NLP Applications II

Introduction to Practical NLP

NLP in applications

John: “How is the weather today?”

Digital assistant: “It is 37 degrees centigrade outside with no rain today.”

John: “What does my schedule look like?”

Digital assistant: “You have a strategy meeting at 4 p.m. and an all-hands at 5:30 p.m. Based on today’s traffic situation, it is recommended you leave for the office by 8:15 a.m.”

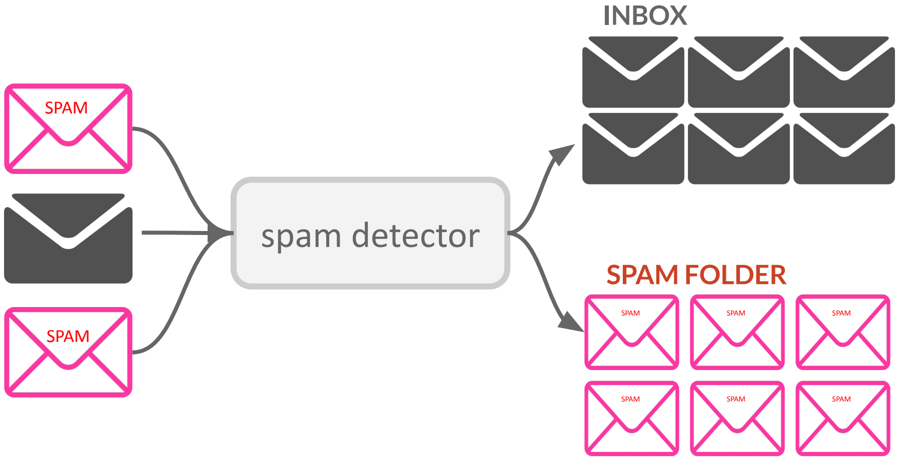
While he’s getting dressed, John probes the assistant on his fashion choices:

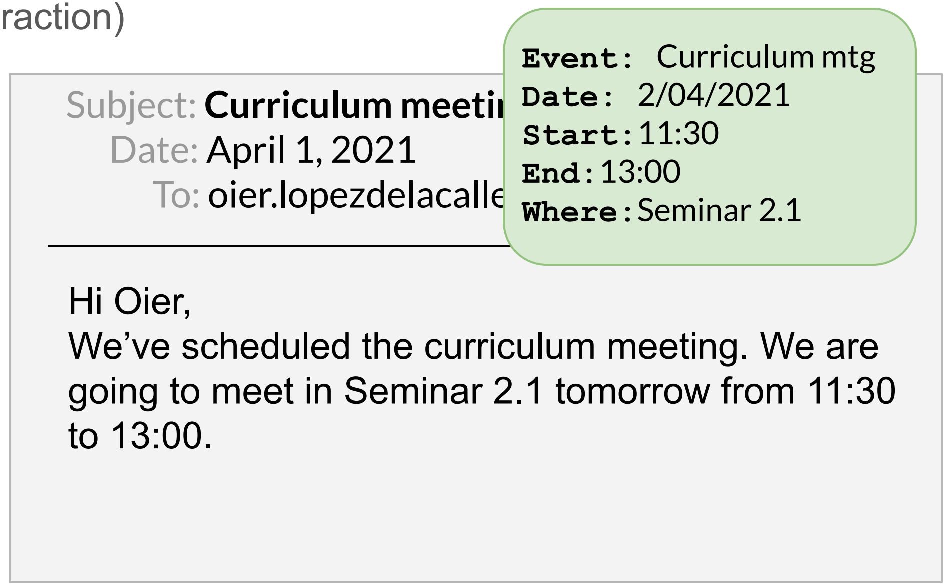
John: “What should I wear today?”

Digital assistant: “White seems like a good choice.”

NLP tasks (in the real world)

- **Email platforms** provide multiple functionalities based on text technology
  - Spam classification (text classification)
  - Priority inbox (text classification)
  - Calendar event extraction (information extraction)
  - Auto-complete (language modeling)





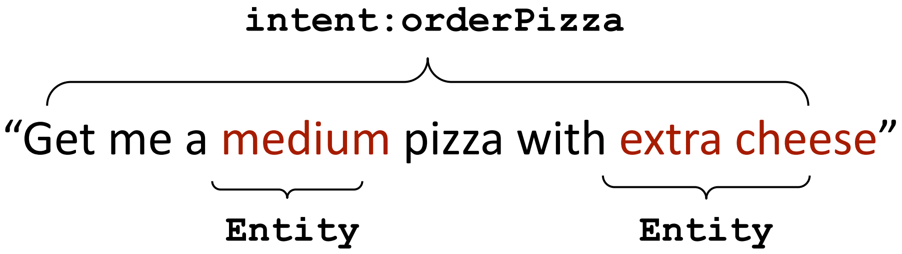
Subject: **Curriculum meeti**
Date: April 1, 2021
To: oier.lopezdelacalle

Hi Oier,
We’ve scheduled the curriculum meeting. We are going to meet in Seminar 2.1 tomorrow from 11:30 to 13:00.

Event: Curriculum mtg
Date: 2/04/2021
Start: 11:30
End: 13:00
Where: Seminar 2.1

NLP tasks (in the real world)

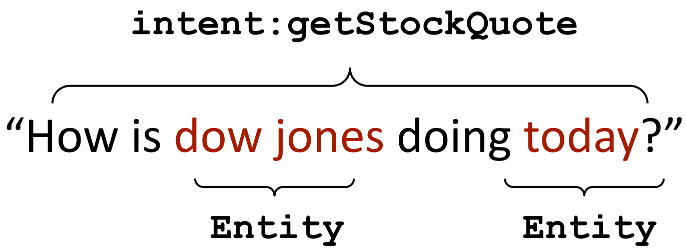
- **Voice based assistant** rely on multiple based on text technology to interact with user
  - Understand user’s commands (intent/act identification, entities)
  - Respond accordingly to user’s commands



`intent:orderPizza`

“Get me a medium pizza with extra cheese”

`Entity` `Entity`



`intent:getStockQuote`

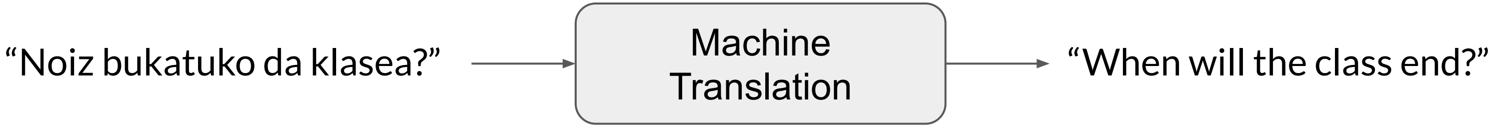
“How is dow jones doing today?”

`Entity` `Entity`

NLP tasks (in the real world)

- **Search engines** use NLP heavily for various subtasks
  - Query understanding
  - Query extension
  - Question Answering
  - Information Retrieval

- **Machine translation** services are increasingly used today.



“Noiz bukatuko da klasea?” → Machine Translation → “When will the class end?”

Domain/Industry specific NLP

- Social Media Analysis
  - Deeper understanding of people in different situations and topics.
  - Topic detection, opinion mining, sentiment analysis, fake news, content filtering
- E-commerce
  - Extract information about product description, understanding user reviews
  - Recommender systems
- NLP in specific domains
  - Healthcare, finance, law
- Automatic Report Generation
  - Generate reports for various domains: Weather forecast, financial services

Domain/Industry specific NLP

- Spelling- / Grammar- correction
  - Language modeling, rule based tools.

- Assessment Tools
  - Automated scoring of student’s exams, plagiarism detection,
  - Intelligent tutoring systems, language learning applications (e.g. Duolingo)

- Knowledge bases
  - Building large knowledge bases useful for QA and information searching

NLP tasks

Applications can be build solving and combining existing **NLP fundamental tasks**:

- *Language modeling*
  - Predicting next word
- *Text classification*
  - Map into set of categories
- *Information extraction*
  - Extract relevant information
- *Information retrieval*
  - Find relevant documents
- *Conversational agent*
  - Dialog systems
- *Text summarization*
  - Generate shorter text
- *Question Answering*
  - Find/extract relevant information
- *Machine translation*
  - Generate text in other languages
- *Topic Modeling*
  - Uncover topical structures of document collections

Source: http://www.practicalnlp.ai/

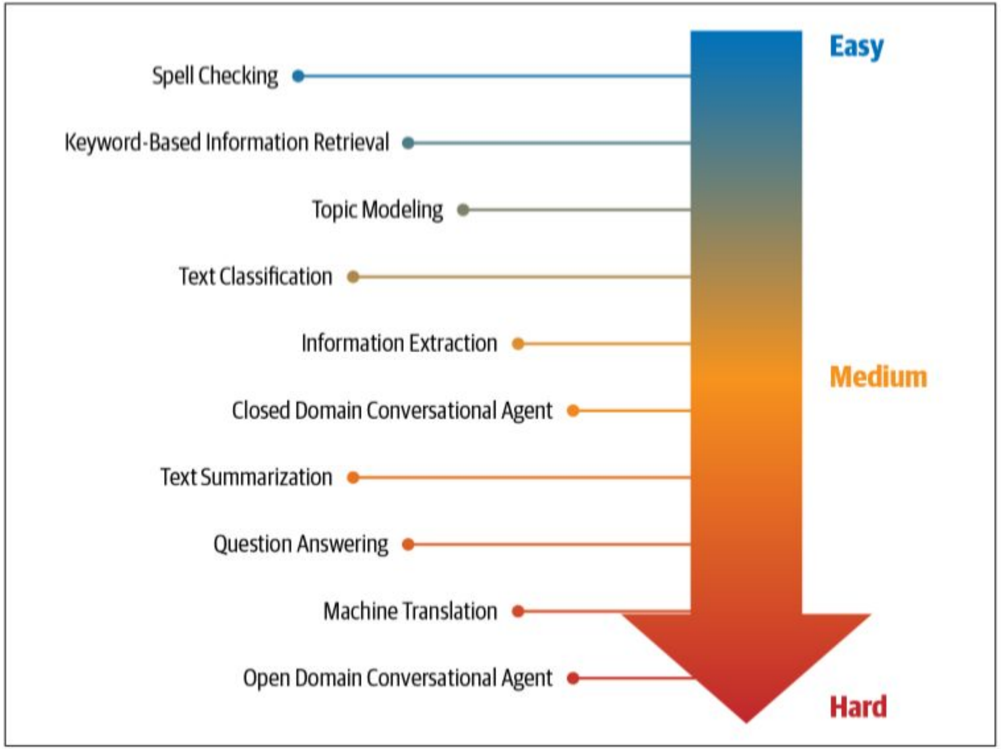
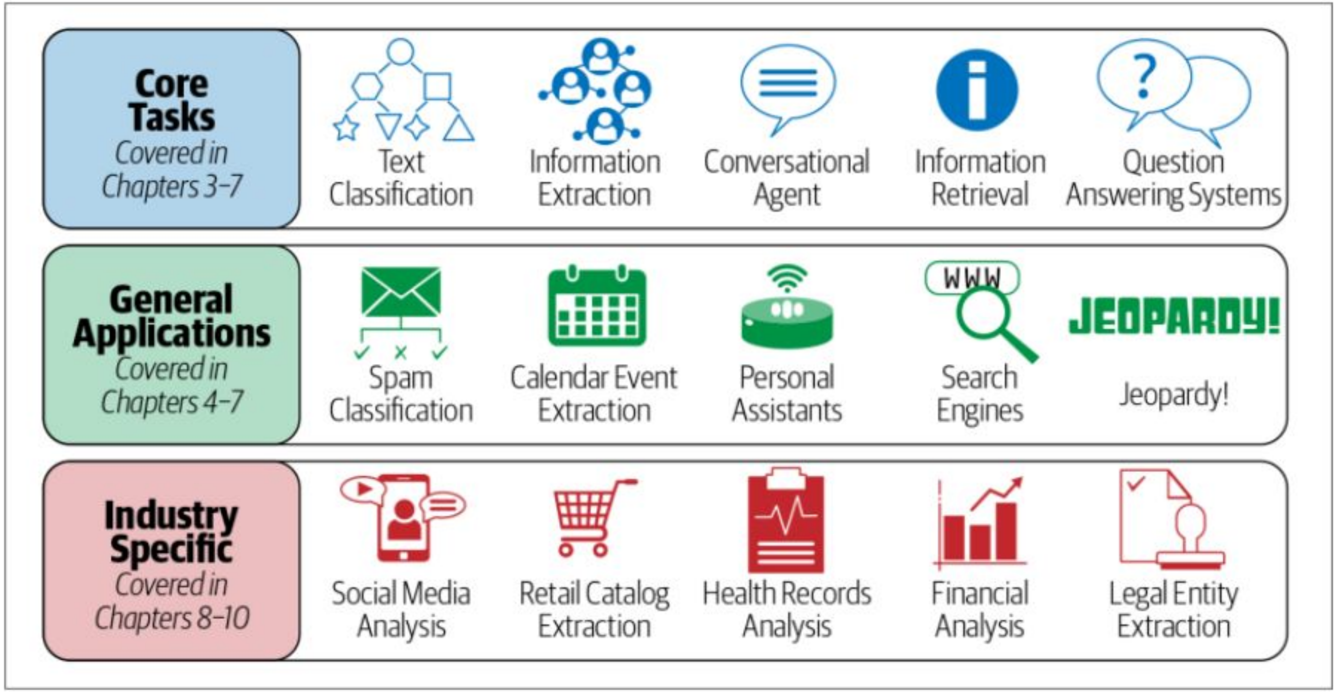


Figure 1-2. NLP tasks organized according to their relative difficulty

NLP tasks and applications



Core Tasks *Covered in Chapters 3-7*	Text Classification	Information Extraction	Conversational Agent	Information Retrieval	Question Answering Systems
General Applications *Covered in Chapters 4-7*	Spam Classification	Calendar Event Extraction	Personal Assistants	Search Engines	Jeopardy!
Industry Specific *Covered in Chapters 8-10*	Social Media Analysis	Retail Catalog Extraction	Health Records Analysis	Financial Analysis	Legal Entity Extraction

Figure 1-1. NLP tasks and applications

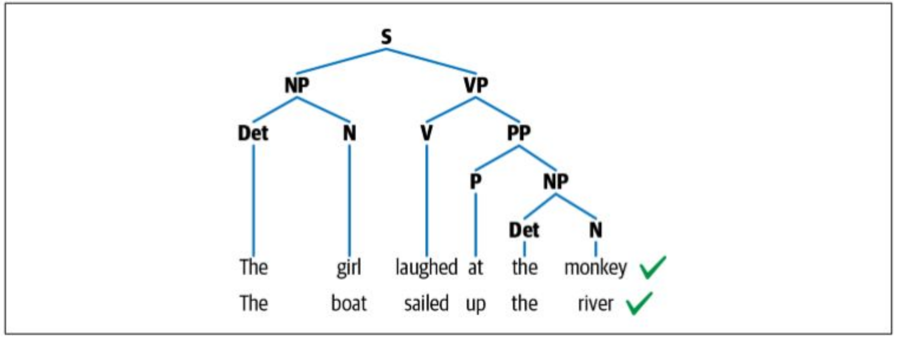
Source: http://www.practicalnlp.ai/

What is a Language?

- Language is a structured system of communication that involves complex combinations of its constituent components, such as characters, words, sentences, etc.
- Linguistics is the systematic study of language.
- Although most of NLP is based in ML, important to understand some concepts of linguistics in NLP.
- Composed of four building blocks:
    - phonemes, morphemes and lexemes, syntax, and context.
    - NLP applications need knowledge of different levels of these building blocks.

Building blocks of Language

- **Phonemes**: Smallest units of sound in a lar
  - Combination of sounds induce meaning (words, s
- **Morphemes**: Smallest unit of language that
  - Words, prefixes, suffixes: Unbreakable = un + brea
  - Prefixes/suffuxes change meaning: media vs mult
- **Lexemes**: Unit of lexical meaning that unde inflection.
  - "run", "running"
- **Syntax**: A set of rules to construct grammat phrases in a language
- **Context**: How various parts in a language c
  - Semantics: Direct meaning from sentence.
    - "The dog is in the pen" vs "The ink is in the p
  - Pragmatics: Need external knowledge.



S
NP
VP
Det
N
V
PP
P
NP
Det
N
The girl laughed at the monkey ✓
The boat sailed up the river ✓

Figure 1-6. Syntactic structure of two syntactically similar sentences

From "The Pink Panther Strikes Again"

Clouseau: Does your dog bite?

Hotel Clerk: No.

Clouseau: [*bowing down to pet the dog*] Nice doggie.

[*Dog barks and bites Clouseau in the hand*]

Clouseau: I thought you said your dog did not bite!

Hotel Clerk: That is not my dog.

Building blocks of Language

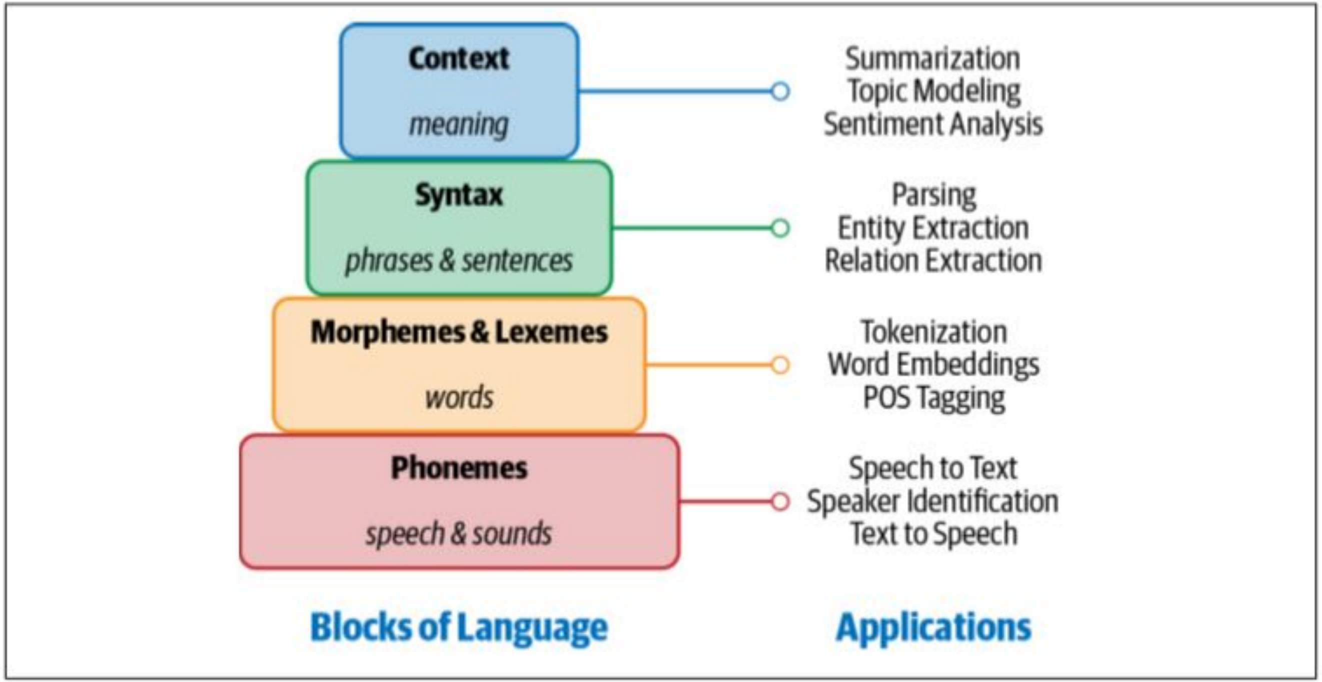


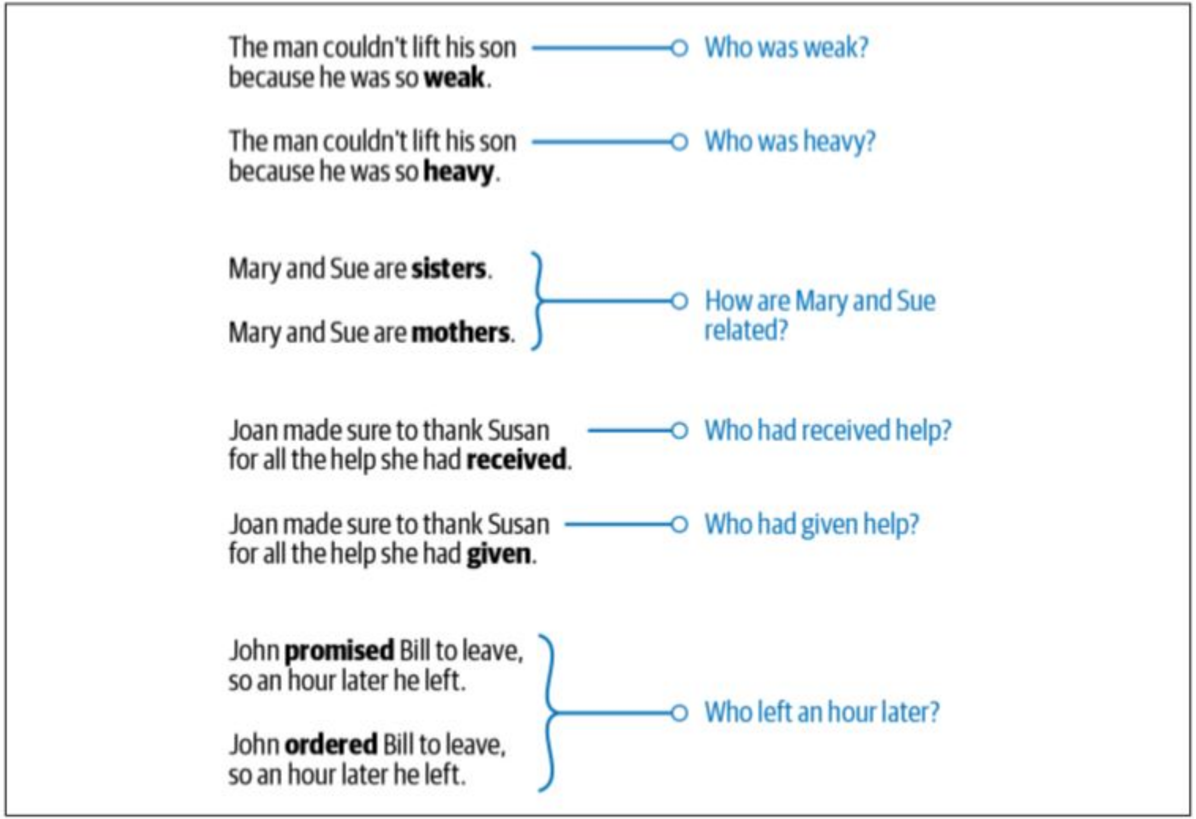
Figure 1-3. Building blocks of language and their applications

Source: http://www.practicalnlp.ai/

How Does Language Make NLP Challenging?

- **Ambiguity**: The uncertainty of meaning.
  - E.g.: “I made her duck.”
  - “made” has two meanings in the context: 1) *cook* 2) *bend down*
- **Common knowledge**: The set of all facts that most humans are aware of.
  - Known facts, not explicitly mentioned.
  - “Man bit dog” vs “Dog bit man”
  - **Key challenge**: How to encode common knowledge in a computational model.
- **Creativity**: Language is not just rule driven.
  - Various styles, dialects, genres, etc.
  - Understanding creativity difficult to AI in general.
- **Diversity across languages**: No direct mapping between any two languages.
  - Porting solution from one language to other is difficult.
  - Solutions: Language agnostic vs Separate solutions per language

Ambiguity



The man couldn't lift his son because he was so **weak**. — Who was weak?

The man couldn't lift his son because he was so **heavy**. — Who was heavy?

Mary and Sue are **sisters**.

Mary and Sue are **mothers**. — How are Mary and Sue related?

Joan made sure to thank Susan for all the help she had **received**. — Who had received help?

Joan made sure to thank Susan for all the help she had **given**. — Who had given help?

John **promised** Bill to leave, so an hour later he left.

John **ordered** Bill to leave, so an hour later he left. — Who left an hour later?

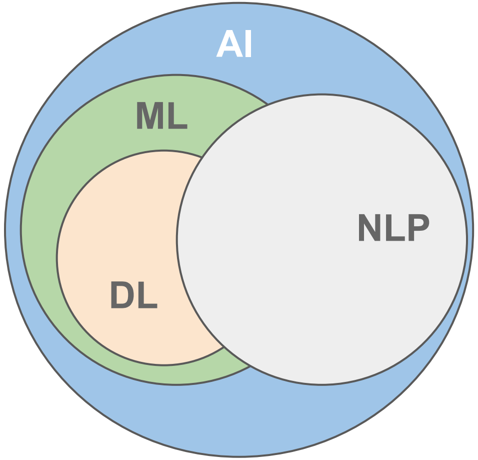
Figure 1-7. Examples of ambiguity in language from the Winograd Schema Challenge

- The meaning of the sentences is often flipped because of this minor change.
- Easy to humans.
- Difficult for machines.

Source: http://www.practicalnlp.ai/

Machine Learning, Deep Learning, and NLP: An Overview

- Artificial Intelligence branch of Computer Science
  - Aims to build systems that can perform tasks that require human intelligence.
- Initial AI (1950s) built on logic-, heuristics-, and rule-based systems.
  - Same for Natural Language Processing applications
- Nowadays: Machine Learning (ML) and Deep Learning (DL) used to build AI systems.
  - Same for Natural Language Processing applications



Machine Learning

- Goal: ***Learn*** to **perform tasks** based on examples (*training data*) without **explicit instruction**
    - Features: Numeric representation of the training data
    - Learning: Learn the patterns in training data.
- Three main groups:
    - **Supervised learning**: Learn mapping function ($f : X \rightarrow Y$) based labeled examples
        - Text classification, sequence labeling
    - **Unsupervised learning**: Find hidden patterns in given input data without any reference output.
        - Topic modelling, semi-supervised
    - **Reinforcement learning**: Learn tasks via trial and error and is characterized by the absence of either labeled or unlabeled data in large quantities.
        - Getting more common in NLP!

Approaches to NLP

Approaches to solve NLP problems fall into three categories:

- **Heuristic based NLP**
- **Machine Learning based NLP**
- **Deep Learning based NLP**

Many applications might combine more than one category

Heuristic based NLP

- Early systems in NLP based in defining **rules** for specific task
- Need domain expertise
- Rely on **structured resources**: Dictionaries, Thesauruses, Knowledge Bases
    - E.g WordNet
        - Concepts are synsets (synonym set)
        - Semantic relationships: Hyper-/Hyponyms, Meronyms...
- **Regular Expressions** (regexp): Pattern that is used to match and find substrings in text.
    - *Find all emails in text*: ‘^([a-zA-Z0-9_\-\.]+)@([a-zA-Z0-9_\-\.]+)\.([a-zA-Z]{2,5})$’
- **Context-free grammar** (CFG): Useful to extract more complex and hierarchical information.

Heuristic based NLP

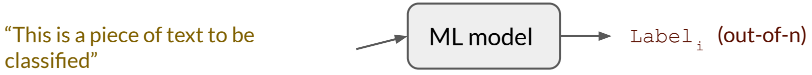
Still useful in many situations:

- **Annotated data**: Widely used in industry and domains with no annotated data
- **Feature engineering**: Useful for defining and extracting ML features
- **Postprocessing**: Filter/correct ML/DL output.

ML based NLP

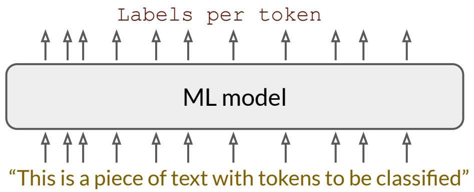
- Extracting features from text.
- Use the feature representation to learn a model.
- Evaluate and improve the model.

Text classification



- Naive Bayes
- Support Vector Machines
- Logistic Regression

Sequence labeling



- Hidden Markov Models
- Conditional Random Fields

Deep Learning for NLP

- Recurrent neural networks (RNN)
  - Language is sequential
- Long short-term memory (LSTM)
  - Perform better than RNN when text is longer
- Convolutional neural networks (CNN)
  - Ability to look at group of words together (~*n-grams*)
- Transformers
  - Given a word, it prefers to look at all the words around it (*self-attention*)
- Autoencoders
  - Learn compressed vector representation of the input for any downstream task.

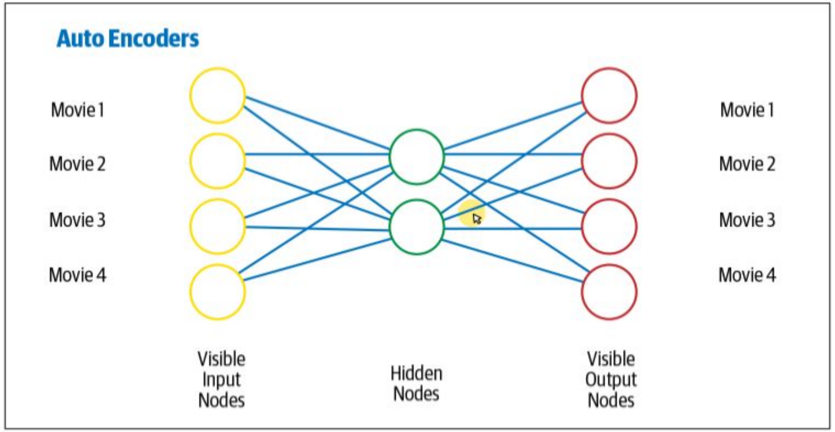


Figure 1-18. Architecture of an autoencoder

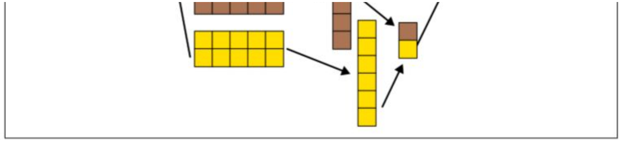


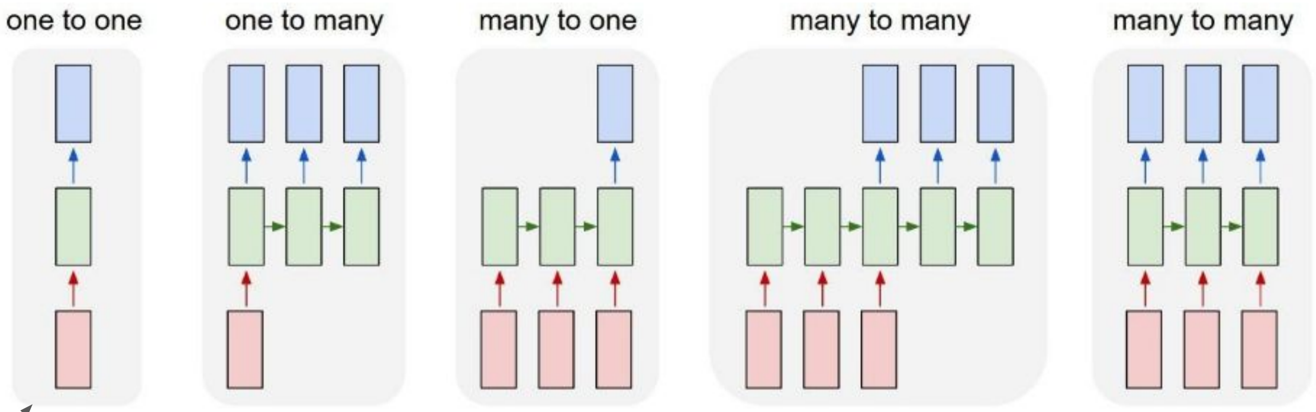
Figure 1-15. CNN model in action [25]

Sources:

- https://colah.github.io/posts/2015-08-Understanding-LSTMs/
- http://jalammar.github.io/illustrated-transformer/
- https://github.com/practical-nlp/practical-nlp/tree/master/Ch1

Sequence modeling

e.g: Machine Translation
Sequence of words → sequence of words



Source: Fei-Fei Li & Andrej Karpathy & Justin Johnson

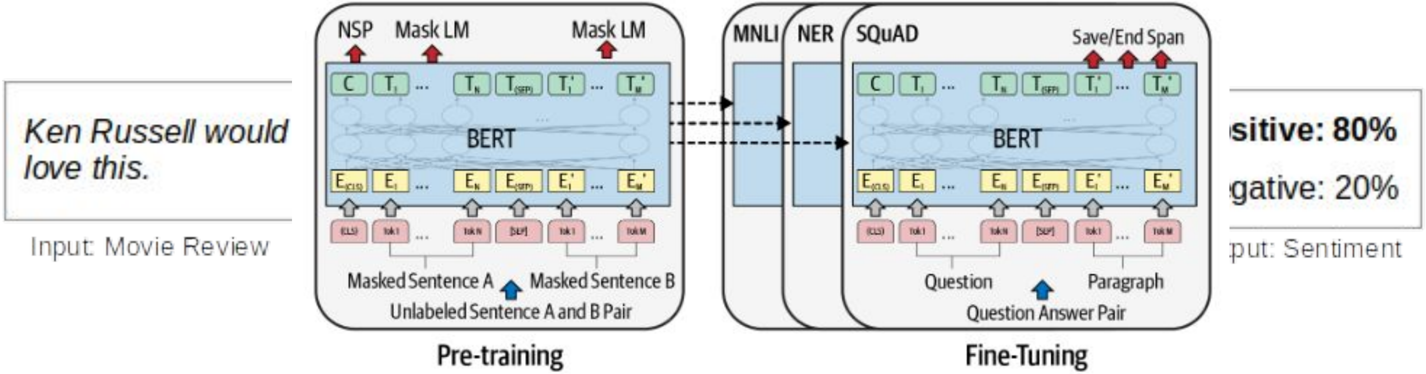
Vanilla neural network

e.g: Sentiment Classification
Sequence of words → sentiment

e.g: Image captioning
Image → sequence of words

e.g: Named Entity Recognition
Sequence of words → labels

Transfer Learning: Train then fine-tune



Pre-training:

- Train a very large transformer based LM (known as pre-training)
- Predict a part of a sentence (masking) given the rest of the content (self-learning)
- Encode the high-level nuances of the language in it.

Fine-tuning:

- Fine-tuned on downstream NLP tasks, such as text classification, entity extraction, question answering

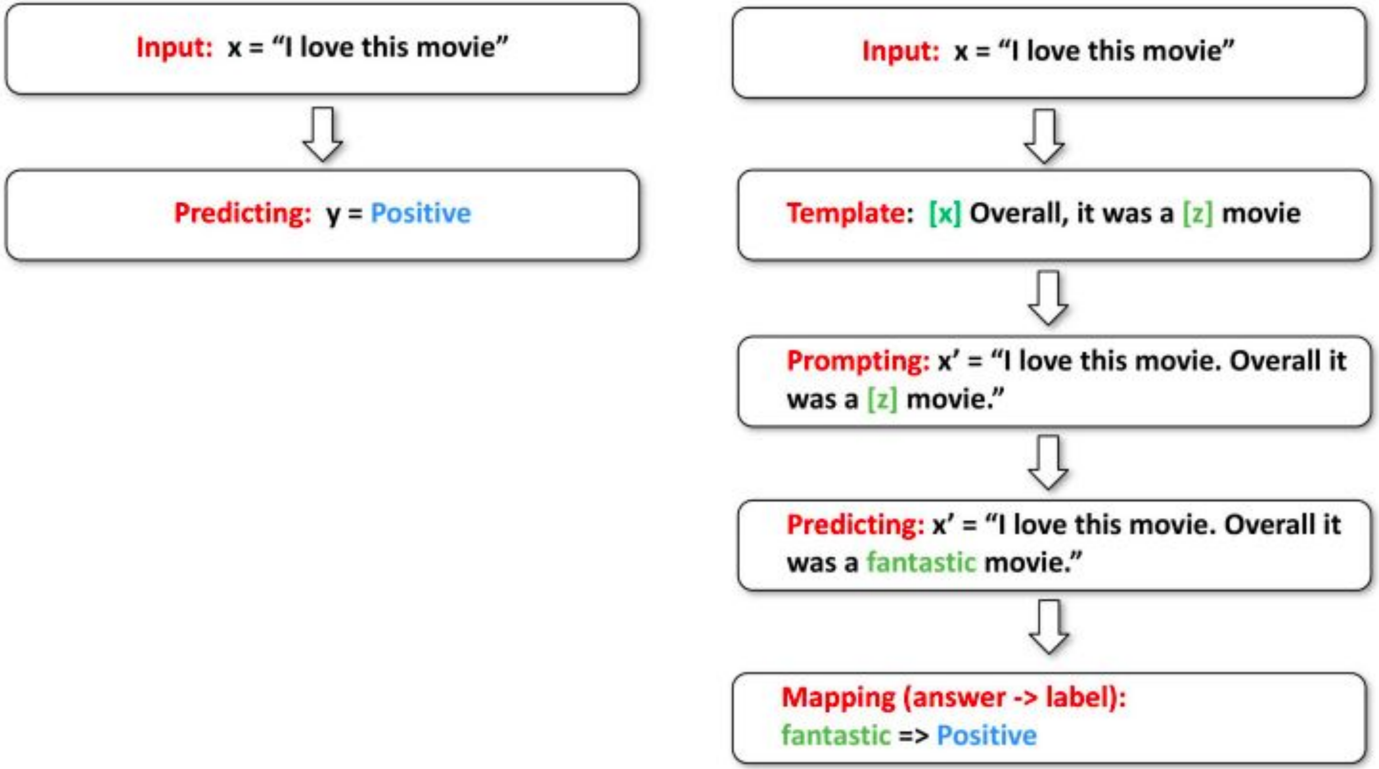
Prompt-based learning

What is Prompting?

- Encourage a pre-trained model to make a particular predictions by providing a “prompt” specifying the task to be done.

Prompt-based learning

Traditional Formulation vs Prompt Formulation



Input: x = “I love this movie”

Predicting: y = Positive

Input: x = “I love this movie”

Template: [x] Overall, it was a [z] movie

Prompting: x’ = “I love this movie. Overall it was a [z] movie.”

Predicting: x’ = “I love this movie. Overall it was a fantastic movie.”

Mapping (answer -> label): fantastic => Positive

Prompt Addition

Answer Prediction

Answer Mapping

DL not a silver bullet (I)

- Overfitting on small datasets
  - Tend to have more parameters and memorize small datasets
  - Poorer generalization properties in production
- Few-shot learning and synthetic data generation
  - Compared to CV, NLP application need more data for few-shot
  - Synthetic data generation is more challenging than in CV
- Domain adaptation
  - DL models may have poor performance when domain changes.
  - Models trained on internet texts and product reviews will not work well in healthcare domain.
  - Syntactic and semantic structure of the language is specific to the domain.

DL not a silver bullet

- Interpretable models
  - DL models are hard to interpret as they are used as black-box.
  - The are few approaches to gain insight of DL model in a particular task.
- Common sense and world knowledge
  - ML/DL lack of reasoning abilities.
  - Most challenging research is to incorporate common sense, world knowledge and reasoning abilities to DL models
  - *"If John walks out of the bedroom and goes to the garden, then John is not in the bedroom anymore, and his current location is the garden."* → Where is John?
- Cost
  - DL-based solutions are very expensive: Time, money, environmental and hardware resources
- On-device deployment
  - DL models are too large to embed in smaller devices (e.g. mobile phones)
  - Good MT need powerful server (+internet connection)

Useful Resources

- **Book**: **Practical Natural Language Processing:** http://www.practicalnlp.ai
  - https://github.com/practical-nlp/practical-nlp
- **Book: Introduction to Natural Language Processing**: https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf
- **NLP surveys**: https://github.com/NiuTrans/ABigSurvey
- **Book**: **Dive into Deep Learning**: https://d2l.ai/index.html
- **ML from scratch repository**: https://github.com/eriklindernoren/ML-From-Scratch?s=03
- **Legal texts**: https://github.com/LexPredict/lexpredict-lexnlp
- **Papers with Code**: https://paperswithcode.com/