CS 4248
Natural Language Processing

Professor NG Hwee Tou

Department of Computer Science
School of Computing
National University of Singapore
nght@comp.nus.edu.sg

Chapter 1: Introduction

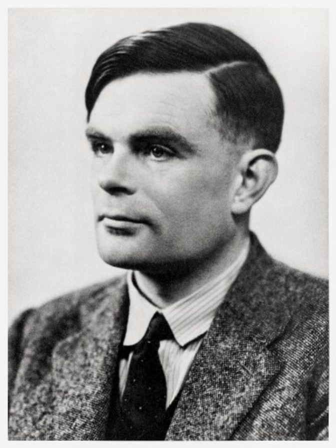
- What is Natural Language Processing (NLP)?
  - The field of designing methods and algorithms that take as input or produce as output human language
- In this course, we focus on processing written text (i.e., excluding speech processing)

Why NLP?

- Theoretical
  - The ability to understand natural language is a hallmark of human intelligence
- Practical
  - Dominance of the Internet, social media
  - Numerous applications
    - Machine translation (Google Translate)
    - Grammar checking (Grammarly)
    - Question answering (Apple Siri)

Language, Thought, & Intelligence

- Conversation
  - Turing Test (1950)



Alan Turing
Image credit: Photograph from Alamy
https://www.newyorker.com/culture/culture-desk/living-in-alan-turings-future

Creating An Intelligent Robot

- Speech recognition
- Natural language understanding
- Natural language generation
- Speech synthesis

Challenges in NLP

- Ambiguous
  - He complained of chest pains.
  - His medical records were packed in a chest for shipping.
- Variable
  - I ate pizza with my friends.
  - My friends and I shared some pizza.
  - “One thought, many expressions”

Properties of Natural Language

- Discrete / symbolic
  - Red vs. pink
- Compositional
  - Letters form words
  - Words form phrases and sentences
  - Sentences form documents
- Sparse
  - Number of possible valid sentences is enormous
  - Very likely to encounter new sentences not seen previously

Knowledge of Language

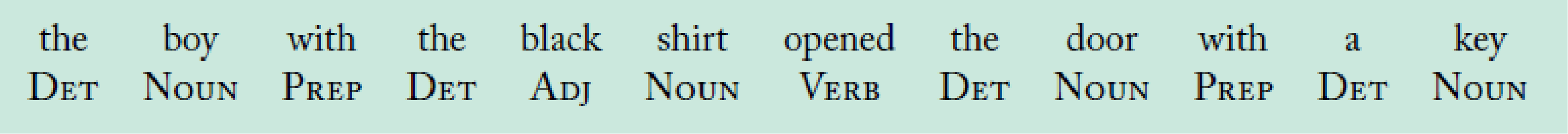
- Phonetics and Phonology
  - Mapping between written words and audio signals
- Morphology
  - Word formation
    - door & doors, goose & geese, eat & ate, …
- Syntax
  - Ordering and grouping of words
  - Sentence structure

Knowledge of Language

- Semantics
  - Meaning of words and sentences
  - Lexical semantics
  - Compositional semantics
- Discourse
  - Multi-sentence processing
    - coreference resolution

NLP Tasks

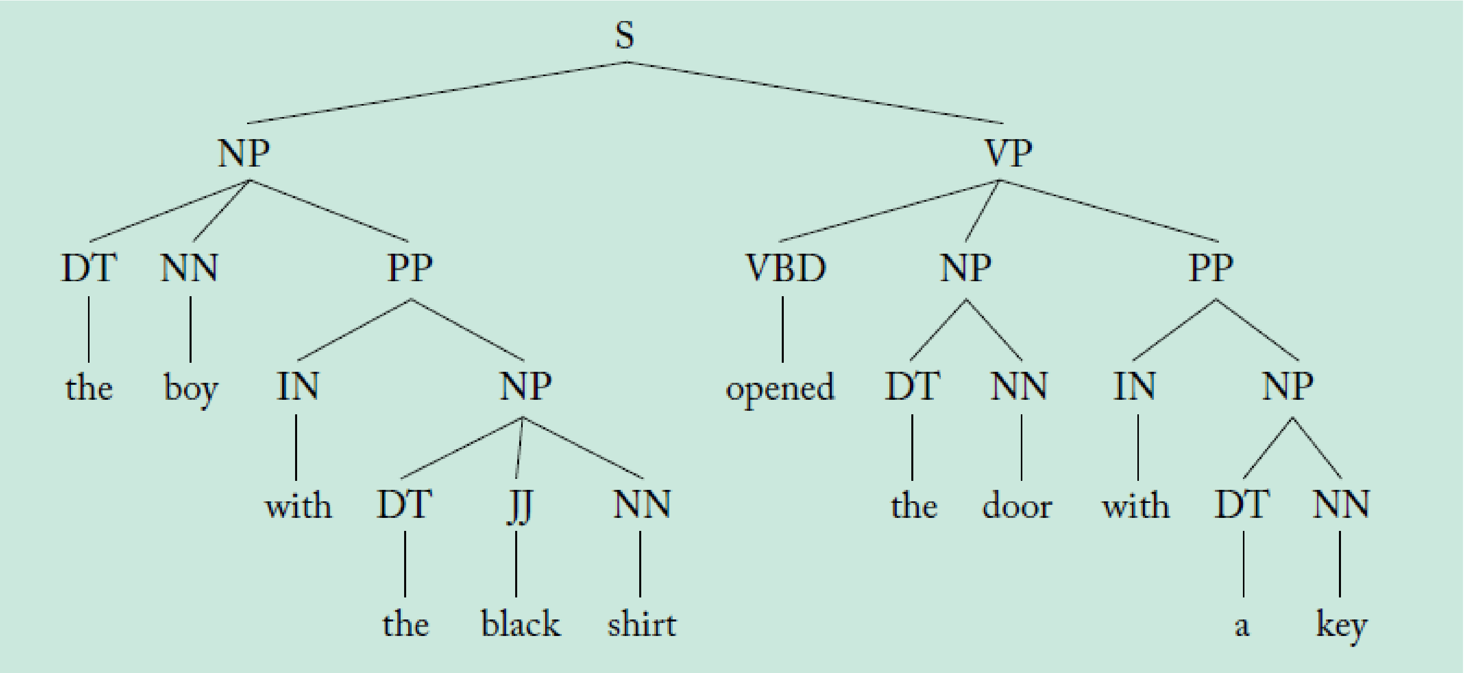
- Part-of-speech tagging



the	boy	with	the	black	shirt	opened	the	door	with	a	key
DET	NOUN	PREP	DET	ADJ	NOUN	VERB	DET	NOUN	PREP	DET	NOUN

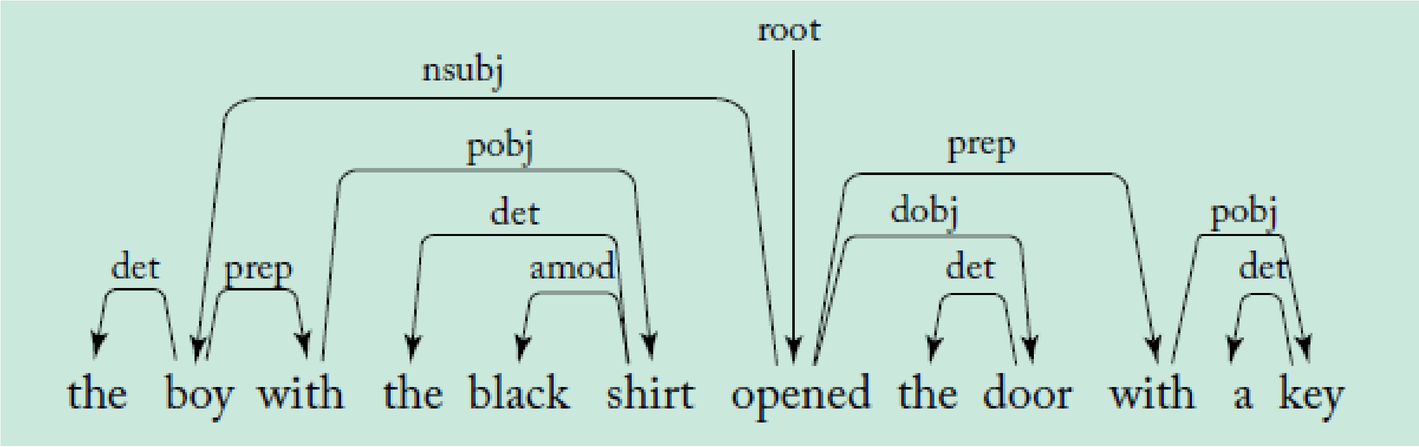
NLP Tasks

- Constituency parsing



NLP Tasks

- Dependency parsing



NLP Tasks

- Word sense disambiguation

The institutions have already consulted the staff concerned through various *channels*, including discussion with the staff representatives.

Sense 1: A path over which electrical signals can pass

Sense 2: A passage for water

Sense 3: A long narrow furrow

Sense 4: A relatively narrow body of water

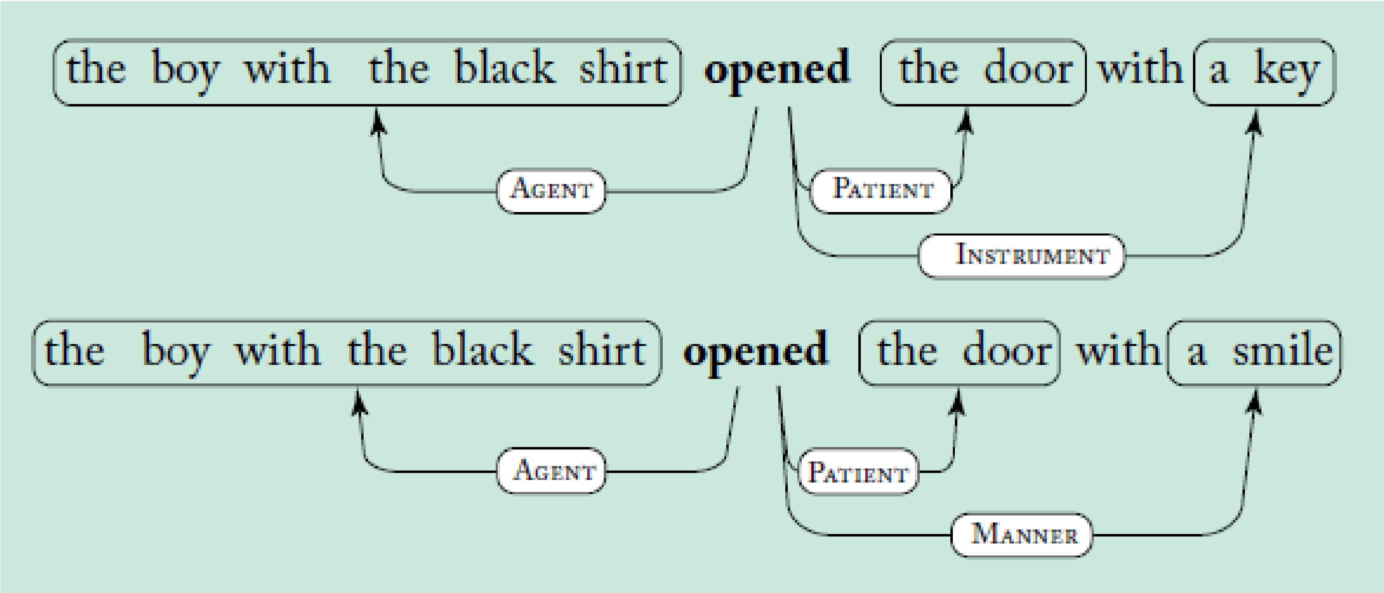
Sense 5: A means of communication or access

Sense 6: A bodily passage or tube

Sense 7: A television station and its programs

NLP Tasks

- Semantic role labeling



NLP Tasks

- Semantic textual similarity (STS)

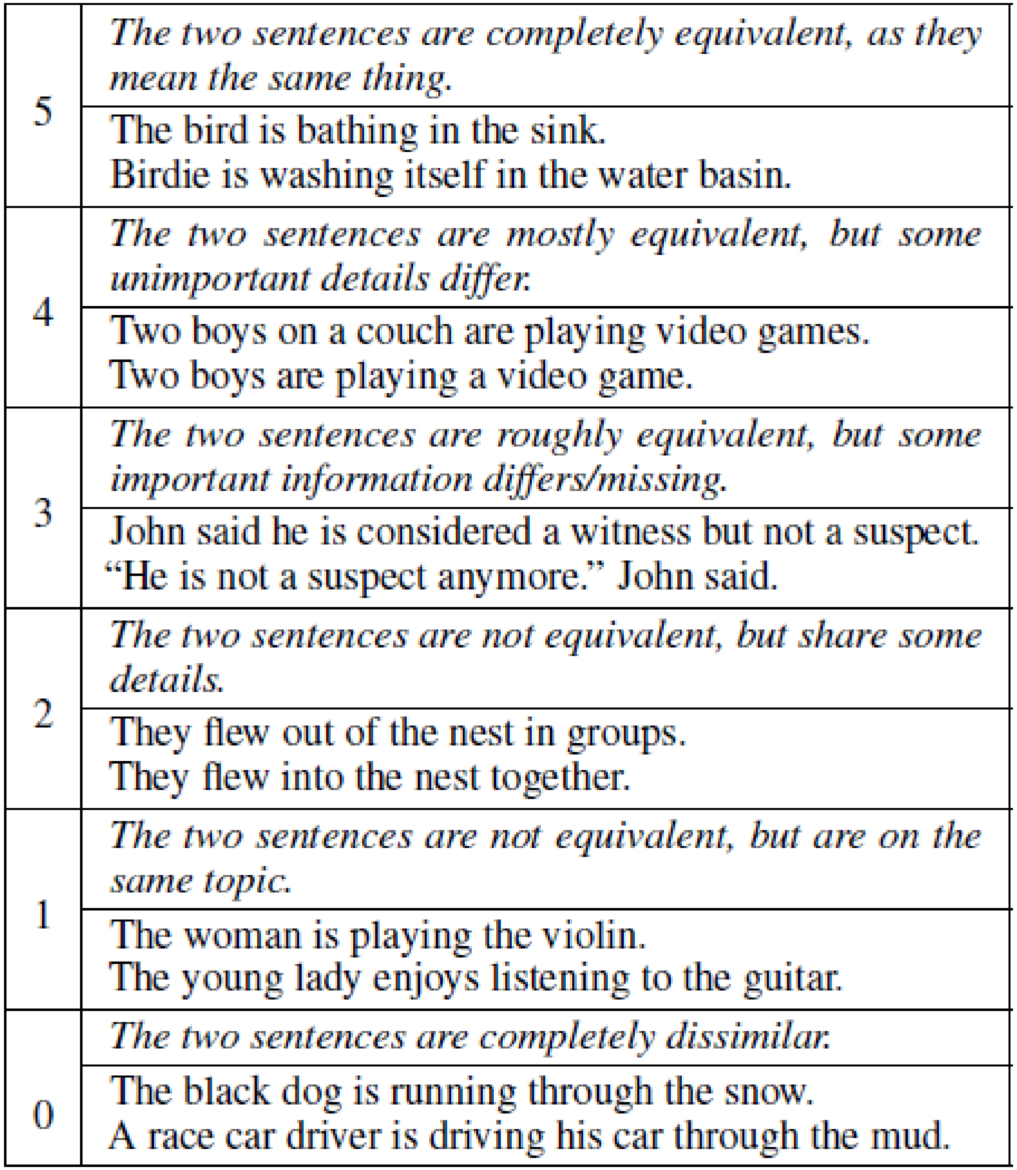
Input:

Two sentences

Output:

A score (0 – 5) indicating the degree of semantic similarity between the two input sentences

Semantic textual similarity (STS)



5	*The two sentences are completely equivalent, as they mean the same thing.*
	The bird is bathing in the sink. Birdie is washing itself in the water basin.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*
	Two boys on a couch are playing video games. Two boys are playing a video game.
3	*The two sentences are roughly equivalent, but some important information differs/missing.*
	John said he is considered a witness but not a suspect. “He is not a suspect anymore.” John said.
2	*The two sentences are not equivalent, but share some details.*
	They flew out of the nest in groups. They flew into the nest together.
1	*The two sentences are not equivalent, but are on the same topic.*
	The woman is playing the violin. The young lady enjoys listening to the guitar.
0	*The two sentences are completely dissimilar.*
	The black dog is running through the snow. A race car driver is driving his car through the mud.

NLP Tasks

- Coreference resolution

Victoria Chen, Chief Financial Officer of Megabucks Banking Corp since 2004, saw her pay jump 20%, to 1.3 million, as the 37-year-old also became the Denver-based financial-services company’s president. It has been ten years since she came to Megabucks from rival Lotsabucks.

NLP Applications

- Text classification
  - Assigning a subject class to a text based on its content
  - Classes are pre-defined beforehand
  - Example:
    - Text → Economy or Military or Sport

Text Classification: Email Spam Filtering

```
From: ``'' <takworlld@hotmail.com>
Subject: real estate is the only way... gem oalvgkay
Anyone can buy real estate with no money down
Stop paying rent TODAY !
There is no need to spend hundreds or even thousands for
similar courses
I am 22 years old and I have already purchased 6 properties
using the
methods outlined in this truly INCREDIBLE ebook.
Change your life NOW !
=================================================
Click Below to order:
http://www.wholesaledaily.com/sales/nmd.htm
=================================================
```

Task: Determine if an email is spam or non-spam

NLP Applications

- Sentiment analysis
  - Analyzes people’s sentiments, opinions, appraisals, attitudes, and emotions toward entities and their attributes
  - Aka Opinion Mining

NLP Applications

- Sentiment analysis
  - Classify a movie review into positive and negative review
    - Positive: “This is the greatest comedy ever filmed.”
    - Negative: “It was pathetic. The worst part about it was the boxing scenes.”

NLP Applications

- Named entity recognition

Input:

John Smith , president of McCormik Industries visited his niece Paris in Milan , reporters say .

Output:

[$_{\text{PER}}$ John Smith] , president of [$_{\text{ORG}}$ McCormik Industries] visited his niece [$_{\text{PER}}$ Paris] in [$_{\text{LOC}}$ Milan] , reporters say .

NLP Applications

- Relation extraction

Input:
John Smith , president of McCormik Industries visited his niece Paris in Milan , reporters say .

Output:
Employee(John Smith, McCormik Industries)
Niece(Paris, John Smith)
Location(Paris, Milan)

NLP Applications

- Event extraction

Input:

… Mr. Murdoch moved to Los Angeles from New York to focus on the filmed entertainment operations that were then under Barry Diller. Mr. Diller was Fox chief executive. …

Output:

Organization:	Fox
Post:	chief executive
Person In:	Murdoch
Person Out:	Barry Diller

NLP Applications

- Question answering
- Example (taken from SQuAD)

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”.

NLP Applications

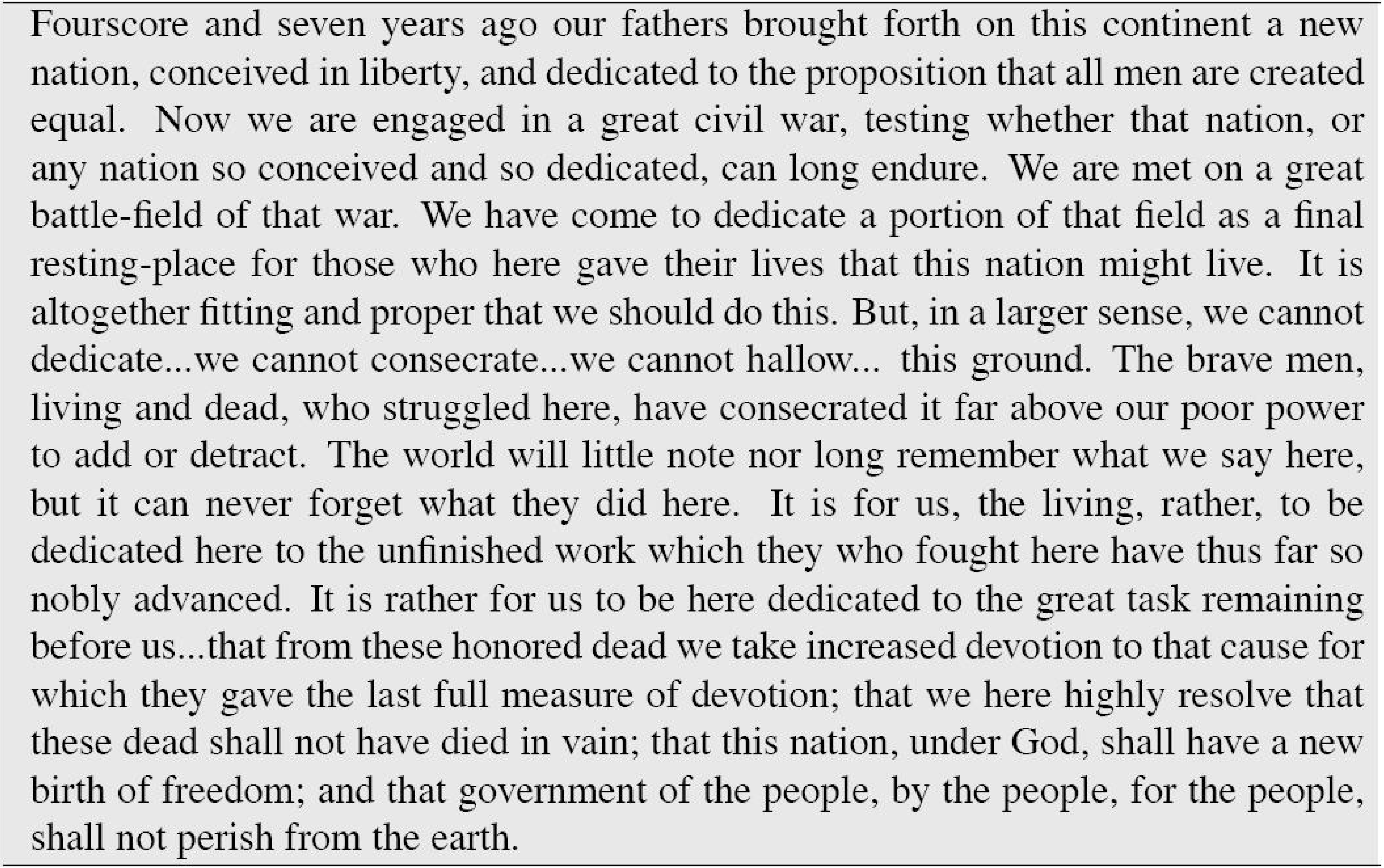
- Question answering

What causes precipitation to fall?
gravity

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

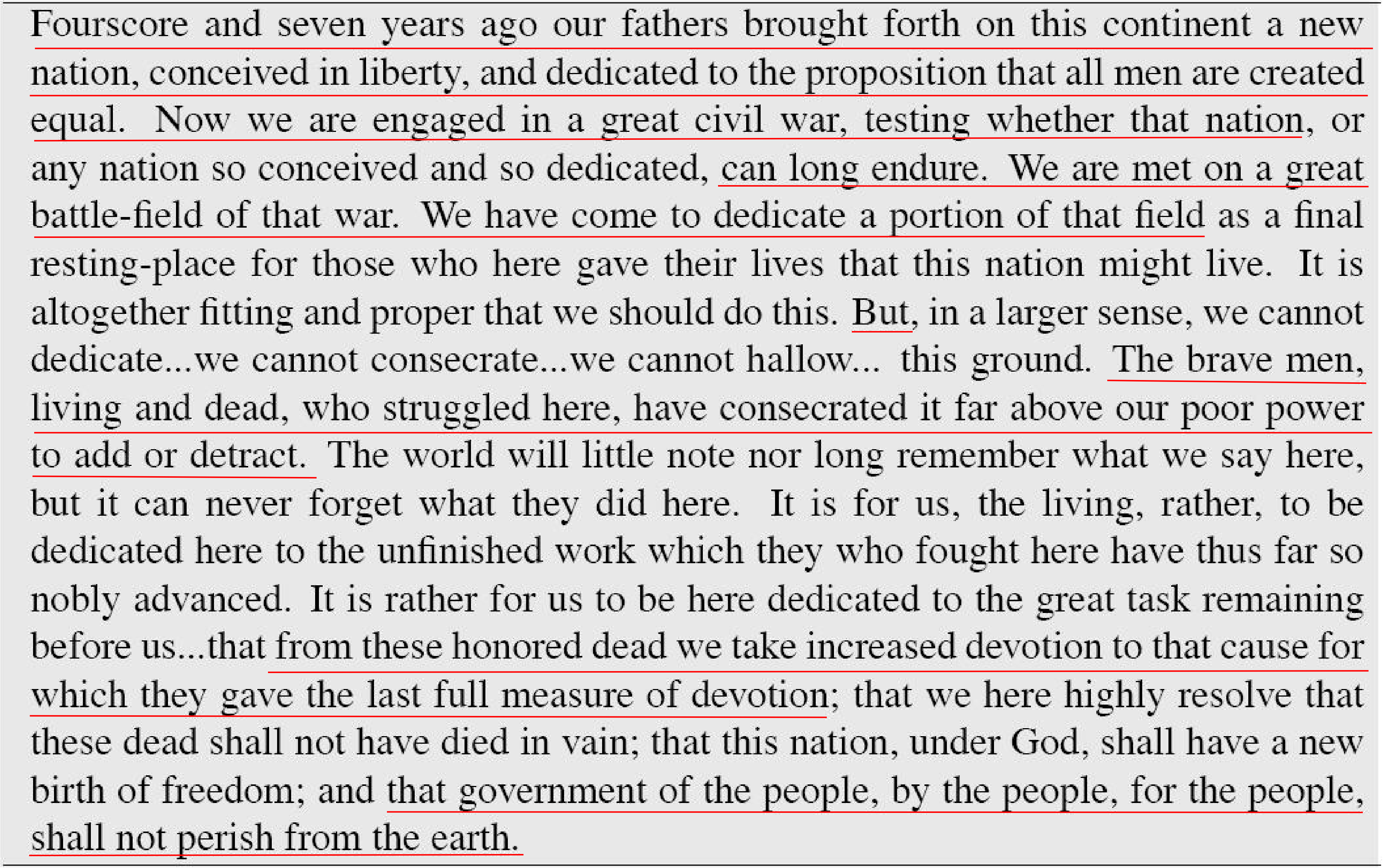
Where do water droplets collide with ice crystals to form precipitation?
within a cloud

NLP Applications: Summarization



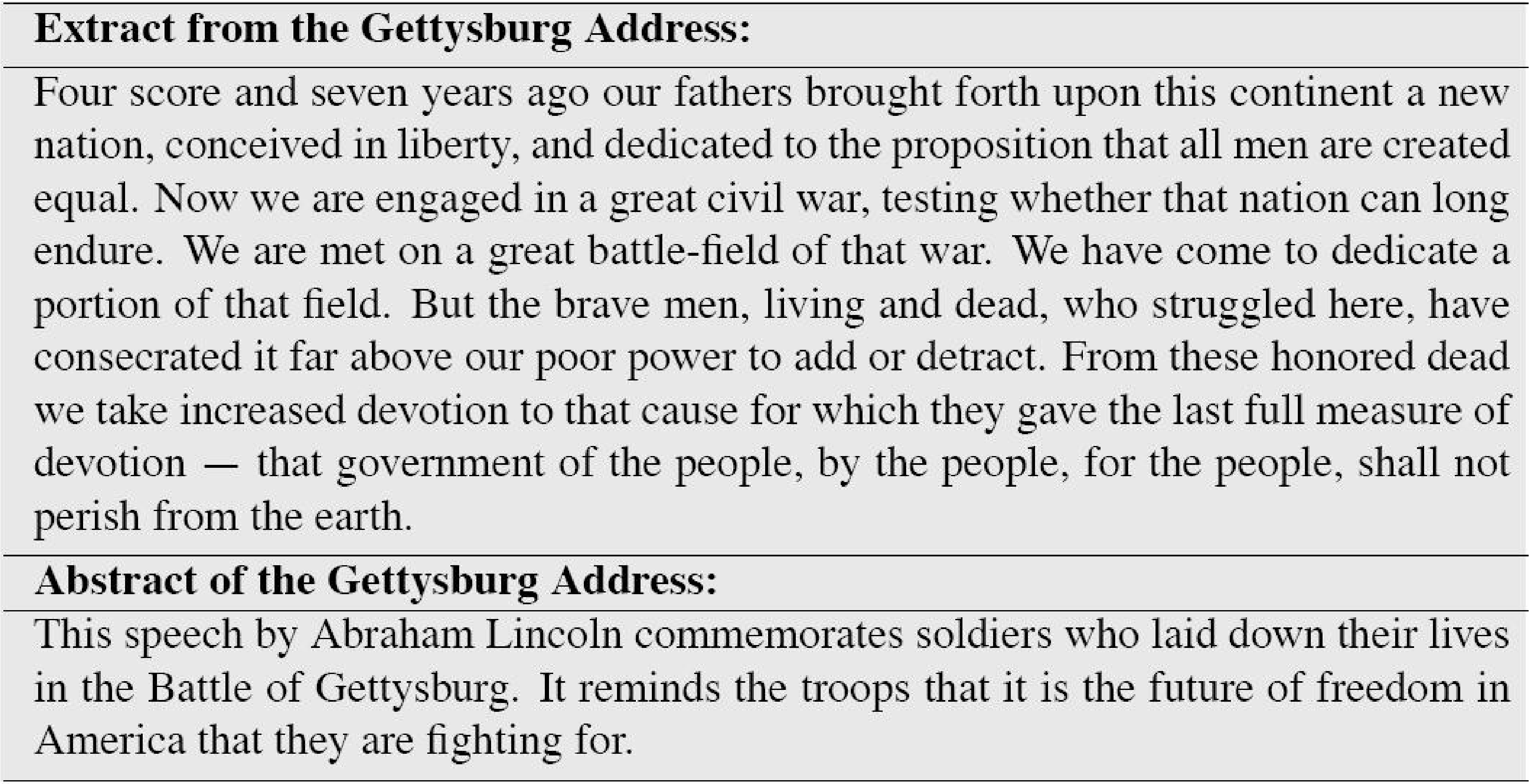
Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field as a final resting-place for those who here gave their lives that this nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we cannot dedicate...we cannot consecrate...we cannot hallow... this ground. The brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. The world will little note nor long remember what we say here, but it can never forget what they did here. It is for us, the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us...that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion; that we here highly resolve that these dead shall not have died in vain; that this nation, under God, shall have a new birth of freedom; and that government of the people, by the people, for the people, shall not perish from the earth.

NLP Applications: Summarization



Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field as a final resting-place for those who here gave their lives that this nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we cannot dedicate...we cannot consecrate...we cannot hallow... this ground. The brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. The world will little note nor long remember what we say here, but it can never forget what they did here. It is for us, the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us...that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion; that we here highly resolve that these dead shall not have died in vain; that this nation, under God, shall have a new birth of freedom; and that government of the people, by the people, for the people, shall not perish from the earth.

Summarization: Extract versus Abstract



Extract from the Gettysburg Address:

Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field. But the brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. From these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion — that government of the people, by the people, for the people, shall not perish from the earth.

Abstract of the Gettysburg Address:

This speech by Abraham Lincoln commemorates soldiers who laid down their lives in the Battle of Gettysburg. It reminds the troops that it is the future of freedom in America that they are fighting for.

NLP Applications

- Machine translation
- E.g., Chinese-to-English translation

Input:
佐科威昨天在北京发表讲话。

Output:
Jokowi made a speech in Beijing yesterday.

NLP Applications

- Grammatical error correction

Input:

The problems bring some effect on engineering design from two aspect, independent innovation and engineering application.

Output:

The problems affect engineering design in two aspects, independent innovation and engineering application.

Ambiguity

- **THE** central problem in NLP
- Multiple alternative interpretations
- Many tasks in NLP can be viewed as resolving ambiguities at various levels

Ambiguity

I made her duck.

- “I cooked waterfowl for her.”
- I made duck for her.
- made: cooked
- her: dative (I made him duck.)
- duck: noun, animal

Ambiguity

I made her duck.

- “I cooked waterfowl belonging to her.”
- made: cooked
- her: possessive (I made his duck.)
- duck: noun, animal

Ambiguity

I made her duck.

- “I created the (plaster?) duck belonging to her.”
- made: created
- her: possessive (I made his duck.)
- duck: noun, artifact/toy

Ambiguity

I made her duck.

- “I caused her to quickly lower her head or body.”
- I caused her to duck.
- made: caused
- her: accusative (I made him duck.)
- duck: verb, lower one’s head or body

Ambiguity

I made her duck.

- “I waved my magic wand and turned her into undifferentiated waterfowl.”
- I turned her into a duck.
- made: turned/transformed
- her: accusative (I made him duck.)
- duck: noun, animal

Resolving Ambiguities

- Lexical disambiguation
  - Part-of-speech (POS) tagging
    - duck: noun or verb
  - Word sense disambiguation (WSD)
    - make: cook or create
- Syntactic disambiguation
  - Parsing
    - (I (made (his duck)))
    - (I (made (him) (duck)))

Machine Learning

- Supervised learning from large annotated corpora (1990 –)
- Learn how to perform part-of-speech tagging, word sense disambiguation, parsing, etc.
- Learning with neural networks (deep learning) (2013 –)

Deep Learning

- Learning with a neural network with many layers (“deep”)
- Pre-train then fine-tune paradigm (2019 –)
  - Build a large pre-trained language model
    - Train on a huge corpus to predict missing (masked) words
  - Fine-tune (train further) on task-specific training data

Ever Larger Language Models

Year	Model	Billions of Parameters
2019	BERT	0.34
2019	GPT-2	1.5
2020	GPT-3	175
2022	PaLM	540