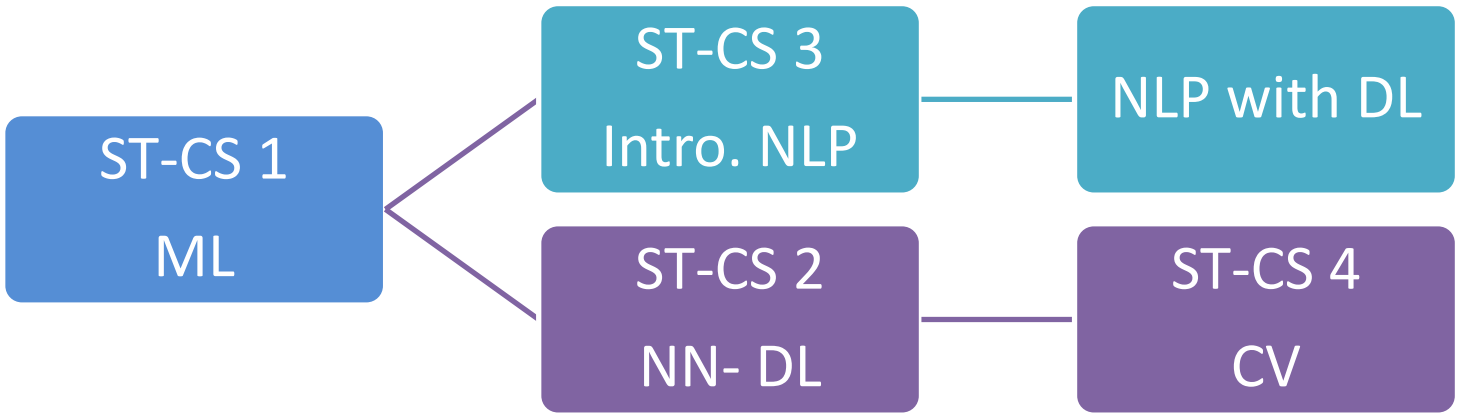
Selected Topics in CS 3

Intro to Natural Language Processing (NLP)

Assoc. Prof. Ensaf Hussein

- Selected Topics Plan
- What we will Learn in this Course?
- Course Prerequisites
- Resources
- Grading Policy
- Tentative Schedule
- What is NLP?
- NLP Applications

Selected Topics Plan



In this Course we will learn to:

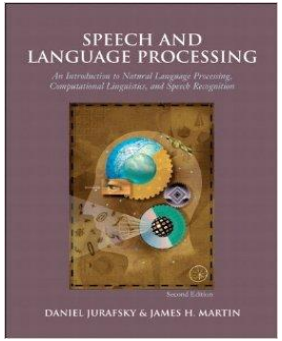
- Learn NLP Preprocessing tasks such as Tokenization, Normalization, Stemming, Lemmatization.
- Perform sentiment analysis of tweets using logistic regression and then naïve Bayes,
- Create a simple auto-correct algorithm using minimum edit distance and dynamic programming,
- Apply the Viterbi Algorithm for part-of-speech (POS) tagging, which is important for computational linguistics,
- Write a better auto-complete algorithm using an N-gram language model.
- Topic Modeling and Classification

Prerequisites

No course prerequisites, but I will assume:

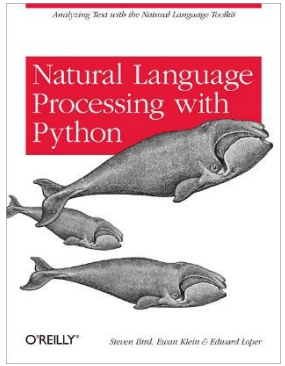
- some programming experience (Python language required)
- familiarity with basics of calculus, linear algebra, and conditional probability
- It will be very helpful to have taken a machine learning course.

Resources:



Books:

- Jurafsky & Martin. *Speech and Language Processing*, 2nd Ed. & 3rd Ed.
- *Natural Language Processing with Python, Bird, Klein & Loper, O'Reilly.*



Course home page:

https://teams.microsoft.com/l/team/19%3agEcXb-rG50datrcAKVeipYBE-uC7EEf16395YKY7AMM1%40thread.tacv2/conversations?groupId=d11742c7-a468-4a4c-97dc-637f77908b53&tenantId=aadc0e0a-65ee-471a-99a1-9f86faecbaed

Code: **zlii10w**

Additional readings:

- *www.nltk.org*

Grading policy

Mid-term Exam	15%
Individual coding assignments	10%
Final group project	15%
Final-Exam (Written Exam)	60 %

Tentative Schedule

- Regular Expression (regex) Basics- NLP Preprocessing Tasks (Week 2)
- Deriving features from text – Similarity Measures - MED (Week 3)
- Machine Learning with Text (Text Classification) (Week 4-5)
- Language Models (Weeks 6)
- Hidden Markov Models-Viterbi Algorithm (Week 8)
- Part-of-Speech Tagging (Week 9)
- Syntax Parsing (PCFG – CKY) (Week 10)
- Topic Modeling (Week 11)

What is Natural Language Processing (NLP)?

By “natural language” we mean a language that is used for everyday communication by humans.

NLP is an Intersection of several fields

- Computer Science
- Artificial Intelligence
- Linguistics

It is basically teaching computers to process human language

Two main components:

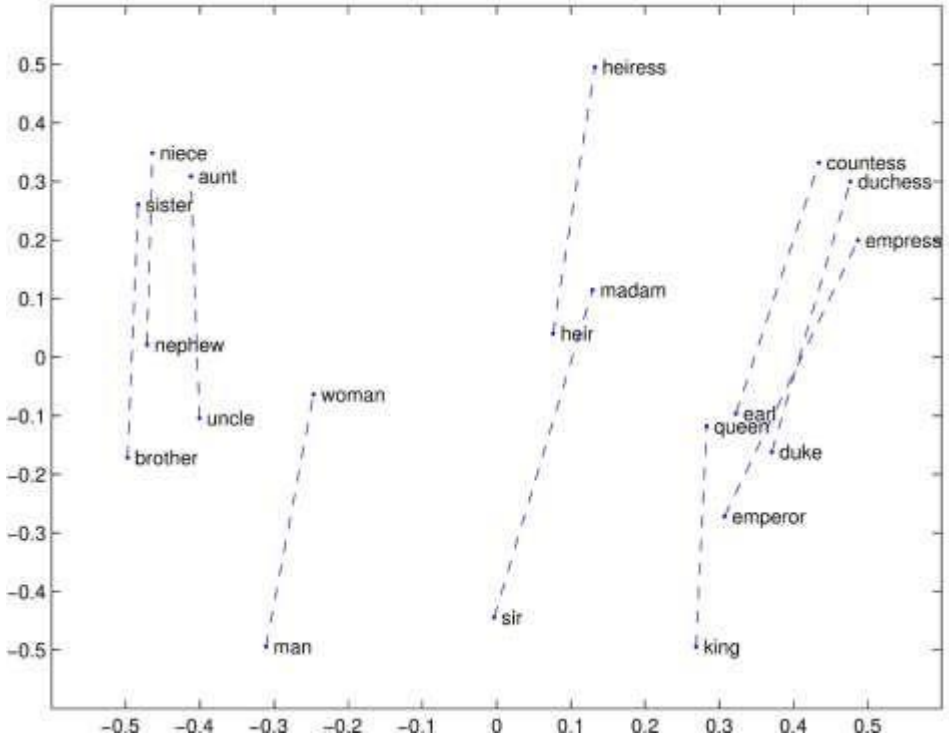
- Natural Language Understanding (NLU)
- Natural Language Generation (NLG)

NLP is AI Complete

- Requires all types of knowledge humans possess → It’s hard!

Natural Language Understanding (NLU)

- Deriving meaning from natural language
- Imagine a Concept (aka Semantic or Representation) space
  - In it, any idea/word/concept has unique computer representation
  - Usually via a vector space
  - NLU → Mapping language into this space

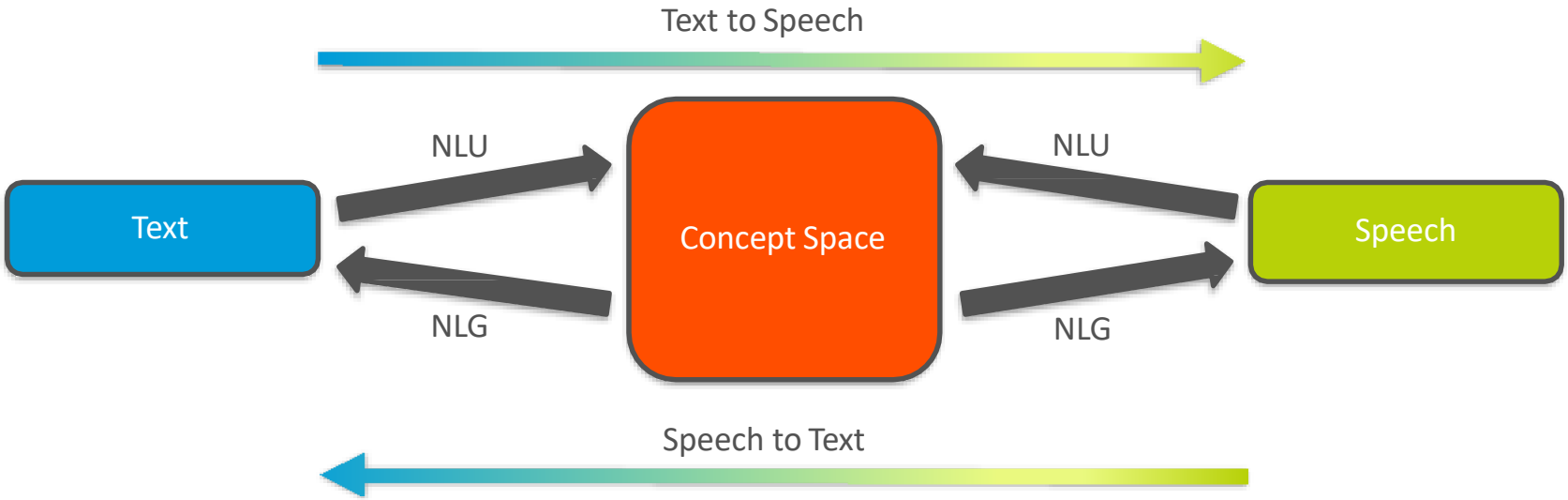


Natural Language Generation (NLG)

- Mapping from computer representation space to language space
- Opposite direction of NLU
  - Usually need NLU to perform NLG!
- NLG is really hard!

NLP: Speech vs Text

- Natural Language can refer to Text or Speech
- Goal of both is the same: translate raw data (text or speech) into underlying concepts (NLU) then possibly into the other form (NLG)



History of NLP

- NLP has been through (at least) 3 major eras:
  - 1950s-1980s: Linguistics Methods and Handwritten Rules
  - 1980s-Now: Corpus/Statistical Methods
  - Now-???: Deep Learning
- Lucky you! You're right near the start of a paradigm shift!

1950s - 1980s: Linguistics/Rule Systems

- NLP systems focus on:
  - Linguistics: Grammar rules, sentence structure parsing, etc
  - Handwritten Rules: Huge sets of logical (if/else) statements
  - Ontologies: Manually created (domain-specific!) knowledge bases to augment rules above
- Problems:
  - Too complex to maintain
  - Can't scale!
  - Can't generalize!

Eliza: 1966

User: *You are like my father in some ways.*
ELIZA: *WHAT RESEMBLANCE DO YOU SEE*
User: *You are not very aggressive but I think you don't want me to notice that.*
ELIZA: *WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE*
User: *You don't argue with me.*
ELIZA: *WHY DO YOU THINK I DON'T ARGUE WITH YOU*
User: *You are afraid of me.*
ELIZA: *DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU*

Weizenbaum (1966)

ELIZA is a simple pattern-based that uses pattern matching to recognize phrases like "You are X" and translate them into suitable outputs like "What makes you think I am X?". ELIZA doesn't actually need to know anything to mimic a Rogerian psychotherapist. modern conversational agents are much more than a diversion; they can answer questions, book flights, or find restaurants, functions for which they rely on a much more sophisticated understanding of the user's intent.

1980s - Now: Corpus/Statistical Methods

- NLP starts using Machine Learning methods
- Use statistical learning over huge datasets of unstructured text
  - Corpus: Collection of text documents
  - e.g. Supervised Learning: Machine Translation
  - e.g. Unsupervised Learning: Deriving Word "Meanings" (vectors)

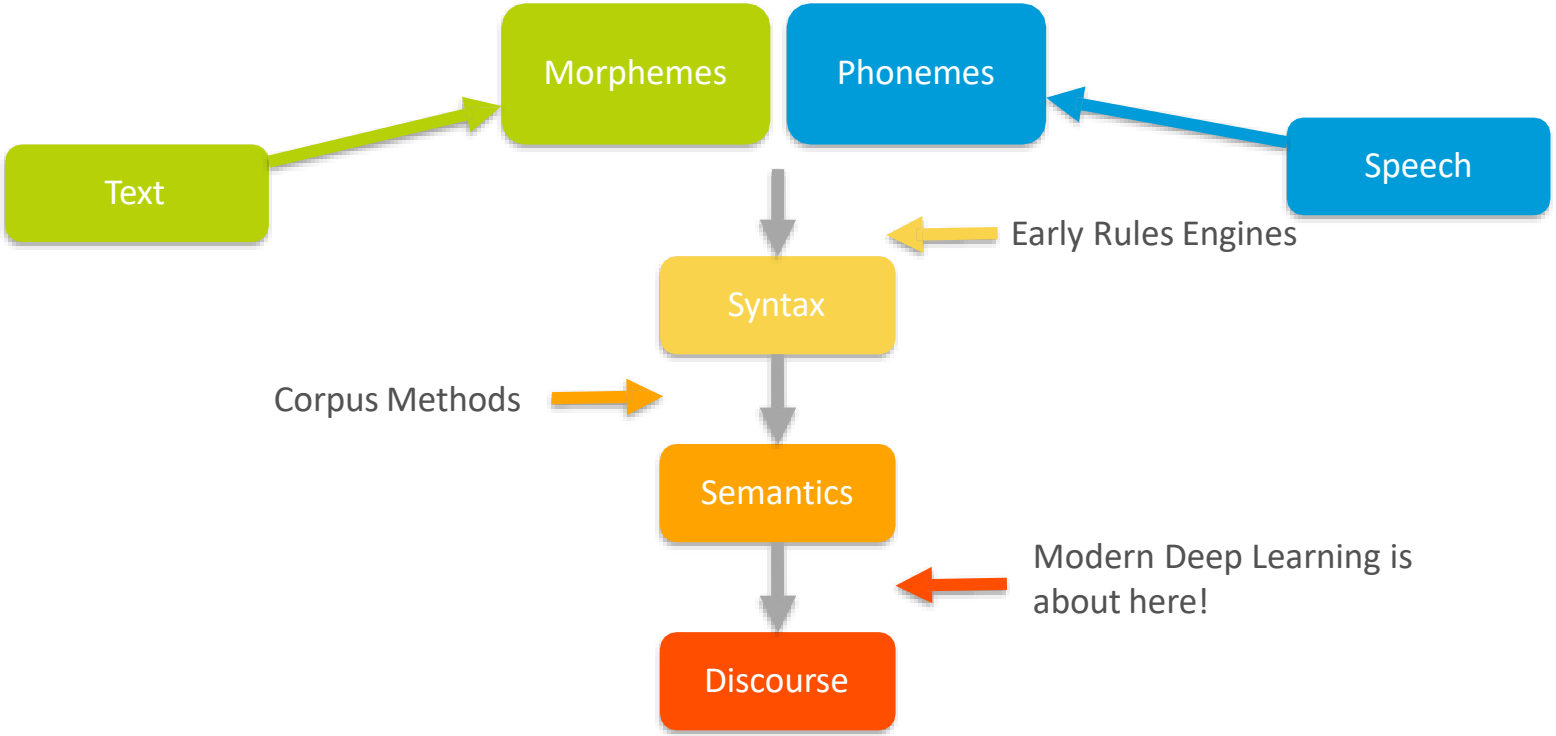
Now - ???: Deep Learning

- Deep Learning made its name with Images first
- 2012: Deep Learning has major NLP breakthroughs
  - Researchers use a neural network to win the Large Scale Visual Recognition Challenge (LSVRC)
  - This state of the art approach beat other ML approaches with half their error rate (26% vs 16%)
- Very useful for unified processing of Language + Images

NLP Definitions

- Phonemes: the smallest *sound* units in a language
- Morphemes: the smallest units of *meaning* in a language
- Syntax: how words and sentences are constructed from these two building blocks
- Semantics: the *meaning* of those words and sentences
- Discourse: semantics *in context*. Conversation, persuasive writing, etc.

Levels of NLP



NLU Applications

- ML on Text (Classification, Regression, Clustering)
- Document Recommendation
- Language Identification
- Natural Language Search
- Sentiment Analysis
- Text Summarization
- Extracting Word/Document Meaning (vectors)
- Relationship Extraction
- Topic Modeling
- …and more!

NLU Application: Document Classification

- Classify "documents" - discrete collections of text - into categories
  - Example: classify emails as spam vs. not spam
  - Example: classify movie reviews as positive vs. negative
  - Example: classify legal documents as relevant vs. not relevant to a topic

NLU Application: Document Recommendation

- Choosing the most relevant document based on some information:
  - Example: show most relevant webpages based on query to search engine
  - Example: recommend news articles based on past articles liked
  - Example: recommend restaurants based on Yelp reviews

NLU Application: Topic Modeling

- Breaking a set of documents into topics at the word level
  - Example: see how prevalence of certain topics covered in a magazine changes over time
  - Example: find documents belonging to a certain topic

NLG Applications

- Image Captioning
- (Better) Text Summarization
- Machine Translation
- Question Answering/Chatbots
- …so much more
- Notice NLU is almost a prerequisite for NLG

Image Captioning

- Automatically generate captions for images

Captions automatically generated.

Source: https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

Machine Translation

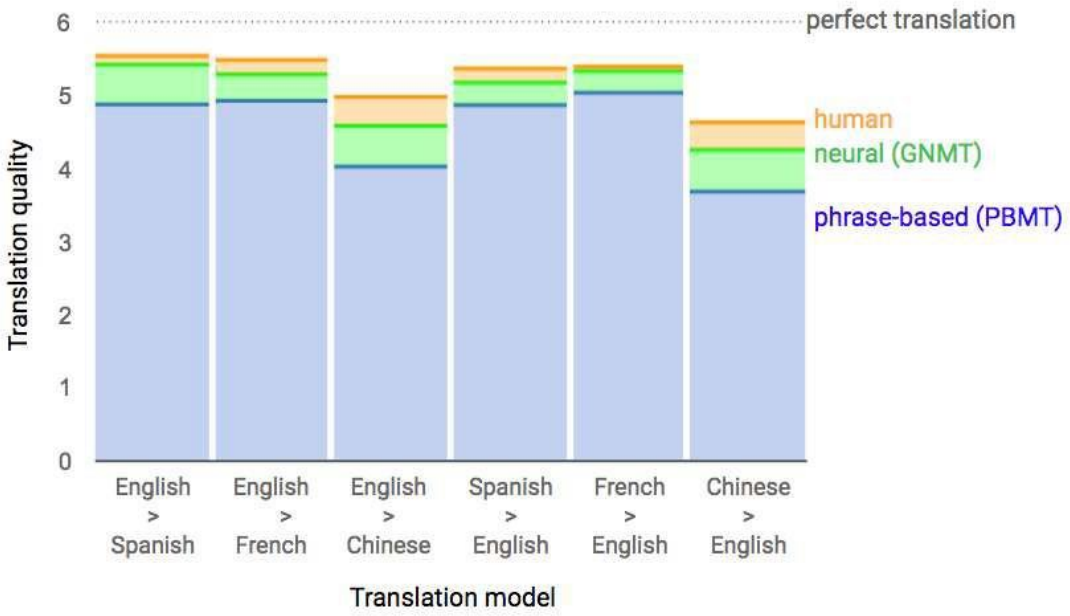
- Automatically translate text between language

Input sentence:	*Translation (PBMT):*	*Translation (GNMT):*	*Translation (human):*
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

Example from Google®'s machine translation system (2016)

Source: https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Machine Translation



Source: https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Text Summarization

- Automatically generate text summaries of documents
  - Example: generate headlines of news articles

Input: Article 1st sentence	Model-written headline
metro-goldwyn-mayer reported a third-quarter net loss of dlrs 16 million due mainly to the effect of accounting rules adopted this year	mgm reports 16 million net loss on higher revenue
starting from july 1, the island province of hainan in southern china will implement strict market access control on all incoming livestock and animal products to prevent the possible spread of epidemic diseases	hainan to curb spread of diseases
australian wine exports hit a record 52.1 million liters worth 260 million dollars (143 million us) in september, the government statistics office reported on monday	australian wine exports hit record high in september

Source: https://ai.googleblog.com/2016/08/text-summarization-with-tensorflow.html

Question Answering

Question Answering

Dialog Systems

figure credit: Phani Marupaka

Part-of-Speech-Tagging

Some questioned if Tim Cook ’s first product

would be a breakaway hit for Apple .

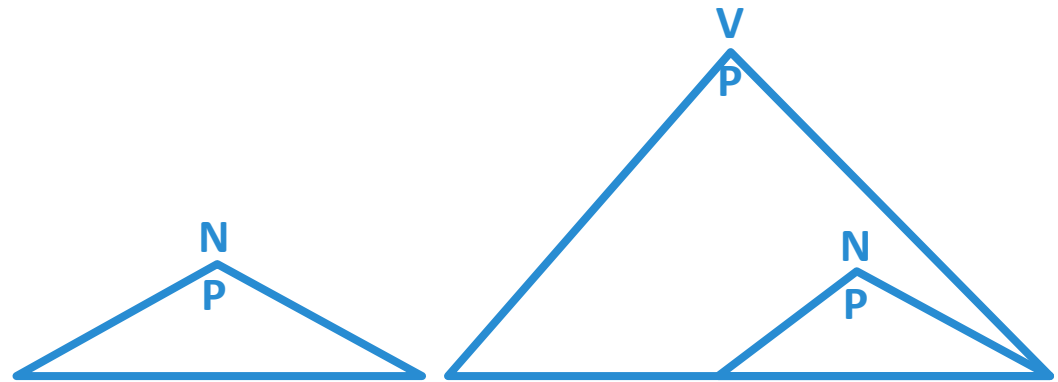
Part-of-Speech-Tagging

determiner	verb (past)	prep.	proper noun	proper noun	poss.	adj.	noun
Some	questioned	if	Tim	Cook	’s	first	product

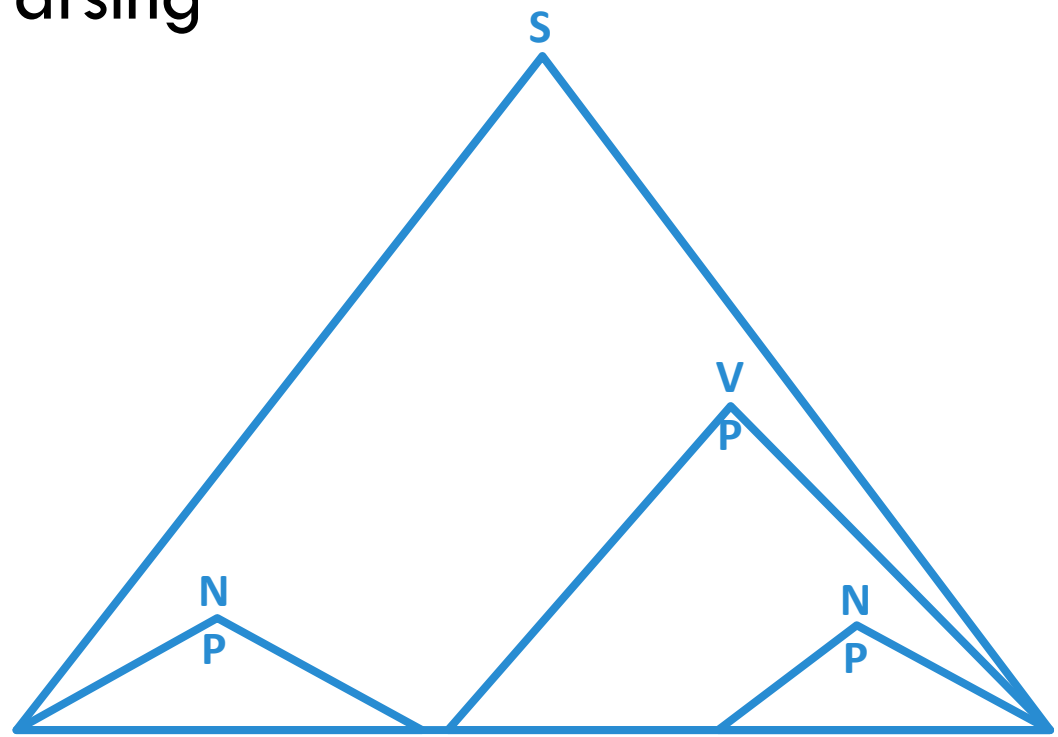
modal	verb	det.	adjective	noun	prep.	proper noun	punc.
would	be	a	breakaway	hit	for	Apple	.

Syntactic Parsing

Syntactic Parsing



Syntactic Parsing



Named Entity Recognition

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

PERSON **ORGANIZATION**

Entity Linking

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

Tim Cook

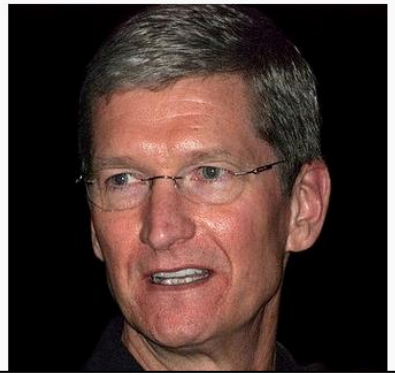
From Wikipedia, the free encyclopedia

For other people named Tim Cook, see Tim Cook (disambiguation).

Timothy Donald Cook (born November 1, 1960) is an American business executive, industrial engineer, and developer. Cook is the Chief Executive Officer of Apple Inc., previously serving as the company's Chief Operating Officer, under its founder Steve Jobs.[4]

Cook joined Apple in March 1998

Tim Cook



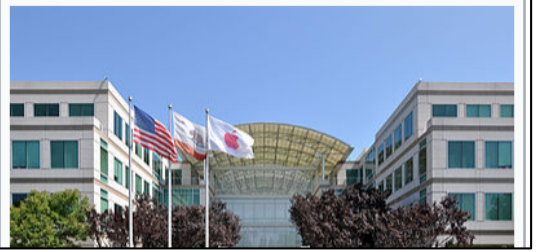
Apple Inc.

From Wikipedia, the free encyclopedia

Coordinates: 37.33182°N 122.03118°W

Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, and online services. The company's hardware products include the iPhone smartphone, the iPad tablet computer, the Mac personal computer, the iPod portable

Apple Inc.



Coreference Resolution

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

It's the company's first new device since he became CEO.

Coreference Resolution

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

It's the company's first new device since he became CEO.

Coreference Resolution

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

It's the company's first new device since he became CEO.

Coreference Resolution

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

??

It's the company's first new device since he became CEO.

Reading Comprehension

Once there was a boy named Fritz who loved to draw. He drew everything. In the morning, he drew a picture of his cereal with milk. His papa said, “Don’t draw your cereal. Eat it!”
After school, Fritz drew a picture of his bicycle. His uncle said, “Don't draw your bicycle. Ride it!”
…

What did Fritz draw first?

A) the toothpaste
B) his mama
C) cereal and milk
D) his bicycle

MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Reading Comprehension

A Turing machine is a mathematical model of a general computing machine. It is a theoretical device that manipulates symbols contained on a strip of tape. Turing machines are not intended as a practical computing technology, but rather as a thought experiment representing a computing machine—anything from an advanced supercomputer to a mathematician with a pencil and paper. It is believed that if a problem can be solved by an algorithm, there exists a Turing machine that solves the problem. Indeed, this is the statement of the Church–Turing thesis. Furthermore, it is known that everything that can be computed on other models of computation known to us today, such as a RAM machine, Conway's Game of Life, cellular automata or any programming language can be computed on a Turing machine. Since Turing machines are easy to analyze mathematically, and are believed to be as powerful as any other model of computation, the Turing machine is the most commonly used model in complexity theory.

What is the term for a mathematical model that theoretically represents a general computing machine?
Ground Truth Answers: A Turing machine | A Turing machine | Turing machine
Prediction: A Turing machine

It is generally assumed that a Turing machine can solve anything capable of also being solved using what?
Ground Truth Answers: an algorithm | an algorithm | an algorithm
Prediction: RAM machine, Conway's Game of Life, cellular automata or any programming language

What is the most commonplace model utilized in complexity theory?
Ground Truth Answers: the Turing machine | the Turing machine | Turing machine
Prediction: Turing machine

What does a Turing machine handle on a strip of tape?
Ground Truth Answers: symbols | symbols | symbols
Prediction: general computing machine

SQuAD
The Stanford Question Answering Dataset

Sentence Similarity

Input	Output
Other ways are needed. We must find other ways.	4.4
Pakistan bomb victims' families end protest Pakistan bomb victims to be buried after protest ends	2.6
I absolutely do believe there was an iceberg in those waters. I don't believe there was any iceberg at all anywhere near the Titanic.	1.2

Word Prediction

he bent down and searched the large container, trying to find anything else hidden in it other than the _____

Word Prediction

he turned to one of the cops beside him. “search the entire coffin.” the man nodded and bustled forward towards the coffin.

he bent down and searched the large container, trying to find anything else hidden in it other than the _____

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Next Lecture

- Upcoming:
  - Regular Expression (regex) Basics- NLP Preprocessing Tasks