

# COMP150


## Natural Language Processing: Introduction

---

Bonan Min

[bonanmin@gmail.com](mailto:bonanmin@gmail.com)

Some slides are based on class materials from Ralph Grishman



# This Course

---

Instructor: Bonan Min ([bonanmin@gmail.com](mailto:bonanmin@gmail.com))

TA: Andy Valenti ([Andrew.Valenti@tufts.edu](mailto:Andrew.Valenti@tufts.edu))

Time: Tuesdays & Thursdays 6:00 – 7:15PM

Location: RM 136, Sci-Tech Center (4 Colby Street, Medford)

Course materials: <https://bnmin.github.io/comp150nlp/>

# Myself

---

Ph.D. in Computer Science with dissertation in NLP

Senior Scientist at BBN, conducting NLP research in a wide range of topics:

- Information extraction
- Cross-lingual NLP
- Information retrieval
- NLP for social science

Homepage: <https://bnmin.github.io/>

# Tentative Syllabus

---

Dates	Topics
1/16	Introduction
1/21, 1/23	Text classification, Bag of Words Models, and Machine Learning (ML) for classification
1/28, 1/30	Part of Speech Tagging and Sequential Labeling (HMM, MEMM, CRF and RNN)
2/4, 2/6	Lexical semantics: WordNet, distributional analysis, and word embeddings
2/11, 2/13	Hands-on ML tutorials (MLP, CNN, RNN for text classification) by Andrew Valenti, Language Modeling
2/18, 2/20	Language Modeling con'd. Syntax, Constituent and Dependency Parsing
2/25, 2/27	Syntax, Constituent and Dependency Parsing con'd. Information Extraction Overview
3/3, 3/5	Named Entity Recognition and Entity Linking
3/10, 3/12	Coreference Resolution, WSD, Contextualized Word Embeddings
3/17, 3/19	Spring break; no class
3/24, 3/26	Relation Extraction
3/31, 4/2	Relation Extraction con'd
4/7, 4/9	Event Extraction
4/14, 4/16	Automatic Knowledge Base Population, Beyond Pipeline Models
4/21, 4/23	Machine Translation
4/28, 4/30	Reading period; no class
5/5 or 5/7	Final project presentation

NLP tasks/topics

Computational/ML tools

# Grading (1)

---

## Homework (50%):

- 4 assignment (10-15% each):
  - written assignments involving topics covered in the classes
  - Programming assignments on NLP problems covered in the classes.
- Must be finished on your own
  
- Will be announced through the course website
  - Roughly 1 assignment in every 2-3 weeks
  - Due by 11:59 pm of the stated due date
- Late Policy
  - Assignments will be accepted for up to three days past the due date with a penalty of 20% for each (calendar) day.

# Grading (2)

---

## Team project (50%):

- Possibilities:
  - Developing a solution to an existing NLP problem
  - Defining a new problem & developing a (proof-of-concept) solution
- Recommended team size: 3-4 students
- Proposal (15%):
  - Due date: TBD; ~4 weeks before the final project is due
  - Demonstrating understanding of the proposed NLP problem
  - Describing your plans for solving the problem
- Final project (30%):
  - Due date: TBD; final exam time of the semester; before the project presentation
  - You will need to submit code, data (if a new dataset is created), and a written report
- Project presentation (5%):
  - A short (TBD; ~5 minutes) presentation will be scheduled for each team during final week of the term

# Text Book & Reading Materials (1)

---

Text book: Speech and Language Processing, by Dan Jurafsky and James H. Martin

- Primary: 3<sup>rd</sup> ed. Draft <https://web.stanford.edu/~jurafsky/slp3/>
- Auxiliary: 2<sup>nd</sup> ed.

Other excellent books:

- Foundations of Statistical Natural Language Processing, by Christopher D. Manning and Hinrich Schütze
- Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, by Steven Bird, Ewan Klein and Edward Loper
- An excellent book for deep learning is [Deep Learning](#), by Ian Goodfellow, Yoshua Bengio and Aaron Courville.

# Text Book & Reading Materials (2)

---

Conferences (most are available on-line through the ACL Anthology <https://www.aclweb.org/anthology/>)

- Meetings of the Association for Computational Linguistics (ACL), including ACL Conferences, European ACL Conferences (EACL), and North American ACL Conferences (NAACL)
- International Conferences on Computational Linguistics (COLING)
- Language Resource and Evaluation Conferences (LREC)

## Journals

- Transactions of the Association for Computational Linguistics (TACL)
- Computational Linguistics (CL)



# What is Natural Language Processing?

---

Natural Language Processing (NLP) is an interdisciplinary field that studies how to process, analyze, or generate natural language text

- Subject: Natural Language (NL)
- Research tools: Computer Science (CS), e.g.,
- Modeling tools: Mathematics and Artificial Intelligence (AI)

Also known as

- Computational Linguistics (CL)
- Natural Language Understanding (NLU)
- Human Language Technology (HLT)

# Why Study NLP?

---

## Centrality of Natural Language

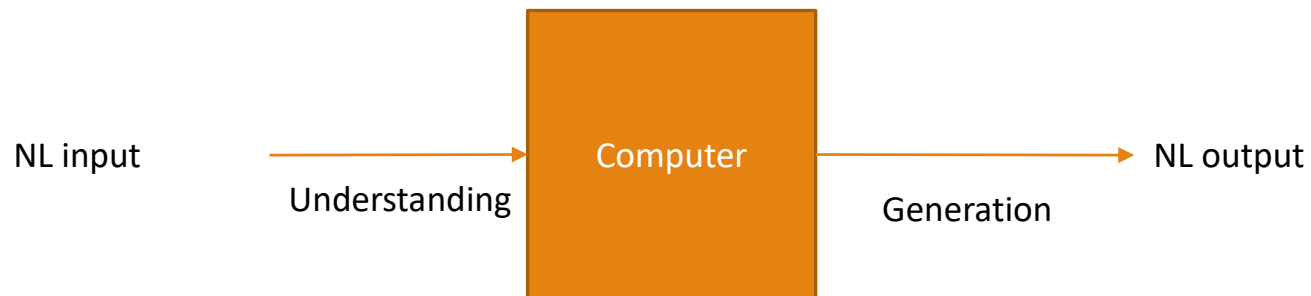
- a primary (and natural) mode of human communication
- representation for most recorded human knowledge
- a very rich and flexible representation (when compared to most formal representations)

# Why Study NLP?

---

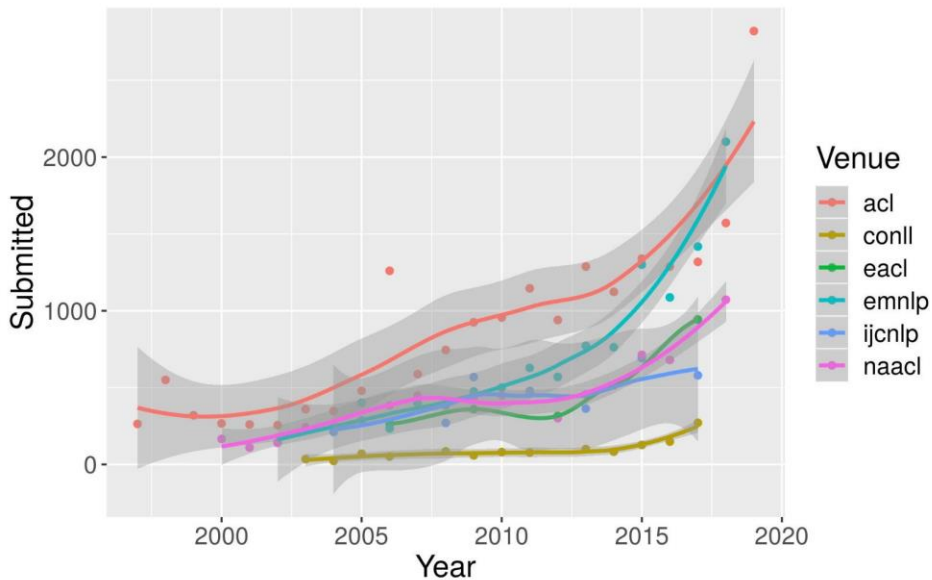
## Language barriers

- Human-human: Machine Translation
- Human-computer: NL human machine interfaces, e.g., question answering and chatbots



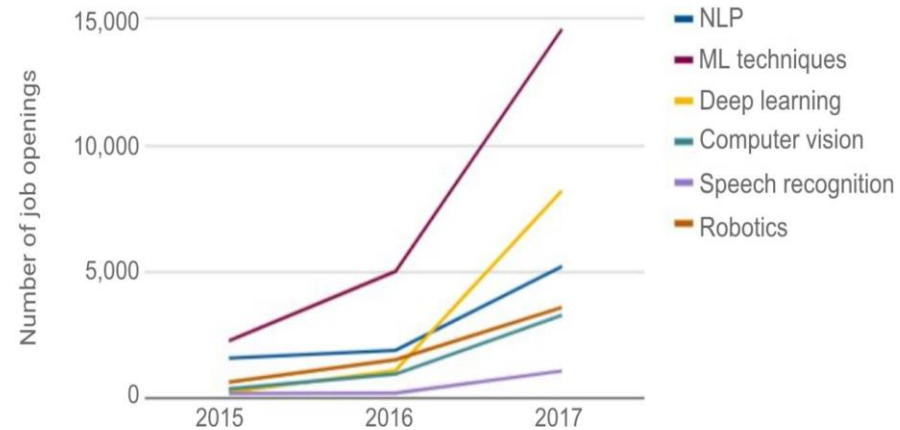
# Why Study NLP?

Rapid growth



Research: ACL Submissions

Job openings by AI skills required (2015 – 2017)  
Source: Monster.com



Industry: Job openings

<http://www.marekrei.com/blog/geographic-diversity-of-nlp-conferences/>

# Why Study NLP?

---

## Many applications

- Machine Translation
- Question Answering
- Interactive Systems
- Information Extraction
- Text Mining
- Spelling and grammar checking
- Writing assessment
- Document (e.g., web page) retrieval
- Automatic summarization
- ...

# Application: Machine Translation

Long history of development (since early 1950's)

How good is MT today?

- Still not good enough for polished translations
- But effective for producing rough drafts for post-editing, or for getting the gist of a text

Quality depends on similarity of language structures

- Chinese to English is much harder than French to English

Systems

- Ex: Systran, Google translate

The attached memorandum on translation from one language to another, and on the possibility of contributing to this process by the use of modern computing devices of very high speed, capacity, and logical flexibility, has been written with one hope only - that it might possibly serve in some small way as a stimulus to someone else, who would have the techniques, the knowledge, and the imagination to do something about it.

I have worried a good deal about the probable naivete of the ideas here presented; but the subject seems to me so important that I am willing to expose my ignorance, hoping that it will be slightly shielded by my intentions.

<http://www.mt-archive.info/Weaver-1949.pdf>

Warren Weaver  
The Rockefeller Foundation  
49 West 49th Street  
New York 20, New York

The screenshot shows the Google Translate interface. The source language is 'ENGLISH - DETECTED' and the target language is 'CHINESE (SIMPLIFIED)'. The input text is 'One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know.' The output text is '一天早上, 我穿着睡衣射杀了一头大象。他永远不会进入我的睡衣。'. A red box highlights the output text, and a red caption below it reads: 'I shot an elephant wearing my pajamas. He will never get into my pajamas.'

I shot an elephant wearing my pajamas.  
He will never get into my pajamas.

# Application: Question Answering

Originally passage retrieval systems, gradually enriched with NLP

- MIT Start system (<http://start.csail.mit.edu/>), Answers.com (<http://www.answers.com/bb/>), Ask.com

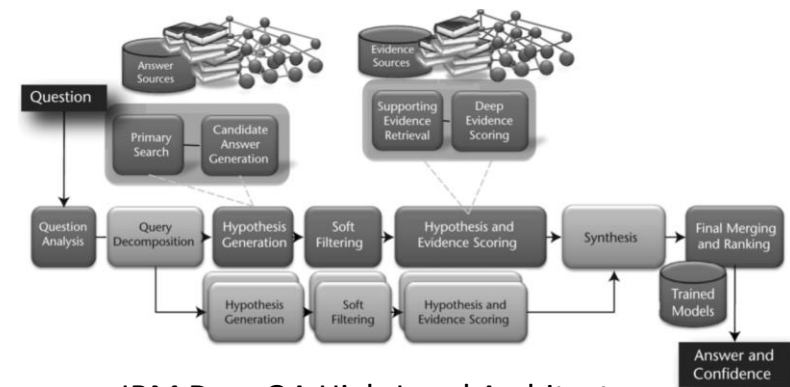
Major web search tools originally operated as page retrieval using word-based strategies but now seek to answer questions directly

Hybrid systems, e.g., IBM Watson

- Requires many NLP techniques, e.g.,
  - Parsing (questions & paragraphs)
  - Information Extraction (entities, relations, events)
  - Text retrieval/matching
- Other components
  - Knowledge representation
  - Answering scoring/re-ranking
  - Distributed processing



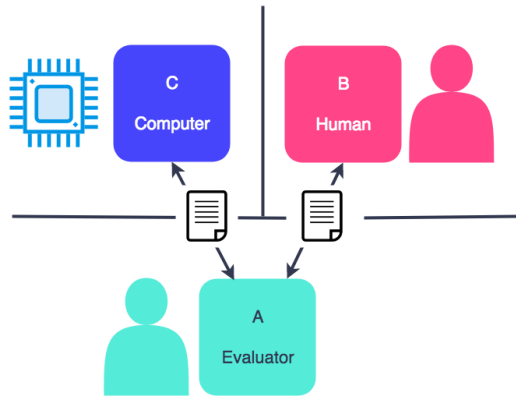
IBM Watson in Jeopardy!



IBM DeepQA High-Level Architecture

# Application: Interactive Systems

---



## The Turing test

- Test a machine's ability to exhibit intelligent behavior indistinguishable from, that of a human
- A human evaluator would judge NL conversations between a human and a machine
- Text-only channel

One of the first interactive applications was NL data base query, but that had limited appeal with written input: people don't like to type a lot; GUIs have been more effective

Chatbots (like the old [Eliza](#) program) provide the impression of intelligent conversation

Conversational agents support simple conversations (using text or speech) for order taking, information

Smartphones with speech recognition (e.g., SIRI) have greatly increased opportunities for speech input



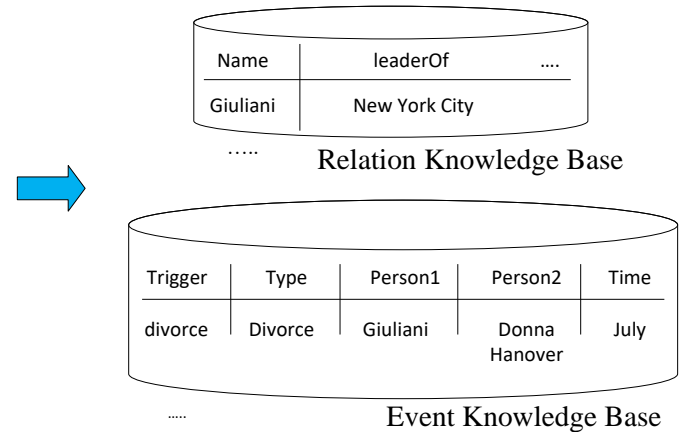
# Application: Information Extraction

Information extraction: automatic conversion of unstructured (or semi-structured) data to a structured form

*Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris\_ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.*

*In interviews last year, Giuliani said Nathan gave him "tremendous emotional support" through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.*

- Resumes: monster.com
- Name taggers, e.g., Gate: <https://gate.ac.uk/>
- Wikification: Name tagging + linking to a standard data base, e.g., DBpedia Spotlight
- Enable better search and general news monitoring
  - Google news
  - European Media Monitor
- For specialized report monitoring
  - Infectious disease reports, e.g., NYU Proteus-Bio
  - Electronic health records
  - Scientific literature
- Assist decision making: Tracking real-world events and their causal/temporal relationships
  - BBN Hume: <https://github.com/BBN-E/Hume>



# Application: Text Mining

---

Text mining (text analytics) discovers patterns from large text collections

- First generation: using "bag of words" representation of documents
- Second generation: information extraction + data mining = text mining

Take advantage of social media (e.g., [Dataminr](#))

Product monitoring

- Based on sentiment analysis for collecting detailed feedback from customers

Situational awareness

- Rapid detection of emergency situations

Finance

- Rapid response to financial news ([Reuters NewsScope](#))

Scientific research

- Mine large collections of research papers for trends and correlations (e.g., treatments and adverse reactions)

Legal

- Document review for compliance

# Applications: Many More

---

## Spelling/Grammar checking

- [grammarly.com](http://grammarly.com)

## Writing assessment

- Educational Testing Service [e-rater](#)

## Dictation (requires accurate prediction of expected words)

- [Dragon Naturally Speaking](#); Apple Dictation

## Document / web page / passage retrieval

- Word-based for English, with simple stemming
- May require morphological analysis for other languages

## Automatic summarization

- <http://newsblaster.cs.columbia.edu/>

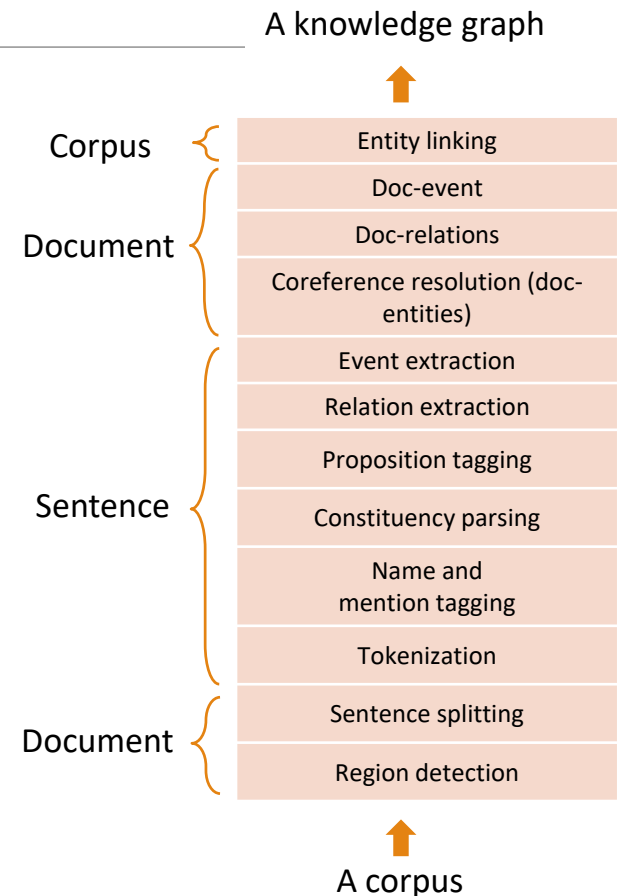
# Our Goal

Create systems which can perform such applications (i.e., understand and generate NL)

- Develop algorithms and formal models which reflect aspects of the structure of language
- Also an engineering problem: the term "language engineering" is sometimes used to reflect this orientation

NLP systems are complex, and require good design techniques

- Modular approaches to break the problem up at appropriate points



An automatic knowledge base construction system as participated in NIST's Text Analysis Conference (TAC) 's Knowledge Base Population (KBP) evaluation.

# Challenges

---

Large vocabularies, large corpora, many languages

- Oxford English Dictionary (2019): more than 600K words

There are always exceptions to any rules: no complete, finite rule sets for defining & processing an NL

Common sense, common (evolving) world knowledge is pervasive in NL

- Need a massive & complete knowledge base which doesn't exist (as least expensive to curate & maintain)

Ambiguity, ambiguity, ambiguity

# Challenge: Ambiguity

---

## Ambiguity

- e.g., two senses of the same word, e.g., *attack*
- Intrinsic to NL itself (how about computer language?)
- Exist at vary level of units, e.g., lexical, phrasal, syntactic

## Examples

- *I saw the man with the telescope*
  - Who has the telescope?
- *At last, a computer understands you like your mother*
  - The computer understands you as well as your mother understands you.
  - The computer understands that you like your mother.
  - The computer understands you as well as it understands your mother.

# Challenge: Ambiguity

NLP can't achieve satisfactory performance without resolving ambiguity

- Some ambiguity can be solved by looking at wider context, or considering common sense.
- Others may be intrinsically ambiguous
  - Interactive processing: more context helps with resolving ambiguity
  - Restricted languages: sublanguages, domains
  - Constrained vocabulary & grammars: reduce ambiguity

ENGLISH - DETECTED    ENGLISH    v    ↔    CHINESE (SIMPLIFIED)    ENGLISH

One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know.    X

一天早上, 我穿着睡衣射杀了一头大象。他永远不会进入我的睡衣。

I shot an elephant wearing my pajamas. He will never get into my pajamas.

*in my pajamas* is ambiguous  
an elephant can't be fit into your pajamas, so the reader will assume **you are wearing the pajamas**.  
However, the following line confirms that **the elephant was in fact in the pajamas**

Many jokes are constructed this way!

# Analyzing Our Needs: Setting the Agenda

---

What functionality do we require to address NLP applications?

Take Machine Translation (MT) as an example

- At first, people imagined that machine translation is mostly a "data processing" task
  - A system looks up the words one at a time in a bilingual dictionary, and then maybe has to fix up the translation a bit.
- However, there is a lot more to do for machine translation



# What Makes Good MT? (1)

Google Translate

ENGLISH - DETECTED

ENGLISH



CHINESE (SIMPLIFIED)

ENGLISH

One morning I shot an elephant in  
my pajamas. How he got into my  
pajamas I'll never know.



一天早上, 我穿着睡衣射杀了一  
头大象。他永远不会进入我的睡  
衣。

I shot an elephant wearing my pajamas.

He will never get into my pajamas.

## Word segmentation

- For some languages (such as Japanese and Chinese) there are no spaces between words, so it's not clear what the words are

## Morphology

- Words appear in different forms, indicating singular vs. plural (for nouns), present tense vs. past tense (for verbs), nominative vs. accusative case, etc.
- English has only a few morphological forms, so it's possible to put them all in a dictionary.
- This isn't true of most Western languages; for example, a Spanish verb could have over 50 forms.

## Syntax

- Word-for-word translation only works if the word order in the two languages is about the same; if it's not, we need to understand enough about the structure of the two languages (their syntax) to change from one word order to another.
- English has a rather fixed subject-verb-object order ("SVO"), while many more inflected languages have more variable word order and some languages have basically different word order (e.g., SOV for Japanese).

# What Makes Good MT? (2)

## Lexical semantics

- Many words are polysemous. They have multiple meanings.
- A word will have to be translated differently depending on its meaning in a particular context; otherwise the translation is likely to make little sense.
  - (1) I **attack** the math problem. vs. (2) A cat **attacked** Bonan.
  - (1) I went to the **bank** to withdraw cash. vs. (2) Cars slide down river **bank**.

## Discourse

- In order to create a proper translation we sometimes have to look beyond the individual sentence.
- That can be true in selecting word senses
- The need also arises in translating into English from languages where subject pronouns can be omitted
  - We need to figure out what the subject actually is, so that we can supply a "he" or "she" or "it" in English.

## World knowledge

- *I shot an elephant in my pajamas*
- *I shot an elephant in my house*
- *I shot an elephant in my dream*

ENGLISH - DETECTED    ENGLISH    ▾    ↔    CHINESE (SIMPLIFIED)    ENGLISH

One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know. ×

一天早上, 我穿着睡衣射杀了一头大象。他永远不会进入我的睡衣。

I shot an elephant wearing my pajamas.  
He will never get into my pajamas.

# Similarly for Information Extraction

---

IBM hired Fred Smith as president.



Person	Company	Position
Fred Smith	IBM	president

## Name recognition

- A company name may be several words ("General Motors"), a person may have a title or middle name ("Mr. Smith", "Fred X. Smith")

## Syntax

- The information may appear in the passive ("Fred Smith was hired by IBM") or in a relative clause ("Fred Smith, who was hired by IBM"); also, there may be extra modifiers ("IBM yesterday hired Fred Smith as president")

## Lexical semantics

- There may be lots of synonyms for hired ("appointed", "named", ...) which the system should recognize

## Discourse – pronouns

- If a pronoun appears in a relevant sentence, the system has to figure out what the pronoun refers to ("Fred Smith left Compaq last week. IBM hired him yesterday as president.")

# Analyzing Our Needs

---

What functionality do we require to address NLP applications?

NL are rich and ambiguous. To build a good system, we need to analyze NL at several levels:

- Syntax: what is the structure of a sentence?
- Semantics: what is the meaning of a word (lexical) and a sentence (compositional) in isolation?
- Discourse: how can a sentence be interpreted in context?
- Dialog: how is language used to exchange information?

# Sneak Peek of NLP Methods

---

Rule-based methods

Statistical, data-driven methods

Hybrid methods

# NLP methods – Rule-based Methods

---

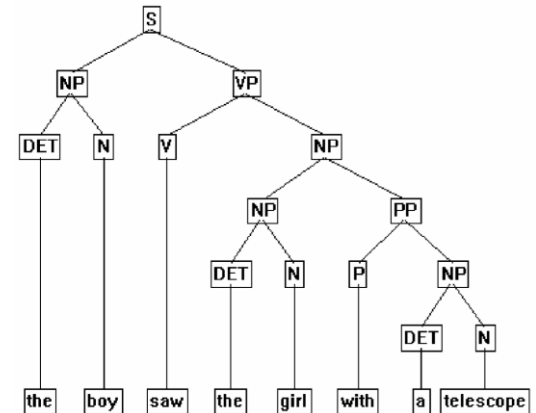
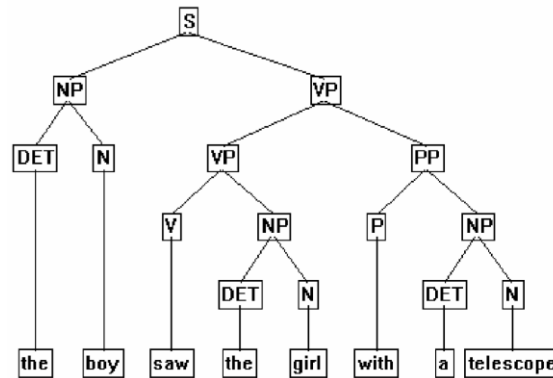
Rule-based methods (symbolic): Researchers (i.e., Computational linguists)

- Summarize the commonalities in language, and write rules
- Develop rule-based algorithms to process NL
- Adjust rules based on analyzing results to improve performance

# NLP methods – Rule-based Methods

An example grammar

$S \rightarrow NP VP$   
 $NP \rightarrow DET N$   
 $NP \rightarrow NP PP$   
 $VP \rightarrow VP PP$   
 $VP \rightarrow V NP$   
 $PP \rightarrow P NP$



*The boy saw the girl with a telescope*

## Problems

- All grammars leak, Sapir 1921
- It's impossible to write a comprehensive list of rules

# NLP methods – Statistical Methods

---

## Statistical/Data-Driven Methods

- Built (annotated) corpora
- Develop a statistical model that is useful for the tasks of interest
- Train the model with the corpora
- Apply the model to NL
- Adjust the model based on analyzing results to improve performance

## An example

- N-gram Language models:  $P(\textit{barks} | a, \textit{dog}) > P(\textit{moos} | a, \textit{dog})$

## Problems

- Data intensive
- It ignores the intrinsic syntactic/semantic structures of language



# NLP methods – Hybrid Methods

---

## Hybrid Methods

- Examples: statistical PCFG parsers
- Best of both worlds: active research areas

Hybrid methods or some ML/statistical methods are dominating nowadays

# Relation to Other Fields: Linguistics

---

Goal of linguistics is to describe language

- Provide simple models which can predict language behavior
- Understand what is universal about language
- Through these formal models, understand how language can be acquired

Formal models from linguistics have been of value in NLP, but its goals are not the same as NLP:

- A single counterexample can invalidate a model as a linguistic theory, but would not significantly lessen its value for NLP
- NLP must address all phenomena which arise in an application, while linguistics may focus on select phenomena which give insight into solutions

# Relation to Other Fields: (Symbolic) Artificial Intelligence (AI)

---

Classical 'symbolic' AI is concerned primarily with generic problem solving strategies & suitable knowledge representations

There is an inherent link between AI and NLP: some NLP problems require the sort of deep reasoning addressed by these AI methods

But NLP (and AI) has found increasing success through avoiding deep reasoning and turning instead to Machine Learning

# Relation to Other Fields: Statistics and Machine Learning

---

Early NLP systems (before 1990) were purely symbolic and hand crafted

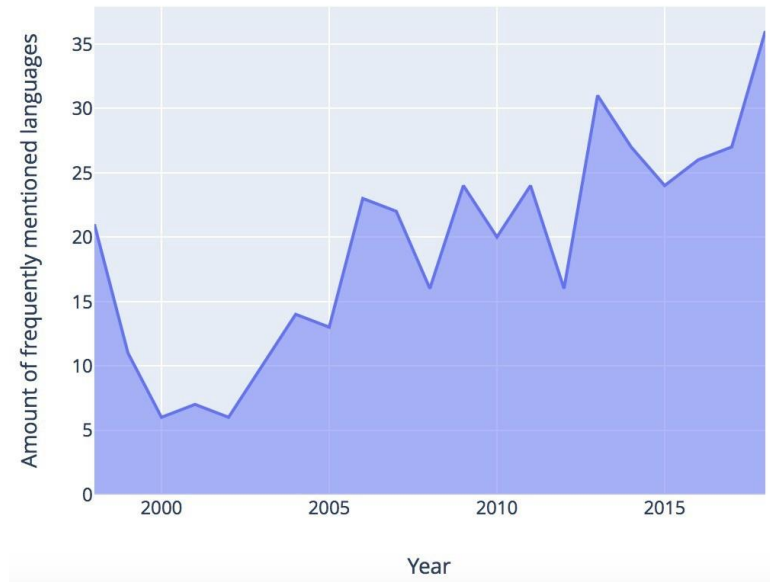
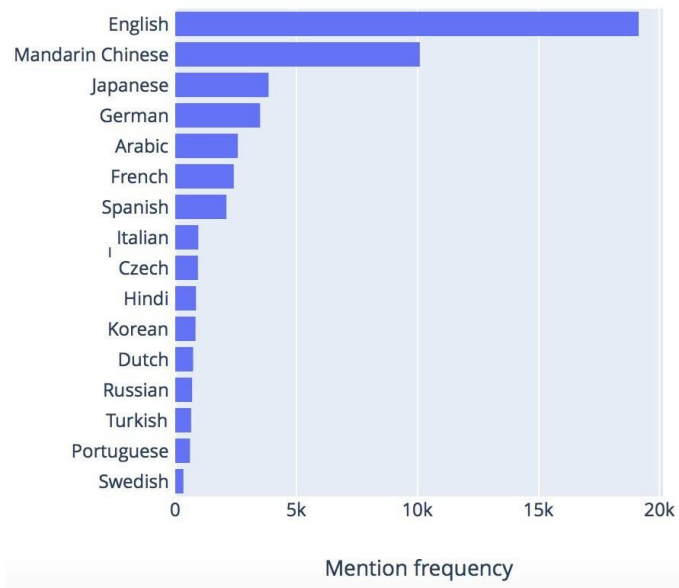
Statistical methods and models have become more widely used in NLP since the mid 1990's

*Easily trainable* and easily computable models have for some NLP tasks proven much more effective than more complex hand-crafted models

- Furthermore, they have become more attractive now that lots of training data is available ("Big Data" initiatives)

The past few years have seen the rapid growth of Neural Network (Deep Learning) models for NLP, achieving better performance than earlier models (e.g., MaxEnt, SVM)

# New Trends and Understudied: Many Languages



Numbers of languages mentioned in ACL conference papers

<https://towardsdatascience.com/major-trends-in-nlp-a-review-of-20-years-of-acl-research-56f5520d473>

# New Trends and Understudied: New Domains, New Tasks

---

Information extraction for an open domain

- New entities: product names
- New events: agricultural activities
- New relations: various causal relations

Problems:

- No large-scale labeled training corpus available
- Even the schema can be defined on-demand
  - Humans (non-experts) are horrible at defining & labeling vague, socio-economic concepts consistently

Directions

- Weakly supervised methods, e.g., bootstrapping
- Distant supervision
- Active learning

# A Quick Survey

---

How many of you are familiar with

- Python
- Java
- ?

Have you taken a Machine Learning course before?

Have you taken an intro to NLP course (or linguistics courses) before?