Natural Language Processing

Lecture I. Introduction and Syllabus

Forrest Sheng Bao, Ph.D.

Dept. of Computer Science
Iowa State University
Ames, IA 50011

Aug. 24, 2021

Outline

The Instructor

Natural Language Processing

The Instructor

My research

My PhD dissertation is not on NLP, ML nor CV. My PhD dissertation is about GOFAI.

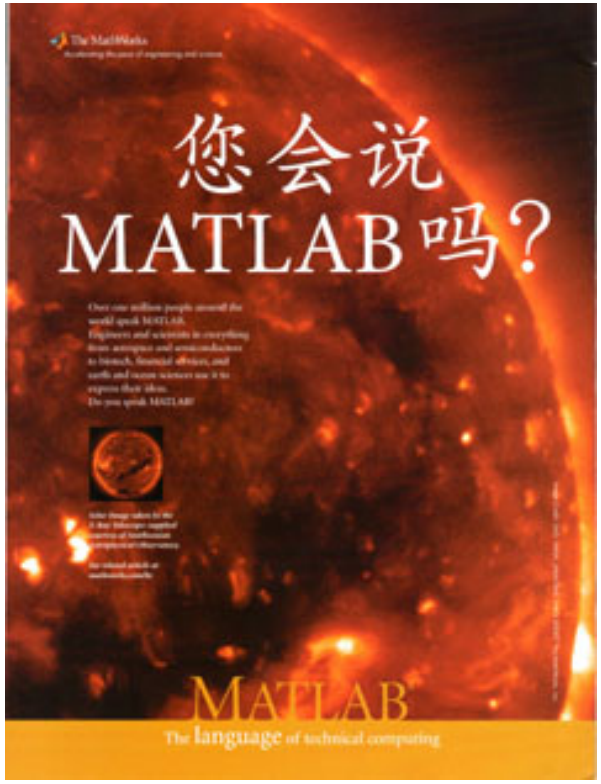
- Artificial Intelligence
  - Knowledge Representation and Reasoning
  - Natural Language Processing
- AI's applications in other sciences
  - Routing and palcements in IC and PCB designs
  - Reinforcement learning for storage systems
  - NLP for biomedical papers

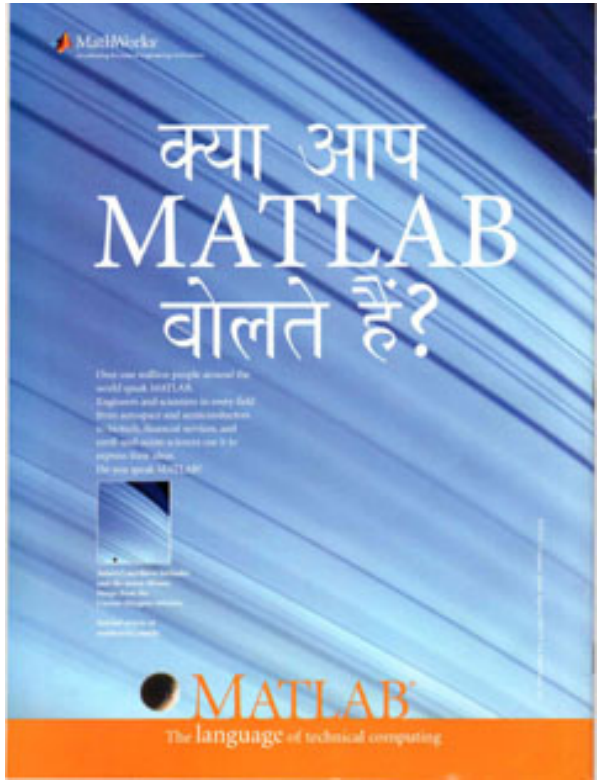
Carl Gauss, Letter to Bolyai (1808)

When I have clarified and exhausted a subject,
then I turn away from it,
in order to go into darkness again.

* I am a tenth generation academic descendant of Carl Gauss.

What (Un)Natural Languages are





- Lots of data, e.g., Amazon reviews
- Against computer programming languages
- Very easy to handle: discretized objects
- Very difficult to handle: unstructured, ambiguity, variety, etc.

Why do we study NLP...instead of other hot areas in AI

- Unlike other animals, we have complicated langauges.
- Languages make us great:

> “a group of people get together and exist as an institution that we call a company so they are able to accomplish something collectively which they could not accomplish separately”– The HP Way.

- When we spoke the same langauge, even God was afraid.

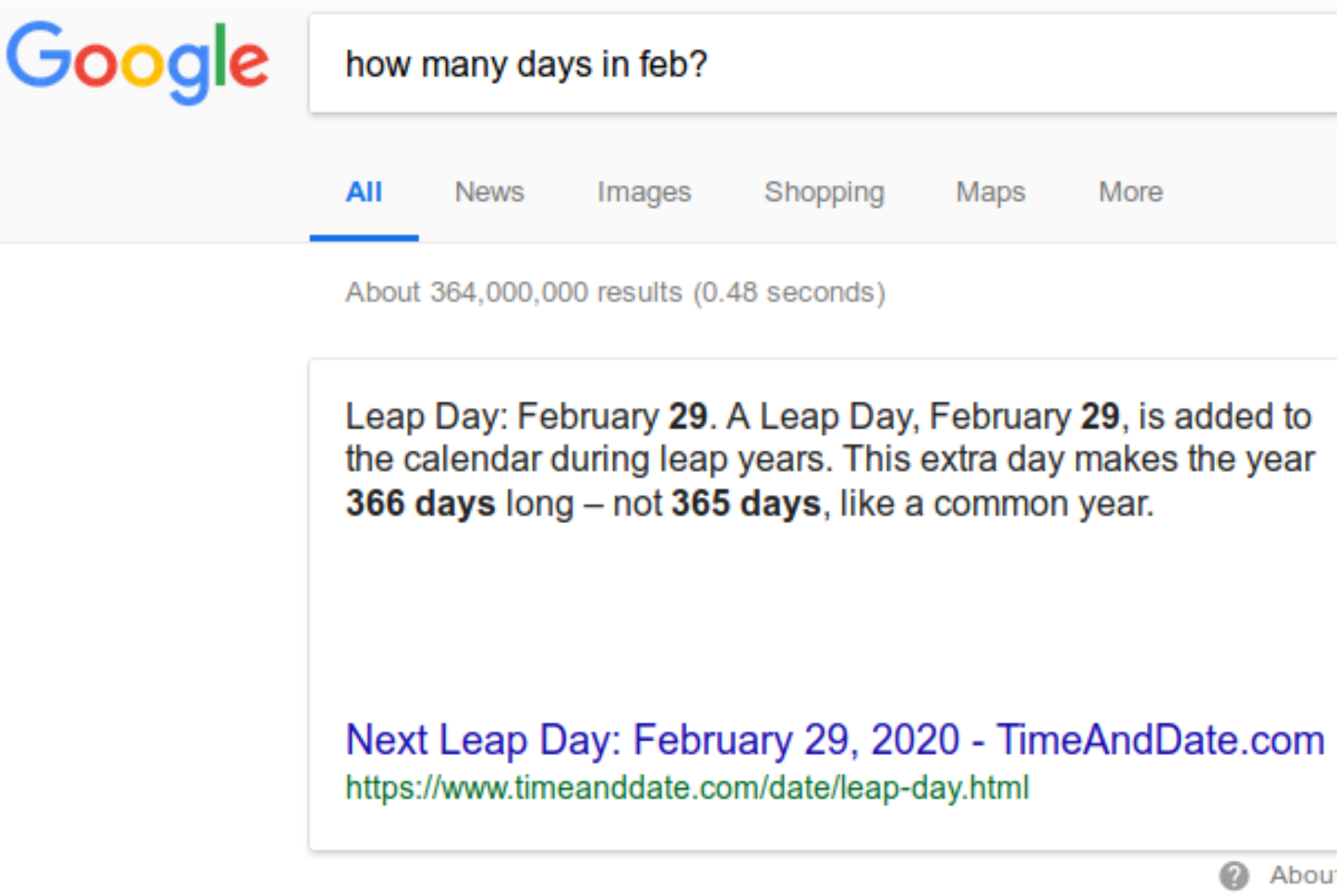
> “ And the LORD said, Behold, the people is one, and they have all one language; and this they begin to do: and now nothing will be restrained from them, which they have imagined to do.” –Genesis 1:6

- NLP is about understanding ourselves.

NLP vs. speech-related research

- NLP usually does not cover speech-related research, such as automatic speech recognition (ASR, speech-to-text) or speech synthesis (text-to-speech, TTS)
- Speech is about vocal time series, acoustic signals, or waves.
- NLP usually deals with the written form of languages, i.e., the scripts.
- For example, Siri, Cortana, Alexa have both speech part and NLP part.

NLP is far from expectation (circa. 2018)



NLP is far from expectation (circa. 2018)

Google

how many days per week

Search

All News Images Maps Videos More Settings Tools

About 2,600,000,000 results (0.61 seconds)

If you're going for the full **five days** per week, **three days** should focus on strength training, **two days** should focus on cardio, and two should be active rest. If you only want to work out **four days** a week, think about your goals: If you want to add muscle tone, cut a cardio day. Jan 2, 2018

How Often Should You Work Out? The Perfect Weekly Workout Routine
https://www.self.com/story/heres-what-a-perfect-week-of-working-out-looks-like

About this result Feedback

People also ask

How many days do you run a week?

How many days a week should you work out?

How many days a week should I rest?

Can you work out every day?

Feedback

Videos

Winograd Schema Challenge (WCS)

Let's fill a blank

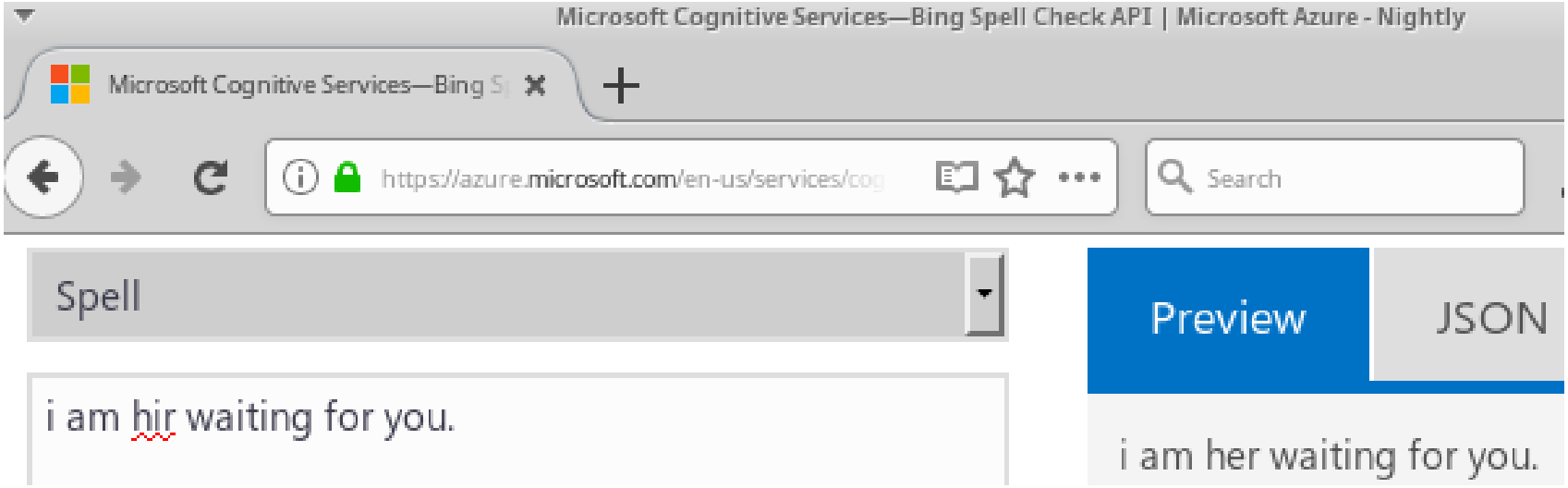
The city councilmen refused the demonstrators a permit because they [ ] violence.

A. feared

B. advocated

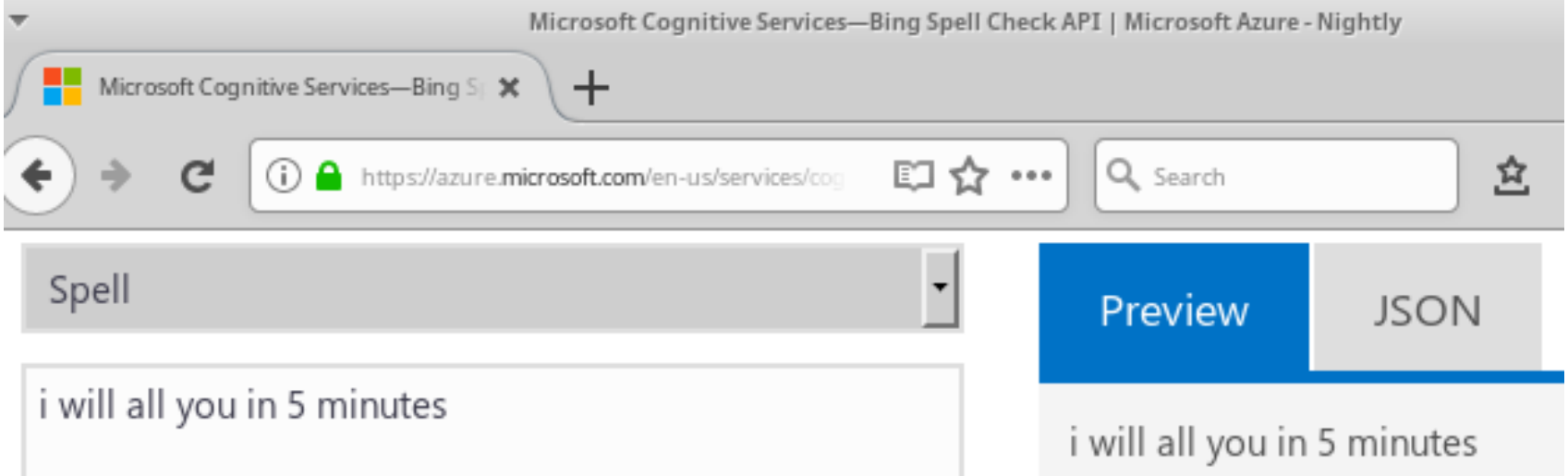
NLP is far from expectation

Spell checking is NLP.



NLP is far from expectation (circa. 2018)

Spell checking is NLP.



NLP in AI

- Alan Turing 1950 paper “Computing Machinery and Intelligence”
- Early sci-fi movies all have what today we call “question answering” (QA) or “automatic speech recognition” (ASR): Startrek communicator, 2001: A Space Odyssey (HAL can even read our lips!)
- Machine translation as at the center of early AI research
- “the spirit is willing but the flesh is weak” $->$ Russian and back $->$ “the vodka is good but the meat is rotten.”
- Almost no AI funding by 1974 – first AI winter.
- Reading assignment: “Technology; The Computer As Translator”, New York Times, April 28, 1983. https://www.nytimes.com/1983/04/28/business/technology-the-computer-as-translator.html

American academia should stop chasing buzzwords and imminent applications

- Almost half a century later, after two AI winters, machine translation final becomes plausible.
- Automatic speech recognition (ASR) also comes into our daily life: Google Now, Amazon Alexa, Apple Siri, Microsoft Cortana, etc.
- Persistent research into neural networks (or in today's buzzword, "deep learning") made those task tractable.
- Ironically, "neural network" was almost dead in the US between mid 1990s and mid 2000s. Little funding. Few papers. Everyone was going after SVM.
- Hence, the deep learning wave started from Canada: Hinton (Toronto) and Y. Bengio (Montreal).
- Yann LeCun said in one of his Facebook posts that he was nobody until DL became popular.

Why NLP is hard

- Ambiguity 1:
  - “This book is mine.”
  - “I live next to the coal mine.”
  - “Millions of mines along the border is a result of the war.”
- Ambiguity 2:
  - “You are so sweet.”
  - “The soup is too sweet.”
  - “Sweet, let’s play the game!”
- Morphology
  - “Carbon dioxide” vs. CO_2
  - “Iowa State University” vs. “ISU”
  - “Step 1” vs. “First step”
- We are not aware that we know so much.
  - This guy has two legs. – is it informative?

Typical tasks in or closely related to NLP (not mutually exclusive)

- Spell checking, next-work prediction
- Parsing and syntax
- Named entity recognition (NER)
- Relation/knowledge/information extraction, knowledge base/graph, e.g., MeasEval
- Information retrieval (historically considered a separate topic), e.g., Google search
- Semantics (you may count word embedding into this)
- Text classification (e.g., was this review helpful?)
- Machine translation
- Automatic summarization (abstractive and extractive)
- Question answering (QA)
- Machine Reading Comprehension (MRC)
- Text generation (a way to achieve many tasks aboves)
- Sentiment analysis
- Topic modeling

Rule-based vs. statistical NLP

- There are general two approaches to NLP problems: Rule-based (expert system) vs. statistical (machine learning).
- Tokenization is a typical task that rules may work without any problem.

```
>>> text = 'That U.S.A. poster-print costs $12.40...'
>>> pattern = r'''(?x)     # set flag to allow verbose regexps
...     ([A-Z]\.)+         # abbreviations, e.g. U.S.A.
...   | \w+(-\w+)*         # words with optional internal hyphens
...   | \$?\d+(\.\d+)?%?   # currency and percentages, e.g. $12.40,
...   | \.\.\.             # ellipsis
...   | [][.,;"'?():-_`]   # these are separate tokens; includes ],
... '''
>>> nltk.regexp_tokenize(text, pattern)
['That', 'U.S.A.', 'poster-print', 'costs', '$12.40', '...']
```

- However, there are many cases that rules cannot cover varieties.
- There shouldn't be a clear distinction between the two approaches. For example, is BOW model rule-based or statistical?

KR needed in ML for NLP

Pure statistical language model is insufficient to judge whether the following statements makes a review helpful.

- “a waste of money”
- “Amazon shipping and return is so easy ”
- “This car has 4 wheels ”

NLP is language-dependent

- Some tasks are particularly difficult in some languages, e.g., segmentation in Chinese and Arabic.
- Some tasks are not “symmetric” between languages, e.g., in machine translation,
  - a sentence in Chinese may be required by grammar to be broken into multiple sentences in English, or
  - the order of phrases needs to be rearranged after translation.

Computational Linguistics or NLP?

- Apparently, mission-oriented research will not produce a long-lasting impact and/or a big breakthrough.
- “Physics is like sex.” Richard Feynman.
- However, many problems in linguistics are not well defined, such as writing style. No ground truth. Hard to quantitatively measure.
- “To create an artificial being has been the dream of man since the birth of science.”
- Computer science is about processing information, maybe for human use.
- Hence, the instructor prefer the name NLP over CL.

What I am working on in NLP

- Review analysis (“was this review helpful?”, ACL 2015, ICTAI 2016, EACL 2017, NAACL 2018)
- Summarization (EACL 2017, and on going)
- Sentence purpose understanding (e.g., whether a sentence is about experimental conditions and yields, funded by NSF)

About the class

- Introductory: letting you know something about a lot of topics in NLP.
- NLP is so rapidly evolving so papers or online materials will be the main source of knowledge.
- No HW or exams to periodically push you to learn. Be self-motivated.
- Projects are optional. Bonus points. Potential papers.

Outline of the class

Let's go over the syllabus.

Prerequisite

- Linear algebra, calculus (up to multivariate calculus or calculus III), Probability theory – in era of deep learning, it is very hard to avoid such math basics.
- Machine Learning – however, we will spend some time to cover due to different background of students.
- Computer programming, e.g., you can solve Leetcode easy-level problems under 10 minutes.
- The patience and desire to to think systematically and structurally