CS 6770 Natural Language Processing

Introduction

Yangfeng Ji

Information and Language Processing Lab
Department of Computer Science
University of Virginia

UNIVERSITY of VIRGINIA | ENGINEERING

Overview

1. Course Information

2. A Brief History of NLP

Course Information

Course Webpage

`https://yangfengji.net/uva-nlp-grad/`

Instructors

- Instructor
  - Yangfeng Ji
  - Office hour: Friday 11 AM – 12 PM
  - Location: Rice 510

Instructors

- Instructor
    - Yangfeng Ji
    - Office hour: Friday 11 AM – 12 PM
    - Location: Rice 510
- TAs:
    - Elizabeth Palmieri
        - Office hour: Monday 11 AM - 12 PM
        - Location: Rice 414
    - Caroline Gihlstorf
        - Office hour: TBA
        - Location: TBA

Prerequisites

Some requirements

- Proficiency in Python
- Basic Calculus and Linear Algebra
- Basic Probability and Statistics
- Foundations of Machine Learning

Goal of This Course

1. Explain the fundamental NLP techniques
   - Text classification
   - Word embeddings
   - Language modeling
   - Sequence-to-sequence modeling
   - Pre-trained large language models
   - Fine-tuning, prompt and context engineering

Class Schedule

Goal of This Course

1. Explain the fundamental NLP techniques
    - Text classification
    - Word embeddings
    - Language modeling
    - Sequence-to-sequence modeling
    - Pre-trained large language models
    - Fine-tuning, prompt and context engineering
2. Opportunities of working on some NLP problems
    - Homework Assignments

Class Schedule

Goal of This Course

1. Explain the fundamental NLP techniques
   - Text classification
   - Word embeddings
   - Language modeling
   - Sequence-to-sequence modeling
   - Pre-trained large language models
   - Fine-tuning, prompt and context engineering
2. Opportunities of working on some NLP problems
   - Homework Assignments
3. Prepare for research in NLP
   - Read and discuss recent papers

Class Schedule

Assignments

- No exam

Assignments

- No exam
- Four homeworks
  - $18\% \times 4 = 72\%$

Assignments

- No exam
- Four homeworks
  - $18\% \times 4 = 72\%$
- Four reading assignments
  - $7\% \times 4 = 28\%$

Four Homeworks

- Two tracks: each homework will have a specific track
  - Track 1: text classification
  - Track 2: text generation

Homework 0

Four Homeworks

- Two tracks: each homework will have a specific track
  - Track 1: text classification
  - Track 2: text generation
- Each student will choose one track for all homeworks
  - It is possible to switch track after the first homework, but not recommended

Homework 0

Four Homeworks

- Two tracks: each homework will have a specific track
  - Track 1: text classification
  - Track 2: text generation
- Each student will choose one track for all homeworks
  - It is possible to switch track after the first homework, but not recommended
- In each track, some example tasks and datasets will be provided
  - The instruction will be released later
  - Students can also choose their own tasks and datasets

Homework 0

Four Homeworks (II)

Each track has four steps to implement the entire NLP pipeline, one step for each homework

- Homework 1: Data Exploration and Preprocessing
- Homework 2: Classical Model Baselines
- Homework 3: Word Embeddings and Basic Neural Models
- Homework 4: LLM-based Methods

Reading Assignments

Along with the topics discussed in class, there will be four reading assignments.

Generative AI Examples

Students can leverage various tools for different aspects of their coursework:

- Google Colab: Coding
- VS Code + Github Copilot: Coding
- Google NotebookLM: Reading

Generative AI Examples

Students can leverage various tools for different aspects of their coursework:

- Google Colab: Coding
- VS Code + Github Copilot: Coding
- Google NotebookLM: Reading

Reminders:

- Always verify the output and make necessary adjustments.
- Students should be responsible for the correctness of their submissions.

Policy: late penalty

Homework submission will be accepted up to 72 hours late, with 20% deduction per 24 hours on the points as a penalty.

For example,

- Deadline: Sept. 15th, 11:59 PM
- Submission timestamp: Sept. 17th, 9:00 AM ($\leq$ 48 hours)
- Original points of a homework: 10
- Actual points:

$$10 \times (1 - 40\%) = 6.0 \tag{1}$$

It is usually better if students just turn in what they have in time.

Policy: late penalty

Homework submission will be accepted up to 72 hours late, with 20% deduction per 24 hours on the points as a penalty.

For example,

- Deadline: Sept. 15th, 11:59 PM
- Submission timestamp: Sept. 17th, 9:00 AM (≤ 48 hours)
- Original points of a homework: 10
- Actual points:

$$10 \times (1 - 40\%) = 6.0 \tag{1}$$

It is usually better if students just turn in what they have in time.

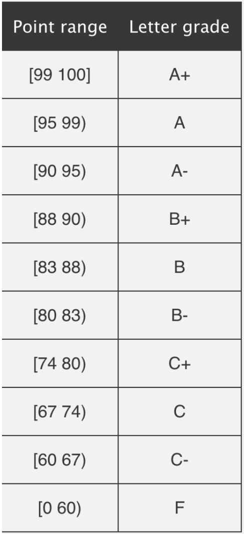
- It's the students' responsibility to double check their submission and make sure you submit the correct file.
- If a student submits one homework via multiple files/times, we will use the latest timestamp for deciding and calculating the late penalty.

Policy: collaboration

On the course webpage

`https://yangfengji.net/uva-nlp-grad/`

Policy: grades



Point range	Letter grade
[99 100]	A+
[95 99)	A
[90 95)	A-
[88 90)	B+
[83 88)	B
[80 83)	B-
[74 80)	C+
[67 74)	C
[60 67)	C-
[0 60)	F

Textbooks

- Textbook
  - Eisenstein, *Natural Language Processing*, 2018

Textbooks

- Textbook
  - Eisenstein, *Natural Language Processing*, 2018
- Additional textbooks
  - Jurafsky and Martin, *Speech and Language Processing*, 3rd Edition, 2020
  - Shalev-Shwartz and Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, 2014
  - Goodfellow, Bengio and Courville, *Deep Learning*, 2016

Textbooks

- Textbook
  - Eisenstein, *Natural Language Processing*, 2018
- Additional textbooks
  - Jurafsky and Martin, *Speech and Language Processing*, 3rd Edition, 2020
  - Shalev-Shwartz and Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, 2014
  - Goodfellow, Bengio and Courville, *Deep Learning*, 2016

All free online

Question?

A Brief History of NLP

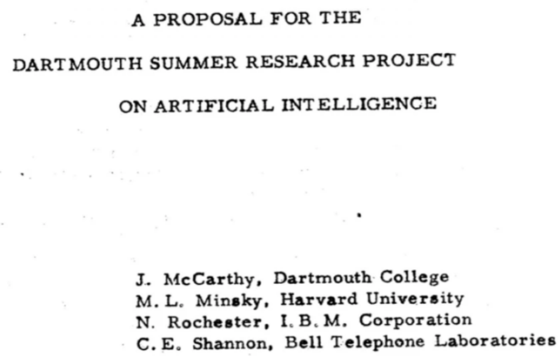
1954: The Georgetown-IBM Demonstrations

The first public demonstration of non-numeric applications of digital computers, which also encouraged the government investment of NLP research

Fig. 2: Hurd, Dostert and Watson at the demonstration

1955: The Proposal of Dartmouth AI Workshop

Where the term “Artificial Intelligence” was coined



A PROPOSAL FOR THE

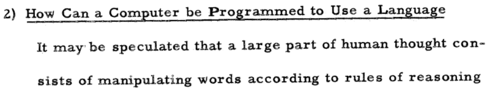
DARTMOUTH SUMMER RESEARCH PROJECT

ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I. B. M. Corporation
C. E. Shannon, Bell Telephone Laboratories

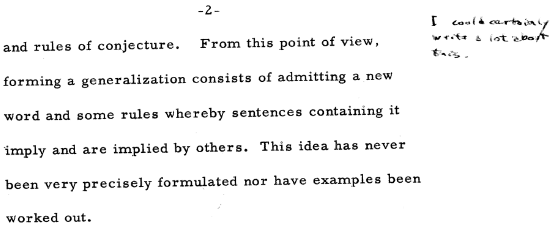
1955: The Proposal of Dartmouth AI Workshop (II)

Seven aspects of AI problem were listed in the proposal, and the second one is about language



2) <u>How Can a Computer be Programmed to Use a Language</u>

It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning



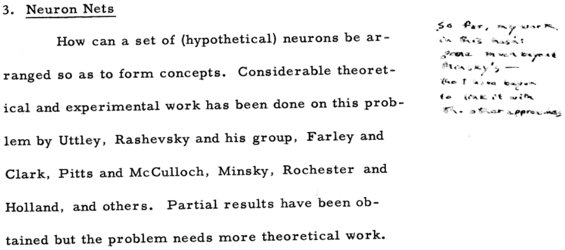
-2-

and rules of conjecture. From this point of view, forming a generalization consists of admitting a new word and some rules whereby sentences containing it imply and are implied by others. This idea has never been very precisely formulated nor have examples been worked out.

I could certainly write a lot about this.

1955: The Proposal of Dartmouth AI Workshop (III)

... and the third one is about neuron nets



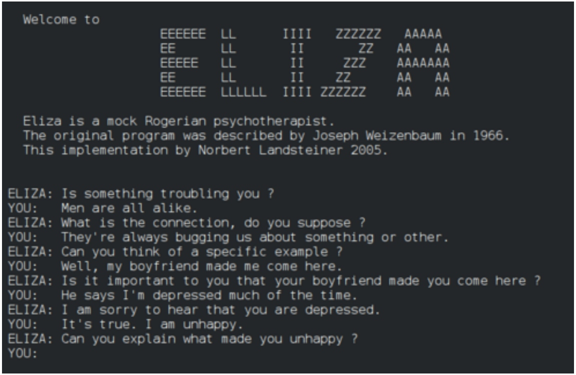
3. Neuron Nets

How can a set of (hypothetical) neurons be arranged so as to form concepts. Considerable theoretical and experimental work has been done on this problem by Uttley, Rashevsky and his group, Farley and Clark, Pitts and McCulloch, Minsky, Rochester and Holland, and others. Partial results have been obtained but the problem needs more theoretical work.

So far, my work
in this hasn't
gone much beyond
Minsky's —
tho I have begun
to link it with
the other approaches

1964: ELIZA

ELIZA, created by Joseph Weizenbaum at MIT, simulated conversation by using pattern matching and substitution methodology

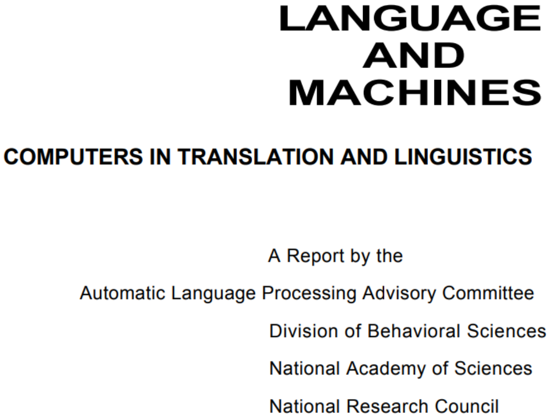


```
  Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II      ZZ    AA   AA
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

ELIZA is considered as the first chatbot.

1966: The ALPAC Report

ALPAC refers to Automatic Language Processing Advisory Committee, led by John R. Pierce, an information theory expert

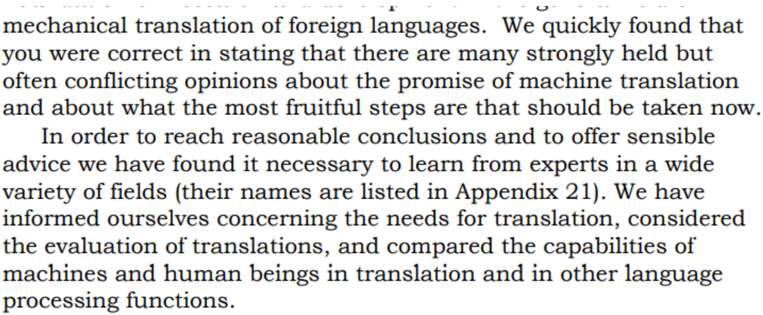


LANGUAGE
AND
MACHINES

COMPUTERS IN TRANSLATION AND LINGUISTICS

A Report by the
Automatic Language Processing Advisory Committee
Division of Behavioral Sciences
National Academy of Sciences
National Research Council

1966: The ALPAC Report

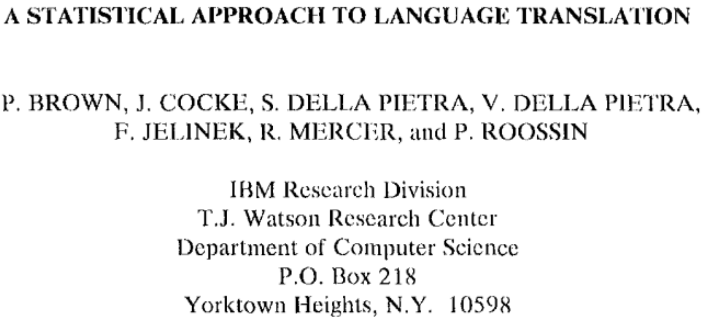


mechanical translation of foreign languages. We quickly found that you were correct in stating that there are many strongly held but often conflicting opinions about the promise of machine translation and about what the most fruitful steps are that should be taken now.

In order to reach reasonable conclusions and to offer sensible advice we have found it necessary to learn from experts in a wide variety of fields (their names are listed in Appendix 21). We have informed ourselves concerning the needs for translation, considered the evaluation of translations, and compared the capabilities of machines and human beings in translation and in other language processing functions.

The beginning of AI winter ...

1980s: Statistical Machine Translation

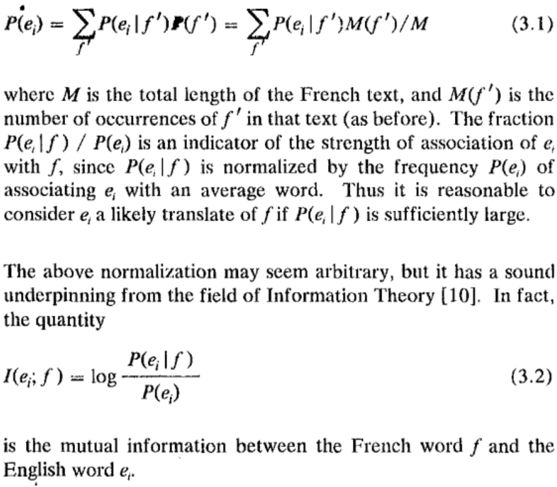


A STATISTICAL APPROACH TO LANGUAGE TRANSLATION

P. BROWN, J. COCKE, S. DELLA PIETRA, V. DELLA PIETRA,
F. JELINEK, R. MERCER, and P. ROOSSIN

IBM Research Division
T.J. Watson Research Center
Department of Computer Science
P.O. Box 218
Yorktown Heights, N.Y. 10598

1980s: Statistical Machine Translation (II)



$$P(e_i) = \sum_{f'} P(e_i \mid f')P(f') = \sum_{f'} P(e_i \mid f')M(f')/M \qquad (3.1)$$

where M is the total length of the French text, and $M(f')$ is the number of occurrences of f' in that text (as before). The fraction $P(e_i \mid f) \;/\; P(e_i)$ is an indicator of the strength of association of e_i with f, since $P(e_i \mid f)$ is normalized by the frequency $P(e_i)$ of associating e_i with an average word. Thus it is reasonable to consider e_i a likely translate of f if $P(e_i \mid f)$ is sufficiently large.

The above normalization may seem arbitrary, but it has a sound underpinning from the field of Information Theory [10]. In fact, the quantity

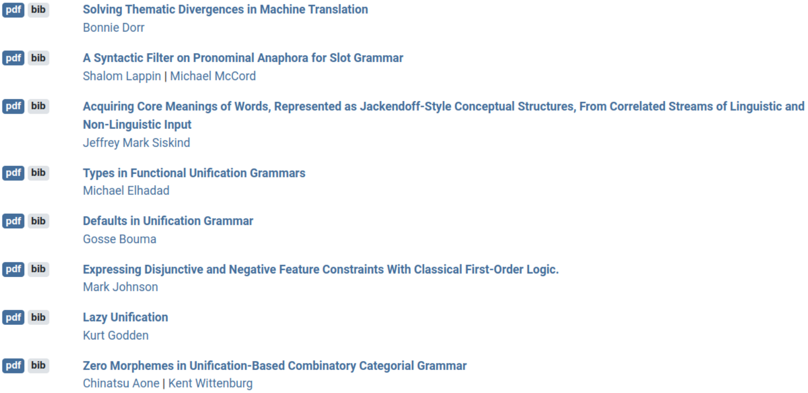
$$I(e_i; f) = \log \frac{P(e_i \mid f)}{P(e_i)} \qquad (3.2)$$

is the mutual information between the French word f and the English word e_i.

1988: The Famous Jelinek Quote

- “Every time I fire a linguist, the performance of our speech recognition system goes up.”
- “Anytime a linguist leaves the group the recognition rate goes up” (recalled by himself)

ACL 1990: What NLP looked like?



pdf bib **Solving Thematic Divergences in Machine Translation**
Bonnie Dorr

pdf bib **A Syntactic Filter on Pronominal Anaphora for Slot Grammar**
Shalom Lappin | Michael McCord

pdf bib **Acquiring Core Meanings of Words, Represented as Jackendoff-Style Conceptual Structures, From Correlated Streams of Linguistic and Non-Linguistic Input**
Jeffrey Mark Siskind

pdf bib **Types in Functional Unification Grammars**
Michael Elhadad

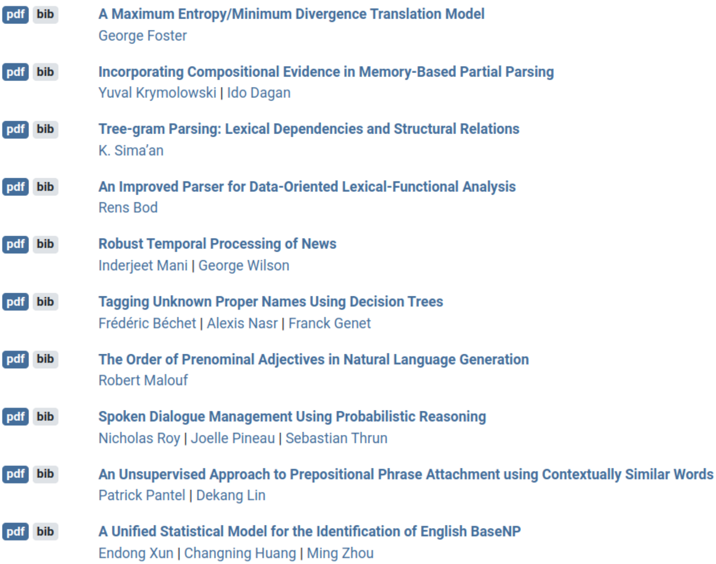
pdf bib **Defaults in Unification Grammar**
Gosse Bouma

pdf bib **Expressing Disjunctive and Negative Feature Constraints With Classical First-Order Logic.**
Mark Johnson

pdf bib **Lazy Unification**
Kurt Godden

pdf bib **Zero Morphemes in Unification-Based Combinatory Categorial Grammar**
Chinatsu Aone | Kent Wittenburg

ACL 2000: What NLP looked like?



A Maximum Entropy/Minimum Divergence Translation Model
George Foster

Incorporating Compositional Evidence in Memory-Based Partial Parsing
Yuval Krymolowski | Ido Dagan

Tree-gram Parsing: Lexical Dependencies and Structural Relations
K. Sima'an

An Improved Parser for Data-Oriented Lexical-Functional Analysis
Rens Bod

Robust Temporal Processing of News
Inderjeet Mani | George Wilson

Tagging Unknown Proper Names Using Decision Trees
Frédéric Béchet | Alexis Nasr | Franck Genet

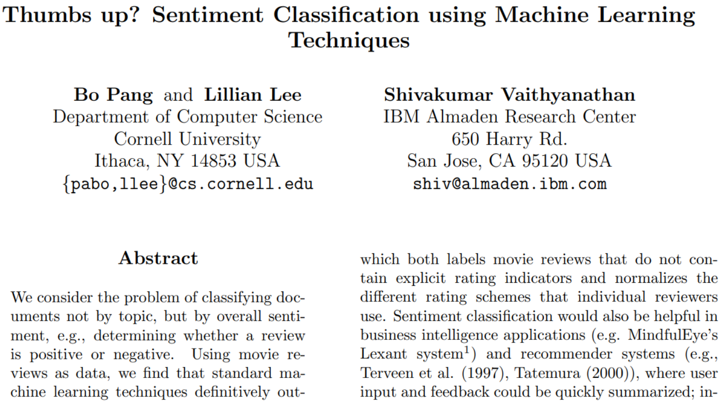
The Order of Prenominal Adjectives in Natural Language Generation
Robert Malouf

Spoken Dialogue Management Using Probabilistic Reasoning
Nicholas Roy | Joelle Pineau | Sebastian Thrun

An Unsupervised Approach to Prepositional Phrase Attachment using Contextually Similar Words
Patrick Pantel | Dekang Lin

A Unified Statistical Model for the Identification of English BaseNP
Endong Xun | Changning Huang | Ming Zhou

2002: Sentiment Analysis



Thumbs up? Sentiment Classification using Machine Learning Techniques

Bo Pang and **Lillian Lee**
Department of Computer Science
Cornell University
Ithaca, NY 14853 USA
{pabo,llee}@cs.cornell.edu

Shivakumar Vaithyanathan
IBM Almaden Research Center
650 Harry Rd.
San Jose, CA 95120 USA
shiv@almaden.ibm.com

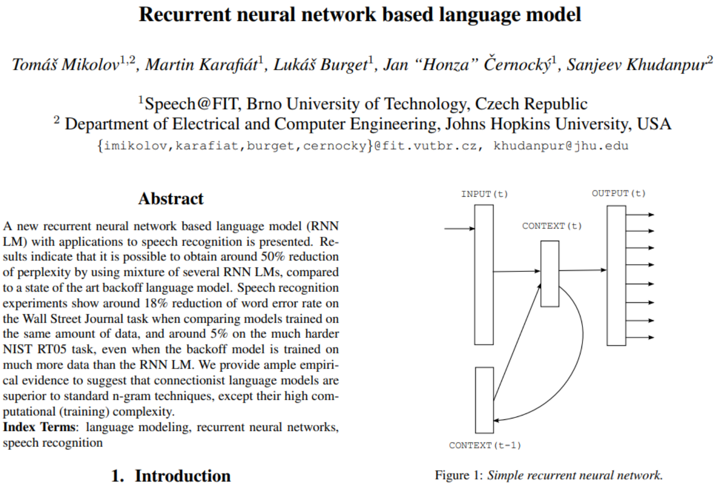
Abstract

We consider the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, we find that standard machine learning techniques definitively out-

which both labels movie reviews that do not contain explicit rating indicators and normalizes the different rating schemes that individual reviewers use. Sentiment classification would also be helpful in business intelligence applications (e.g. MindfulEye's Lexant system[1]) and recommender systems (e.g., Terveen et al. (1997), Tatemura (2000)), where user input and feedback could be quickly summarized; in-

A strong use case of how classification can be used in NLP.

INTERSPEECH 2010: RNN Language Models



Recurrent neural network based language model

Tomáš Mikolov[1,2], *Martin Karafiát*[1], *Lukáš Burget*[1], *Jan "Honza" Černocký*[1], *Sanjeev Khudanpur*[2]

[1]Speech@FIT, Brno University of Technology, Czech Republic
[2] Department of Electrical and Computer Engineering, Johns Hopkins University, USA
{imikolov,karafiat,burget,cernocky}@fit.vutbr.cz, khudanpur@jhu.edu

Abstract

A new recurrent neural network based language model (RNN LM) with applications to speech recognition is presented. Results indicate that it is possible to obtain around 50% reduction of perplexity by using mixture of several RNN LMs, compared to a state of the art backoff language model. Speech recognition experiments show around 18% reduction of word error rate on the Wall Street Journal task when comparing models trained on the same amount of data, and around 5% on the much harder NIST RT05 task, even when the backoff model is trained on much more data than the RNN LM. We provide ample empirical evidence to suggest that connectionist language models are superior to standard n-gram techniques, except their high computational (training) complexity.

Index Terms: language modeling, recurrent neural networks, speech recognition

1. Introduction

Figure 1: *Simple recurrent neural network.*

NeurIPS 2013: Word Embeddings

The idea of pre-training realized in NLP

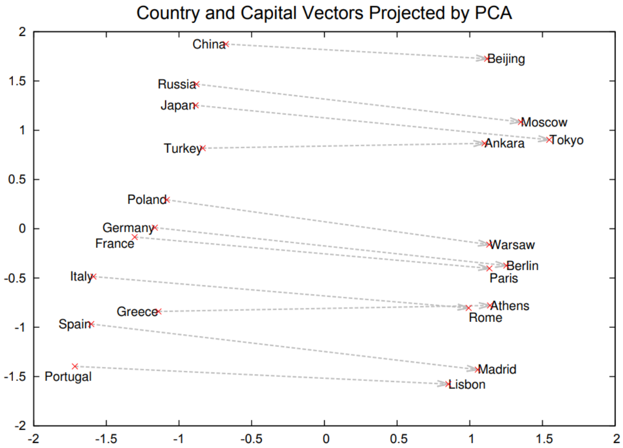
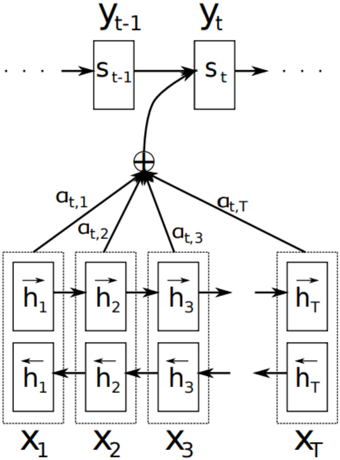


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

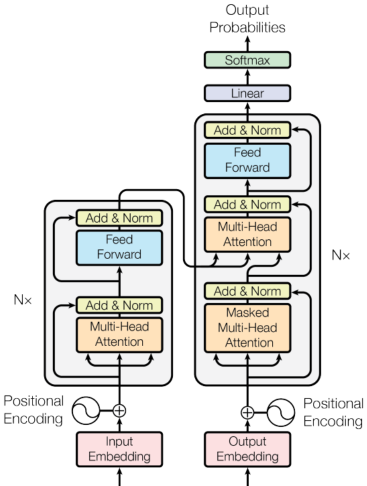
ICLR 2015: Neural MT and Attention Mechanism



The attention mechanism allows the model to focus on different parts of the input sequence, improving translation quality.

2017: Attention is All You Need

The Transformer model was introduced, revolutionizing NLP by enabling more efficient and effective processing of sequential data.

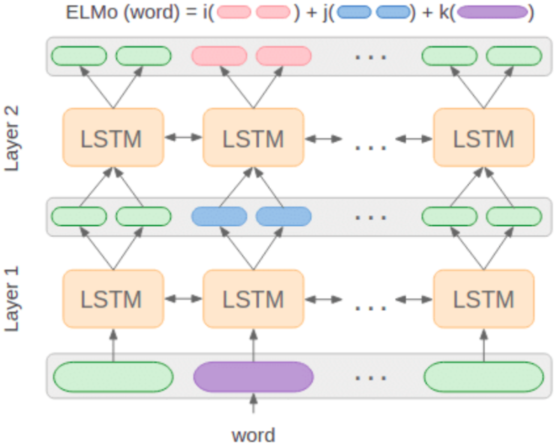


2018

The beginning of large-scale pre-trained models in NLP

NAACL 2018: Embeddings from Language Model (ELMo)

The idea of pre-training came back, together with language models



2018: Generative Pre-trained Transformer (GPT)

This is the first model of the GPT family

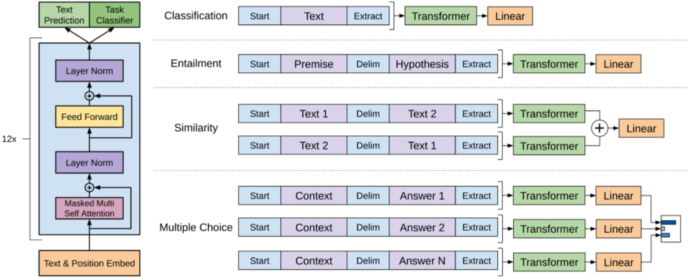


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

2018: Pretrained Deep Bidirectional Transformers (BERT)

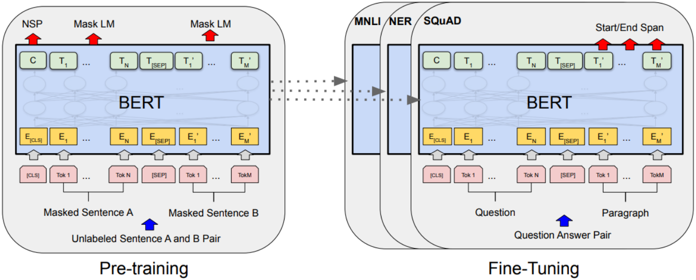
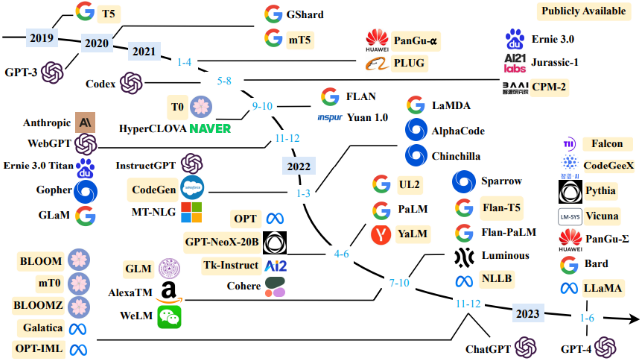


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

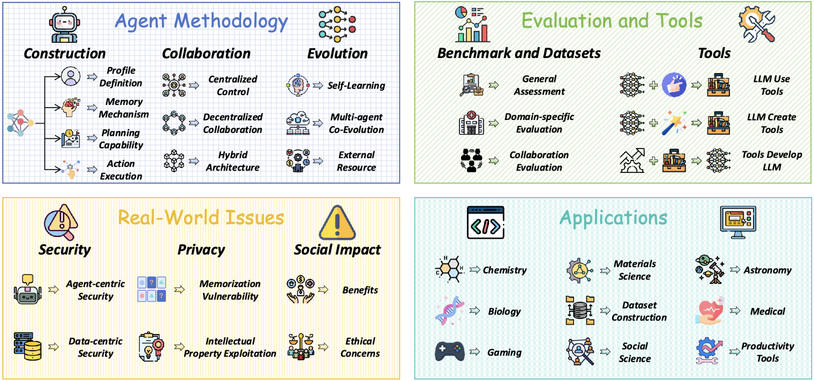
LLM Arms Race (2019 – 2023)

A period marked by rapid advancements and competitive development in large language models.



The Boom of LLM Agents (2024 –)

The rapid development and deployment of LLM agents have transformed the landscape of natural language processing, enabling a wide range of applications from chatbots to automated content generation.



Agent Methodology

Construction: Profile Definition; Memory Mechanism; Planning Capability; Action Execution

Collaboration: Centralized Control; Decentralized Collaboration; Hybrid Architecture

Evolution: Self-Learning; Multi-agent Co-Evolution; External Resource

Evaluation and Tools

Benchmark and Datasets: General Assessment; Domain-specific Evaluation; Collaboration Evaluation

Tools: LLM Use Tools; LLM Create Tools; Tools Develop LLM

Real-World Issues

Security: Agent-centric Security; Data-centric Security

Privacy: Memorization Vulnerability; Intellectual Property Exploitation

Social Impact: Benefits; Ethical Concerns

Applications

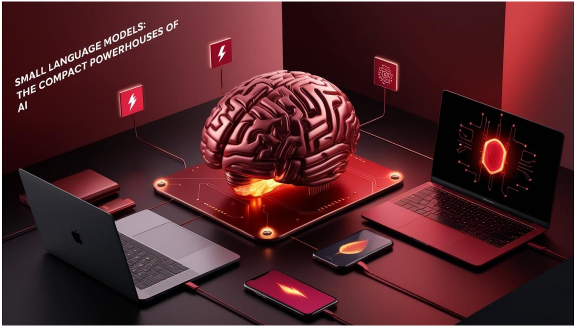
Chemistry; Materials Science; Astronomy

Biology; Dataset Construction; Medical

Gaming; Social Science; Productivity Tools

The Rise of Small LLMs (2025 –)

The emergence of smaller, efficient language models has made advanced NLP capabilities more accessible, allowing for deployment in resource-constrained environments and fostering innovation in various applications.



Link