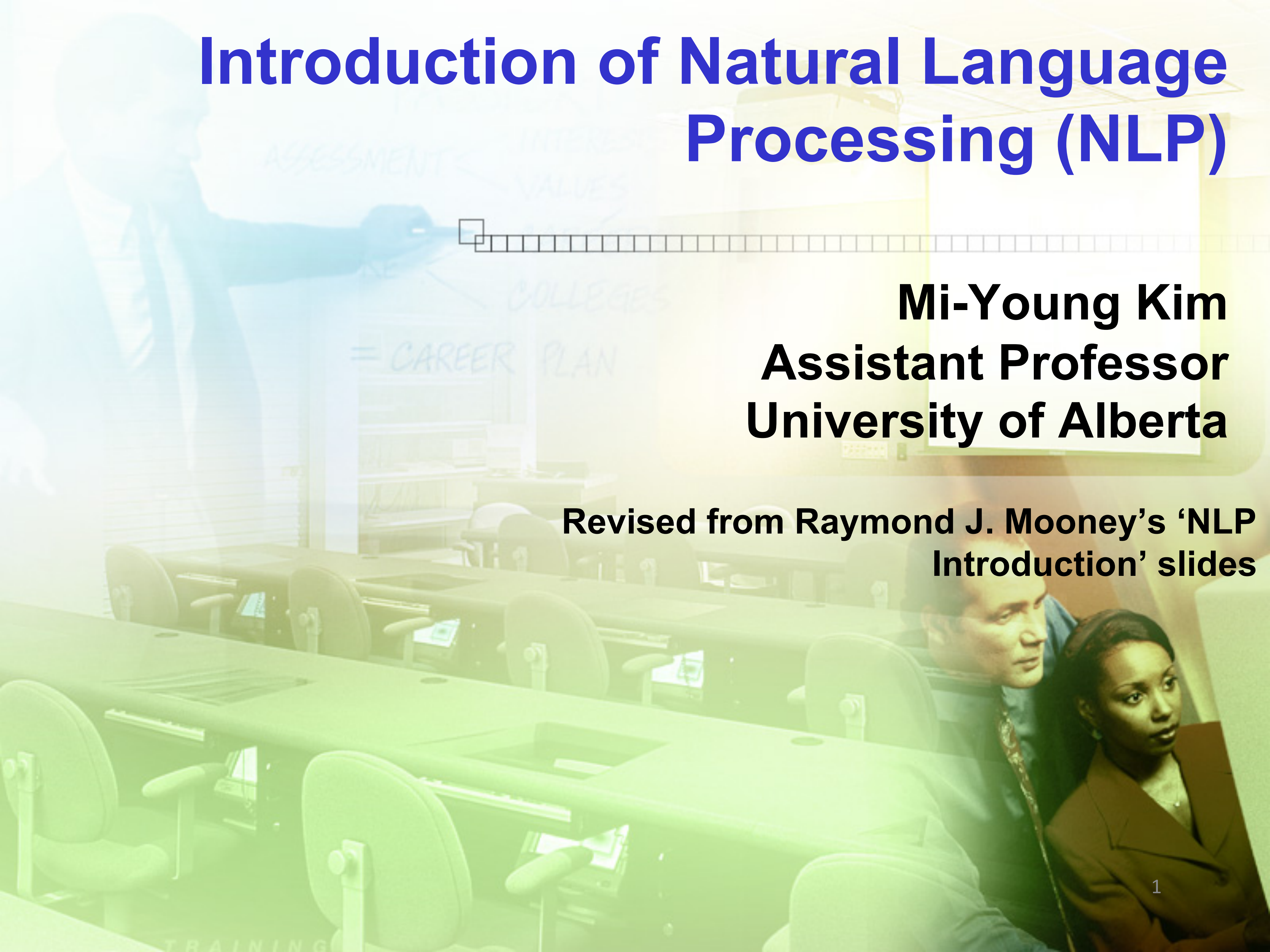


Introduction of Natural Language Processing (NLP)

Mi-Young Kim
Assistant Professor
University of Alberta

Revised from Raymond J. Mooney's 'NLP Introduction' slides

Natural Language Processing

- NLP is the branch of computer science focused on developing systems that allow computers to communicate with people using everyday language.
- Also called Computational Linguistics
  - concerns how computational methods can aid the understanding of human language

Related Areas

- Artificial Intelligence
- Machine Learning
- Linguistics
- Psycholinguistics
- Cognitive Science
- Philosophy of Language

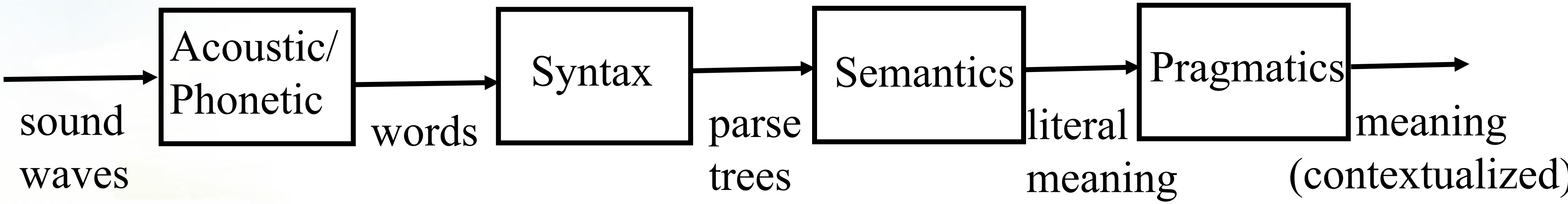
Communication

- The goal in the generation and comprehension of natural language is **communication**.

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.
- 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.

Modular Comprehension



Ambiguity

- **Natural language is highly ambiguous and must be *disambiguated*.**
  - **I saw the man on the hill with a telescope.**
  - **I saw the Grand Canyon flying to LA.**
  - **Time flies like an arrow.**
  - **Horse flies like a sugar cube.**
  - **Time runners like a coach.**
  - **Time cars like a Porsche.**

Ambiguity is Ubiquitous

- Speech Recognition
  - “recognize speech” vs. “wreck a nice beach”
  - “youth in Asia” vs. “euthanasia”
- Syntactic Analysis
  - “I ate spaghetti with chopsticks” vs. “I ate spaghetti with meatballs.”
- Semantic Analysis
  - “The dog is in the pen.” vs. “The ink is in the pen.”
  - “I put the plant in the window” vs. “Ford put the plant in Mexico”
- Pragmatic Analysis
  - 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.

Humor and Ambiguity

- Many jokes rely on the ambiguity of language:
  - She criticized my apartment, so I knocked her flat.
  - Noah took all of the animals on the ark in pairs. Except the worms, they came in apples.

Ambiguous Language

- Language relies on people's ability to use their knowledge and inference abilities to properly resolve ambiguities.
- Infrequently, disambiguation fails.
- Ambiguity is the primary difference between natural and computer languages.

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.

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

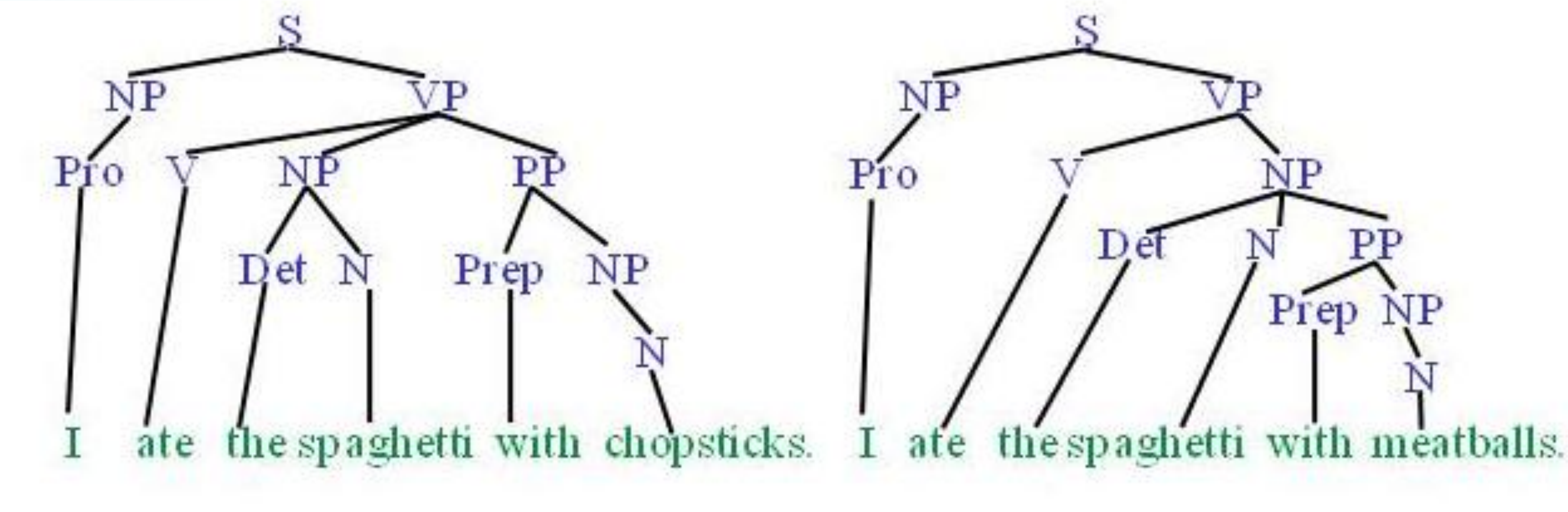
- Useful for subsequent syntactic parsing and word sense disambiguation.

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.

Semantic Parsing

- A *semantic parser* maps a natural-language sentence to a complete, detailed semantic representation (*logical form*).

- Example: Microsoft purchases Powerset.

- BUY(Microsoft, Powerset)

Textual Entailment (Natural Language Inference)

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

TEXT	HYPOTHESIS	ENTAIL MENT
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

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)

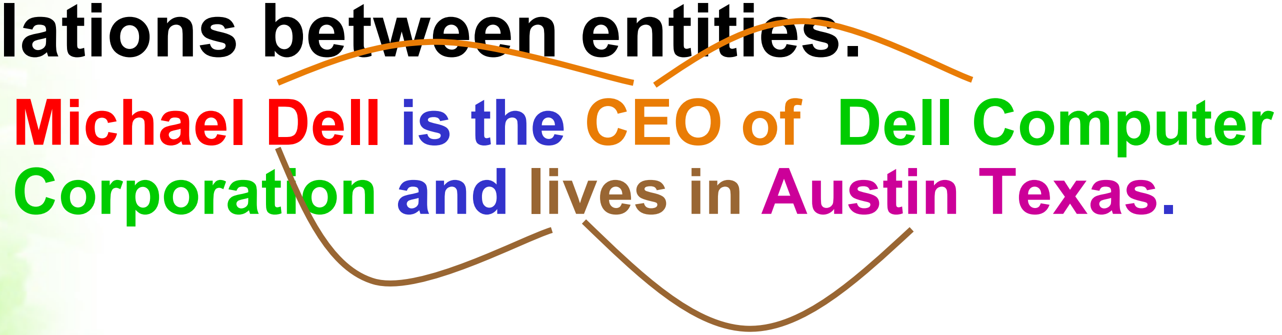
Other Tasks

Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is a 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.



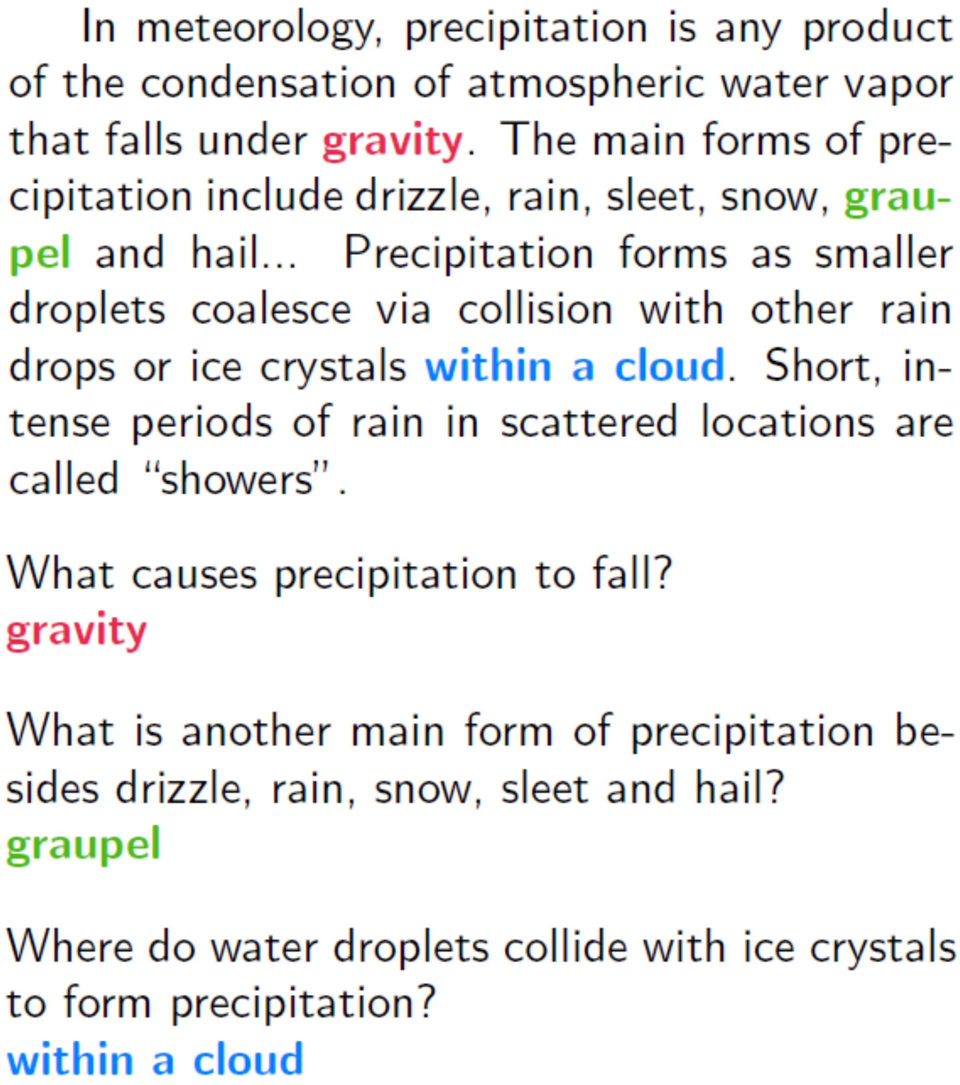
  – Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

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?**
    - **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 (Spanish)

    Until we see each other again, baby.

Ambiguity Resolution is Required for Translation

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
  - “John plays the guitar.” → “John toca la guitarra.”
  - “John plays soccer.” → “John juega el fútbol.”
- An apocryphal story is that an early MT system gave the following results when translating from English to Russian and then back to English:
  - “The spirit is willing but the flesh is weak.” → “The liquor is good but the meat is spoiled.”
  - “Out of sight, out of mind.” → “Invisible idiot.”

Resolving Ambiguity

- Choosing the correct interpretation of linguistic utterances requires knowledge of:
  - Syntax
    - An agent is typically the subject of the verb
  - Semantics
    - Michael and Ellen are names of people
    - Austin is the name of a city (and of a person)
    - Toyota is a car company and Prius is a brand of car
  - Pragmatics
  - World knowledge
    - Credit cards require users to pay financial interest
    - Agents must be animate and a hammer is not animate

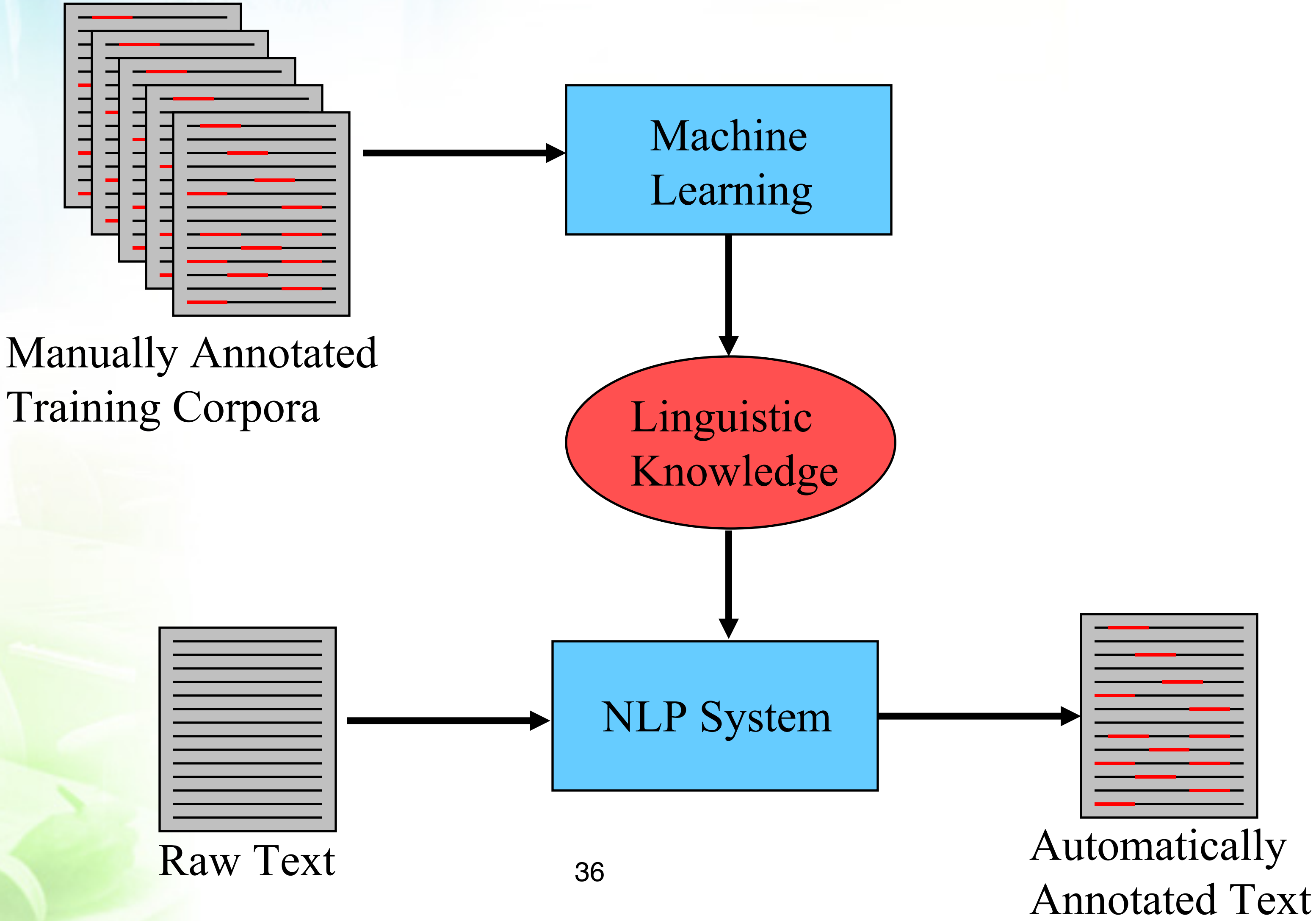
Manual Knowledge Acquisition

- Traditional, “rationalist,” approaches to language processing require human specialists to specify and formalize the required knowledge.
- Manual knowledge engineering, is difficult, time-consuming, and error prone.
- “Rules” in language have numerous exceptions and irregularities.
  - “All grammars leak.”: Edward Sapir (1921)
- Manually developed systems were expensive to develop and their abilities were limited and “brittle” (not robust).

Automatic Learning Approach

- **Use machine learning methods to automatically acquire the required knowledge from appropriately annotated text corpora.**
- **Variously referred to as the "corpus based," "statistical," or "empirical" approach.**
- **Statistical learning methods were first applied to speech recognition in the late 1970's and became the dominant approach in the 1980's.**
- **During the 1990's, the statistical training approach expanded and came to dominate almost all areas of NLP.**

Learning Approach



Manually Annotated Training Corpora

Machine Learning

Linguistic Knowledge

Raw Text

NLP System

Automatically Annotated Text

Advantages of the Learning Approach

- Large amounts of electronic text are now available.
- Annotating corpora is easier and requires less expertise than manual knowledge engineering.
- Learning algorithms have progressed to be able to handle large amounts of data and produce accurate probabilistic knowledge.
- The probabilistic knowledge acquired allows robust processing that handles linguistic regularities as well as exceptions.

The Importance of Probability

- Some combinations of words are more likely than others:
  - “vice president Gore” vs. “dice precedent core”
- Statistical methods allow computing the most likely interpretation by combining probabilistic evidence from a variety of uncertain knowledge sources.

Early History: 1950's

- Shannon (the father of information theory) explored probabilistic models of natural language (1951).
- Chomsky (the extremely influential linguist) developed formal models of syntax, i.e. finite state and context-free grammars (1956).
- First computational parser developed at U Penn as a cascade of finite-state transducers (Joshi, 1961; Harris, 1962).
- Bayesian methods developed for *optical character recognition* (OCR) (Bledsoe & Browning, 1959).

History: 1960's

- Work at MIT AI lab on question answering (BASEBALL) and dialog (ELIZA).
- Semantic network models of language for question answering (Simmons, 1965).
- First electronic corpus collected, Brown corpus, 1 million words (Kucera and Francis, 1967).
- Bayesian methods used to identify document authorship (*The Federalist* papers) (Mosteller & Wallace, 1964).

History: 1970's

- **"Natural language understanding" systems developed that tried to support deeper semantic interpretation.**
  - Schank *et al*. (1972, 1977) developed systems for conceptual representation of language and for understanding short stories using hand-coded knowledge of scripts, plans, and goals.
- **Prolog programming language developed to support logic-based parsing (Colmeraurer, 1975).**
- **Initial development of hidden Markov models (HMMs) for statistical speech recognition (Baker, 1975; Jelinek, 1976).**

History: 1980's

- Development of more complex (mildly context sensitive) grammatical formalisms, e.g. unification grammar, HPSG, tree-adjoning grammar.
- Symbolic work on discourse processing and NL generation.
- Initial use of statistical (HMM) methods for syntactic analysis (POS tagging) (Church, 1988).

History: 1990’s

- Rise of statistical methods and empirical evaluation causes a “scientific revolution” in the field.
- Initial annotated corpora developed for training and testing systems for POS tagging, parsing, WSD, information extraction, MT, etc.
- First statistical machine translation systems developed at IBM for Canadian Hansards corpus (Brown *et al*., 1990).
- First robust statistical parsers developed (Magerman, 1995; Collins, 1996; Charniak, 1997).
- First systems for robust information extraction developed (e.g. MUC competitions).

History: 2000's

- Increased use of a variety of ML methods, SVMs, logistic regression (i.e. max-ent), CRF's, etc.
- Continued developed of corpora and competitions on shared data.
  - TREC Q/A
  - SENSEVAL/SEMEVAL
  - CONLL Shared Tasks (NER, SRL…)
- Increased emphasis on unsupervised, and semi-supervised, as alternatives to purely supervised learning.
- Shifting focus to semantic tasks such as WSD, SRL, and semantic parsing.

History: 2010's

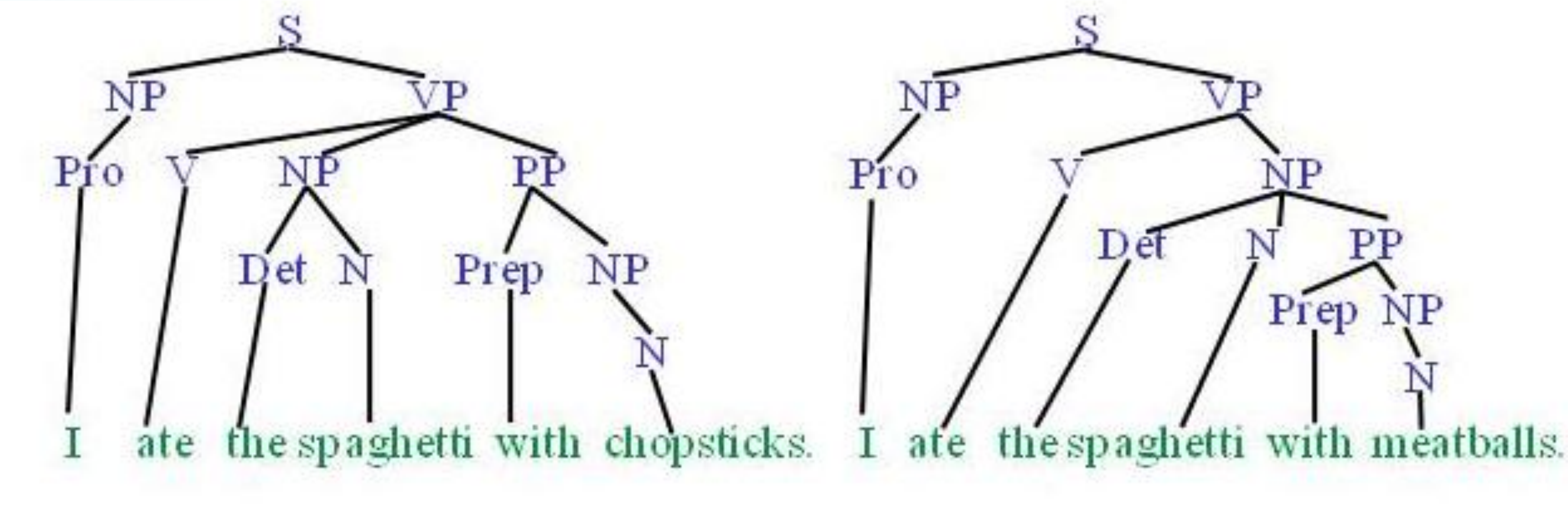
- Grounded Language: Connecting language to perception and action.
  - Image and video description
  - Visual question answering (VQA)
  - Human-Robot Interaction (HRI) in NL
- Deep Learning: Neural network learning with many layers or recurrence.
  - Long Short Term Memory (LSTM) recurrent neural networks using encoder/decoder sequence-to-sequence mapping.
  - Neural Machine Translation (NMT)
  - Spreading to syntactic/semantic parsing and most other NLP tasks.

NLP Tools

- Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP/)
  - Parser, POS-Tagger, NER, Co-referencncr resolution, Word Segmenter, Information Extraction, Relation Extractor
- NLTK (https://www.nltk.org)
  - The most popular Python NLP Library
- AllenNLP (https://demo.allennlp.org/reading-comprehension)
  - NER, Open IE, Sentiment Analysis, Dependency parsing, Constituency parsing, SRL, Co-reference resolution, Semantic parser, Text To SQL, Textual Entailment, Language Modeling

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.

Question : Syntactic parsing

- Create passible syntactic trees for the sentence:
  - I saw the man on the hill with a telescope.

- How many syntactic ambiguities exist in the sentence?