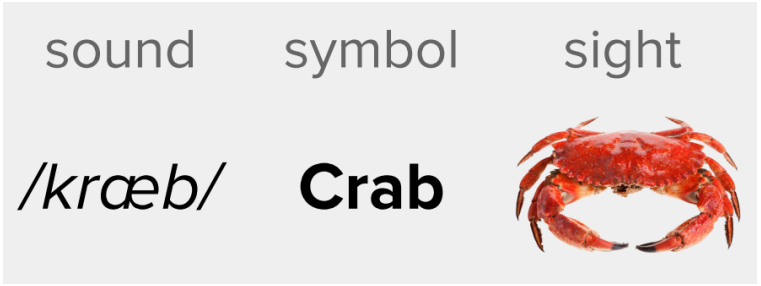
Words == Meaning?

When you read the word, say, "Crab", what does this mean to you?
Bunch of letters? or something else?

Words == Meaning?

Is the symbol representative of its meaning?

Words == Meaning?
Now, which of these you can understand?



Or is it just a mapping in a mental lexicon/vocabulary/dictionary (word, picture, understanding)?

These symbols are arbitrary. Do the words, embed knowledge by themselves?
But words is the only thing we have for processing? Then how what can be done?
Verbal and Written communication is primarily though words.

Why should we analyze words?
Words/Language is:

- Main channel of Communication
- Knowledge Acquisition

Knowledge and Communication in Language

- Human knowledge, human communication, is expressed in language
- Language technologies: process human language automatically
- Examples:
  - Hand-held devices: predictive text, handwriting recognition
  - Web search engines: access to information locked up in text

Questions

- Which Language systems you have recently used?
- What is impressive about these systems? (Imagine you stepped out of time machine from the 1970s or 1870s)
- How could these systems be improved?

Why Language Processing, and why now?

- Data: There is huge wealth of human knowledge that has been digitized. A Twitter firehouse of current events and trending topics.
- Compute: Our computers are very fast and there are ways to scale out.
- Algorithms: Using a combination of CS, ML, linguistic technique we can get the insight quickly.

Leveraging Big Data

- Examples are easier to create than rules.
- Rules and logic miss frequency and language dynamics
- More data is better for machine learning, relevance is in the long tail
- Knowledge engineering is not scalable
- Computational linguistics methodologies are stochastic

Question: What do computers like more numbers or words?

Questions

- How do we write programs to manipulate natural language?
- What questions about language could we answer?
- How would the programs work?
- What data would they need?
- First: what do they look like?

What is Natural Language Processing?

What is NLP?

Natual Language Processing

- Natural Language: languages spoken by people (English, French, German, etc.) as opposed to artificial languages, also called as Formal, (C++, Java, Python, etc.) built for computer manipulation
- Natural Language Processing: computer applications that automatically analyze natural language

What is NLP?

- Language = Words + Rules + Exceptions + ...
- Formally: dictionary (vocabulary) + grammar + ...
- Dictionary : set of words defined in the language (static or dynamic)
- Grammar: set of rules which describe what is allowable in a language

Formal vs. Natural Languages

Formal Languages

- Strict, unchanging rules defined by grammars and parsed by regular expressions
- Generally application specific (chemistry, math)
- Literal: exactly what is said is meant.
- No ambiguity
- Parsable by regular expressions
- Inflexible: no new terms or meaning.

Natural Languages

- Flexible, evolving language that occurs naturally in human communication
- Unspecific and used in many domains and applications
- Redundant and verbose in order to make up for ambiguity
- Expressive
- Difficult to parse
- Very flexible even in narrow contexts

The Richness of Natual Language

- Basic needs and lofty aspirations; technical know-how and flights of fantasy
- Ideas are shared over great separations of distance and time

Examples (think of processing these!!!):

- Overhead the day drives level and grey, hiding the sun by a flight of grey spears. (William Faulkner, *As I Lay Dying*, 1935)
- When using the toaster please ensure that the exhaust fan is turned on. (sign in dormitory kitchen)
- Amiodarone weakly inhibited CYP2C9, CYP2D6, and CYP3A4-mediated activities with Ki values of 45.1-271.6 μM (Medline)
- Iraqi Head Seeks Arms (spoof headline, http://www.snopes.com/humor/nonsense/head97.htm
- The earnest prayer of a righteous man has great power and wonderful results. (James 5:16b)
- Twas brillig, and the slithy toves did gyre and gimble in the wabe (Lewis Carroll, *Jabberwocky*, 1872)
- There are two ways to do this, AFAIK :smile: (internet discussion archive)

History of NLP/Text/Linguistics, a western view

History of NLP

Antiquity

What is language?

- Plato (~400 BCE)
  - *Cratylus* – dialogue about language
  - Word mappings are innate, not learned
  - Words are intrinsically related to their meanings
- Aristotle (~350 BCE)
  - Language is deductive, like physics
  - Abstracted impressions

Plato[1]

Aristotle[2]

12

1: https://www.britannica.com/biography/Plato 2: https://www.britannica.com/biography/Aristotle 3: (Tomasello, 2010) 4: (Miner, 2012)

Note: A deductive language is a computer programming language in which the program is a collection of predicates ('facts') and rules that connect them. Such a language is used to create knowledge based systems or expert systems which can deduce answers to problem sets by applying the rules to the facts they have been given.

(Ref: Text Mining - Jeff Shaul)

History of NLP

1950s

Automatic abstracting

- H.P. Luhn, 1958[1,2]
  - *The Automatic Creation of Literature Abstracts*
  - Pioneer of auto-abstracting for business intelligence
  - Used a room-sized computer: IBM 704
  - Statistical method for automatic abstract generation of technical articles
- Noam Chomsky, 1959[4]
  - *Syntactic Structures*
  - "generative grammar"
  - Rule-based descriptions of syntactic structures
  - Extremely influential in computational linguistics

H.P. Luhn[3]

Noam Chomsky[5]

14

1: (Luhn, 1958a) 2: (Luhn, 1958b) 3: https://www.asist.org/wp-content/uploads/2015/01/Hans_Luhn.png 4: (Chomsky, 1959) 5: http://radioopensource.org/vietnam-worst-war/

(Ref: Text Mining - Jeff Shaul)

History of NLP

Pre-1950s

Cataloguing

- Thomas Hyde, 1674[3]
  - Cataloguing by classification
  - Before Hyde, catalogued by author or title
- Melvil Dewey, 1876[4]
  - Index card catalog
  - Dewey decimal classification
- Claude Shannon, 1948[6]
  - *A Mathematical Theory of Communication*
  - "father of information theory"
  - Source > transmitter > channel > receiver > destination

Thomas Hyde[1]

Melvil Dewey[2]

Claude Shannon[5]

13

1: http://www.mindsacollection.org/lockes-library/ 2: http://alchetron.com/Melvil-Dewey-1187480-W 3: (Marshall, 2012) 4: (Satija, 2013)
5: http://www.newyorker.com/tech/elements/claude-shannon-the-father-of-the-information-age-turns-1100100 6: (Shannon, 1948)

(Ref: Text Mining - Jeff Shaul)

History of NLP

1960s

Hard-coded rules

- Lauren B. Doyle, 1961[2]
  - *Semantic Road Maps for Literature Searchers*
  - Classify documents based on word frequencies and associations
- Weizenbaum, 1966[3]
  - *ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine*
  - First system to detect medical terms in text with pattern matching
- Pratt & Pacak, 1969[4]
  - *Identification and Transformation of Terminal Morphemes in Medical English*
  - Outlined a system for automated processing of medical English
  - Introduced "kernel phrases" (later entities)

The first chatterbot[1]

15

1: http://aseelghazal.weebly.com/uploads/4/0/4/5/40450643/1420310801.png 2: (Doyle, 1961) 3: (Weizenbaum, 1966) 4: (Pratt, 1969)

(Ref: Text Mining - Jeff Shaul)

History of NLP

1970s

New methods and metrics

- Codd, 1970[1]
  - *A Relational Model of Data for Large Shared Data Banks*
  - Revolutionary paper that invented relational databases
- Earl, 1970[2]
  - *Experiments in Automatic Extracting and Indexing*
  - Combines statistics and syntax to find key sentences
- Salton *et al*, 1974[3]
  - *A Theory of Term Importance in Automatic Text Analysis*
  - Invented discrimination value analysis
  - Metric of how well a term can identify a document
- Salton *et al*, 1975[5]
  - *A Vector Space Model for Automatic Indexing*
  - Introduced the vector space model and TF-IDF scoring

Gerard Salton[4]

Edgar Codd[6]

1: (Codd, 1970) 2: (Earl, 1970) 3: (Salton, 1975a) 4: http://www.ef.org/fellows/all-fellows/gerard-salton/ 5: (Salton, 1975b)
6: https://www.nap.edu/read/12473/chapter/15

16

(Ref: Text Mining - Jeff Shaul)

History of NLP

Early 1990s

Machine learning

- Cutting *et al*, 1992[2]
  - *Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections*
  - Improved clustering algorithms from $O(n^2)$ to $O(n)$
- Hafner *et al*, 1994[3]
  - *Creating a Knowledge Base of Biological Research Papers*
  - Realization by genomic scientists that NLP could be useful
- Apté *et al*, 1994[5]
  - *Automated Learning of Decision Rules for Text Categorization*
  - First use of decision trees for text categorization

Douglass Cutting[1]

1: https://spark-summit.org/2016/speakers/doug-cutting/ 2: (Cutting, 1992) 3: (Hafner, 1994) 4: (Salton, 1994) 5: (Apté, 1994)

18

(Ref: Text Mining - Jeff Shaul)

History of NLP

1980s

Shifting to statistical NLP

- Shapiro, 1980[1]
  - *A System for Conceptual Analysis of Medical Practices* (SCAMP)
  - SCAMP – retrieved patient records matching concepts
- Friedman *et al*, 1983[2]
  - *Computer Structuring of Free-Text Patient Data*
  - Automated database-building using semantic analysis of patient records
- Swanson, 1986[3]
  - *Undiscovered Public Knowledge*
  - Articulation of the growing realization that knowledge is embedded within unstructured text, waiting to be discovered by computers
- Deerwester, 1988[5]
  - *Using Latent Semantic Analysis to Improve Access to Textual Information*
  - invented latent semantic indexing (LSI) – uses singular value decomposition to identify patterns

Carol Friedman[7]

Don R. Swanson[6]

1: (Shapiro, 1980) 2: (Friedman, 1983) 3: (Swanson, 1986) 4: (Hayes, 1988) 5: (Deerwester, 1988)
6: http://www.the-scientist.com/?articles.view/articleNo/18922/title/Program-Uncovers-Hidden-Connections-In-The-Literature/
7: https://www.dbmi.columbia.edu/people/carol-friedman/

17

(Ref: Text Mining - Jeff Shaul)

History of NLP

Mid 1990s

Biologists adopt NLP

- Cortes & Vapnik, 1995[3]
  - *Support-Vector Networks*
  - First to use support-vector machines (SVM) for text classification
- Grishman & Sundheim, 1996[4]
  - Presented at Message Understanding Conference 6
  - Coined "named entity recognition"
- Fukuda *et al*, 1998[5]
  - *Toward Information Extraction: Identifying Protein Names from Biological Papers*
  - Early example of named entity recognition of proteins
- Landauer *et al*, 1998[6]
  - *An Introduction to Latent Semantic Analysis*
  - Statistical approach to resolve contextual meanings of words

Ralph Grishman[1]

Corinna Cortes[2]

1: http://www.tsdconference.org/tsd2014/gallery/friday.html 2: http://research.google.com/pubs/author121.html
3: (Cortes, 1995) 4: (Grishman, 1996) 5: (Fukuda, 1998) 6: (Landauer, 1998) 7: (Andrade, 1998)

19

(Ref: Text Mining - Jeff Shaul)

History of NLP

1999

Explosion

- Blaschke *et al*[3]
  - *Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions*
  - First example of relationship extraction in biological context
- Tanabe *et al*[5]
  - *MedMiner: An Internet Tool for Filtering and Organizing Biomedical Information*
  - Easily-accessible tool for exploring gene-gene and gene-drug relationships
- Smalheiser & Swanson[6]
  - *Implicit Text Linkages between Medline Records: Using Arrowsmith as an Aid to Scientific Discovery*
  - Improved literature-based discovery tool that found much more use by medical scientists

Christian Blaschke[1]

1: http://www.tsdconference.org/tsd2014/gallery/friday.html 2: http://dblp.uni-trier.de/pers/hd/d/D=ouml=rre:Jochen
3: (Blaschke, 1999) 4: (Craven, 1999) 5: (Tanabe, 1999) 6: (Swanson, 1999) 7: (Dörre, 1999)

20

(Ref: Text Mining - Jeff Shaul)

History of NLP

Early 2000s

Tools, tools, tools

- Friedman *et al*, 2001[3]
  - *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles.*
  - First widespread/successful tool for extracting structured information from genomic literature
- Collins & Duffy, 2002[4]
  - *New ranking algorithms for parsing and tagging*
  - Improved use of decision trees in text mining
- Hotho *et al*, 2003[5]
  - *Wordnet improves Text Document Clustering*
  - Formal use of domain ontologies/lexicons
- Müller *et al*, 2004[6]
  - *Textpresso: An ontology-based information retrieval and extraction system for biological literature*
  - Robust literature-based discovery system

Hans-Michael Müller[1]

Andreas Hotho[2]

1: http://www.jbc.org/content/290/45/27015/suppl/DCAuthor_profile_HM 2: http://www.dmir.uni-wuerzburg.de/staff/hotho/
3: (Friedman, 2001) 4: (Collins, 2002) 5: (Hotho, 2003) 6: (Müller, 2004)

21

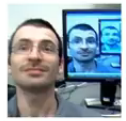
(Ref: Text Mining - Jeff Shaul)

History of NLP

Late 2000s

Statistical NER matures

- Chen *et al*, 2008[4]
  - *Automated Acquisition of Disease – Drug Knowledge from Biomedical and Clinical Documents: An Initial Study*
  - One of the first applications of NLP to finding disease-drug interactions in the biomedical literature
- Krallinger *et al*, 2008[5]
  - *Linking genes to literature: text mining, information extraction, and retrieval applications for biology.*
  - Robust automated annotation system
- Ratinov & Roth, 2009[6]
  - *Design challenges and misconceptions in named entity recognition*
  - Showed that statistical methods are better for NER



Lev-Arie Ratinov[1]

Elizabeth Chen[2]

1: https://www.linkedin.com/in/lev-arie-ratinov-9869201 2: http://www.mbg.ihmi.edu/people/faculty/elizabeth-chen
3: (Senellart, 2008) 4: (Chen, 2008) 5: (Krallinger, 2008) 6: (Ratinov, 2009)

22

(Ref: Text Mining - Jeff Shaul)

History of NLP

2010s

Supervision is for kids

- Collobert *et al*, 2011[3]
  - *Natural Language Processing (Almost) from Scratch*
  - Unified and unsupervised neural network framework for all tasks of NLP
- Das *et al*, 2011[4]
  - *Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections*
  - Widely applicable to languages without established ontologies/lexicons
- Zhang *et al*, 2013[5]
  - *Unsupervised Biomedical Named Entity Recognition: Experiments with Clinical and Biological Texts*
  - Stepwise approach to NER without any training data or heuristics

Ronan Collobert[1]

Dipanjan Das[2]

1: http://ronan.collobert.com/ 2: http://research.google.com/pubs/DipanjanDas.html
3: (Collobert, 2011) 4: (Das, 2011) 5: (Zhang, 2013) 6: (Zhang, 2016)

23

(Ref: Text Mining - Jeff Shaul)

Why NLP is hard?

Why that is hard?

NLP itself is hard. Why?

Brainstorm the reasons NLP is hard.

Textual Queries, can NLP answer?

Example 1

Text Never before had ski racing, a sport dominated by monosyllabic mountain men, seen the likes of Alberto Tomba, the flamboyant Bolognese flatlander who at 21 captured two gold medals at the Calgary Olympics.

Hypothesis Alberto Tomba won a race.

Textual Queries, can NLP answer?

Example 2

Text Researchers at the Harvard School of Public Health say that people who drink coffee may be doing a lot more than keeping themselves awake—this kind of consumption apparently also can help reduce the risk of diseases.

Hypothesis Coffee drinking has health benefits.

Why NLP is hard?

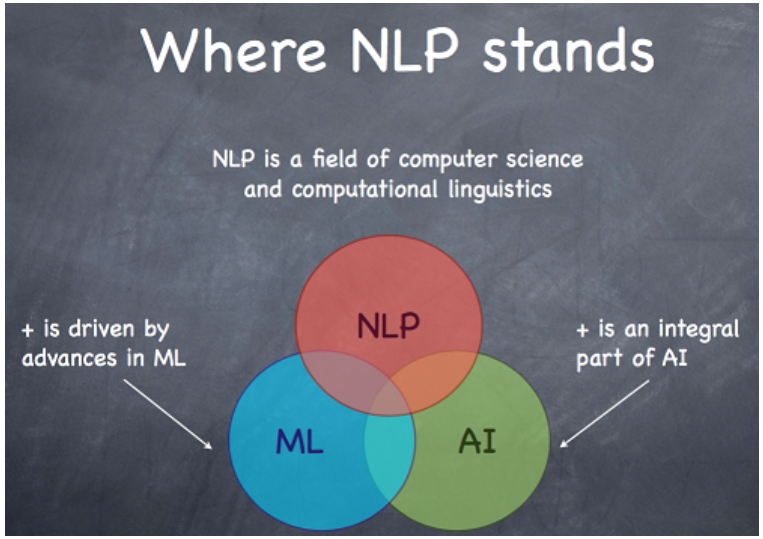
- Ambiguity: There are many different ways to represent the same thing.
- Language is inherently very high dimensional and sparse. There are a lot of rare words.
- Out of sample generalization: New words and new sentences all the time.
- Order and context are extremely important. "Dog bites man" and "Man bites dog" have vastly different meanings even though they differ by a very small amount.
- Anything else?

Where NLP stands?

Where is it useful?

- Information Extraction
- Spoken Dialog
- Question Answering
- Text Summarization
- Etc . . .

Where NLP stands?



Main approaches in NLP

- Rule-based methods: eg Regular expressions.
- Machine learning: eg Linear classifiers.
- Deep Learning: eg Recurrent Neural Networks

Example: rule based approach

Let's say, you want to find slots (City, date, etc) from following sentence:
"Show me flights from Boston to San Francisco on Tuesday"
Context-free grammar is defined by "rules" with substitution syntax such as:

- SHOW: show me — i want — can i see
- FLIGHTS: (a) flight — flights
- ORIGIN: from CITY
- DESTINATION: to CITY
- CITY: Boston — San Francisco — Denver — Washington

Need to parse the sentence and wherever you find any of the right-hand side words, you put left hand side annotation

(Ref: https://web.stanford.edu/ jurafsky/slp3/29.pdf)

Explanation: What is Context Free Grammar?

Context-free grammars are named as such because any of the production rules in the grammar can be applied regardless of context. It does not depend on any other symbols that may or may not be around a given symbol that is having a rule applied to it.

- Context-free Grammars allow a non-terminal to be replaced by a corresponding production rule whenever it appears in a derivation process. The replacement occurs irrespective of what lies before or after the non-terminal. This happens because the LHS of a production rule allows only a single non-terminal of the form :

  $V \rightarrow (V + T)*$, where V is a non-terminal and T is a terminal

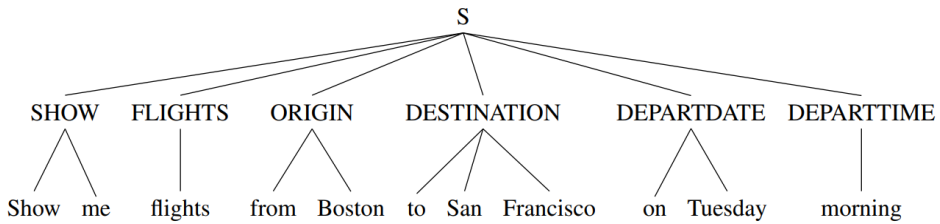
  Hence, the dependence on context is removed owing to this restriction.

- Context-sensitive Grammars being stronger of the two allows replacement of a non-terminal only based on it's left or right context,i.e, depending on the terminals or non-terminals that precedes or succeeds it.

(Ref: https://www.quora.com/What-is-the-meaning-of-context-free-in-context-free-grammar)

Example: rule based approach

"Show me fligths from Boston to San Francisco on Tuesday"
Parsing results in:



(Ref: https://web.stanford.edu/ jurafsky/slp3/29.pdf)

Example: machine learning approach

CRF (Conditional Random Field): probabilistic Named entity recognition.
Training Corpus:

Feature engineering:

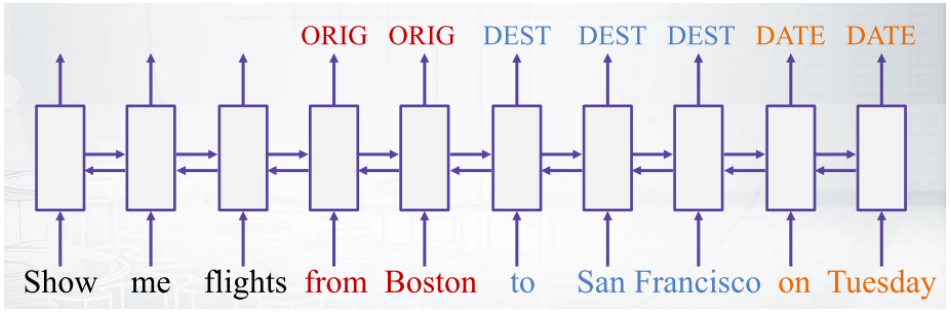
- Is the word capitalized?
- Is the word in a list of city names?
- What is the previous word?
- What is the previous slot?

(Ref: https://www.coursera.org/learn/language-processing/lecture/j8kee/main-approaches-in-nlp)

Example: deep learning approach

LSTM (Long Short Term Memory):

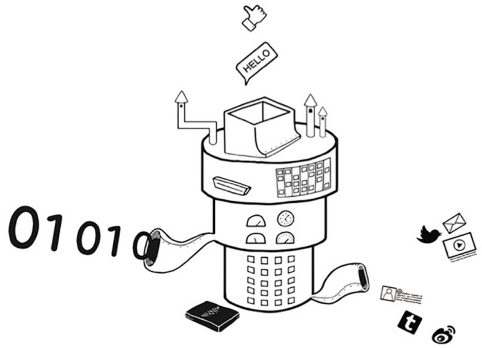
- Big training corpus
- No feature generation
- Defining the model
- Training and inference



(Ref: https://www.coursera.org/learn/language-processing/lecture/j8kee/main-approaches-in-nlp)

Goal of NLP

- To convert letters, words, and ideas into numbers.
- Once we have the numbers we use math and machine learning.



The Promise of NLP

- Importance in scientific, economic, social and cultural arenas
- Growing rapidly as its theories and methods are deployed in new technologies
- Therefore a wide range of people should have a working knowledge of NLP
  - Academia: humanities computing, corpus linguistics, computer science, artificial intelligence
  - Industry: HCI, business information analysis, web software development
- The goal is to open the field of NLP to a broad audience.

NLP and Intelligence

- Long-standing challenge to build intelligent machines
- Chief measure of machine intelligence has been linguistic: Turing test

NLP and Intelligence

- Research on spoken dialog systems, also MT — *integrated NLP systems which future users would regard as highly intelligent*
- Example human-machine dialog:

S: How may I help you?
U: When is Saving Private Ryan playing?
S: For what theater?
U: The Paramount theater.
S: Saving Private Ryan is not playing at the Paramount theater, but it's playing at the Madison theater at 3:00, 5:30, and 10:30.

NLP and Intelligence (cont)

- Today's systems limited to narrowly defined domains
- Couldn't ask above system for other information, e.g.:
  - Driving instructions
  - Details of nearby restaurants
- To add such support we would have to:
  - Store the required information
  - Incorporate suitable questions and answers into the system
- Common-sense reasoning vs business logic
- Need to make progress on natural linguistic interaction without recourse to this unrestricted knowledge and reasoning capability

Recent NLP Applications

- Sentiment Analysis - Deriving sentiments in sentences (positive, negative, neutral), and also in articles (though that will be more appropriate like bag of sentence sentiments). The future is to include emotions (attributes) in that, like the attributes now on Facebook posts - Love, Like, Angry, Surprised, Sad, Hilarious. These attributes make a lot more sense for sentiments going forward.
- Text Summarization - Summarizing a single or many articles according to a particular theme.
- Information Extraction - Find structured information from unstructured data, like entities, relationships, co-reference resolution. This at a basic level is very useful for algorithmic trading. An extension of this is a global form of extracting logic structures (first order and higher order).

Recent NLP Applications

- Topic Segmentation - Topic Extraction (with regions). Normally, there will be overlapping regions.
- Question Answering - Answer the questions to both closed (specific) and open questions (subjective). Answers to subjective questions is the main challenge for the likes of realistic Virtual Assistants.
- Parsing - Parsing natural language generally in the form a tree. This involves hierarchical segmentation of the language involving the grammar rules.
- Prediction - Given a short text, predict what happens next. The prediction problem is beginning to be targeted in vision, but it has never ever gained paths for realistic products. For closed and deterministic prediction (not innovative else that would fall under the paradigm of creative writing), this can be a useful task for prediction of future events based on past evidences and analysis. This can be then very useful for finance sectors.

Recent NLP Applications

- Part of Speech Tagging (POS) - Tagging words whether they are nouns, verbs or adjectives.
- Translation - Translate one language to another. This can be very challenging given the nature of the language, and the grammar. Normally, under probabilistic models, this assumes that the underlying grammar is mostly the same, and thus, models normally fail for Sanskrit.
- Interestingness - Most interesting portion of text in an article. This can be done very much on the same lines as in images, where one ranks the likeness of images.

In nutshell

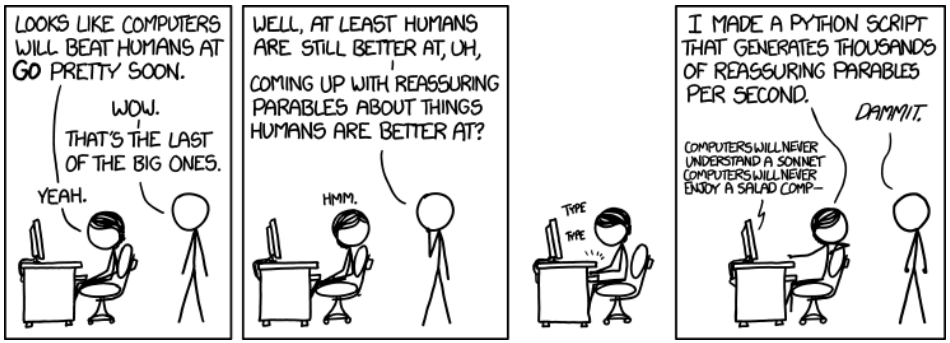
- NLP is an effort to do useful things with the natural language.
- NLP is hard because humans like words and computers like numbers.

In nutshell

Sequence Segmentation

Overview

xkcd



(Parables: verse/short stories with morals)

NLP is AI

Can Machine understand the way humans do?

Can Machine Answer?

(ENGLAND, June, 1989) – Christopher Robin is alive and well. He lives in England. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book. He made up a fairy tale land where Chris lived. His friends were animals. There was a bear called Winnie the Pooh. There was also an owl and a young pig, called a piglet. All the animals were stuffed toys that Chris owned. Mr. Robin made them come to life with his words. The places in the story were all near Cotchfield Farm. Winnie the Pooh was written in 1925. Children still love to read about Christopher Robin and his animal friends. Most people don't know he is a real person who is grown now. He has written two books of his own. They tell what it is like to be famous.

1. Who is Christopher Robin? 2. When was Winnie the Pooh written?
3. What did Mr. Robin do when Chris was three years old?
4. Where did young Chris live? 5. Why did Chris write two books of his own?

(Ref: CS598 DNR FALL 2005 Machine Learning in Natural Language - Dan Roth University of Illinois, Urbana-Champaign)

Understanding Questions

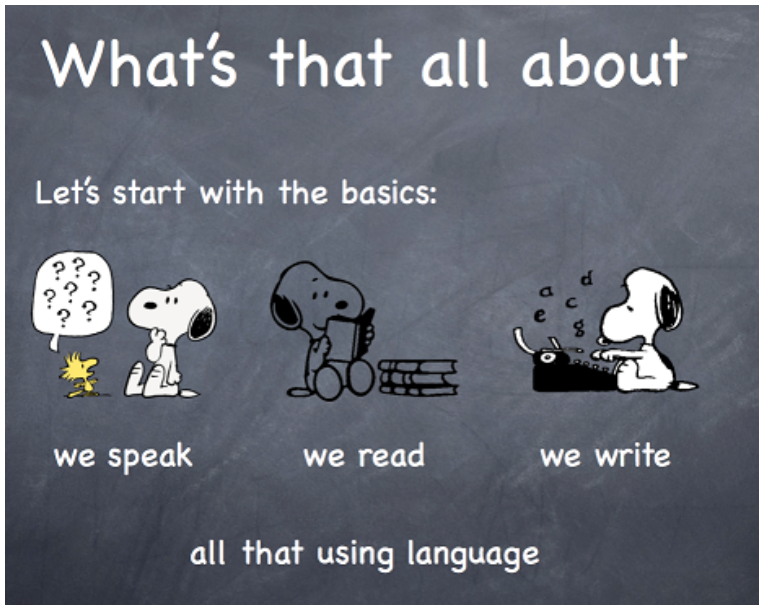
- What is the question asking? (different from Googling)
- Beyond finding candidate passages; choose the right one.
- Say, Q: What is the fastest automobile in the world?
- A1: . . . will stretch Volkswagen's lead in the world's fastest growing vehicle market. Demand for cars is expected to soar . . .
- A2: . . . the Jaguar XJ220 is the dearest (415,000 pounds), fastest (217mph) and most sought after car in the world.
- And, what if the answers require aggregation

Not So Easy

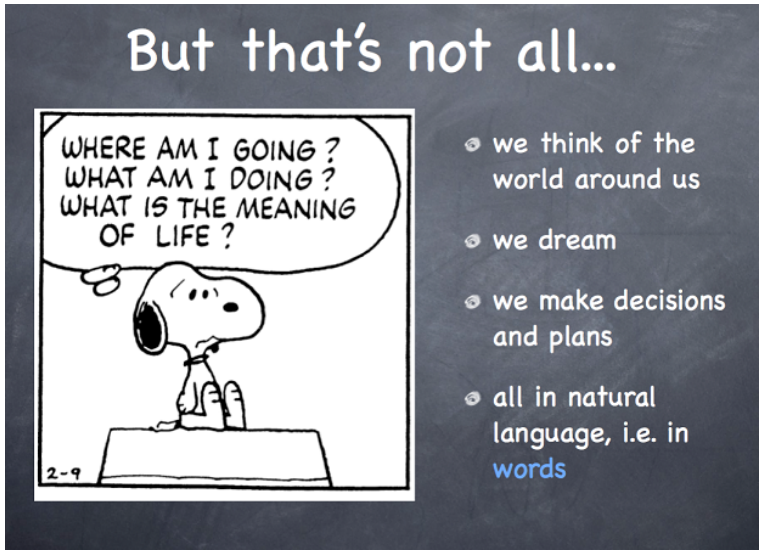
Humans may not follow other humans, then what for the machines. But still we can attempt. Need to study the language well!!!

(Ref: CS598 DNR FALL 2005 Machine Learning in Natural Language - Dan Roth University of Illinois, Urbana-Champaign)

What's a Language?



What's a Language?



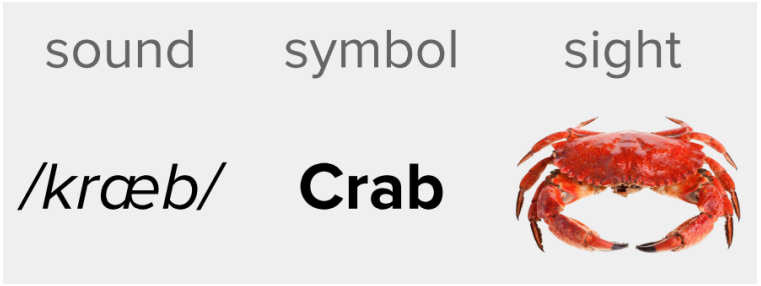
Words == *Meaning*?

When you read the word, say, "Crab", what does this mean to you?
Bunch of letters? or something else?

Words == *Meaning*?

Is the symbol representative of its meaning?

Words == *Meaning*?
Now, which of these you can understand?



Or is it just a mapping in a mental lexicon/vocabulary/dictionary (word, picture, understanding)?

These symbols are arbitrary. Do the words, embed knowledge by themselves?
But words is the only thing we have for processing? Then how what can be done?
Verbal and Written communication is primarily though words.

Why should we analyze words?

Words/Language is:

- Main channel of Communication
- Knowledge Acquisition

Knowledge and Communication in Language

- Human knowledge, human communication, is expressed in language
- Language technologies: process human language automatically
- Examples:
  - Hand-held devices: predictive text, handwriting recognition
  - Web search engines: access to information locked up in text

Questions

- Which Language systems you have recently used?
- What is impressive about these systems? (Imagine you stepped out of time machine from the 1970s or 1870s)
- How could these systems be improved?

Why Language Processing, and why now?

- Data: There is huge wealth of human knowledge that has been digitized. A Twitter firehouse of current events and trending topics.
- Compute: Our computers are very fast and there are ways to scale out.
- Algorithms: Using a combination of CS, ML, linguistic technique we can get the insight quickly.

Leveraging Big Data

- Examples are easier to create than rules.
- Rules and logic miss frequency and language dynamics
- More data is better for machine learning, relevance is in the long tail
- Knowledge engineering is not scalable
- Computational linguistics methodologies are stochastic

Question: What do computers like more numbers or words?

Questions

- How do we write programs to manipulate natural language?
- What questions about language could we answer?
- How would the programs work?
- What data would they need?
- First: what do they look like?

What is Natural Language Processing?

What is NLP?

Natual Language Processing

- Natural Language: languages spoken by people (English, French, German, etc.) as opposed to artificial languages, also called as Formal, (C++, Java, Python, etc.) built for computer manipulation
- Natural Language Processing: computer applications that automatically analyze natural language

What is NLP?

- Language = Words + Rules + Exceptions + ...
- Formally: dictionary (vocabulary) + grammar + ...
- Dictionary : set of words defined in the language (static or dynamic)
- Grammar: set of rules which describe what is allowable in a language

Formal vs. Natural Languages

Formal Languages

- Strict, unchanging rules defined by grammars and parsed by regular expressions
- Generally application specific (chemistry, math)
- Literal: exactly what is said is meant.
- No ambiguity
- Parsable by regular expressions
- Inflexible: no new terms or meaning.

Natural Languages

- Flexible, evolving language that occurs naturally in human communication
- Unspecific and used in many domains and applications
- Redundant and verbose in order to make up for ambiguity
- Expressive
- Difficult to parse
- Very flexible even in narrow contexts

The Richness of Natual Language

- Basic needs and lofty aspirations; technical know-how and flights of fantasy
- Ideas are shared over great separations of distance and time

Examples (think of processing these!!!):

- Overhead the day drives level and grey, hiding the sun by a flight of grey spears. (William Faulkner, *As I Lay Dying*, 1935)
- When using the toaster please ensure that the exhaust fan is turned on. (sign in dormitory kitchen)
- Amiodarone weakly inhibited CYP2C9, CYP2D6, and CYP3A4-mediated activities with Ki values of 45.1-271.6 μM (Medline)
- Iraqi Head Seeks Arms (spoof headline, http://www.snopes.com/humor/nonsense/head97.htm
- The earnest prayer of a righteous man has great power and wonderful results. (James 5:16b)
- Twas brillig, and the slithy toves did gyre and gimble in the wabe (Lewis Carroll, *Jabberwocky*, 1872)
- There are two ways to do this, AFAIK :smile: (internet discussion archive)

History of NLP/Text/Linguistics, a western view

History of NLP

Antiquity

What is language?

- Plato (~400 BCE)
  - *Cratylus* – dialogue about language
  - Word mappings are innate, not learned
  - Words are intrinsically related to their meanings
- Aristotle (~350 BCE)
  - Language is deductive, like physics
  - Abstracted impressions

Plato[1]

Aristotle[2]

12

1: https://www.britannica.com/biography/Plato 2: https://www.britannica.com/biography/Aristotle 3: (Tomasello, 2010) 4: (Miner, 2012)

Note: A deductive language is a computer programming language in which the program is a collection of predicates ('facts') and rules that connect them. Such a language is used to create knowledge based systems or expert systems which can deduce answers to problem sets by applying the rules to the facts they have been given.

(Ref: Text Mining - Jeff Shaul)

History of NLP

1950s

Automatic abstracting

- H.P. Luhn, 1958[1,2]
  - *The Automatic Creation of Literature Abstracts*
  - Pioneer of auto-abstracting for business intelligence
  - Used a room-sized computer: IBM 704
  - Statistical method for automatic abstract generation of technical articles
- Noam Chomsky, 1959[4]
  - *Syntactic Structures*
  - "generative grammar"
  - Rule-based descriptions of syntactic structures
  - Extremely influential in computational linguistics

H.P. Luhn[3]

Noam Chomsky[5]

14

1: (Luhn, 1958a) 2: (Luhn, 1958b) 3: https://www.asist.org/wp-content/uploads/2015/01/Hans_Luhn.png 4: (Chomsky, 1959) 5: http://radioopensource.org/vietnam-worst-war/

(Ref: Text Mining - Jeff Shaul)

History of NLP

Pre-1950s

Cataloguing

- Thomas Hyde, 1674[3]
  - Cataloguing by classification
  - Before Hyde, catalogued by author or title
- Melvil Dewey, 1876[4]
  - Index card catalog
  - Dewey decimal classification
- Claude Shannon, 1948[6]
  - *A Mathematical Theory of Communication*
  - "father of information theory"
  - Source > transmitter > channel > receiver > destination

Thomas Hyde[1]

Melvil Dewey[2]

Claude Shannon[5]

13

1: http://www.mindsacollection.org/lockes-library/ 2: http://alchetron.com/Melvil-Dewey-1187480-W 3: (Marshall, 2012) 4: (Satija, 2013)
5: http://www.newyorker.com/tech/elements/claude-shannon-the-father-of-the-information-age-turns-1100100 6: (Shannon, 1948)

(Ref: Text Mining - Jeff Shaul)

History of NLP

1960s

Hard-coded rules

- Lauren B. Doyle, 1961[2]
  - *Semantic Road Maps for Literature Searchers*
  - Classify documents based on word frequencies and associations
- Weizenbaum, 1966[3]
  - *ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine*
  - First system to detect medical terms in text with pattern matching
- Pratt & Pacak, 1969[4]
  - *Identification and Transformation of Terminal Morphemes in Medical English*
  - Outlined a system for automated processing of medical English
  - Introduced "kernel phrases" (later entities)

The first chatterbot[1]

15

1: http://aseelghazal.weebly.com/uploads/4/0/4/5/40450643/1420310801.png 2: (Doyle, 1961) 3: (Weizenbaum, 1966) 4: (Pratt, 1969)

(Ref: Text Mining - Jeff Shaul)

History of NLP

1970s

New methods and metrics

- Codd, 1970[1]
  - *A Relational Model of Data for Large Shared Data Banks*
  - Revolutionary paper that invented relational databases
- Earl, 1970[2]
  - *Experiments in Automatic Extracting and Indexing*
  - Combines statistics and syntax to find key sentences
- Salton *et al*, 1974[3]
  - *A Theory of Term Importance in Automatic Text Analysis*
  - Invented discrimination value analysis
  - Metric of how well a term can identify a document
- Salton *et al*, 1975[5]
  - *A Vector Space Model for Automatic Indexing*
  - Introduced the vector space model and TF-IDF scoring

Gerard Salton[4]

Edgar Codd[6]

1: (Codd, 1970) 2: (Earl, 1970) 3: (Salton, 1975a) 4: http://www.ef.org/fellows/all-fellows/gerard-salton/ 5: (Salton, 1975b)
6: https://www.nap.edu/read/12473/chapter/15

16

(Ref: Text Mining - Jeff Shaul)

History of NLP

Early 1990s

Machine learning

- Cutting *et al*, 1992[2]
  - *Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections*
  - Improved clustering algorithms from $O(n^2)$ to $O(n)$
- Hafner *et al*, 1994[3]
  - *Creating a Knowledge Base of Biological Research Papers*
  - Realization by genomic scientists that NLP could be useful
- Apté *et al*, 1994[5]
  - *Automated Learning of Decision Rules for Text Categorization*
  - First use of decision trees for text categorization

Douglass Cutting[1]

1: https://spark-summit.org/2016/speakers/doug-cutting/ 2: (Cutting, 1992) 3: (Hafner, 1994) 4: (Salton, 1994) 5: (Apté, 1994)

18

(Ref: Text Mining - Jeff Shaul)

History of NLP

1980s

Shifting to statistical NLP

- Shapiro, 1980[1]
  - *A System for Conceptual Analysis of Medical Practices* (SCAMP)
  - SCAMP – retrieved patient records matching concepts
- Friedman *et al*, 1983[2]
  - *Computer Structuring of Free-Text Patient Data*
  - Automated database-building using semantic analysis of patient records
- Swanson, 1986[3]
  - *Undiscovered Public Knowledge*
  - Articulation of the growing realization that knowledge is embedded within unstructured text, waiting to be discovered by computers
- Deerwester, 1988[5]
  - *Using Latent Semantic Analysis to Improve Access to Textual Information*
  - invented latent semantic indexing (LSI) – uses singular value decomposition to identify patterns

Carol Friedman[7]

Don R. Swanson[6]

1: (Shapiro, 1980) 2: (Friedman, 1983) 3: (Swanson, 1986) 4: (Hayes, 1988) 5: (Deerwester, 1988)
6: http://www.the-scientist.com/?articles.view/articleNo/18922/title/Program-Uncovers-Hidden-Connections-In-The-Literature/
7: https://www.dbmi.columbia.edu/people/carol-friedman/

17

(Ref: Text Mining - Jeff Shaul)

History of NLP

Mid 1990s

Biologists adopt NLP

- Cortes & Vapnik, 1995[3]
  - *Support-Vector Networks*
  - First to use support-vector machines (SVM) for text classification
- Grishman & Sundheim, 1996[4]
  - Presented at Message Understanding Conference 6
  - Coined "named entity recognition"
- Fukuda *et al*, 1998[5]
  - *Toward Information Extraction: Identifying Protein Names from Biological Papers*
  - Early example of named entity recognition of proteins
- Landauer *et al*, 1998[6]
  - *An Introduction to Latent Semantic Analysis*
  - Statistical approach to resolve contextual meanings of words

Ralph Grishman[1]

Corinna Cortes[2]

1: http://www.tsdconference.org/tsd2014/gallery/friday.html 2: http://research.google.com/pubs/author121.html
3: (Cortes, 1995) 4: (Grishman, 1996) 5: (Fukuda, 1998) 6: (Landauer, 1998) 7: (Andrade, 1998)

19

(Ref: Text Mining - Jeff Shaul)

History of NLP

1999

Explosion

- Blaschke *et al*[3]
  - *Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions*
  - First example of relationship extraction in biological context
- Tanabe *et al*[5]
  - *MedMiner: An Internet Tool for Filtering and Organizing Biomedical Information*
  - Easily-accessible tool for exploring gene-gene and gene-drug relationships
- Smalheiser & Swanson[6]
  - *Implicit Text Linkages between Medline Records: Using Arrowsmith as an Aid to Scientific Discovery*
  - Improved literature-based discovery tool that found much more use by medical scientists

Christian Blaschke[1]

1: http://www.tsdconference.org/tsd2014/gallery/friday.html 2: http://dblp.uni-trier.de/pers/hd/d/D=ouml=rre:Jochen
3: (Blaschke, 1999) 4: (Craven, 1999) 5: (Tanabe, 1999) 6: (Swanson, 1999) 7: (Dörre, 1999)

20

(Ref: Text Mining - Jeff Shaul)

History of NLP

Early 2000s

Tools, tools, tools

- Friedman *et al*, 2001[3]
  - *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles.*
  - First widespread/successful tool for extracting structured information from genomic literature
- Collins & Duffy, 2002[4]
  - *New ranking algorithms for parsing and tagging*
  - Improved use of decision trees in text mining
- Hotho *et al*, 2003[5]
  - *Wordnet improves Text Document Clustering*
  - Formal use of domain ontologies/lexicons
- Müller *et al*, 2004[6]
  - *Textpresso: An ontology-based information retrieval and extraction system for biological literature*
  - Robust literature-based discovery system

Hans-Michael Müller[1]

Andreas Hotho[2]

1: http://www.jbc.org/content/290/45/27015/suppl/DCAuthor_profile_HM 2: http://www.dmir.uni-wuerzburg.de/staff/hotho/
3: (Friedman, 2001) 4: (Collins, 2002) 5: (Hotho, 2003) 6: (Müller, 2004)

21

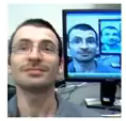
(Ref: Text Mining - Jeff Shaul)

History of NLP

Late 2000s

Statistical NER matures

- Chen *et al*, 2008[4]
  - *Automated Acquisition of Disease – Drug Knowledge from Biomedical and Clinical Documents: An Initial Study*
  - One of the first applications of NLP to finding disease-drug interactions in the biomedical literature
- Krallinger *et al*, 2008[5]
  - *Linking genes to literature: text mining, information extraction, and retrieval applications for biology.*
  - Robust automated annotation system
- Ratinov & Roth, 2009[6]
  - *Design challenges and misconceptions in named entity recognition*
  - Showed that statistical methods are better for NER



Lev-Arie Ratinov[1]

Elizabeth Chen[2]

1: https://www.linkedin.com/in/lev-arie-ratinov-9869201 2: http://www.mbg.ihmi.edu/people/faculty/elizabeth-chen
3: (Senellart, 2008) 4: (Chen, 2008) 5: (Krallinger, 2008) 6: (Ratinov, 2009)

22

(Ref: Text Mining - Jeff Shaul)

History of NLP

2010s

Supervision is for kids

- Collobert *et al*, 2011[3]
  - *Natural Language Processing (Almost) from Scratch*
  - Unified and unsupervised neural network framework for all tasks of NLP
- Das *et al*, 2011[4]
  - *Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections*
  - Widely applicable to languages without established ontologies/lexicons
- Zhang *et al*, 2013[5]
  - *Unsupervised Biomedical Named Entity Recognition: Experiments with Clinical and Biological Texts*
  - Stepwise approach to NER without any training data or heuristics

Ronan Collobert[1]

Dipanjan Das[2]

1: http://ronan.collobert.com/ 2: http://research.google.com/pubs/DipanjanDas.html
3: (Collobert, 2011) 4: (Das, 2011) 5: (Zhang, 2013) 6: (Zhang, 2016)

23

(Ref: Text Mining - Jeff Shaul)

Why NLP is hard?

Why that is hard?

NLP itself is hard. Why?

Brainstorm the reasons NLP is hard.

Textual Queries, can NLP answer?

Example 1

Text Never before had ski racing, a sport dominated by monosyllabic mountain men, seen the likes of Alberto Tomba, the flamboyant Bolognese flatlander who at 21 captured two gold medals at the Calgary Olympics.

Hypothesis Alberto Tomba won a race.

Textual Queries, can NLP answer?

Example 2

Text Researchers at the Harvard School of Public Health say that people who drink coffee may be doing a lot more than keeping themselves awake—this kind of consumption apparently also can help reduce the risk of diseases.

Hypothesis Coffee drinking has health benefits.

Why NLP is hard?

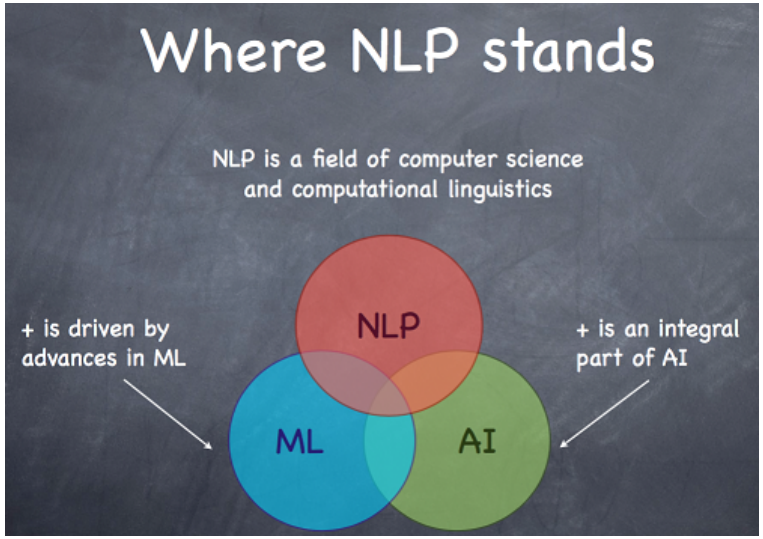
- Ambiguity: There are many different ways to represent the same thing.
- Language is inherently very high dimensional and sparse. There are a lot of rare words.
- Out of sample generalization: New words and new sentences all the time.
- Order and context are extremely important. "Dog bites man" and "Man bites dog" have vastly different meanings even though they differ by a very small amount.
- Anything else?

Where NLP stands?

Where is it useful?

- Information Extraction
- Spoken Dialog
- Question Answering
- Text Summarization
- Etc . . .

Where NLP stands?



Main approaches in NLP

- Rule-based methods: eg Regular expressions.
- Machine learning: eg Linear classifiers.
- Deep Learning: eg Recurrent Neural Networks

Example: rule based approach

Let's say, you want to find slots (City, date, etc) from following sentence:
"Show me flights from Boston to San Francisco on Tuesday"
Context-free grammar is defined by "rules" with substitution syntax such as:

- SHOW: show me — i want — can i see
- FLIGHTS: (a) flight — flights
- ORIGIN: from CITY
- DESTINATION: to CITY
- CITY: Boston — San Francisco — Denver — Washington

Need to parse the sentence and wherever you find any of the right-hand side words, you put left hand side annotation

(Ref: https://web.stanford.edu/ jurafsky/slp3/29.pdf)

Explanation: What is Context Free Grammar?

Context-free grammars are named as such because any of the production rules in the grammar can be applied regardless of context. It does not depend on any other symbols that may or may not be around a given symbol that is having a rule applied to it.

- Context-free Grammars allow a non-terminal to be replaced by a corresponding production rule whenever it appears in a derivation process. The replacement occurs irrespective of what lies before or after the non-terminal. This happens because the LHS of a production rule allows only a single non-terminal of the form :

  $V \rightarrow (V + T)*$, where V is a non-terminal and T is a terminal

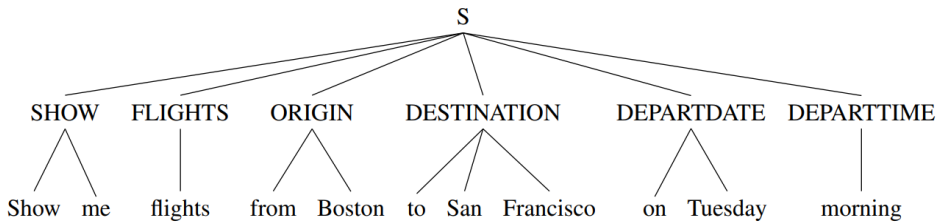
  Hence, the dependence on context is removed owing to this restriction.

- Context-sensitive Grammars being stronger of the two allows replacement of a non-terminal only based on it's left or right context,i.e, depending on the terminals or non-terminals that precedes or succeeds it.

(Ref: https://www.quora.com/What-is-the-meaning-of-context-free-in-context-free-grammar)

Example: rule based approach

"Show me fligths from Boston to San Francisco on Tuesday"
Parsing results in:



(Ref: https://web.stanford.edu/ jurafsky/slp3/29.pdf)

Example: machine learning approach

CRF (Conditional Random Field): probabilistic Named entity recognition.
Training Corpus:

Feature engineering:

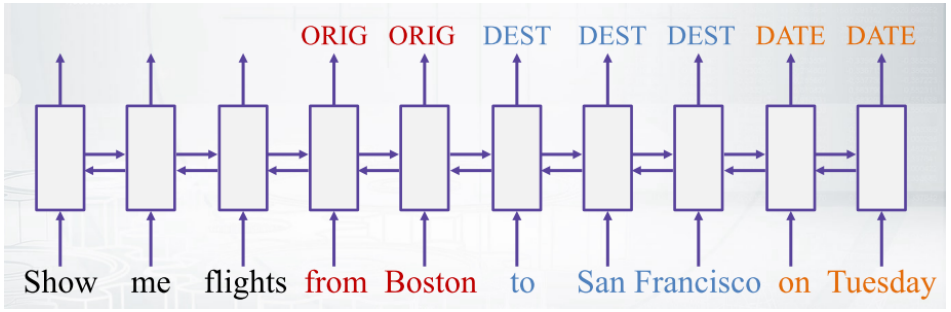
- Is the word capitalized?
- Is the word in a list of city names?
- What is the previous word?
- What is the previous slot?

(Ref: https://www.coursera.org/learn/language-processing/lecture/j8kee/main-approaches-in-nlp)

Example: deep learning approach

LSTM (Long Short Term Memory):

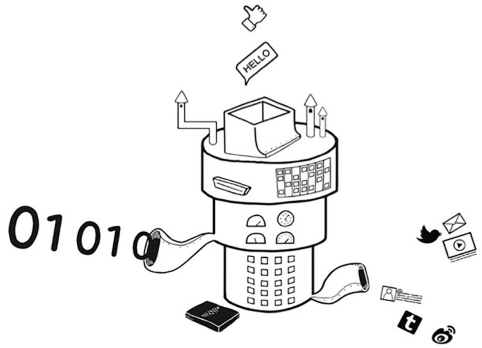
- Big training corpus
- No feature generation
- Defining the model
- Training and inference



(Ref: https://www.coursera.org/learn/language-processing/lecture/j8kee/main-approaches-in-nlp)

Goal of NLP

- To convert letters, words, and ideas into numbers.
- Once we have the numbers we use math and machine learning.



The Promise of NLP

- Importance in scientific, economic, social and cultural arenas
- Growing rapidly as its theories and methods are deployed in new technologies
- Therefore a wide range of people should have a working knowledge of NLP
  - Academia: humanities computing, corpus linguistics, computer science, artificial intelligence
  - Industry: HCI, business information analysis, web software development
- The goal is to open the field of NLP to a broad audience.

NLP and Intelligence

- Long-standing challenge to build intelligent machines
- Chief measure of machine intelligence has been linguistic: Turing test

NLP and Intelligence

- Research on spoken dialog systems, also MT — *integrated NLP systems which future users would regard as highly intelligent*
- Example human-machine dialog:

S: How may I help you?
U: When is Saving Private Ryan playing?
S: For what theater?
U: The Paramount theater.
S: Saving Private Ryan is not playing at the Paramount theater, but it's playing at the Madison theater at 3:00, 5:30, and 10:30.

NLP and Intelligence (cont)

- Today's systems limited to narrowly defined domains
- Couldn't ask above system for other information, e.g.:
  - Driving instructions
  - Details of nearby restaurants
- To add such support we would have to:
  - Store the required information
  - Incorporate suitable questions and answers into the system
- Common-sense reasoning vs business logic
- Need to make progress on natural linguistic interaction without recourse to this unrestricted knowledge and reasoning capability

Recent NLP Applications

- Sentiment Analysis - Deriving sentiments in sentences (positive, negative, neutral), and also in articles (though that will be more appropriate like bag of sentence sentiments). The future is to include emotions (attributes) in that, like the attributes now on Facebook posts - Love, Like, Angry, Surprised, Sad, Hilarious. These attributes make a lot more sense for sentiments going forward.
- Text Summarization - Summarizing a single or many articles according to a particular theme.
- Information Extraction - Find structured information from unstructured data, like entities, relationships, co-reference resolution. This at a basic level is very useful for algorithmic trading. An extension of this is a global form of extracting logic structures (first order and higher order).

Recent NLP Applications

- Topic Segmentation - Topic Extraction (with regions). Normally, there will be overlapping regions.
- Question Answering - Answer the questions to both closed (specific) and open questions (subjective). Answers to subjective questions is the main challenge for the likes of realistic Virtual Assistants.
- Parsing - Parsing natural language generally in the form a tree. This involves hierarchical segmentation of the language involving the grammar rules.
- Prediction - Given a short text, predict what happens next. The prediction problem is beginning to be targeted in vision, but it has never ever gained paths for realistic products. For closed and deterministic prediction (not innovative else that would fall under the paradigm of creative writing), this can be a useful task for prediction of future events based on past evidences and analysis. This can be then very useful for finance sectors.

Recent NLP Applications

- Part of Speech Tagging (POS) - Tagging words whether they are nouns, verbs or adjectives.
- Translation - Translate one language to another. This can be very challenging given the nature of the language, and the grammar. Normally, under probabilistic models, this assumes that the underlying grammar is mostly the same, and thus, models normally fail for Sanskrit.
- Interestingness - Most interesting portion of text in an article. This can be done very much on the same lines as in images, where one ranks the likeness of images.

In nutshell

- NLP is an effort to do useful things with the natural language.
- NLP is hard because humans like words and computers like numbers.

In nutshell

Approaches

References

References

Many publicly available resources have been refereed for making this presentation. Some of the notable ones are:

- Nltk - Steven Bird, Ewan Klein, Edward Loper
- Machine Learning for Natural Language Processing - Traian Rebedea, Stefan Ruseti - LeMAS 2016 - Summer School
- Natural Language Processing with Python - District Data Labs
- Natural Language Processing+ Python - Ann C. Tan-Pohlmann
- Applied Natural Language Processing - Barbara Rosario
- Named Entity Recognition - Stephan Lesch
- Galvanize NLP
- Text Mining - Behrang QasemiZadeh
- Natural Language Processing - Information Extractoin, Christopher Manning
- Topic Modeling - (William Bert, Megan R Brett, Ted Underwood, Algobeans, Shivam Bansal)
- Word2Vec - (Girish K.,Sivan Biham & Adam Yaari, Adrian Colyer)