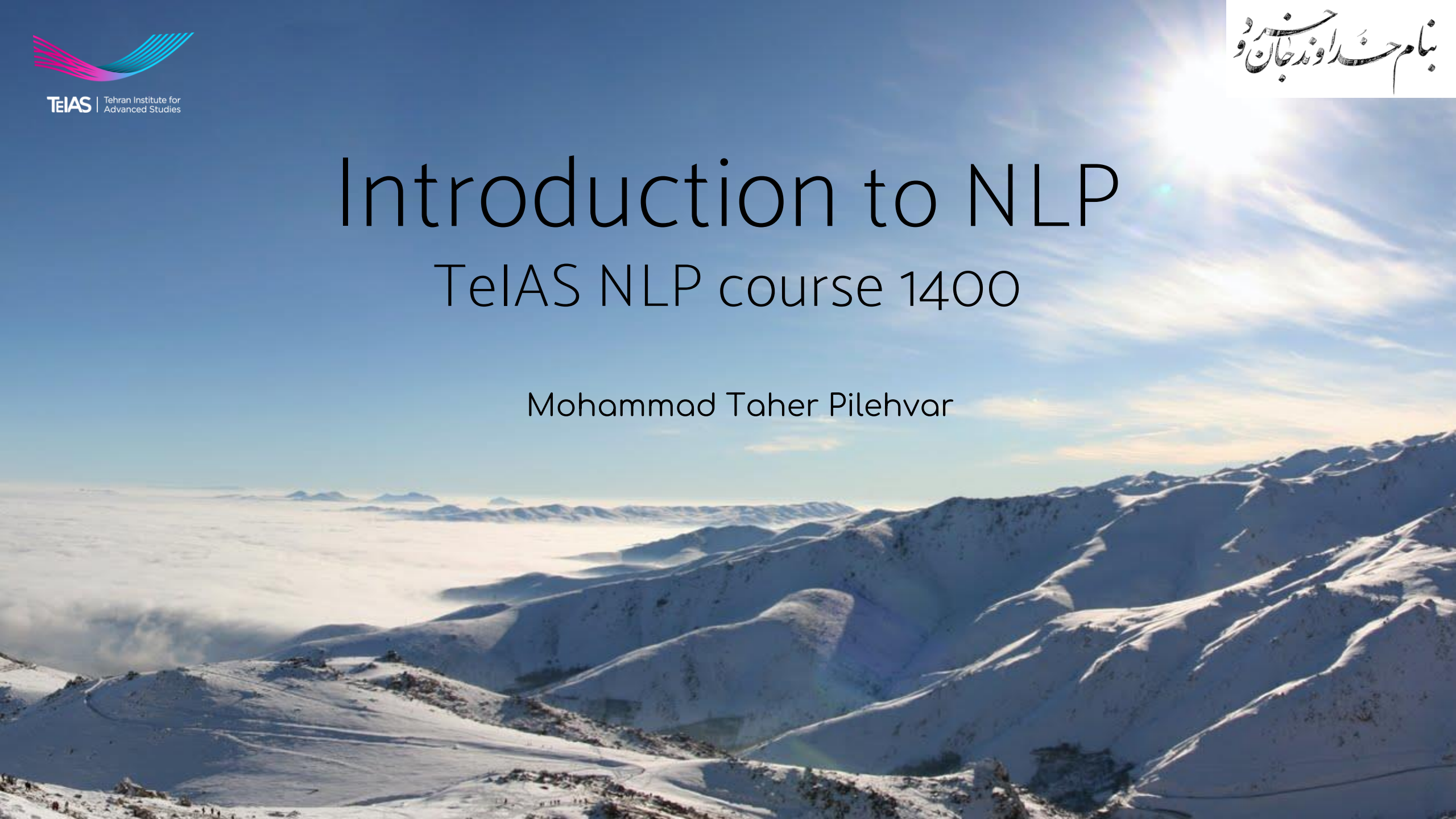
TeIAS | Tehran Institute for Advanced Studies

بنام خداوند جان و خرد

Introduction to NLP

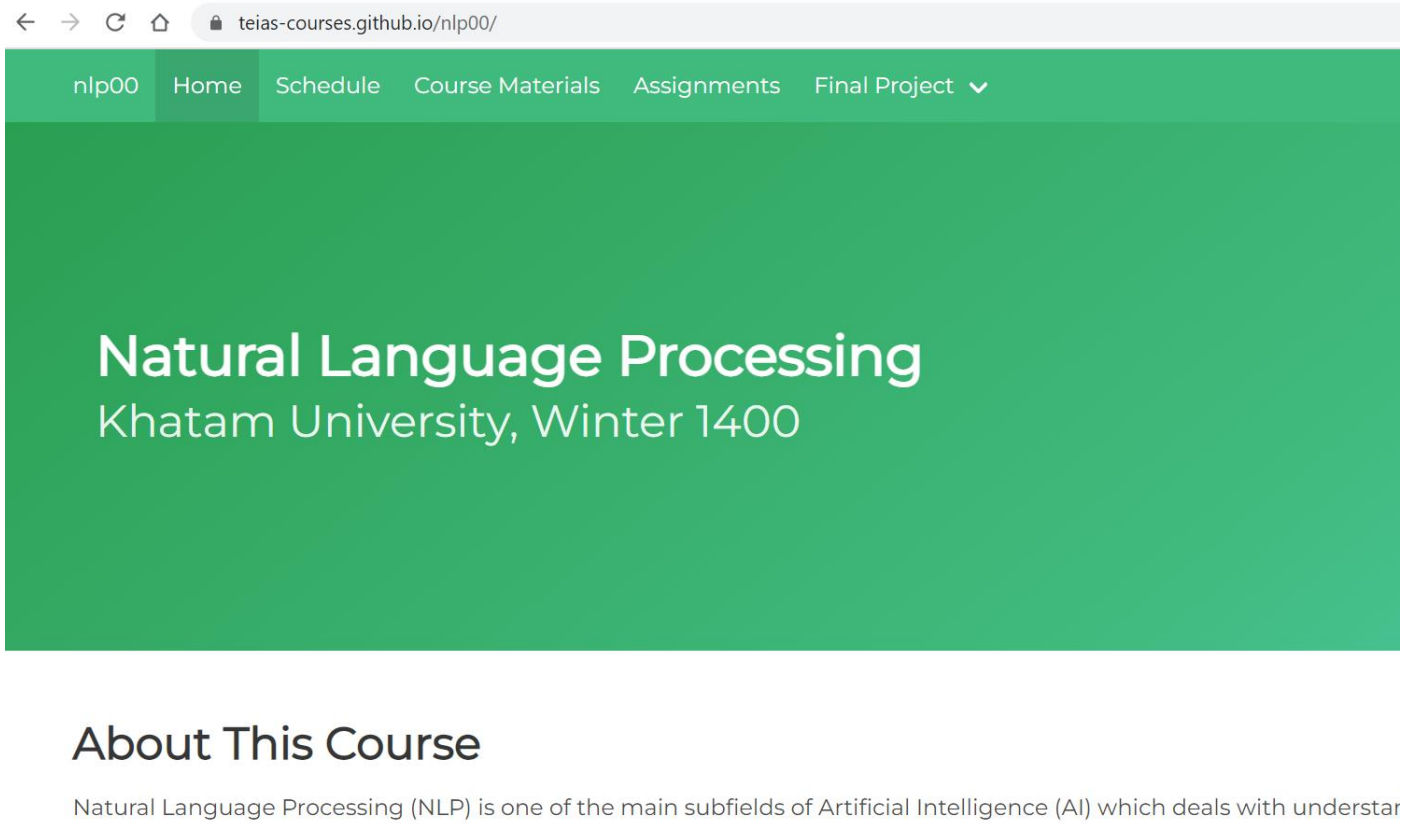
TeIAS NLP course 1400

Mohammad Taher Pilehvar



Website

https://teias-courses.github.io/nlp00/



Overview of the course

- Semantic representation and word embeddings (4 sessions)
- Language models (2 sessions)
- Recurrent Neural Networks (2 sessions)
- Transformers and BERT (3 sessions)
- Machine Translation, Question Answering, model analysis, prompt-based learning, knowledge integration, generation, etc. (each 1 session)
- Practical: 3 to 4 sessions
- Progress reports: 2 sessions
- Research talks: 3 sessions

Survey

https://ahaslides.com/NLP001

Deep Learning background

https://www.coursera.org/specializations/deep-learning

Andrew Ng's Deep Learning specialization

Online quiz
Esfand 7th

COURSE 1

Neural Networks and Deep Learning

4.9 111,875 ratings • 22,162 reviews

In the first course of the Deep Learning Specialization, you will study the foundational concept of neural networks and deep learning.

By the end, you will be familiar with the significant technological trends driving the rise of deep learning; build, train, and apply fully connected

COURSE 2

Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization

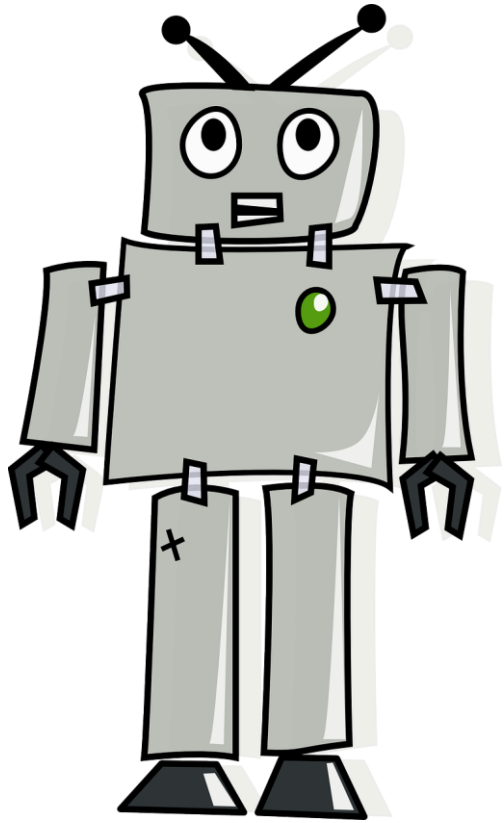
4.9 60,227 ratings • 6,971 reviews

In the second course of the Deep Learning Specialization, you will open the deep learning black box to understand the processes that drive performance and generate good results systematically.

Artificial Intelligence

Mimic “cognitive” functions

- Planning
- Learning
- Reasoning
- Perception
- …
- Vision
- Natural Language Processing



AI - Planning

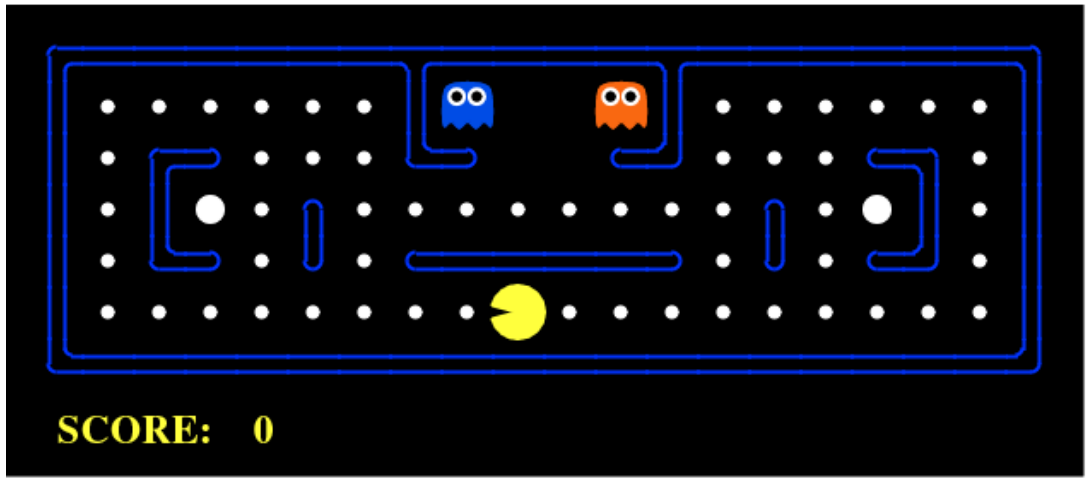


Image Credit: DeepMind

AI - Planning (Game)

AlphaGo

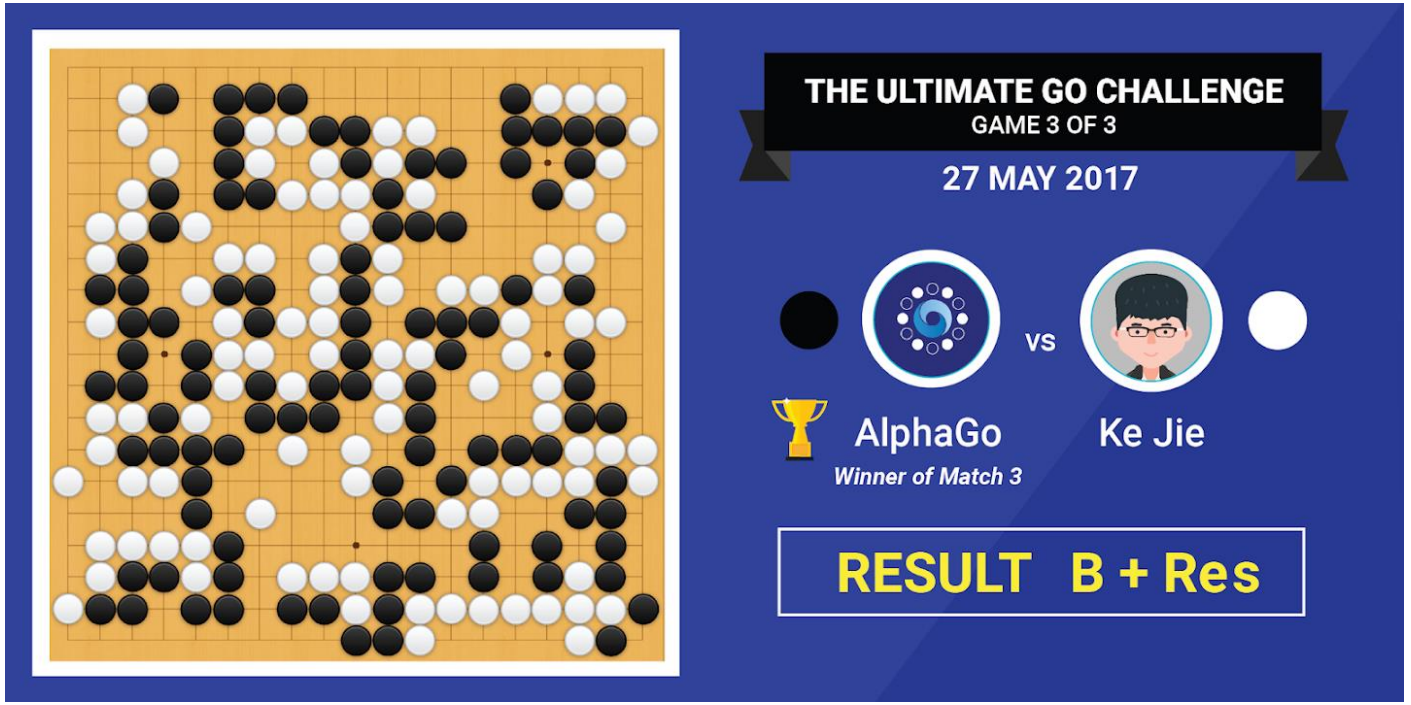


Image Credit: DeepMind

AI - Computer Vision

AI - Biology

AlphaFold, protein structure prediction

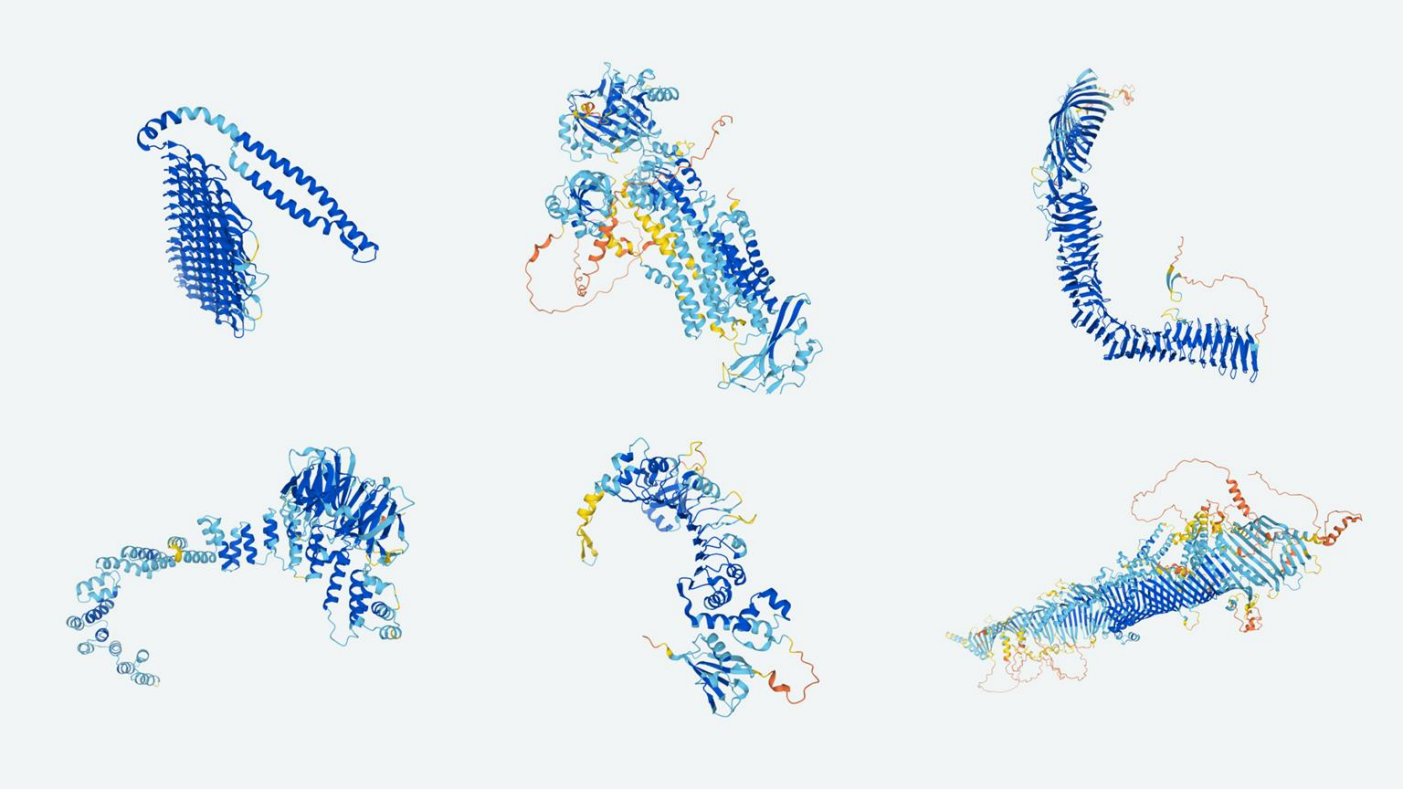


Image Credit: DeepMind

Natural Language Processing

Natural Language Understanding (NLU)

Natural Language Generation (NLG)

NLP: Challenges

NLU

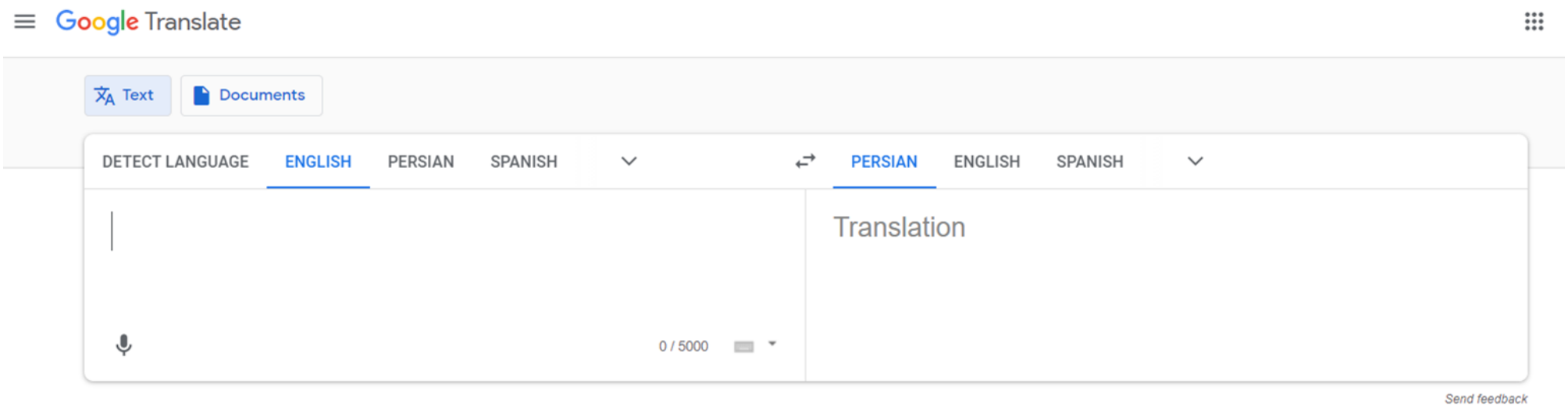
- Ambiguity
  - Lexical — Time flies, pilot flies
  - Syntactic — I saw a man on the hill with binoculars
  - Metonymic — NY voted for Biden
  - …
- Common sense knowledge — The tablet does not fit into my bag because it is too large.
- Figurative language — all ears, fingers crossed

NLG

- Ambiguity
- Word order
- Fluency

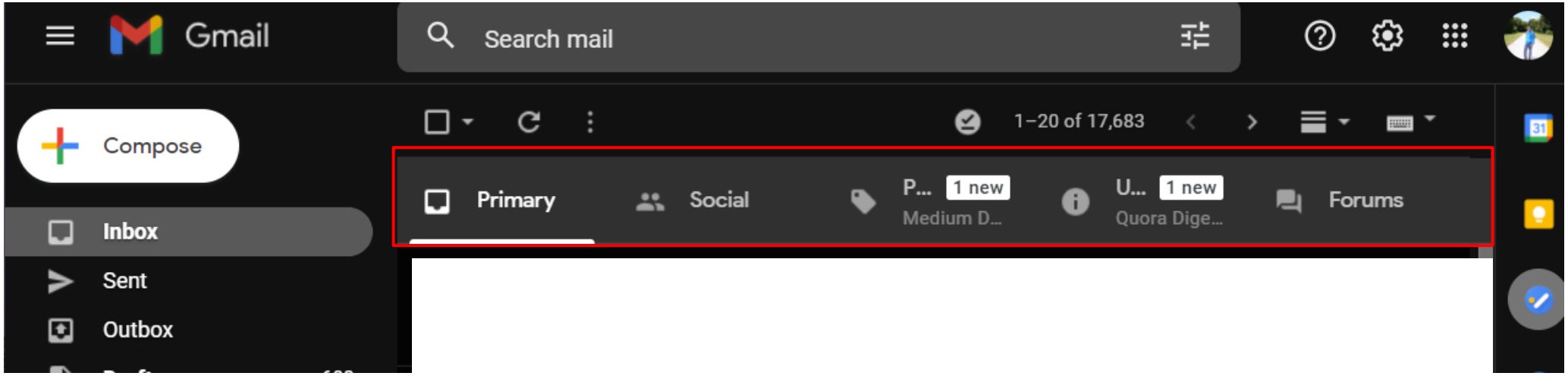
NLP Applications

Machine Translation



NLP Applications

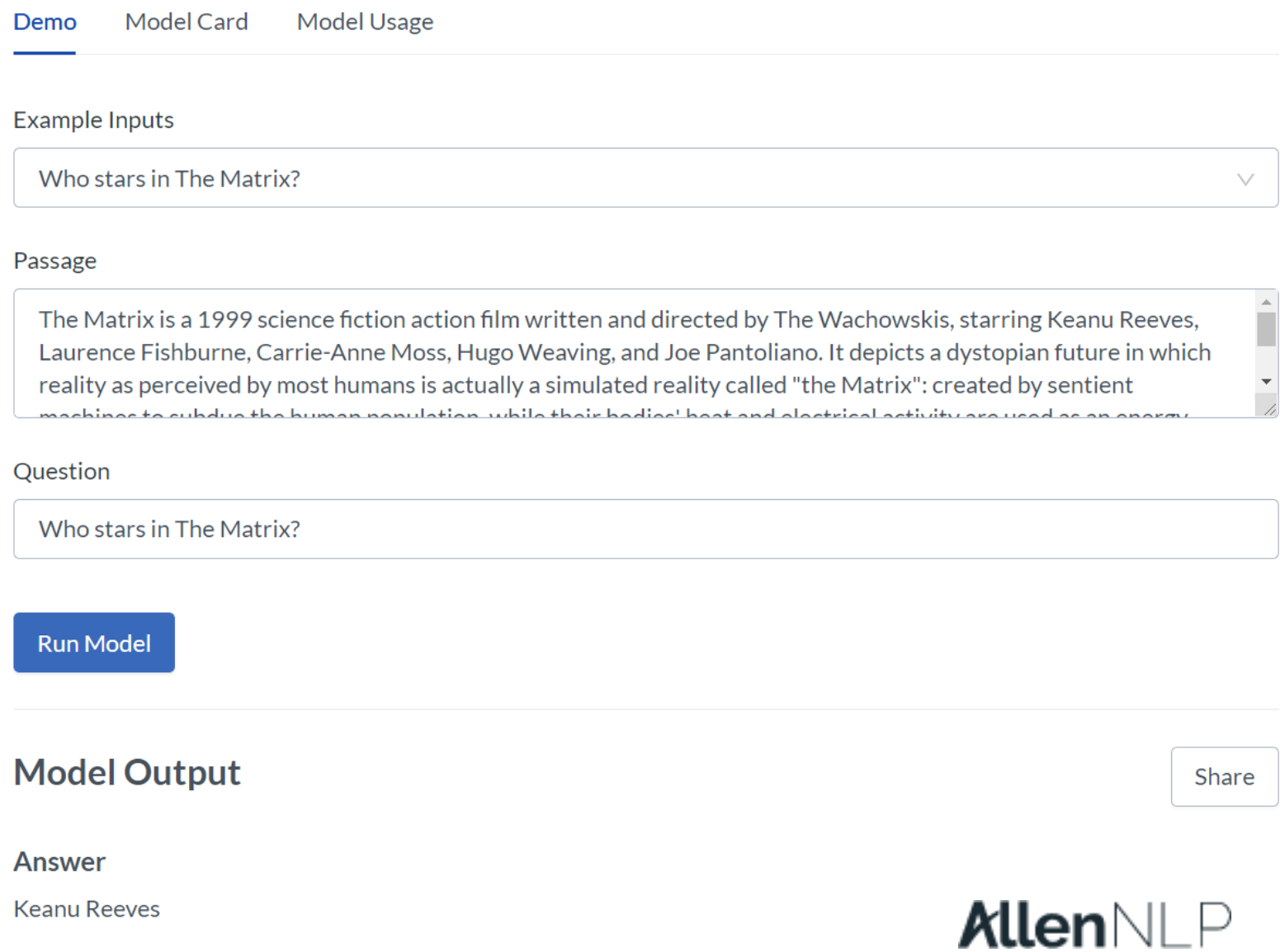
Text classification



NLP Applications

NLP Applications

Question Answering



NLP Applications

Visual QA

Example Inputs

Bus Stop: "What are the people waiting for?"

Image

Question

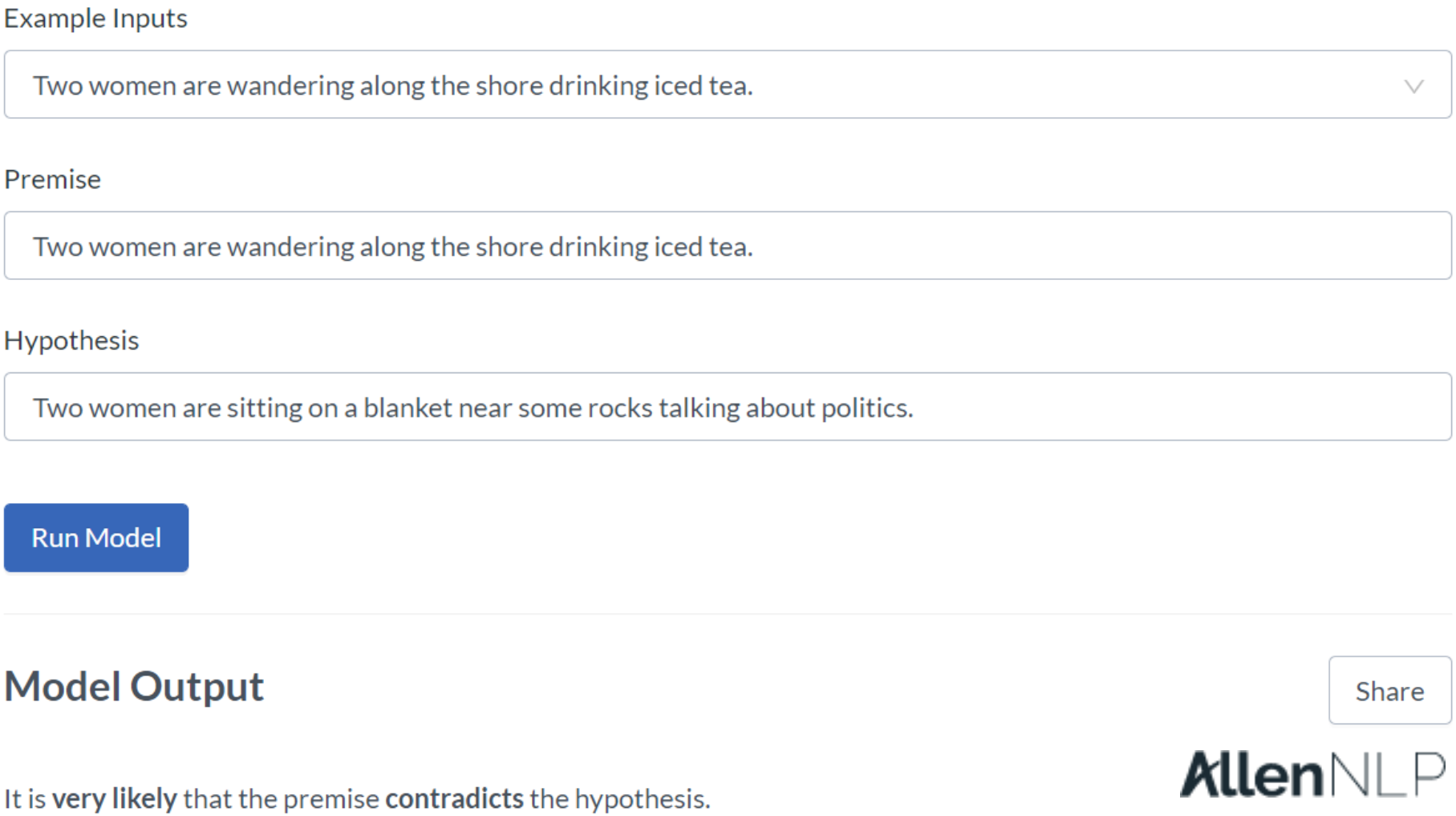
What are the people waiting for?

Run Model

AllenNLP

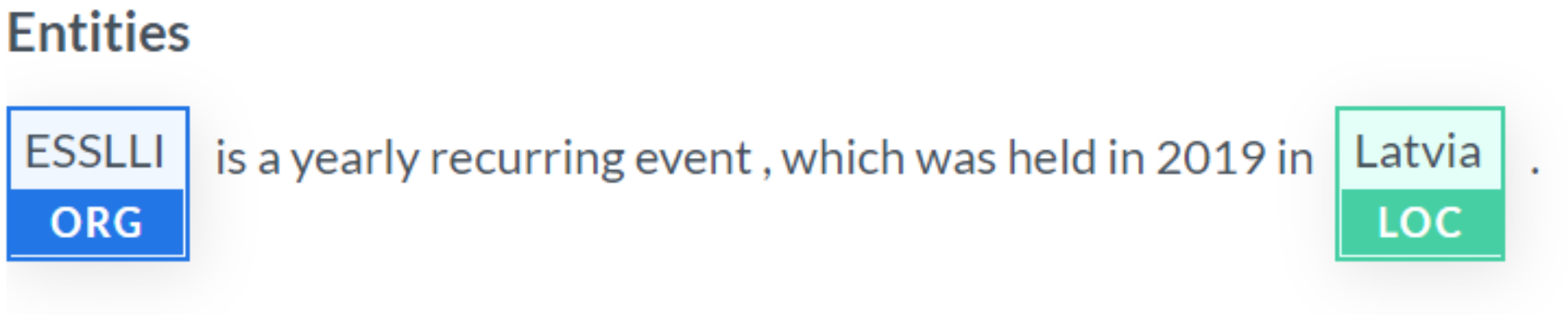
NLP Applications

Entailment (NLI)



NLP Applications

Named Entity Recognition (NER)



AllenNLP

NLP Applications

Coreference resolution

We are looking for [0] a region of central Italy bordering the Adriatic Sea . [0] The area is mostly mountainous and includes Mt. Corno , the highest peak of the mountain range . [0] It also includes [1] many sheep and an Italian entrepreneur has an idea about how to make a little money of [1] them .

AllenNLP

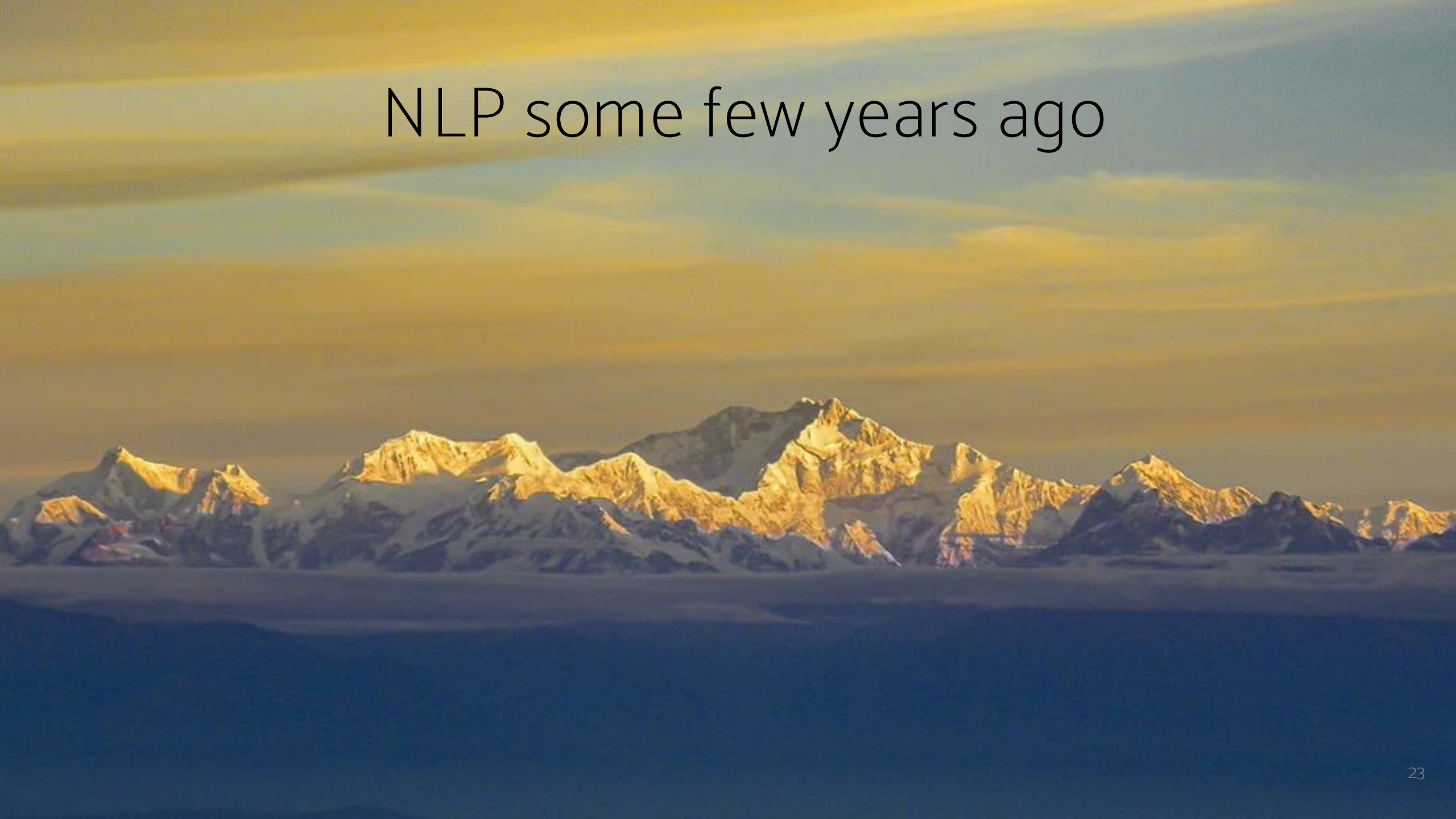
NLP Applications

Many more …

Part of speech tagging
Summarization
Information retrieval
Chatbots
…

NLP some 10 years ago

NLP some few years ago



2017 - Transformers

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

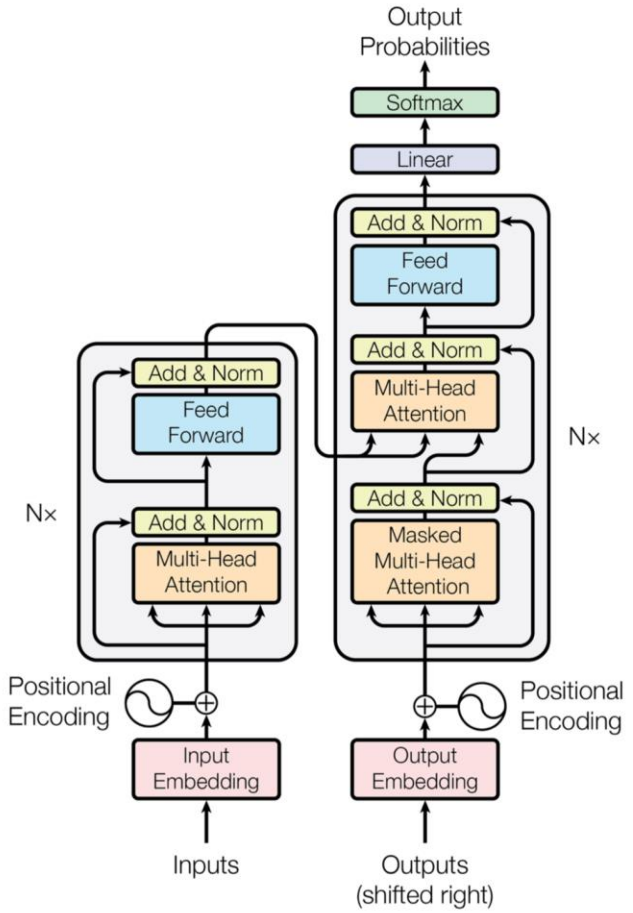
Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

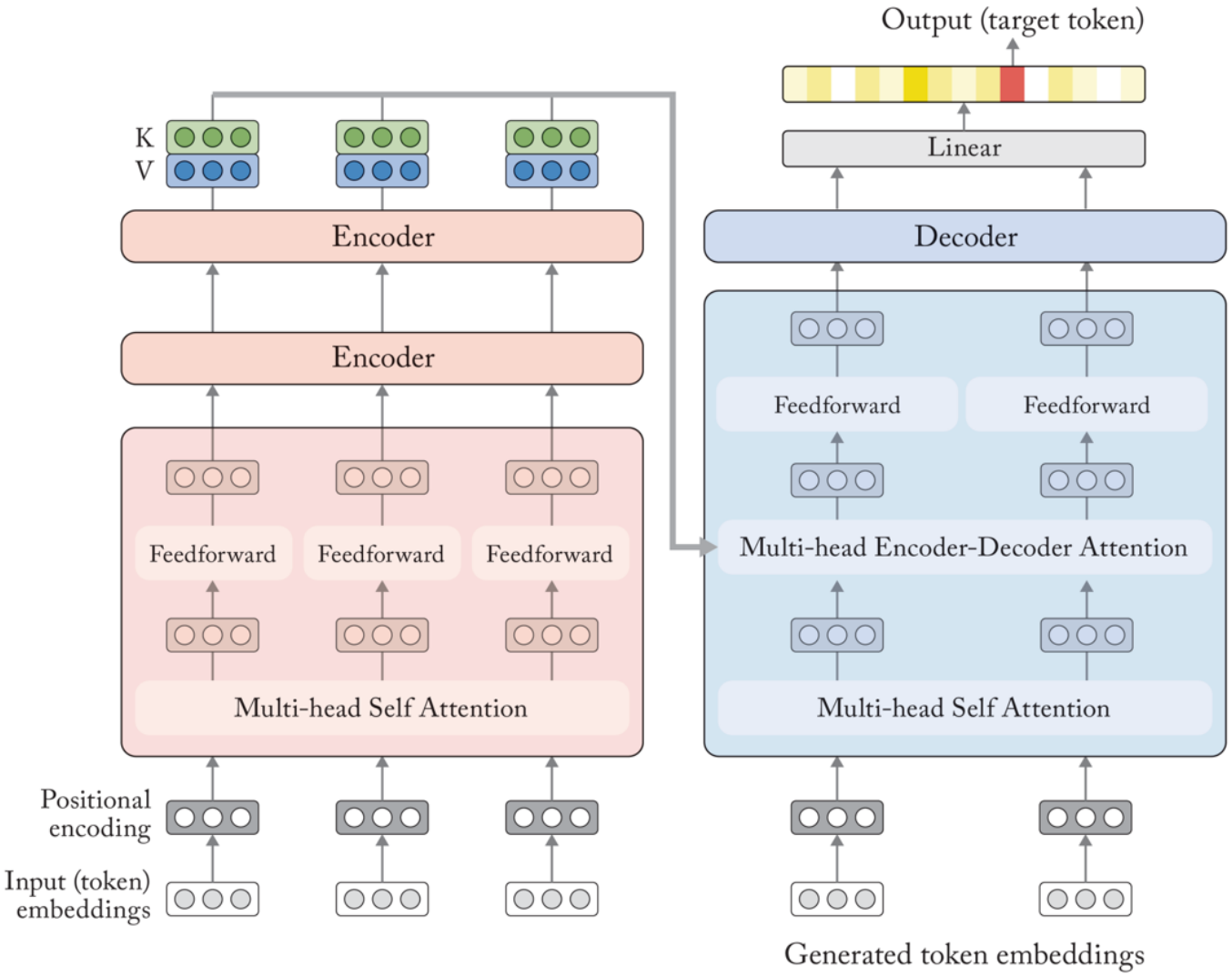
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com



NLP now

Transformers



Teaching Assistants

Kave Eskandari

Mahdi Zakizadeh

Score distribution

- Project: 50%
  - Progress report 1 (10%)
  - Progress report 2 (10%)
  - Final report (30%)
- Homeworks (probably 4): 30%
- Final exam: 20% (minimum of 12/20 to pass the course)

The NLP Research Community

- ACL Anthology has nearly everything, free!
  - Over 70,000 papers!
  - Free-text searchable
    - Great way to learn about current research on a topic
      - Find recent or highly cited work; follow citations
  - Used as a dataset by various projects
    - Analyzing the text of the papers (e.g., parsing it)
    - Extracting a graph of papers, authors, and institutions (Who wrote what? Who works where? What cites what?)

- Google Scholar to sort by citation count / track citations

The NLP Research Community

- Most work in NLP is published as 9-page conference papers with 3 double-blind reviewers.
  Papers are presented via talks, posters, videos

- Also:
  - Conference short papers (5 pages)
  - "Findings" papers (accepted but without presentation)
  - "System demo track" / "Industry track"
  - Journal papers (up to 12 pages, or unlimited)

- Main annual conferences: ACL, EMNLP, NAACL
  + EACL, AACL/IJCNLP, COLING, …; also LREC
  + journals: TACL, CL, …
  + AI/ML journals: JAIR, JMLR, …
  + various specialized conferences and workshops

The NLP Research Community

Pre-COVID, ACL had > 2000 in-person attendees

- ACL 2021 (virtual conference):
  - 3350 papers submitted (710 accepted = 21%)
    - Accepted:
    - 80% “long” (9 pages)
    - 20% “short” (5 pages)
    - + 493 “findings” published outside main conf (➔ total 37%)
  - Awards: Several papers (will be widely read)

“Tracks” at ACL 2021

- Machine Learning for NLP
- Interpretability & Analysis of Models for NLP
- Resources & Evaluation
- Ethics & NLP
- Phonology, Morphology & Word Segmentation
- Syntax: Tagging, Chunking & Parsing
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference & Other areas
- Linguistic Theories, Cognitive Modeling & Psycholinguistics
- Information Extraction
- Information Retrieval & Text Mining
- Question Answering
- Summarization
- Machine Translation & Multilinguality
- Speech & Multimodality
- Discourse & Pragmatics
- Sentiment Analysis, Stylistic Analysis, & Argument Mining
- Dialogue & Interactive Systems
- Language Grounding to Vision, Robotics & Beyond
- Computational Social Science & Cultural Analytics
- NLP Applications
- Special theme: NLP for Social Good

+ 23 workshops on various topics

The NLP Research Community

- arXiv papers
- Twitter accounts
  - NLP researchers with active accounts (grad students, profs, industry folks)
  - Official conference accounts
- "NLP Highlights" podcast
- "NLP News" newsletter

The NLP Research Community - Institutions

- Universities: Many have 2+ NLP faculty
  - Several "big players" with many faculty
  - Some of them also have good linguistics, cognitive science, machine learning, AI
- Companies:
  - Old days: AT&T Bell Labs, IBM
  - Now: Microsoft, Google, FB, Amazon, startups …
    - Many niche markets – online reviews, medical transcription, news summarization, legal search and discovery …

Text Annotation Tasks

- Classify the entire document (“text categorization”)

Sentiment classification

★☆☆☆☆ **An extremely versatile machine!**, November 22, 2006

By **Dr. Nickolas E. Jorgensen "njorgens3"**

This review is from: Cuisinart DGB-600BC Grind & Brew, Brushed Chrome (Kitchen)

This coffee-maker does so much! It makes weak, watery coffee! It grinds beans if you want it to! It inexplicably floods the entire counter with half-brewed coffee when you aren't looking! Perhaps it could be used to irrigate crops... It is time-consuming to clean, but in fairness I should also point out that the stainless-steel thermal carafe is a durable item that has withstood being hurled onto the floor in rage several times. And if all these features weren't enough, it's pretty expensive too. If faced with the choice between having a car door repeatedly slamming into my genitalia and buying this coffee-maker, I'd unhesitatingly choose the Cuisinart! The coffee would be lousy, but at least I could still have children...

Other text categorization tasks

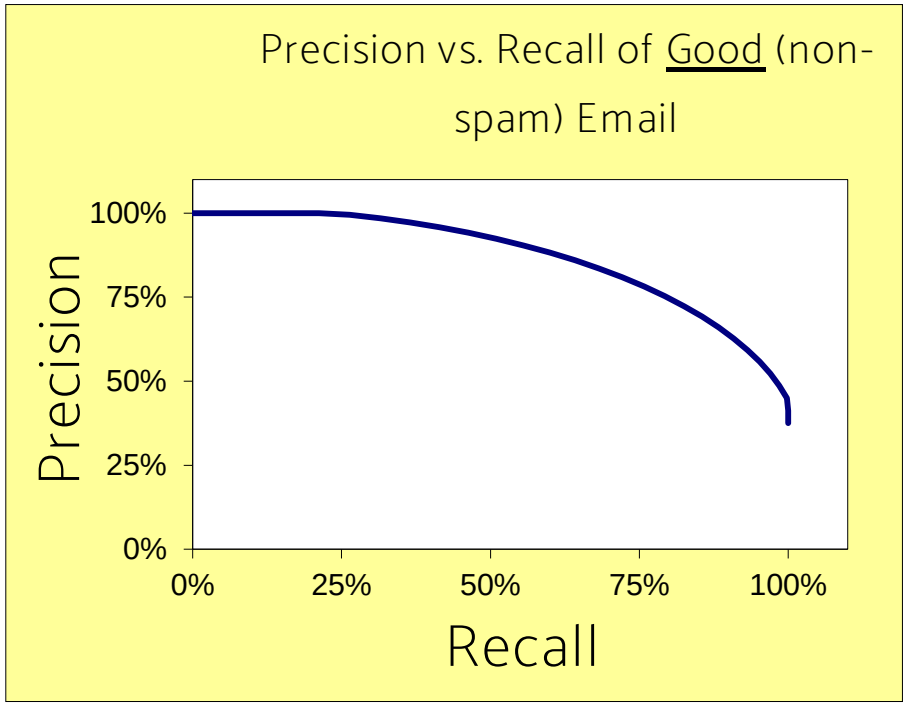
- Is it spam? (see features)
- What grade, as an answer to this essay question?
- Is it interesting to this user?
  - News filtering; helpdesk routing
- Is it interesting to this NLP program?
  - Skill classification for a digital assistant!
  - If it's Spanish, translate it from Spanish
  - If it's subjective, run the sentiment classifier
  - If it's an appointment, run information extraction
- Where should it be filed?
  - Which mail folder? (work, friends, junk, urgent ...)
  - Yahoo! / Open Directory / digital libraries

Measuring Performance

- Classification accuracy: What % of messages were classified correctly?
- Is this what we care about?
- Which system do you prefer?

	Overall accuracy	Accuracy on spam	Accuracy on gen
System 1	95%	99.99%	90%
System 2	95%	90%	99.99%

Measuring Performance



- Precision = $\frac{\text{good messages kept}}{\text{all messages kept}}$
- Recall = $\frac{\text{good messages kept}}{\text{all good messages}}$

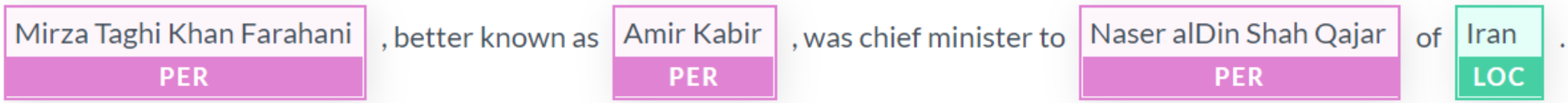
Text Annotation Tasks

- Classify the entire document ("text categorization")
- Classify individual words

Text Annotation Tasks

- Classify the entire document (“text categorization”)
- Classify individual words
- Identify phrases (“chunking”)

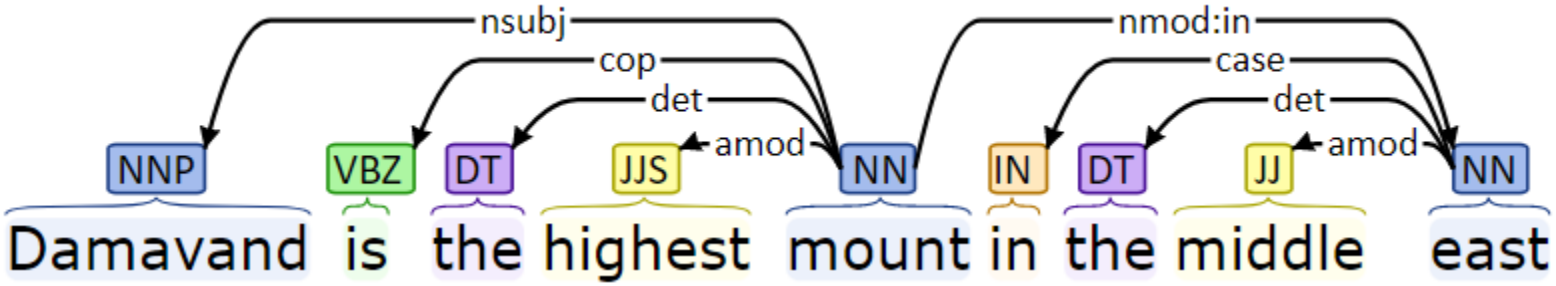
Named Entity Recognition



Text Annotation Tasks

- Classify the entire document (“text categorization”)
- Classify individual words
- Identify phrases (“chunking”)
- Syntactic annotation (parsing)

Labeled Dependency Parsing



Text Annotation Tasks

- Classify the entire document (“text categorization”)
- Classify individual words
- Identify phrases (“chunking”)
- Syntactic annotation (parsing)
- Semantic annotation

Semantic Role Labeling (SRL)

For each predicate (e.g., verb)

1. find its arguments (e.g., NPs)
2. determine their semantic roles

John drove Mary from Austin to Dallas in his Toyota Prius.

The hammer broke the window.

- agent: Actor of an action
- patient: Entity affected by the action
- source: Origin of the affected entity
- destination: Destination of the affected entity
- instrument: Tool used in performing action.
- beneficiary: Entity for whom action is performed