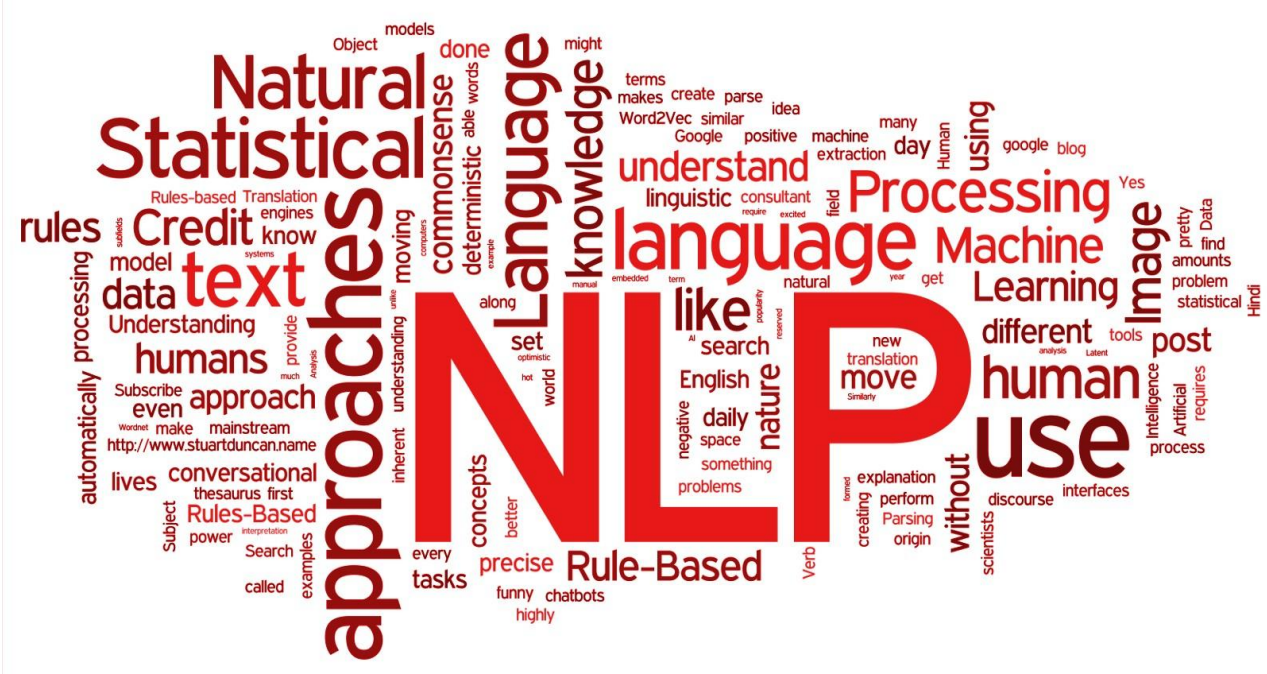
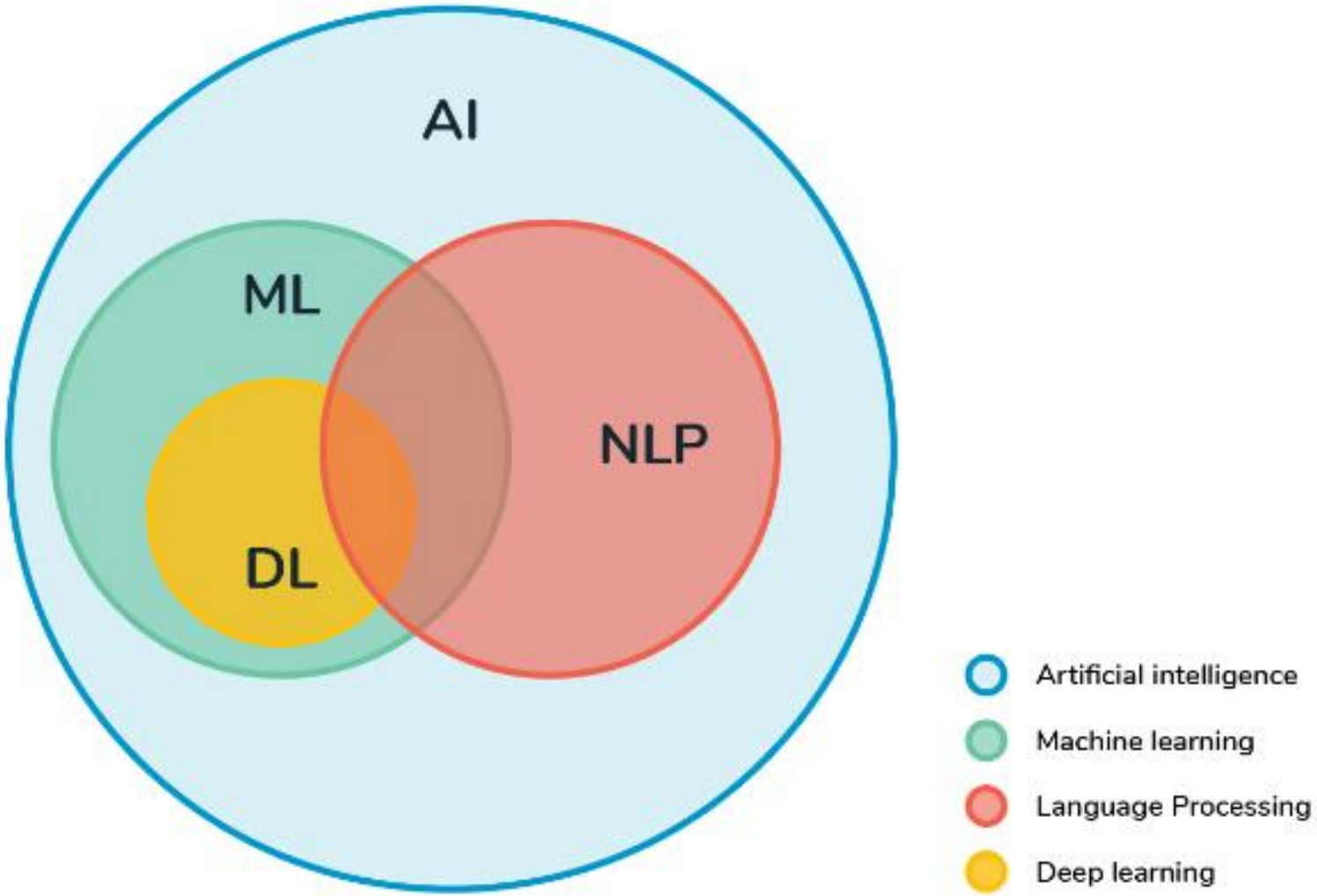
Natural language processing



NLP, AI, ML, DL ……


AI
ML
NLP
DL
Artificial intelligence
Machine learning
Language Processing
Deep learning

About this course

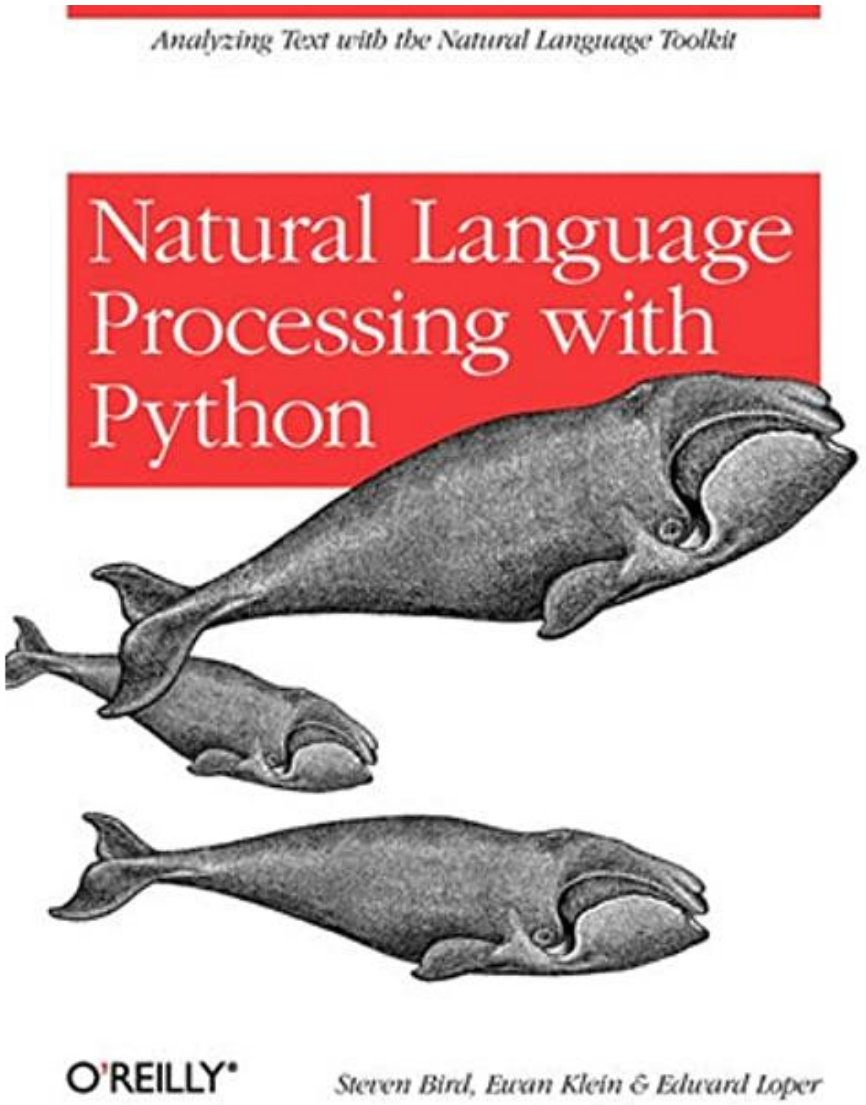
- Class Content
- Reference books
- Class Project
- Grading

Class Content

- Rule-based Automata、Parsing
- Tokenization、Part-of-Speech & NER
- Text Representation and Visualization
- Document Representation and Measure
- Topic Models
- Sequential DL Models（RNN & LSTM）
- Self-attentions, Transformers & BERT
- Chatbots
- TTS (Text-to-speech) & ASR (Automatic Speech Recognition)

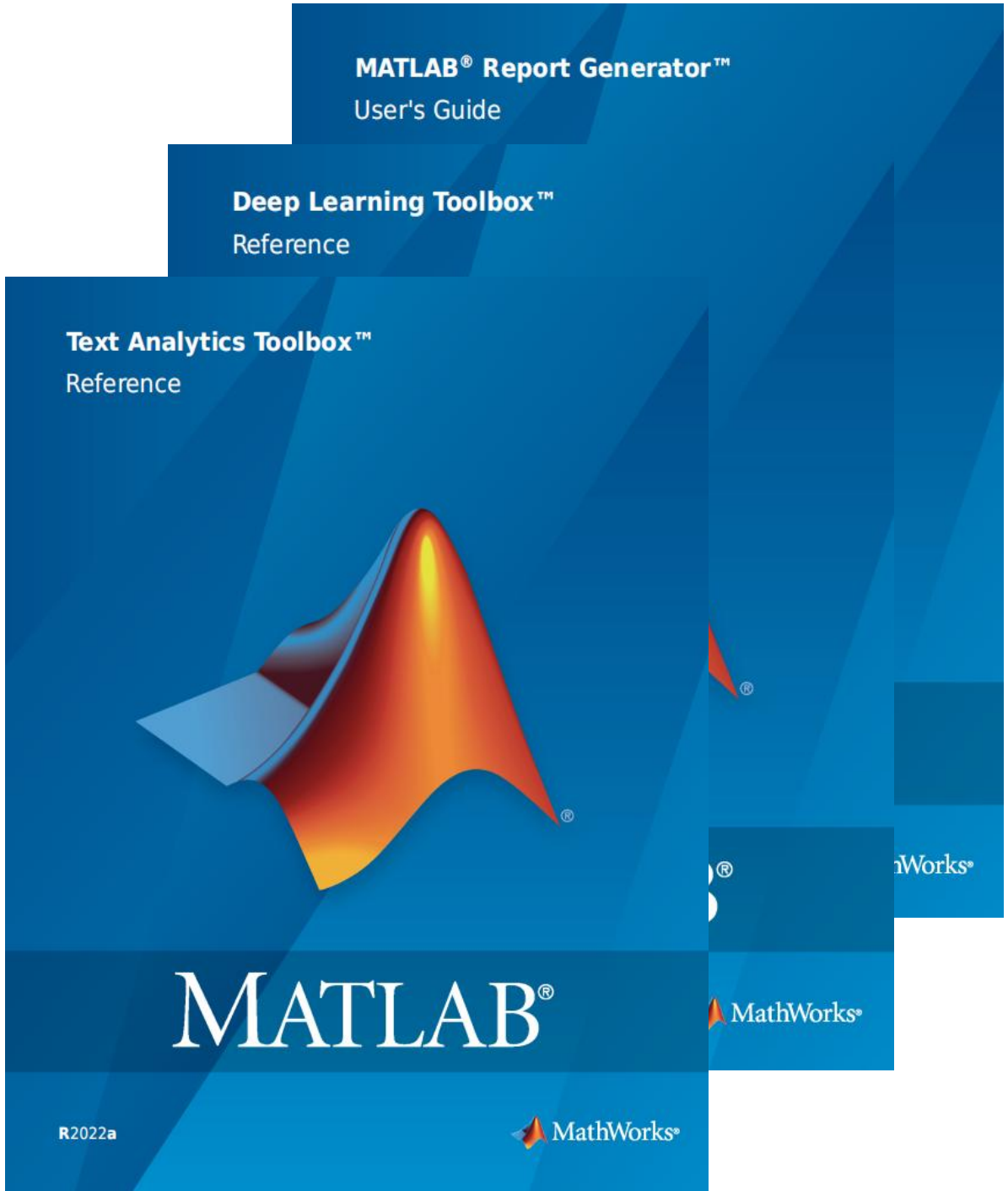
Reference books

- NLTK
- Python



Reference books

- Matlab Docs
- Production level



Grading

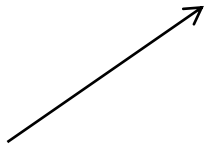
- Three class project reports
  - 70%（20%*3+10%）
- Final Presentation
  - 30%

How to get A+

- Excellent class project reports & presentation
- Preprint a good NLP paper on Arxiv.org
- Enter the final round of NLP contests

How to get A+

- Preprint a **good** NLP paper on Arxiv.org



CCF rank A

https://www.ccf.org.cn/Academic_Evaluation/By_category/

How to get A+

- Enter the final round of NLP contests

AI竞赛网站：https://www.datafountain.cn/

智能人机交互自然语言理解、初赛截止 09/22
https://www.datafountain.cn/competitions/511
预训练模型知识量度量、初赛截止 09/22
https://www.datafountain.cn/competitions/509
基于人工智能的漏洞数据分类、初赛截止 10月10日
https://www.datafountain.cn/competitions/594
Web攻击检测与分类识别、初赛截止 10月10日
https://www.datafountain.cn/competitions/596

Project Scheduling

- Project Proposal
  - The first 3 weeks (Sept. 30)
- Project stage I
  - Oct.
- Project stage II
  - Nov.
- Project stage III & Final Presentation
  - Dec.

Outline

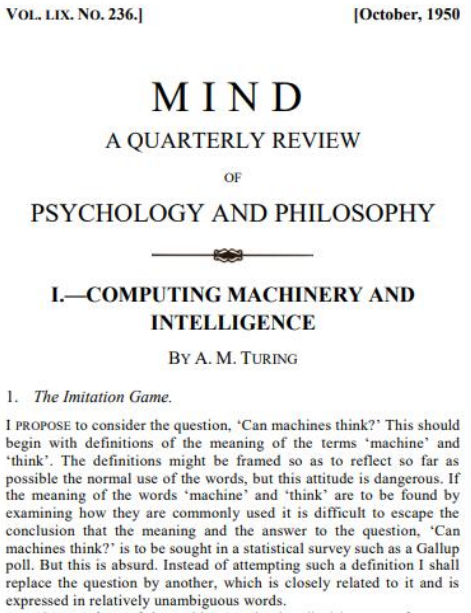
- History
- Methods: Rules, statistics, neural networks
- Common NLP Tasks
  - Text and speech processing
  - Morphological analysis
  - Syntactic analysis
  - Lexical semantics (of individual words in context)
  - Relational semantics (semantics of individual sentences)
  - Discourse (semantics beyond individual sentences)
  - Higher-level NLP applications
- Cognition and NLP

About NLP

- Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.
- Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural-language generation.

History

- Natural language processing has its roots in the 1950s. Already in 1950, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence, a task that involves the automated interpretation and generation of natural language, but at the time not articulated as a problem separate from artificial intelligence.



VOL. LIX. NO. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

History - Symbolic NLP (1950s - early 1990s)

- 1950s: The Georgetown experiment in 1954 involved fully automatic translation of more than sixty Russian sentences into English. The authors claimed that within three or five years, machine translation would be a solved problem. However, real progress was much slower, and after the ALPAC report in 1966, which found that ten-year-long research had failed to fulfill the expectations, funding for machine translation was dramatically reduced. Little further research in machine translation was conducted until the late 1980s when the first statistical machine translation systems were developed.

History - Symbolic NLP (1950s - early 1990s)

- 1960s: Some notably successful natural language processing systems developed in the 1960s were SHRDLU, a natural language system working in restricted "blocks worlds" with restricted vocabularies, and ELIZA, a simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum between 1964 and 1966. Using almost no information about human thought or emotion, ELIZA sometimes provided a startlingly human-like interaction. When the "patient" exceeded the very small knowledge base, ELIZA might provide a generic response, for example, responding to "My head hurts" with "Why do you say your head hurts?".

History - Symbolic NLP (1950s - early 1990s)

Here is a famous demonstration of SHRDLU:

- Person: Pick up a big red block.
- Computer: OK.
- Person: Grasp the pyramid.
- Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
- Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
- Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
- Computer: OK.
- Person: What does the box contain?
- Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.

History - Symbolic NLP (1950s - early 1990s)

```
  Welcome to
                EEEEEE  LL      IIII  ZZZZZZZ   AAAAA
                EE      LL       II       ZZ   AA   AA
                EEEEE   LL       II     ZZZ    AAAAAAA
                EE      LL       II    ZZ      AA   AA
                EEEEEE  LLLLLL  IIII  ZZZZZZZ  AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

History - Symbolic NLP (1950s - early 1990s)

1970s: During the 1970s, many programmers began to write "conceptual ontologies", which structured real-world information into computer-understandable data.

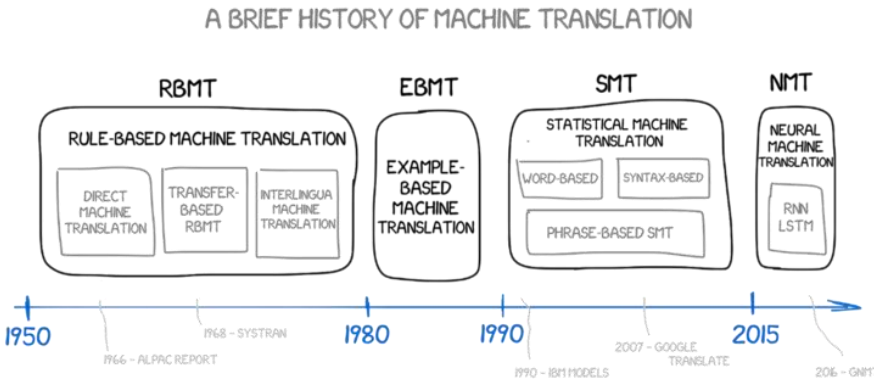
1980s: The 1980s and early 1990s mark the hey-day of symbolic methods in NLP. Focus areas of the time included research on rule-based, morphology, semantics, reference and other areas of natural language understanding. Other lines of research were continued, e.g., the development of chatterbots. An important development (that eventually led to the statistical turn in the 1990s) was the rising importance of quantitative evaluation in this period.

History - Statistical NLP (1990s - 2010s)

Up to the 1980s, most natural language processing systems were based on complex sets of hand-written rules. Starting in the late 1980s, however, there was a revolution in natural language processing with the introduction of machine learning algorithms for language processing. This was due to both the steady increase in computational power (see Moore's law) and the gradual lessening of the dominance of Chomskyan theories of linguistics (e.g. transformational grammar), whose theoretical underpinnings discouraged the sort of corpus linguistics that underlies the machine-learning approach to language processing.

History - Statistical NLP (1990s - 2010s)

1990s: Many of the notable early successes on statistical methods in NLP occurred in the field of machine translation, due especially to work at IBM Research. These systems were able to take advantage of existing multilingual textual corpora. However, most other systems depended on corpora specifically developed for the tasks implemented by these systems, which was a major limitation in the success of these systems. As a result, a great deal of research has gone into methods of more effectively learning from limited amounts of data.

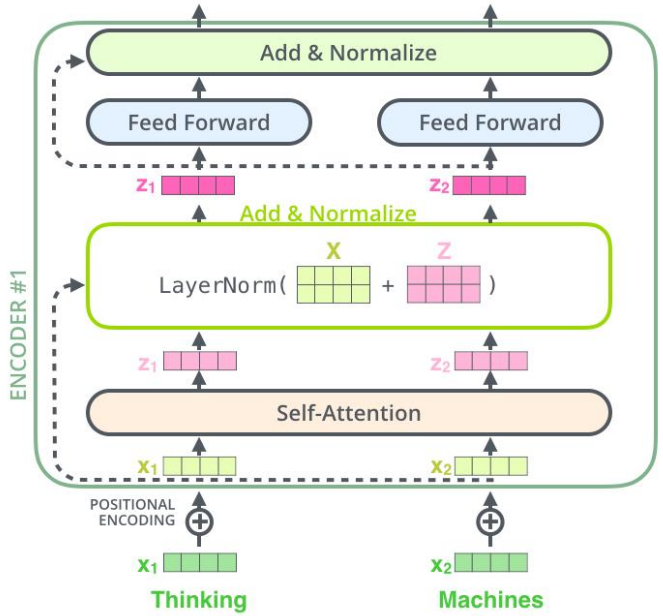


History - Statistical NLP (1990s - 2010s)

2000s: With the growth of the web, increasing amounts of raw (unannotated) language data has become available since the mid-1990s. Research has thus increasingly focused on unsupervised and semi-supervised learning algorithms. Such algorithms can learn from data that has not been hand-annotated with the desired answers or using a combination of annotated and non-annotated data. Generally, this task produces less accurate results for a given amount of input data. However, there is an enormous amount of non-annotated data available, which can often make up for the inferior results if the algorithm is practical.

History - DL Neural NLP (present)

In the 2010s, representation learning and deep neural network-style machine learning methods became widespread in natural language processing, due in part to a flurry of results showing that such techniques can achieve state-of-the-art results in many natural language tasks, for example in language modeling, parsing, and many others.



Methods: Rules, statistics, neural networks

In the early days, many language-processing systems were designed by symbolic methods, i.e., the hand-coding of a set of rules, coupled with a dictionary lookup: such as by writing grammars or devising heuristic rules for stemming.

More recent systems based on machine-learning algorithms have many advantages over hand-produced rules:

- The learning procedures used during machine learning automatically focus on the most common cases, whereas when writing rules by hand it is often not at all obvious where the effort should be directed.

Methods: Rules, statistics, neural networks

- Automatic learning procedures can make use of statistical inference algorithms to produce models that are robust to unfamiliar input and to erroneous input. Generally, handling such input gracefully with handwritten rules, or, more generally, creating systems of handwritten rules that make soft decisions, is extremely difficult, error-prone and time-consuming.
- Systems based on automatically learning the rules can be made more accurate simply by supplying more input data. However, systems based on handwritten rules can only be made more accurate by increasing the complexity of the rules, which is a much more difficult task.

Methods: Rules, statistics, neural networks

Despite the popularity of machine learning in NLP research, symbolic methods are still (2020) commonly used

- when the amount of training data is insufficient to successfully apply machine learning methods, e.g., for the machine translation of low-resource languages such as provided by the Apertium system,
- for preprocessing in NLP pipelines, e.g., tokenization
- for postprocessing and transforming the output of NLP pipelines, e.g., for knowledge extraction from syntactic parses.

Statistical methods

Since the so-called "statistical revolution“ in the late 1980s and mid-1990s, much natural language processing research has relied heavily on machine learning. The machine-learning paradigm calls instead for using statistical inference to automatically learn such rules through the analysis of large corpora (the plural form of corpus, is a set of documents, possibly with human or computer annotations) of typical real-world examples.

Statistical methods

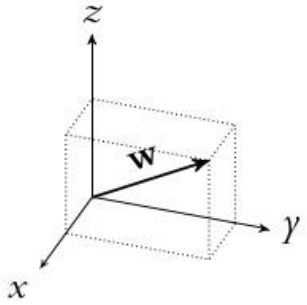
Many different classes of machine-learning algorithms have been applied to natural-language-processing tasks. These algorithms take as input a large set of "features" that are generated from the input data. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. Such models have the advantage that they can express the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system.

Statistical methods

Some of the earliest-used machine learning algorithms, such as decision trees, produced systems of hard if-then rules similar to existing hand-written rules. However, part-of-speech tagging introduced the use of hidden Markov models to natural language processing, and increasingly, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to the features making up the input data. The cache language models upon which many speech recognition systems now rely are examples of such statistical models

Neural networks

A major drawback of statistical methods is that they require elaborate feature engineering. Since the early 2010s, the field has thus largely abandoned statistical methods and shifted to neural networks for machine learning. Popular techniques include the use of word embeddings to capture semantic properties of words, and an increase in end-to-end learning of a higher-level task (e.g., question answering) instead of relying on a pipeline of separate intermediate tasks (e.g., part-of-speech tagging and dependency parsing).



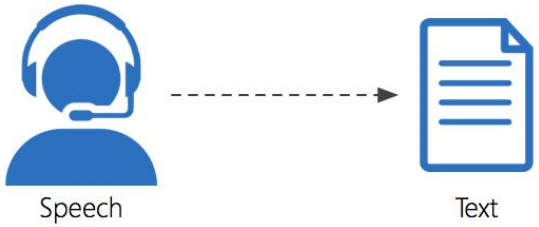
Neural networks

In some areas, this shift has entailed substantial changes in how NLP systems are designed, such that deep neural network-based approaches may be viewed as a new paradigm distinct from statistical natural language processing. For instance, the term neural machine translation (NMT) emphasizes the fact that deep learning-based approaches to machine translation directly learn sequence-to-sequence transformations, obviating the need for intermediate steps such as word alignment and language modeling that was used in statistical machine translation (SMT).

Common NLP Tasks

The following is a list of some of the most commonly researched tasks in natural language processing. Some of these tasks have direct real-world applications, while others more commonly serve as subtasks that are used to aid in solving larger tasks.

Though natural language processing tasks are closely intertwined, they can be subdivided into categories for convenience. A coarse division is given below.



Text and speech processing - Speech recognition

Given a sound clip of a person or people speaking, determine the textual representation of the speech. This is the opposite of text to speech and is one of the extremely difficult problems colloquially termed "AI-complete". In natural speech there are hardly any pauses between successive words, and thus speech segmentation is a necessary subtask of speech recognition.

Also, given that words in the same language are spoken by people with different accents, the speech recognition software must be able to recognize the wide variety of input as being identical to each other in terms of its textual equivalent.

Text and speech processing - Others

Speech segmentation

- Given a sound clip of a person or people speaking, separate it into words. A subtask of speech recognition and typically grouped with it.

Optical character recognition (OCR)

- Given an image representing printed text, determine the corresponding text.

Text-to-speech

- Given a text, transform those units and produce a spoken representation. Text-to-speech can be used to aid the visually impaired.

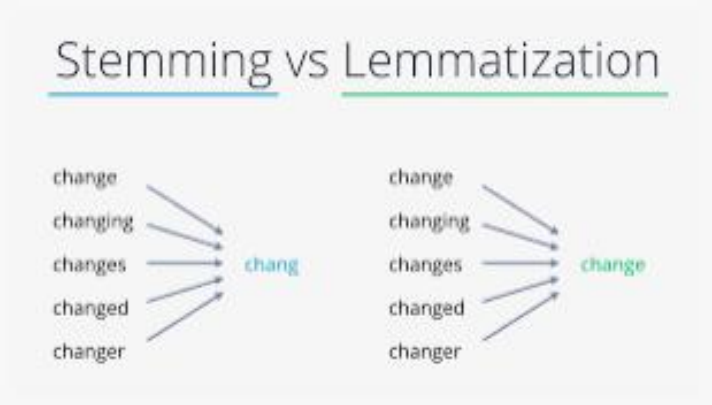
Morphological analysis

Stemming

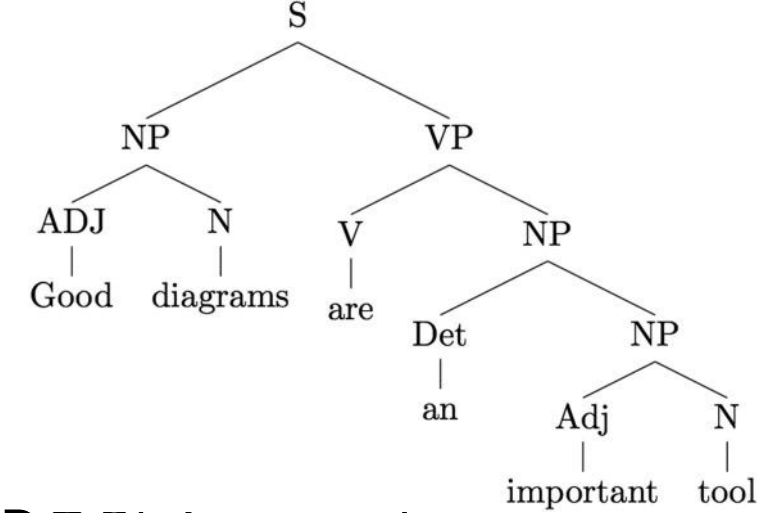
- The process of reducing inflected (or sometimes derived) words to their root form. (e.g., "clos" will be the root for "closed", "closing", "close", "closer" etc.).

Lemmatization

- The task of removing inflectional endings only and to return the base dictionary form of a word which is also known as a lemma.



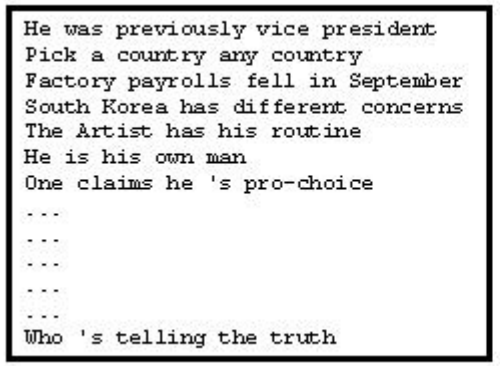
Morphological analysis

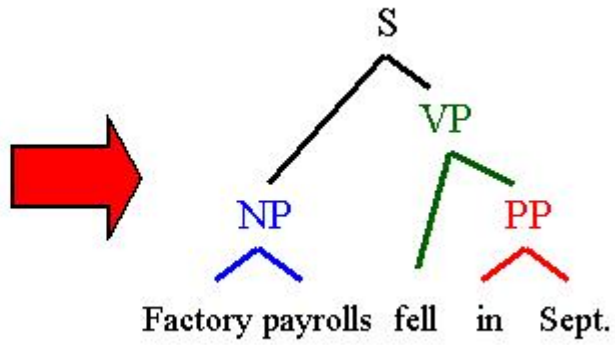


Part-of-speech tagging

- Given a sentence, determine the part of speech (POS) for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a noun ("the book on the table") or verb ("to book a flight"); "set" can be a noun, verb or adjective; and "out" can be any of at least five different parts of speech. Some languages have more such ambiguity than others. Languages with little inflectional morphology, such as English, are particularly prone to such ambiguity.

Syntactic analysis





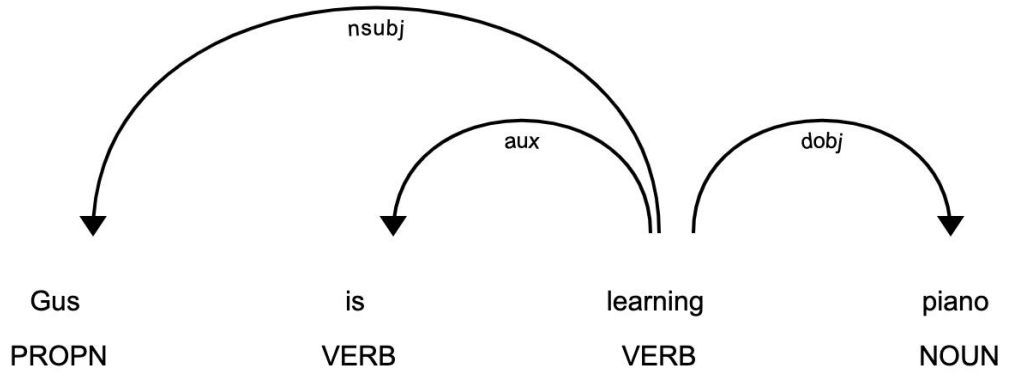
Grammar induction

- Generate a formal grammar that describes a language's syntax.

Sentence breaking

- Given a chunk of text, find the sentence boundaries. Sentence boundaries are often marked by periods or other punctuation marks, but these same characters can serve other purposes (e.g., marking abbreviations).

Syntactic analysis



Parsing

- Determine the parse tree (grammatical analysis) of a given sentence. The grammar for natural languages is ambiguous and typical sentences have multiple possible analyses: perhaps surprisingly, for a typical sentence there may be thousands of potential parses (most of which will seem completely nonsensical to a human). There are two primary types of parsing: dependency parsing and constituency parsing. Dependency parsing focuses on the relationships between words in a sentence (marking things like primary objects and predicates), whereas constituency parsing focuses on building out the parse tree using a probabilistic context-free grammar (PCFG) (see also stochastic grammar).

Lexical semantics (of individual words in context)

Lexical semantics

- What is the computational meaning of individual words in context?

Distributional semantics

- How can we learn semantic representations from data?

Sentiment analysis (see also multimodal sentiment analysis)

- Extract subjective information usually from a set of documents, often using online reviews to determine "polarity" about specific objects. It is especially useful for identifying trends of public opinion in social media, for marketing.

Lexical semantics (of individual words in context)

Terminology extraction

- The goal of terminology extraction is to automatically extract relevant terms from a given corpus.

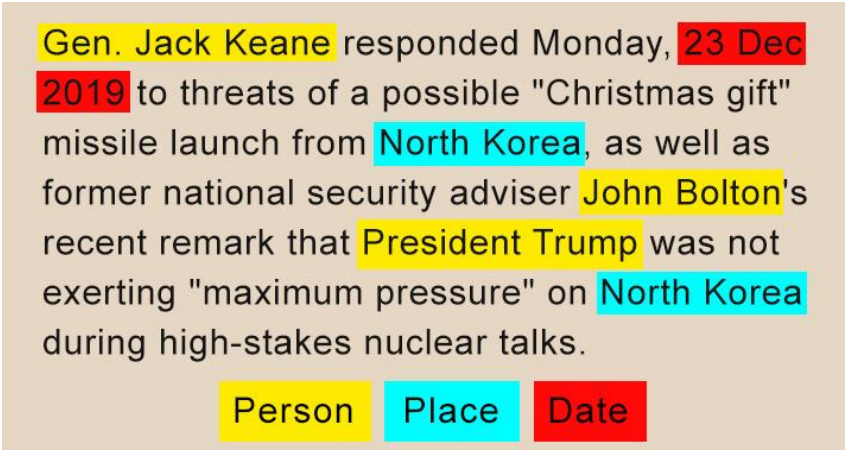
Word sense disambiguation

- Many words have more than one meaning; we have to select the meaning which makes the most sense in context. For this problem, we are typically given a list of words and associated word senses, e.g. from a dictionary or an online resource such as WordNet.

Lexical semantics (of individual words in context)

Named entity recognition (NER)

- Given a stream of text, determine which items in the text map to proper names, such as people or places, and what the type of each such name is (e.g. person, location, organization). Although capitalization can aid in recognizing named entities in languages such as English, this information cannot aid in determining the type of named entity, and in any case, is often inaccurate or insufficient.



Relational semantics (semantics of individual sentences)

Relationship extraction

- Given a chunk of text, identify the relationships among named entities (e.g. who is married to whom).

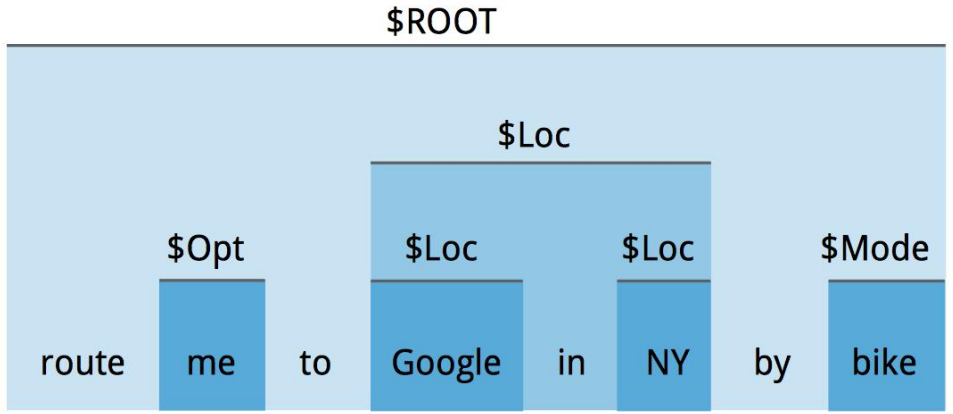
Semantic Role Labelling

- Given a single sentence, identify and disambiguate semantic predicates (e.g., verbal frames), then identify and classify the frame elements (semantic roles).

Relational semantics (semantics of individual sentences)

Semantic Parsing

- Given a piece of text (typically a sentence), produce a formal representation of its semantics, either as a graph (e.g., in AMR parsing) or in accordance with a logical formalism (e.g., in DRT parsing). This challenge typically includes aspects of several more elementary NLP tasks from semantics (e.g., semantic role labelling, word sense disambiguation) and can be extended to include full-fledged discourse.



Discourse (semantics beyond individual sentences)

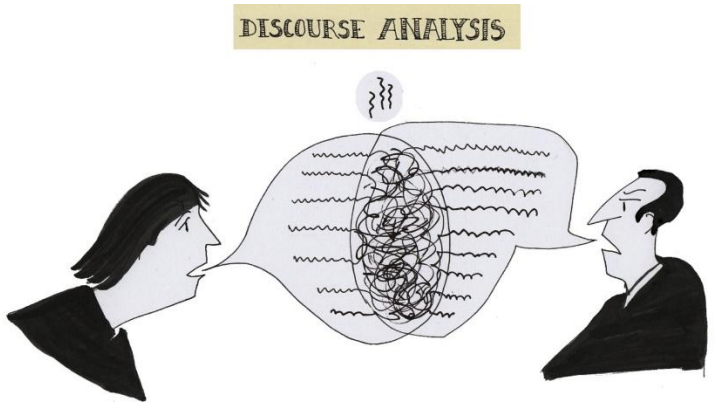
Coreference resolution

- Given a sentence or larger chunk of text, determine which words ("mentions") refer to the same objects ("entities"). Anaphora resolution is a specific example of this task, and is specifically concerned with matching up pronouns with the nouns or names to which they refer. The more general task of coreference resolution also includes identifying so-called "bridging relationships" involving referring expressions. For example, in a sentence such as "He entered John's house through the front door", "the front door" being referred to is the front door of John's house.

Discourse (semantics beyond individual sentences)

Discourse analysis

- This rubric includes several related tasks. One task is discourse parsing, i.e., identifying the discourse structure of a connected text, i.e. the nature of the discourse relationships between sentences (e.g. elaboration, explanation, contrast). Another possible task is recognizing and classifying the speech acts in a chunk of text (e.g. yes-no question, content question, statement, assertion, etc.).



Discourse (semantics beyond individual sentences)

Recognizing Textual entailment

- Given two text fragments, determine if one being true entails the other, entails the other's negation, or allows the other to be either true or false.

Topic segmentation and recognition

- Given a chunk of text, separate it into segments each of which is devoted to a topic, and identify the topic of the segment.

Higher-level NLP applications

Automatic summarization (text summarization)

- Produce a readable summary of a chunk of text. Often used to provide summaries of the text of a known type, such as research papers, articles in the financial section of a newspaper.

Dialogue management

- Computer systems intended to converse with a human.

Machine translation

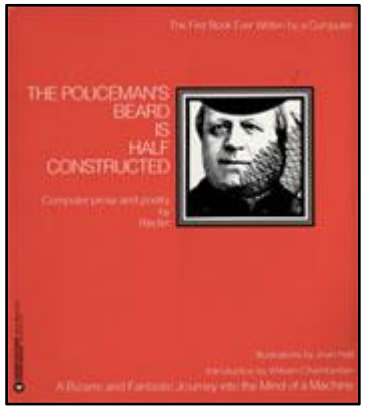
- Automatically translate text from one human language to another. This is one of the most difficult "AI-complete" problems, i.e. requiring all of the different types of knowledge that humans possess (grammar, semantics, facts about the real world, etc.) to solve properly.

Higher-level NLP applications

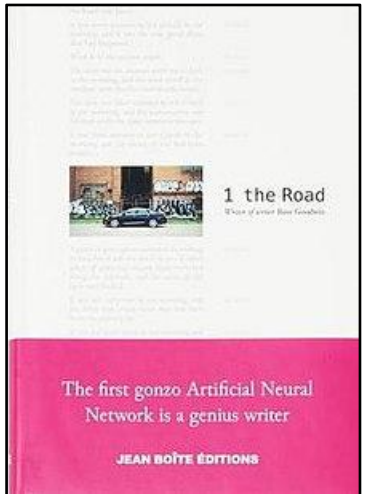
Book generation

- Not an NLP task proper but an extension of Natural Language Generation and other NLP tasks is the creation of full-fledged books. The first machine-generated book was created by a rule-based system in 1984 (**Racter, The policemen's beard is half-constructed**). The first published work by a neural network was published in 2018, **1 the Road**, marketed as a novel, contains sixty million words. Both these systems are basically elaborate but non-sensical (semantics-free) language models. The first machine-generated science book was published in 2019 (Beta Writer, Lithium-Ion Batteries, Springer, Cham). Unlike Racter and 1 the Road, this is grounded on factual knowledge and based on text summarization.

Higher-level NLP applications



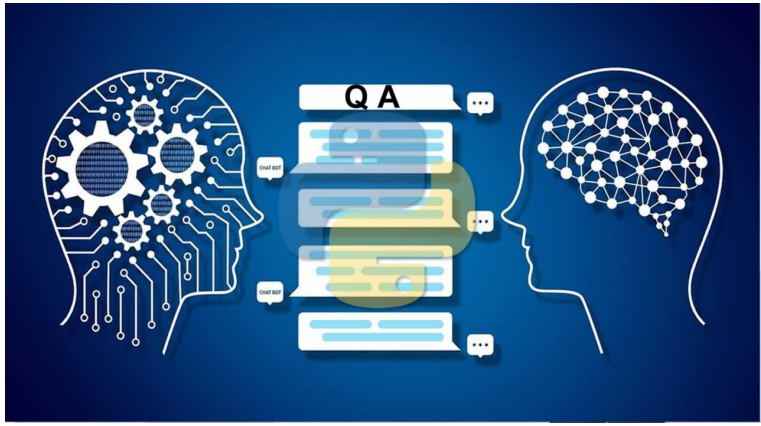
With the exception of this introduction, the writing in this book was all done by a computer. The book has been proofread for spelling but otherwise is completely unedited. The fact that a computer must somehow communicate its activities to us, and that frequently it does so by means of programmed directives in English, does suggest the possibility that we might be able to compose programming that would enable the computer to find its way around a common language “on its own” as it were. ……



1 the Road is an experimental novel composed by artificial intelligence (AI). Emulating Jack Kerouac's On the Road, Ross Goodwin drove from New York to New Orleans in March 2017 with an AI in a laptop hooked up to various sensors, whose output the AI turned into words that were printed on rolls of receipt paper. The novel was published in 2018 by Jean Boîte Éditions.

Goodwin left the text unedited. Although he felt the prose was "choppy", and contained typographical errors, he wanted to present the machine-generated text verbatim, for future study. The story begins: "It was nine seventeen in the morning, and the house was heavy".

Higher-level NLP applications



Question answering

- Given a human-language question, determine its answer. Typical questions have a specific right answer (such as "What is the capital of Canada?"), but sometimes open-ended questions are also considered (such as "What is the meaning of life?"). Recent works have looked at even more complex questions.

Cognition and NLP

Cognition refers to "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses." Cognitive science is the interdisciplinary, scientific study of the mind and its processes. Cognitive linguistics is an interdisciplinary branch of linguistics, combining knowledge and research from both psychology and linguistics. George Lakoff offers a methodology to build Natural language processing (NLP) algorithms through the perspective of Cognitive science, along with the findings of Cognitive linguistics:

The first defining aspect of this cognitive task of NLP is the application of the theory of Conceptual metaphor, explained by Lakoff as “the understanding of one idea, in terms of another” which provides an idea of the intent of the author.

Cognition and NLP

For example, consider some of the meanings, in English, of the word “big”. When used as a Comparative, as in “That is a big tree,” a likely inference of the intent of the author is that the author is using the word “big” to imply a statement about the tree being ”physically large” in comparison to other trees or the authors experience. When used as a Stative verb, as in ”Tomorrow is a big day”, a likely inference of the author’s intent it that ”big” is being used to imply ”importance”. These examples are not presented to be complete, but merely as indicators of the implication of the idea of Conceptual metaphor. The intent behind other usages, like in ”She is a big person” will remain somewhat ambiguous to a person and a cognitive NLP algorithm alike without additional information.

Cognition and NLP

This leads to the second defining aspect of this cognitive task of NLP, namely Probabilistic context-free grammar (PCFG) which enables cognitive NLP algorithms to assign relative measures of meaning to a word, phrase, sentence or piece of text based on the information presented before and after the piece of text being analyzed.

Web resources

Class materials:

https://github.com/info-ruc/nlp22

Class Projects Submission:

https://github.com/info-ruc/nlp22projects

info-ruc / nlp22 Public

Pin Unwatch 1 Fork 0 Star 0

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags Go to file Add file Code

info-ruc Update README.md e96cf46 7 days ago 2 commits

README.md Update README.md 7 days ago

README.md

nlp22

Class material for Natural Language Processing Course 2022

Class Slides

#1 Overview slide

About

No description, website, or topics provided.

Readme
0 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

info-ruc / nlp22projects Public

Pin Unwatch 1 Fork 0 Star 0

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags Go to file Add file Code

info-ruc Update README.md 0c8b489 8 hours ago 2 commits

README.md Update README.md 8 hours ago

README.md

nlp22projects

reports collected for NLP class project @2022

Report Naming Convension

```
student_id (folder)
  |
  └ student_id.md / student_id.docx / ...
  |
  └ src (folder)
  |  |
  |  └ scripts (folder) for codes.
  |  └ dataset (folder) for datasets > 50M, please use pan.baidu.com, and leave url here.
  |
  └ extra (folder) extra files used for the project
```

Pick a task

TBA

About

No description, website, or topics provided.

Readme
0 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

准备环境

SWI Prolog

Python

Matlab