Natural Language Processing Problems and Solutions – A Machine Learning Perspective

Dr. Bal Krishna Bal, bal@ku.edu.np

Associate Professor & Head, Department of Computer Science & Engineering

Kathmandu University, Dhulikhel, Kavre, Nepal

Slide materials have been adopted from different sources. Special thanks to Raymond J. Mooney, University of Texas at Austin.

Contents

- What is Natural Language Processing(NLP)
- Related Areas
- NLP Applications in Real Life
- Components of NLP
- Syntax, Semantics, Pragmatics
- Modular Comprehension
- Classic NLP Problems
- Natural Language Tasks
  - Syntactic, Semantic, Pragmatic, Discourse, Others
- Important Machine Learning Concepts for Building NLP solution
- Machine Learning Approach to Building NLP Solution
- Text Classification
- Tools and Libraries
- Relevant Scientific Conferences
- Top books on NLP
- References

What is Natural Language Processing (NLP)

- Short form for Natural Language Processing.
- A sub-discipline of Artificial Intelligence in Computer Science.
- *According to Wikipedia, NLP is a field of Computer Science and Linguistics concerned with the interactions between computers and human (natural) languages.*
- *Also called Computational Linguistics.*

Related Areas

- Artificial Intelligence
- Formal Language (Automata) Theory
- Machine Learning
- Linguistics
- Psycholinguistics
- Cognitive Science
- Philosophy of Language

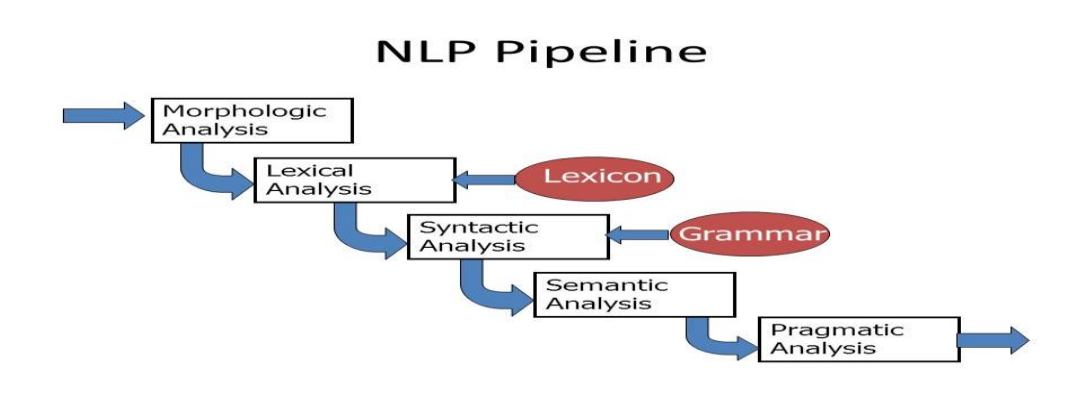
NLP Applications in Real Life

- Information Retrieval
- Information Extraction
- Machine Translation
- Sentiment Analysis
- Text Summarization
- Spam Filter
- Auto-Predict
- Auto-Correct

- Speech Recognition
- Text-to-Speech
- Optical Character Recognition
- Handwriting Recognition
- Question Answering
- Natural Language Generation
- Named-Entity Recognition
- Word Sense Disambiguation

Components of NLP

- Natural Language Understanding (NLU)
  - Input in Natural Language → Useful representations
- Natural Language Generation (NLG)
  - Internal representations → meaningful phrases and sentences in the form of natural language

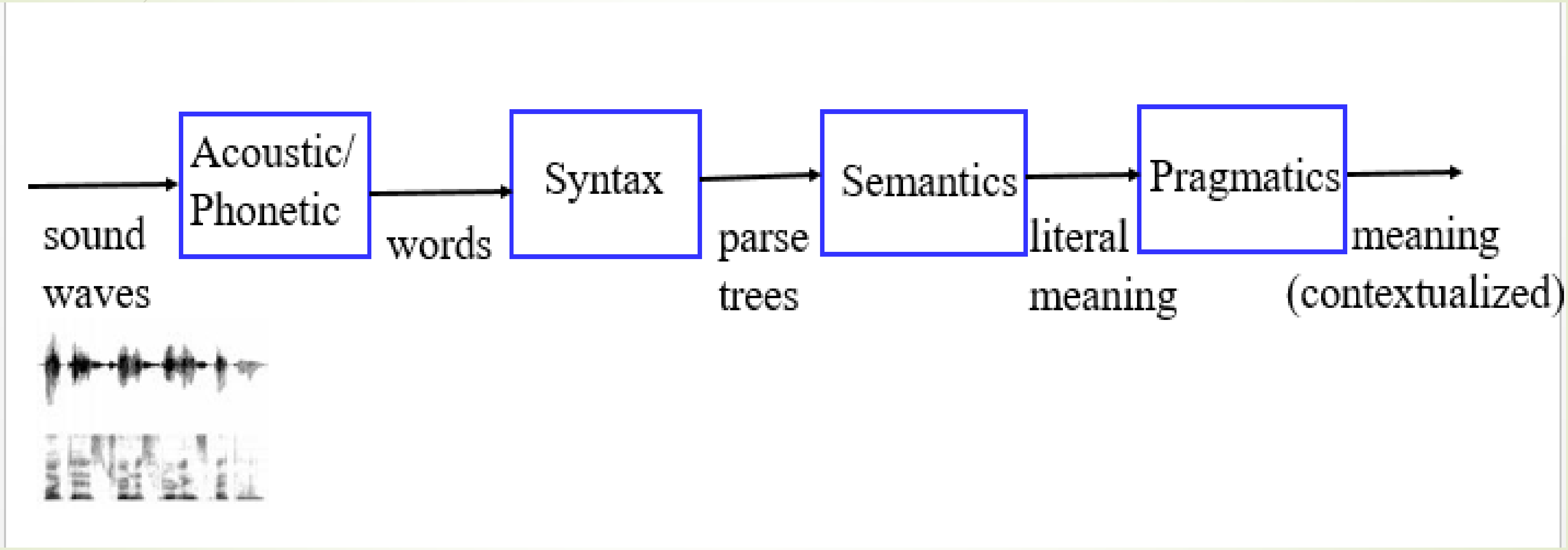


- NLU
  - Different levels of analysis involved:
    - □ Morphological analysis
    - □ Syntactic analysis
    - □ Semantic analysis
    - □ Discourse analysis
- NLG
  - Different levels of synthesis involved:
    - □ Deep planning
    - □ Syntactic generation
- In general, NLU is much harder than NLG, although both are hard problems.

Syntax, Semantic, Pragmatics

- Syntax concerns the proper ordering of words and its affect on meaning.
  - ❑ The dog bit the boy.
  - ❑ The boy bit the dog.
  - ❑ * Bit boy dog the the.
  - ❑ Colorless green ideas sleep furiously.
- Semantics concerns the (literal) meaning of words, phrases, and sentences.
  - ❑ "plant" as a photosynthetic organism
  - ❑ "plant" as a manufacturing facility
  - ❑ "plant" as the act of sowing
- Pragmatics concerns the overall communicative and social context and its effect on interpretation.
  - ❑ The ham sandwich wants another beer. (co-reference, anaphora)
  - ❑ John thinks vanilla. (ellipsis)

Modular Comprehension



Classic NLP Problems

- Linguistically-motivated: segmentation, tagging, parsing
- Analytical: classification, sentiment analysis
- Transformation: translation, correction, generation
- Conversation: question-answering, dialog

Issues in Natural Language Processing:

- Ambiguity
  - ❑ Lexical ambiguity: "bank"
  - ❑ Scope ambiguity: "Every man loves a woman."
  - ❑ Structural ambiguity: "I saw the boy with a telescope."
- Non-standard use of the language
  - ❑ Shorthands: "c u", "b4 u", "want 2 go"
- Variability: "diabetes", "dm", "diab"
- Segmentation issues
- Idioms
- Coining of new words over time: "google" as a verb.
- World knowledge

Natural Language Tasks

- Processing natural language text involves many various syntactic, semantic and pragmatic tasks in addition to other problems.

Syntactic Tasks

Word Segmentation

- Breaking a string of characters (graphemes) into a sequence of words.
- In some written languages (e.g. Chinese) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ( )]
- Examples from English URLs:
  - jumptheshark.com ⇒ jump the shark .com
  - myspace.com/pluckerswingbar

    ⇒ myspace .com pluckers wing bar

    ⇒ myspace .com plucker swing bar

Word Segmentation

- Breaking a string of characters (graphemes) into a sequence of words.
- In some written languages (e.g. Chinese) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ( )]
- Examples from English URLs:
  - jumptheshark.com ⇒ jump the shark .com
  - myspace.com/pluckerswingbar

    ⇒ myspace .com pluckers wing bar

    ⇒ myspace .com plucker swing bar

Morphological Analysis

- ***Morphology*** is the field of linguistics that studies the internal structure of words. (Wikipedia)
- A ***morpheme*** is the smallest linguistic unit that has semantic meaning (Wikipedia)
  - e.g. “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried ⇒ carry + ed (past tense)
  - independently ⇒ in + (depend + ent) + ly
  - Googlers ⇒ (Google + er) + s (plural)
  - unlockable ⇒ un + (lock + able) ?

    ⇒ (un + lock) + able ?

Morphological Analysis

- ***Morphology*** is the field of linguistics that studies the internal structure of words. (Wikipedia)
- A ***morpheme*** is the smallest linguistic unit that has semantic meaning (Wikipedia)
  - e.g. “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried ⇒ carry + ed (past tense)
  - independently ⇒ in + (depend + ent) + ly
  - Googlers ⇒ (Google + er) + s (plural)
  - unlockable ⇒ un + (lock + able) ?

    ⇒ (un + lock) + able ?

Part Of Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech.

I	ate	the	spaghetti	with	meatballs.
Pro	V	Det	N	Prep	N

John	saw	the	saw	and	decided	to	take	it	to	the	table.
PN	V	Det	N	Con	V	Part	V	Pro	Prep	Det	N

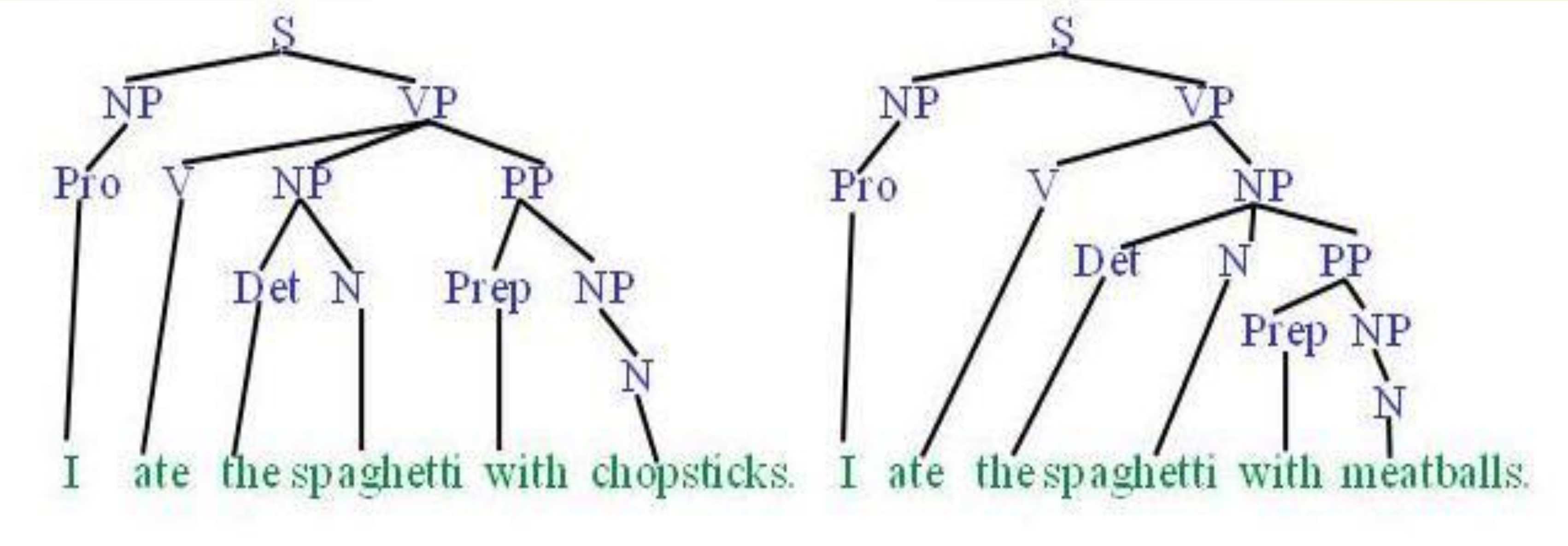
- Useful for subsequent syntactic parsing and word sense disambiguation.

Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.
  - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
  - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

Syntactic Parsing

- Produce the correct syntactic parse tree for a sentence.


S
NP
VP
Pro
V
NP
PP
Det
N
Prep
NP
N
I ate the spaghetti with chopsticks.
S
NP
VP
Pro
V
NP
Det
N
PP
Prep
NP
N
I ate the spaghetti with meatballs.

Semantic Tasks

Word Sense Disambiguation(WSD)

- Words in natural language usually have a fair number of different possible meanings.
  - Ellen has a strong interest in computational linguistics.
  - Ellen pays a large amount of interest on her credit card.
- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

  agent patient source destination instrument

  - John drove Mary from Austin to Dallas in his Toyota Prius.
  - The hammer broke the window.
- Also referred to a "case role analysis," "thematic analysis," and "shallow semantic parsing"

Semantic Parsing

- A ***semantic parser*** maps a natural-language sentence to a complete, detailed semantic representation (***logical form***).
- For many applications, the desired output is immediately executable by another program.
- Example: Mapping an English database query to Prolog:

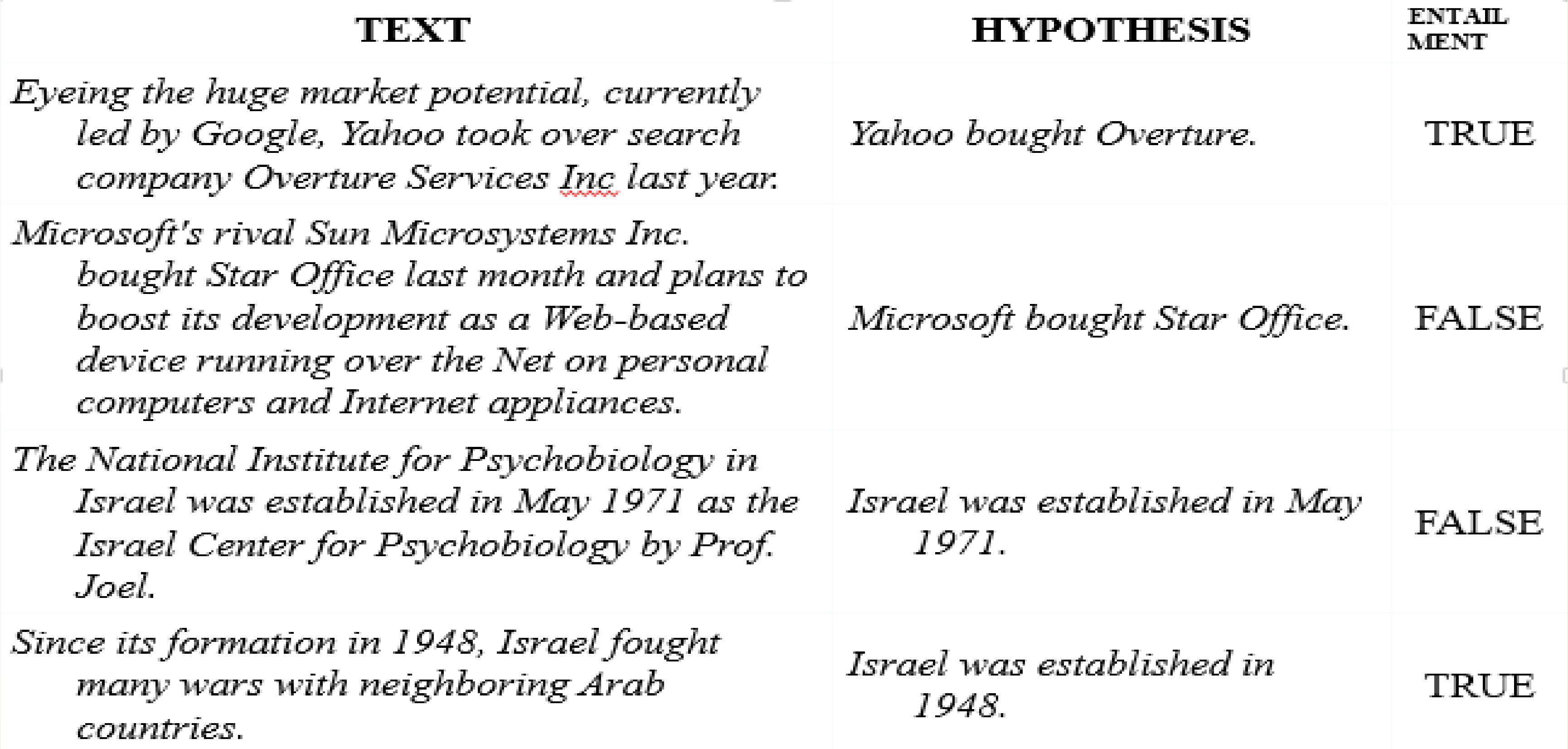
  How many cities are there in the US?

```
answer(A, count(B, (city(B), loc(B, C),
                    const(C, countryid(USA))),
                A))
```

Textual Entailment

- Determine whether one natural language sentence entails (implies) another under an ordinary interpretation.

Textual Entailment Problems from PASCAL Challenge



TEXT	HYPOTHESIS	ENTAILMENT
Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.	*Yahoo bought Overture.*	TRUE
Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.	*Microsoft bought Star Office.*	FALSE
The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.	*Israel was established in May 1971.*	FALSE
Since its formation in 1948, Israel fought many wars with neighboring Arab countries.	*Israel was established in 1948.*	TRUE

Pragmatics/Discourse Tasks

Anaphora resolution/Co-reference

- Determine which phrases in a document refer to the same underlying entity.
  - John put the carrot on the plate and ate it.
  - Bush started the war in Iraq. But the president needed the consent of Congress.
- Some cases require difficult reasoning.
  - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Ellipsis Resolution

- Frequently words and phrases are omitted from sentences when they can be inferred from context.

"Wise men talk because they have something to say;
fools, because they have to say something.“ (Plato)

"Wise men talk because they have something to say;
fools talk because they have to say something.“ (Plato)

Other Tasks

Information Extraction(IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is task of identifying names of people, places, organizations, etc. in text.

  people organizations places

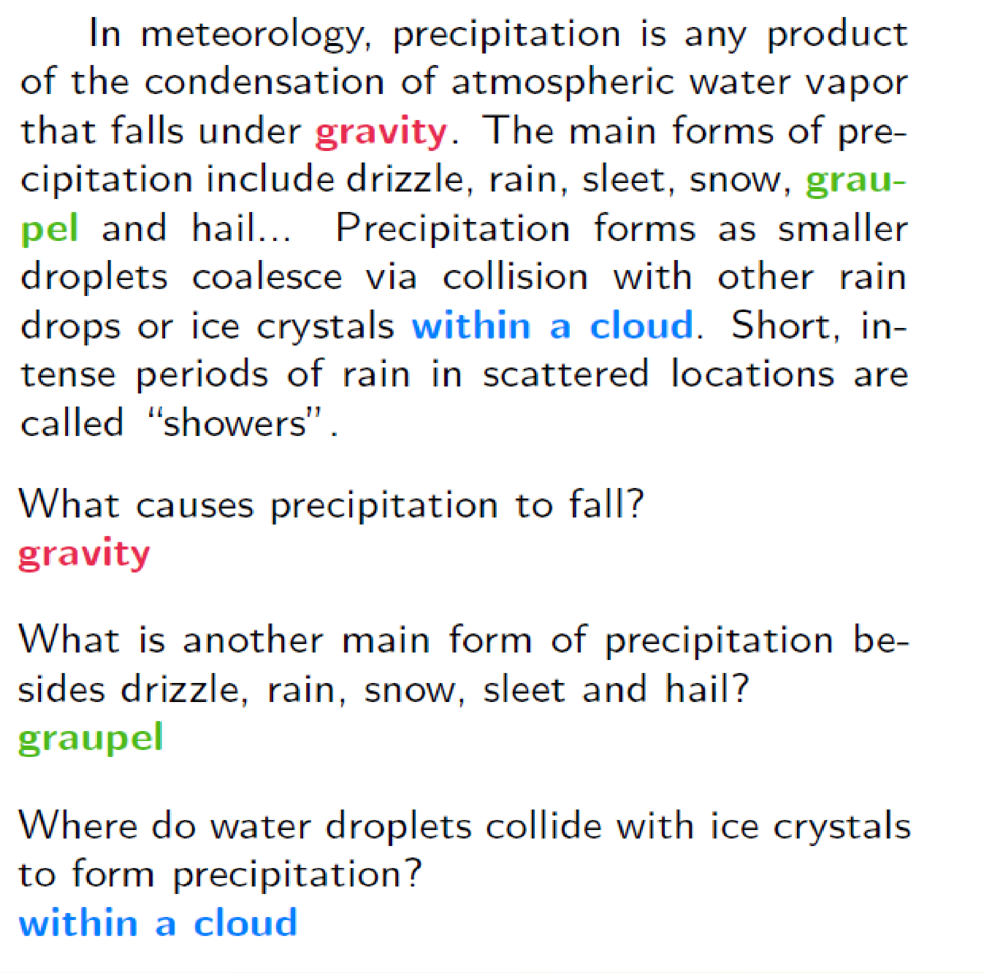
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
- Relation extraction identifies specific relations between entities.
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

Question Answering

- Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the web).
    - When was Barack Obama born? (*factoid*)
        - August 4, 1961
    - Who was president when Barack Obama was born?
        - John F. Kennedy
    - How many presidents have there been since Barack Obama was born?
        - 9

Reading Comprehension

- Read a passage of text and answer questions about it.
- Example from Stanford SQuAD dataset.



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Text Summarization

- Produce a short summary of a longer document or article.
  - Article: With a split decision in the final two primaries and a flurry of superdelegate endorsements, Sen. Barack Obama sealed the Democratic presidential nomination last night after a grueling and history-making campaign against Sen. Hillary Rodham Clinton that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against Sen. John McCain, the presumptive Republican nominee….
  - Summary: Senator Barack Obama was declared the presumptive Democratic presidential nominee.

Machine Translation (MT)

- Translate a sentence from one natural language to another.
  - Hasta la vista, bebé ⇒
    Until we see each other again, baby.

Important Machine Learning Concepts for Building NLP Solutions

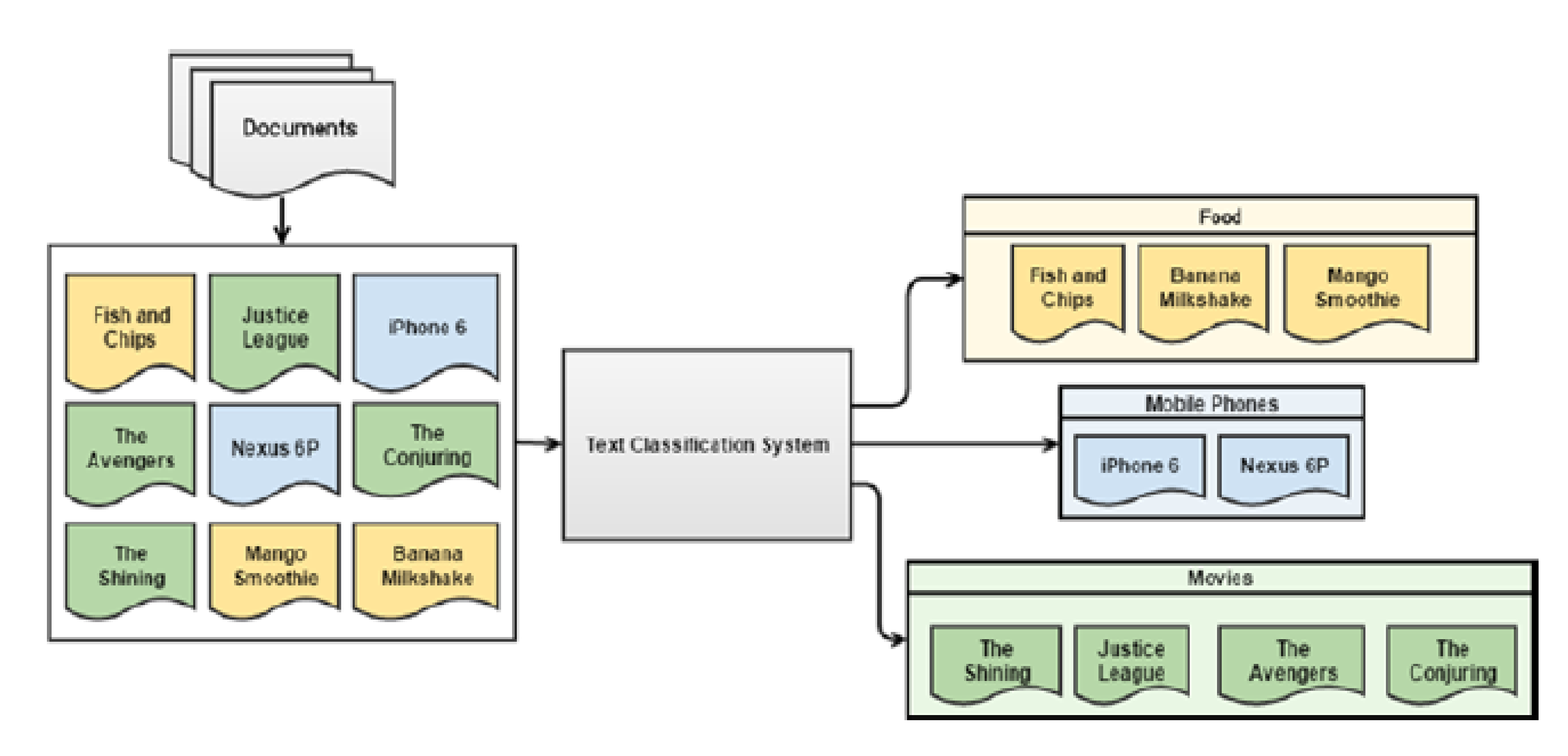
- Data preparation: Usually consists of pre-processing the data before extracting features and training.
- Feature extraction: The process of extracting useful features from raw data that are used to train machine learning models.
- Features: Various useful attributes of the data (examples could be age, weight, and so on for personal data)
- Training data: A set of data points used to train a model.
- Testing/validation data: A set of data points on which a pre-trained model is tested and evaluated to see how well it performs.
- Model: Built using a combination of data/features and a machine learning algorithm that could be supervised or unsupervised.
- Accuracy: How well the model predicts something (also has other detailed evaluation metrics like precision, recall, and F1-score)

Machine Learning Approach to Building NLP Solution

Text Classification

- Process of assigning text documents into one or more classes or categories, assuming that we have a predefined set of classes.
- A text classification system would successfully be able to classify each document to its correct class(es) based on the inherent properties of the document.
- Mathematically, we can define it like this: given some description and attributes *d* for a document *D*, where $d \in D$, and we have a set of predefined classes or categories, $C = \{C_1, C_2, C_3, \ldots, C_n\}$.
- The actual document *D* can have many inherent properties and attributes that lead it to being an entity in a high-dimensional space.
- Using a subset of that space with a limit set of descriptions and features depicted by *d*, we should be able to successfully assign the original document *D* to its correct class C_x using a text classification system *T*.
- This can be represented by $T:D \rightarrow C_x$.

Text Classification



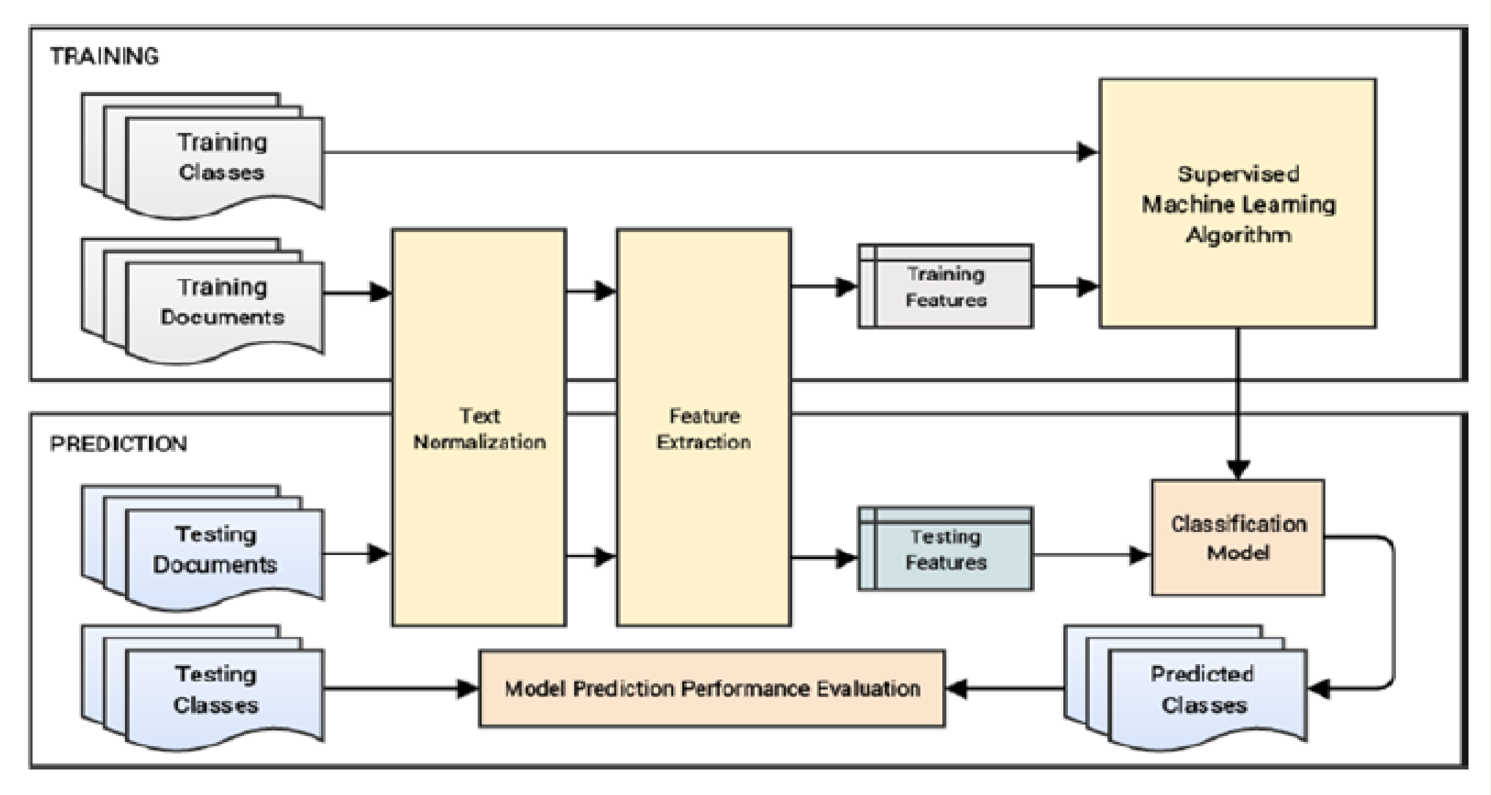
Conceptual overview of text classification

Automated Text Classification

- To automate text classification, we can make use of several ML techniques and concepts.
- There are mainly two types of ML techniques that are relevant to solving this problem:
  - Supervised machine learning
  - Unsupervised machine learning
- There are two main processes in the supervised classification process:
  - Training
  - Prediction

Text Classification Blueprint

- Typical workflow for a text classification system
  - Prepare train and test datasets
  - Text normalization
  - Feature extraction
  - Model training
  - Model prediction and evaluation
  - Model deployment



Text classification blueprint

Some Text Classification Algorithms

- Multinomial Naïve Bayes
- Support Vector Machines
- Logistic Regression
- Decision Trees
- Neural Networks
- Deep Learning-based Techniques

Evaluating Classification Models

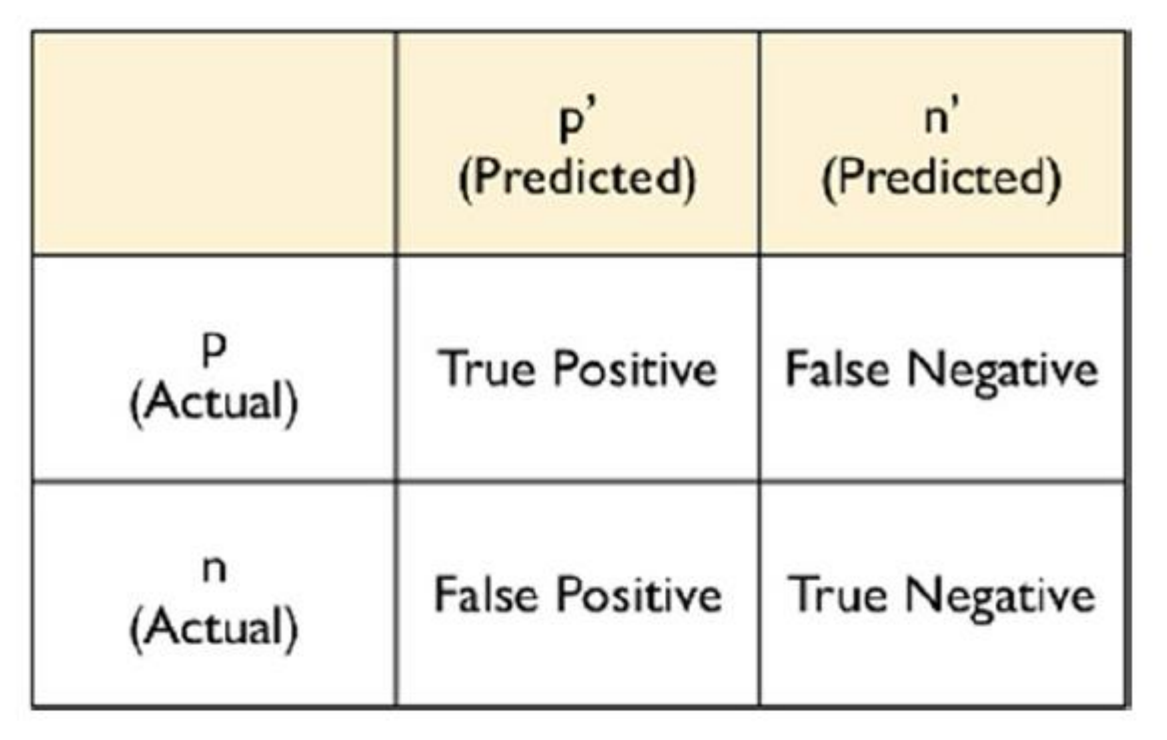
- Accuracy
- Precision
- Recall
- F1 score

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\,Score = \frac{2 \times\ Precision \times Recall}{Precision + Recall}$$



	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

A confusion matrix for a two-class classification problem

Tools and Libraries

- Stanford's Core NLP Suite
- Natural Language Toolkit
- Apache Lucene and Solr
- Apache OpenNLP
- GATE and Apache UIMA

Relevant Scientific Conferences

- Association for Computational Linguistics (ACL)
- North American Association for Computational Linguistics (NAACL)
- International Conference on Computational Linguistics (COLING)
- Empirical Methods in Natural Language Processing (EMNLP)
- Conference on Computational Natural Language Learning (CoNLL)
- International Association for Machine Translation (IMTA)

Top Books on NLP

- Natural Language Processing with Python, Steven Bird, Ewan Klein and Edward Loper.
- Taming text, Grant Ingersoll, Thomas Morton and Drew Farris.
- Text Mining with R, Julia Silge and David Robinson.
- Foundations of Statistical Natural Language Processing, Christopher Manning and Hinrich Shutze.
- Speech and Language Processing, Daniel Jurafsky and James Martin.
- Statistical Machine Translation, Philipp Koehn
- Statistical Methods for Speech Recognition, Frederick Jelinek.
- Neural Network Methods in Natural Language Processing
- The Oxford Handbook of Computational Linguistics

References

- Text Analytics with Python – A Practical Real-World Approach to Gaining Actionable Insights from Your Data, Dipanjan Sarkar, 2016.