Natural Language Processing

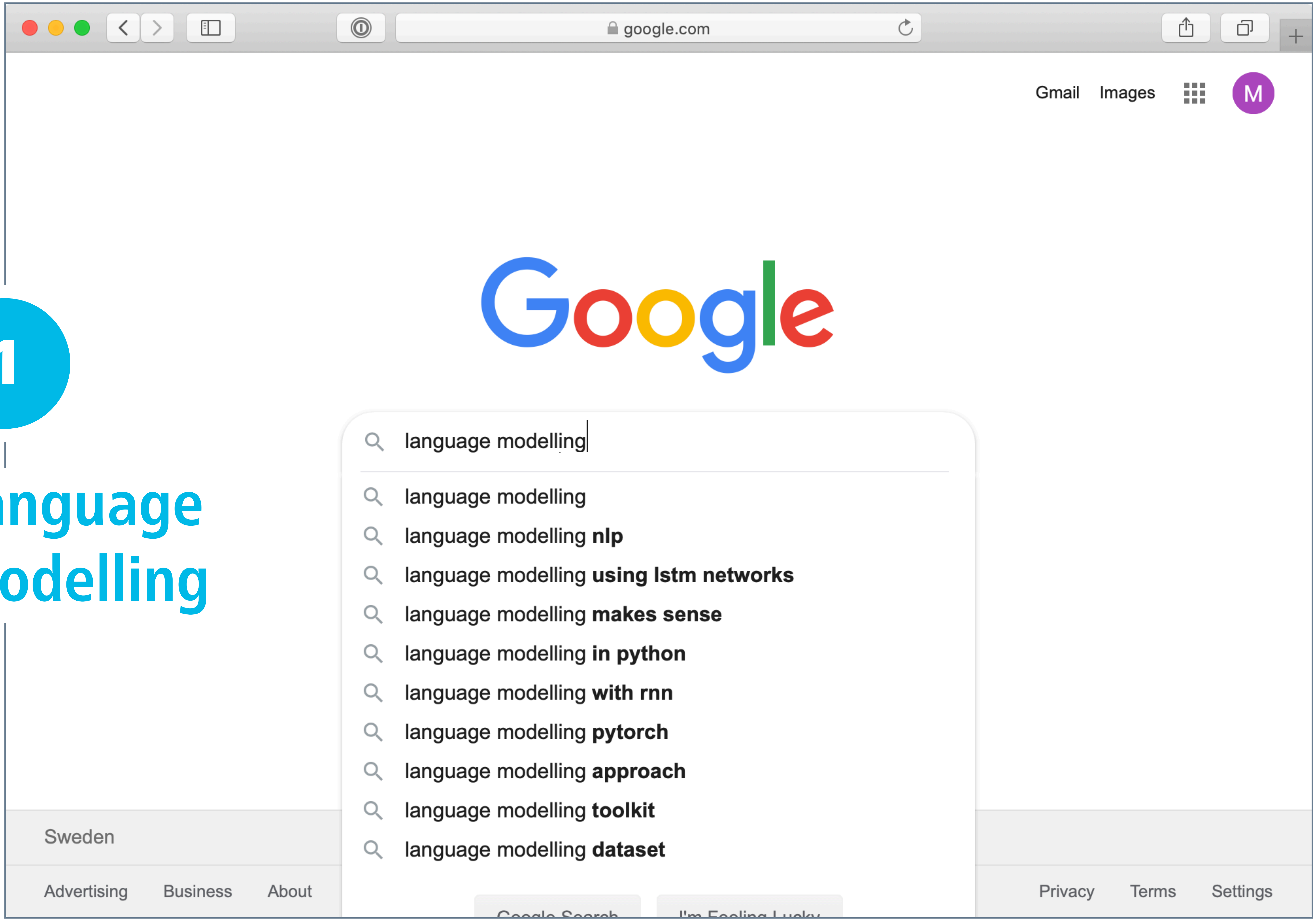
Course overview

Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

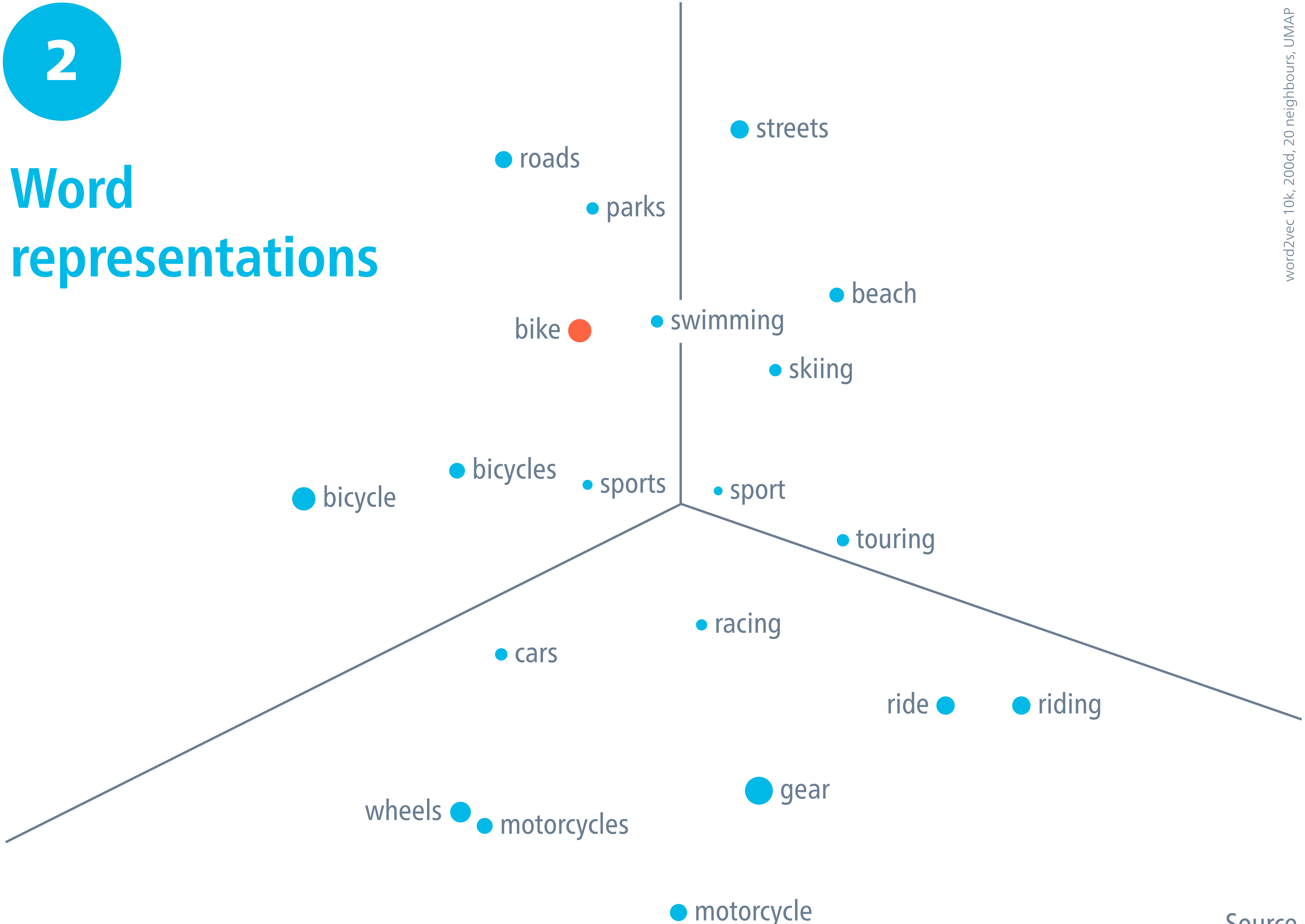
1

Language modelling



2

Word representations

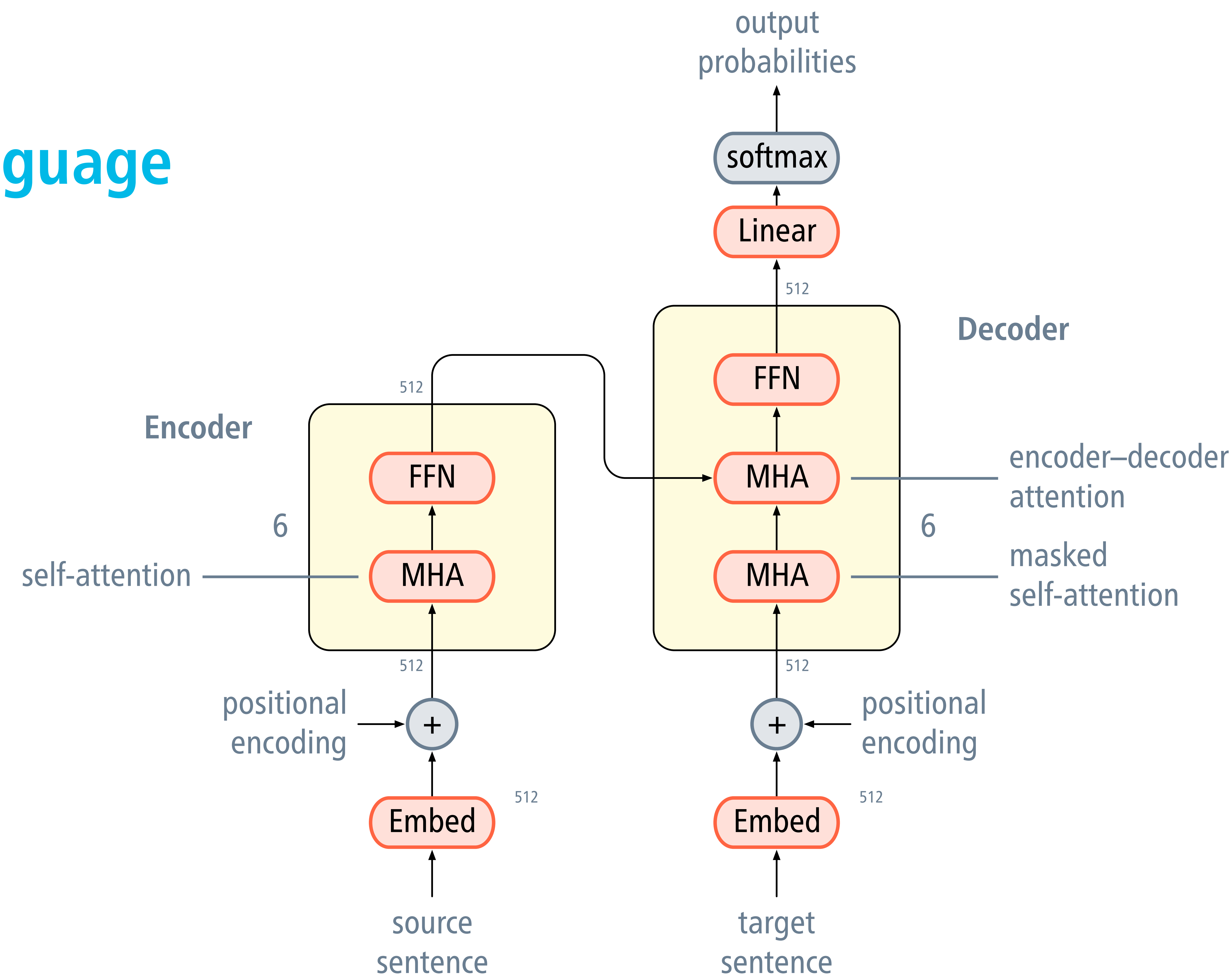


word2vec 10k, 200d, 20 neighbours, UMAP

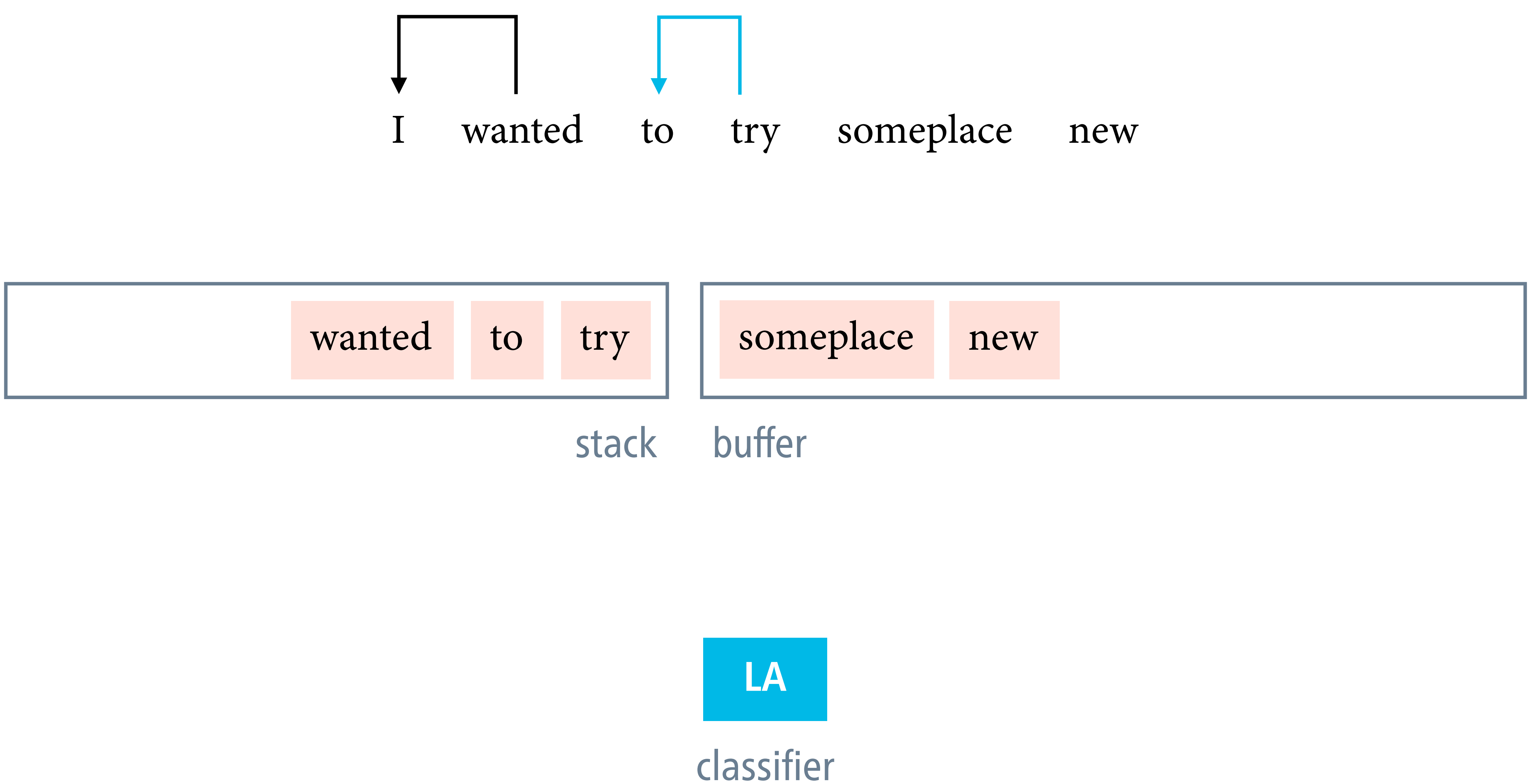
Source

3

Large language models



4 Structured prediction



5 Current research

Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabiliti[…] LLMs through a Global Scale Prompt Hacking Competition

Sander Schulhoff[1*] Jeremy Pinto[2*] Anaum Khan[1] Louis-François Bouchard[2,3] […] Svetlina Anati[5**] Valen Tagliabue[6**] Anson Liu Kost[7**] Christopher Carnahan[…] Jordan Boyd-Graber[1]

Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning

Lean Wang[†,§], Lei Li[†], Damai Dai[†], Deli Chen[§], Hao Zhou[§], Fandong Meng[§], Jie Zhou[§], Xu Sun[†]
[†]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[§]Pattern Recognition Center, WeChat AI, Tencent Inc., China
{lean,daidamai,xusun}@pku.edu.cn nlp.lilei@gmail.com
victorchen@deepseek.com {tuxzhou,fandongmeng,withtomzhou}@tencent.com

Abstract

In-context learning (ICL) emerges as a promising capability of large language models (LLMs) by providing them with demonstration examples to perform diverse tasks. However, the underlying mechanism of how LLMs learn from the provided context remains under-explored. In this paper, we investigate the working mechanism of ICL through an information flow lens. Our findings reveal that label words in the demonstration examples function as anchors: (1) semantic information aggregates into label word representations during the shallow computation layers' processing; (2) the consolidated information in label words serves as a reference for LLMs' final predictions. Based on these insights, we introduce an anchor re-weighting method to improve ICL performance, a demonstration compression technique to expedite inference, and an analysis framework for diagnosing ICL errors in GPT2-XL. The promising applications of our findings again validate the uncovered ICL working mechanism and pave the way for future studies.[1]

Figure 1: Visualization of the information flow in a GPT model performing ICL. The line depth reflects the significance of the information flow from the right word to the left. The flows involving label words are highlighted. Label words gather information from demonstrations in shallow layers, which is then extracted in deep layers for final prediction.

1 Introduction

In-context Learning (ICL) has emerged as a powerful capability alongside the development of scaled-up large language models (LLMs) (Brown et al., 2020). By instructing LLMs using few-shot demonstration examples, ICL enables them to perform a wide range of tasks, such as text classification (Min et al., 2022a) and mathematical reasoning (Wei et al., 2022). Since ICL does not require updates to millions or trillions of model parameters and relies on human-understandable natural language instructions (Dong et al., 2023), it has become a promising approach for harnessing the full potentiality of LLMs. Despite its significance, the inner working mechanism of ICL remains an open question, garnering considerable interest from research communities (Xie et al., 2022; Dai et al., 2022; Akyürek et al., 2022; Li et al., 2023b).

In this paper, we find that the label words serve as anchors that aggregate and distribute information in ICL. We first visualize the attention interactive pattern between tokens with a GPT model (Brown et al., 2020) on sentiment analysis (Figure 1). Initial observations suggest that label words aggregate information in shallow layers and distribute it in deep layers.[2] To draw a clearer picture of this phenomenon, we design two metrics based on saliency

[1]https://github.com/lancopku/label-words-are-anchors

[2]In this paper, "shallow" or "first" layers refer to those closer to the input, while "deep" or "last" layers are closer to the output. Here, "deep layers" include those around the midpoint, e.g., layers 25-48 in a 48-layer GPT2-XL.

9840

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9840–9855
December 6-10, 2023 ©2023 Association for Computational Linguistics

[…]ster Minimum Bayes Risk Decoding with Confidence-based Pruning

Julius Cheng, Andreas Vlachos
Department of Computer Science and Technology
University of Cambridge
{jncc3,av308}@cam.ac.uk

Search and learning

predicted output

input

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\ \mathrm{score}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

candidate output

model parameters

Eisenstein (2019), § 1.2.2

Search and learning

Eisenstein (2019), § 1.2.2

- **Search module**

  The search module is responsible for finding the candidate output $\boldsymbol{y}$ with the highest score relative to the input $\boldsymbol{x}$.

  requires efficient algorithms

- **Learning module**

  The learning module is responsible for finding the model parameters $\boldsymbol{\theta}$ that maximize the predictive performance.

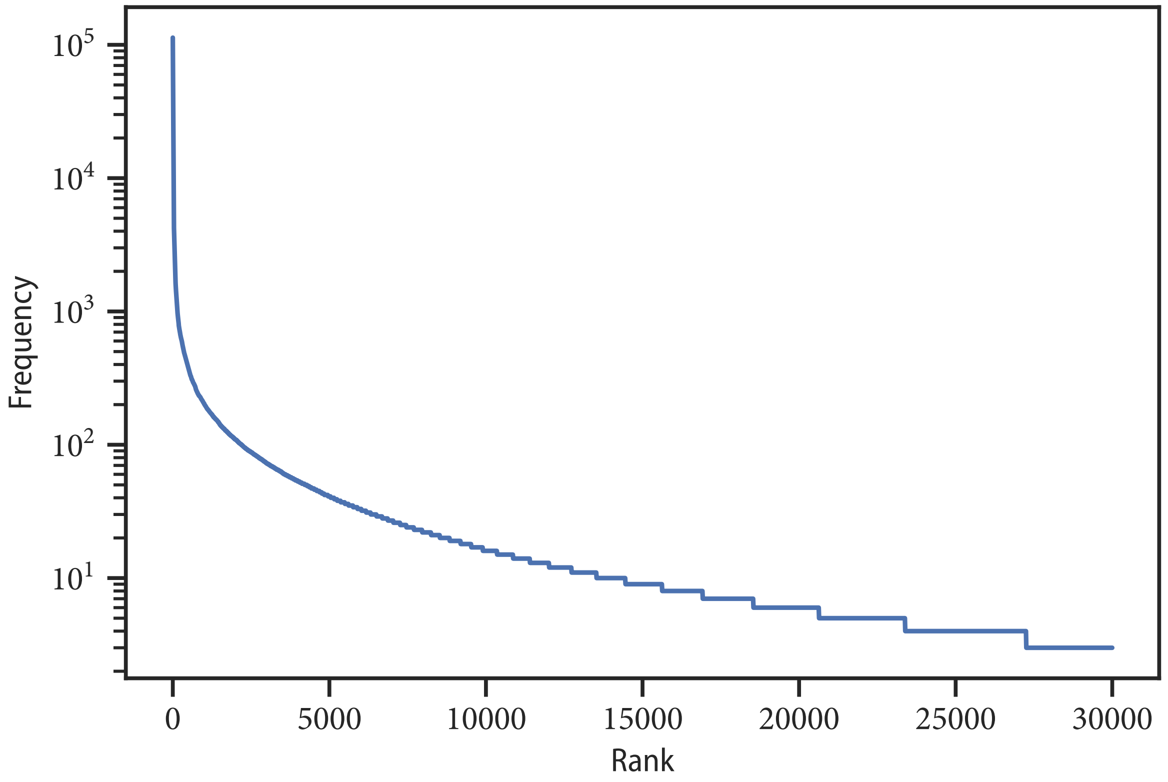
  for example, using supervised machine learning

Language is special

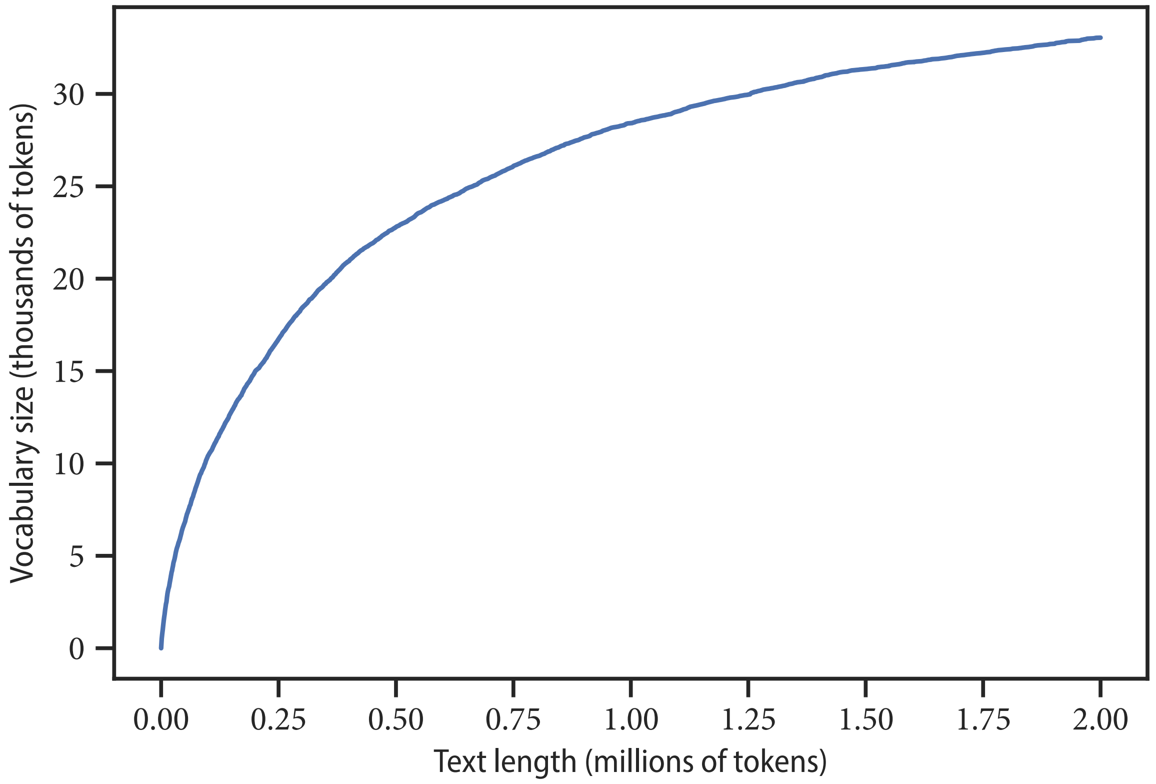
Eisenstein (2019), § 1.1

- Unlike images or audio, text data is fundamentally discrete, with meaning created by combinatorial arrangement.
- Even though text appears as a sequence, machine learning methods must account for its implicit hierarchical structure.
- The distribution of linguistic elements follows a power law – algorithms must be robust to unobserved events.

Zipf’s law and Heaps’ law



Zipf’s law



Heaps’ law