Deep Learning for Natural Language Processing

Introduction to Natural Language Processing

Marco Kuhlmann
Department of Computer and Information Science

What is natural language processing?

- **Natural language processing** develops methods for making human language accessible to computers.
- Some well-known example applications are smart search engines, machine translation, and dialogue systems.
- These diverse applications are based on a common set of ideas from algorithms, machine learning, and other disciplines.

Eisenstein (2019), § 1

JEOPARDY!®

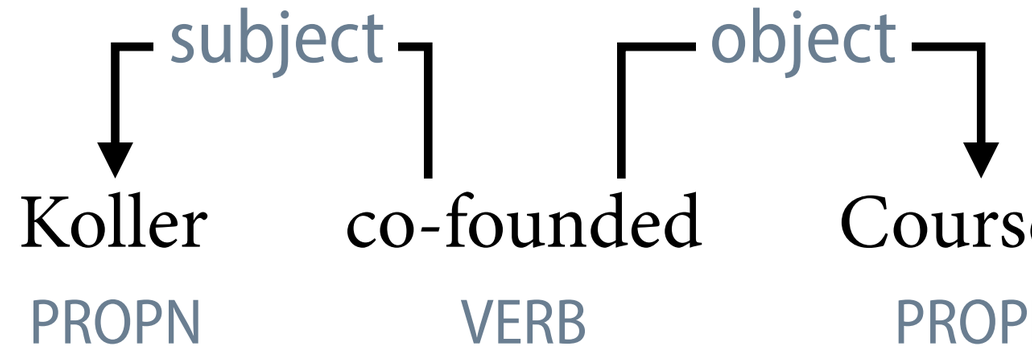
This Stanford University alumna co-founded educational technology company Coursera.

Source: Wikipedia

SPARQL query against DBPedia

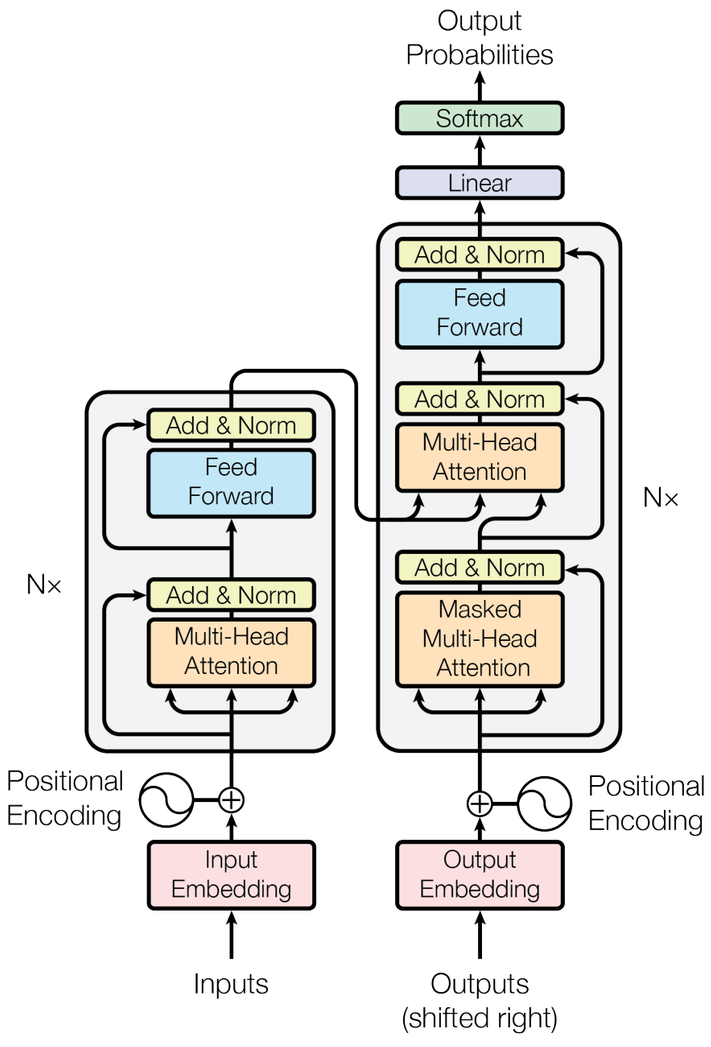
```
SELECT DISTINCT ?x WHERE {
  ?x dbo:almaMater dbr:Stanford_University.
  dbr:Coursera dbo:foundedBy ?x.
}
```

General-purpose linguistic representations

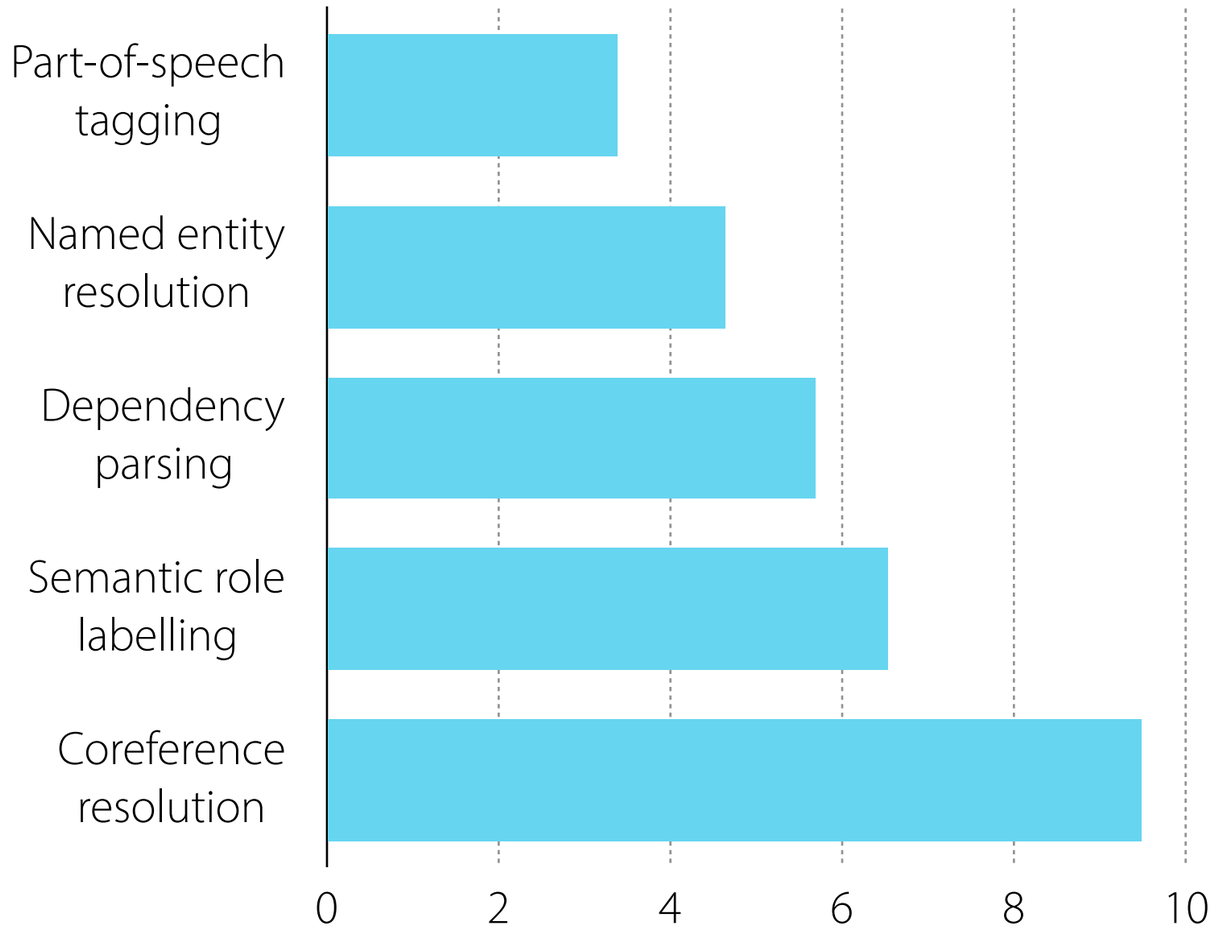


```
dbr:Coursera dbo:foundedBy dbr:Daphne_Koller
```

‘Natural language processing from scratch’



Vaswani et al. (2017)



Tenney et al. (2019)

Two paradigms

Eisenstein (2019), § 1.2.1

- **Linguistic knowledge**

  Build pipelines of modular components that produce general-purpose representations grounded in linguistic knowledge.

  morphemes, parts-of-speech, dependency trees, meaning representations

- **Deep learning**

  Train end-to-end neural networks that directly transmute raw text into whatever structure the desired application requires.

Search and learning

predicted output

input

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

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.

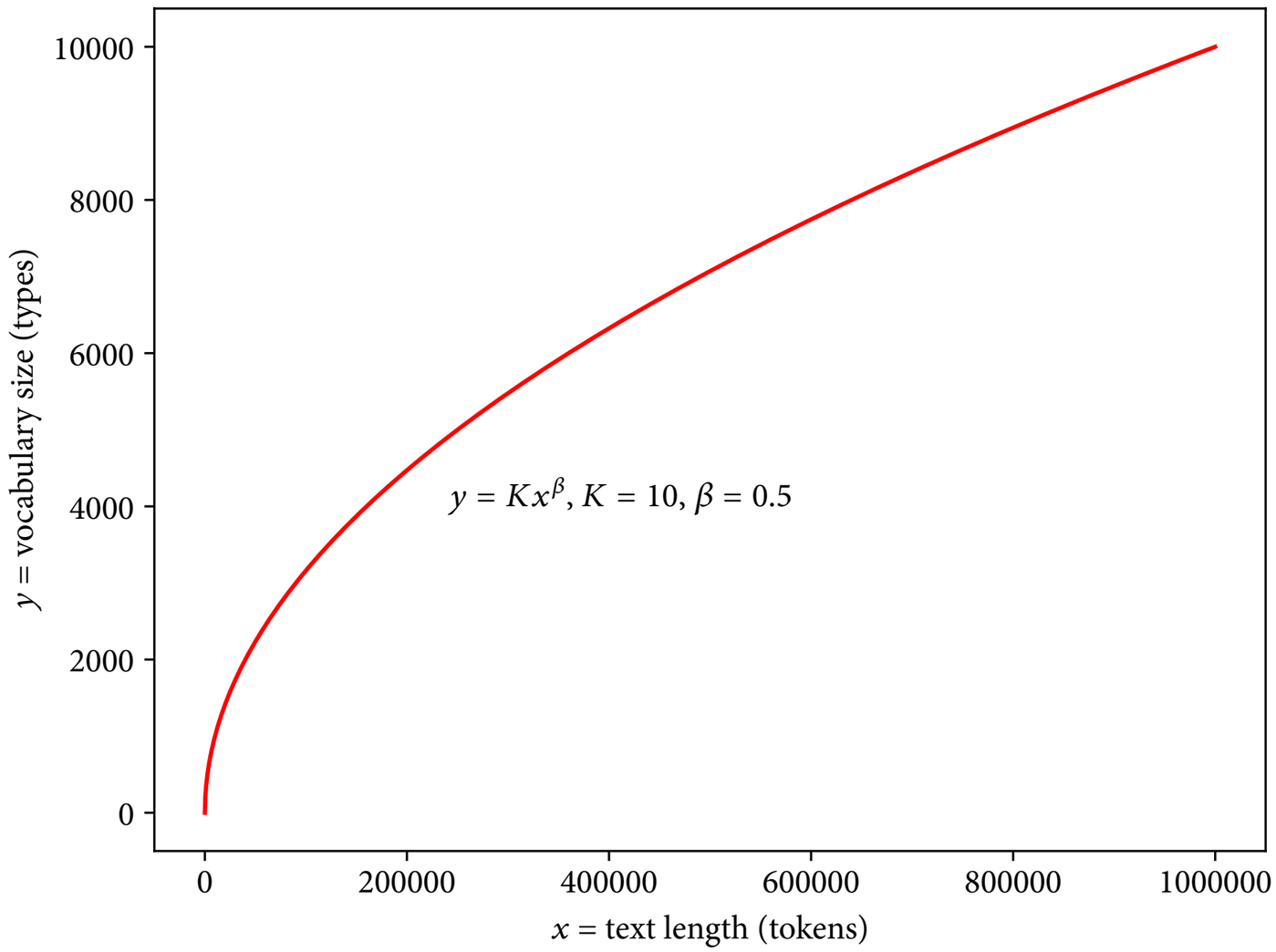
  for example, using supervised machine learning

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 recursive structure.
- The distribution of linguistic elements resembles that of a power law – algorithms must be robust to unobserved events.

Heaps' law


10000
8000
6000
4000
2000
0
y = vocabulary size (types)
$y = Kx^{\beta}$, $K = 10$, $\beta = 0.5$
0
200000
400000
600000
800000
1000000
x = text length (tokens)