

Constructing Knowledge Graphs from Web Data

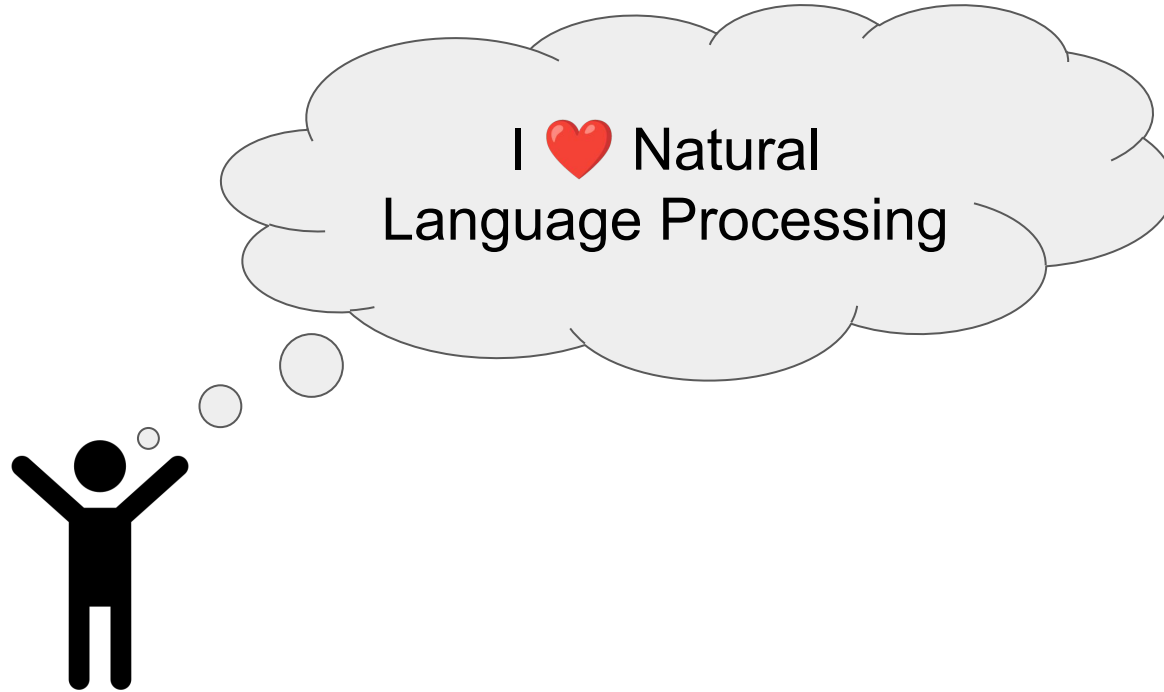
Aidan San

Motivation

A student is trying to build their own online learning curriculum



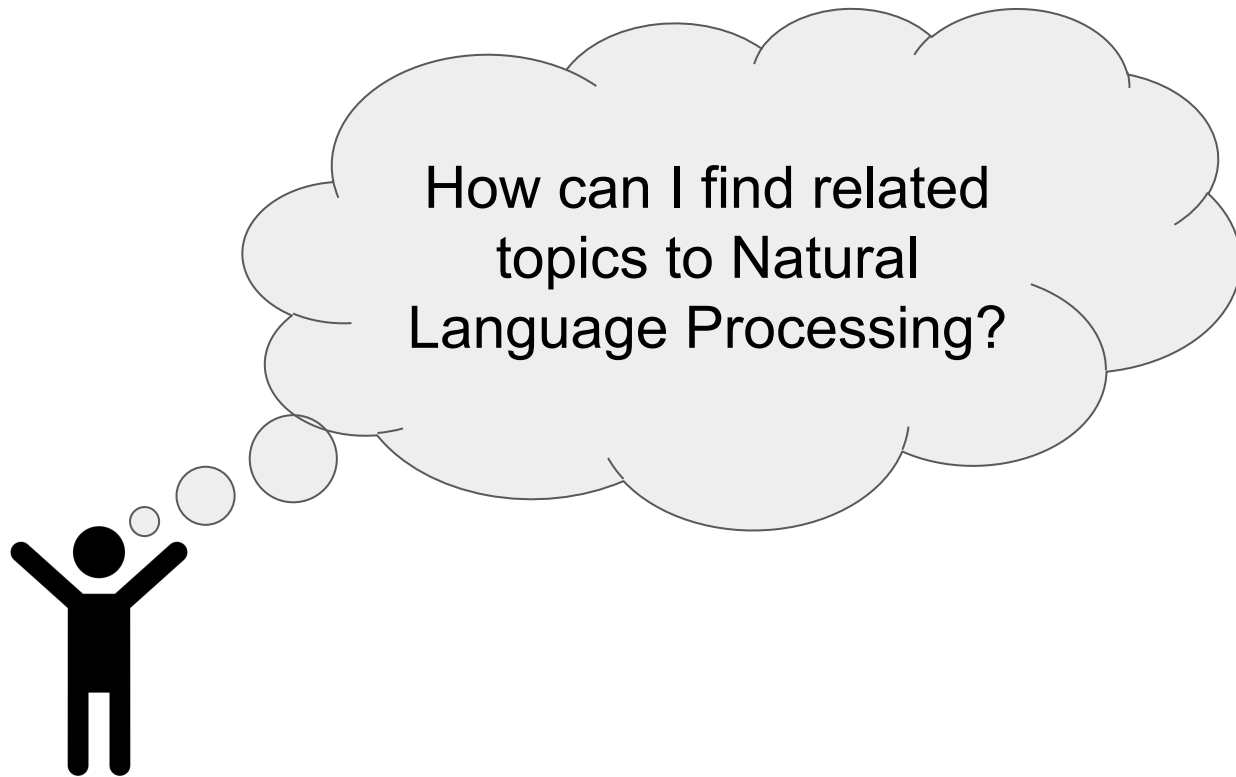
Motivation



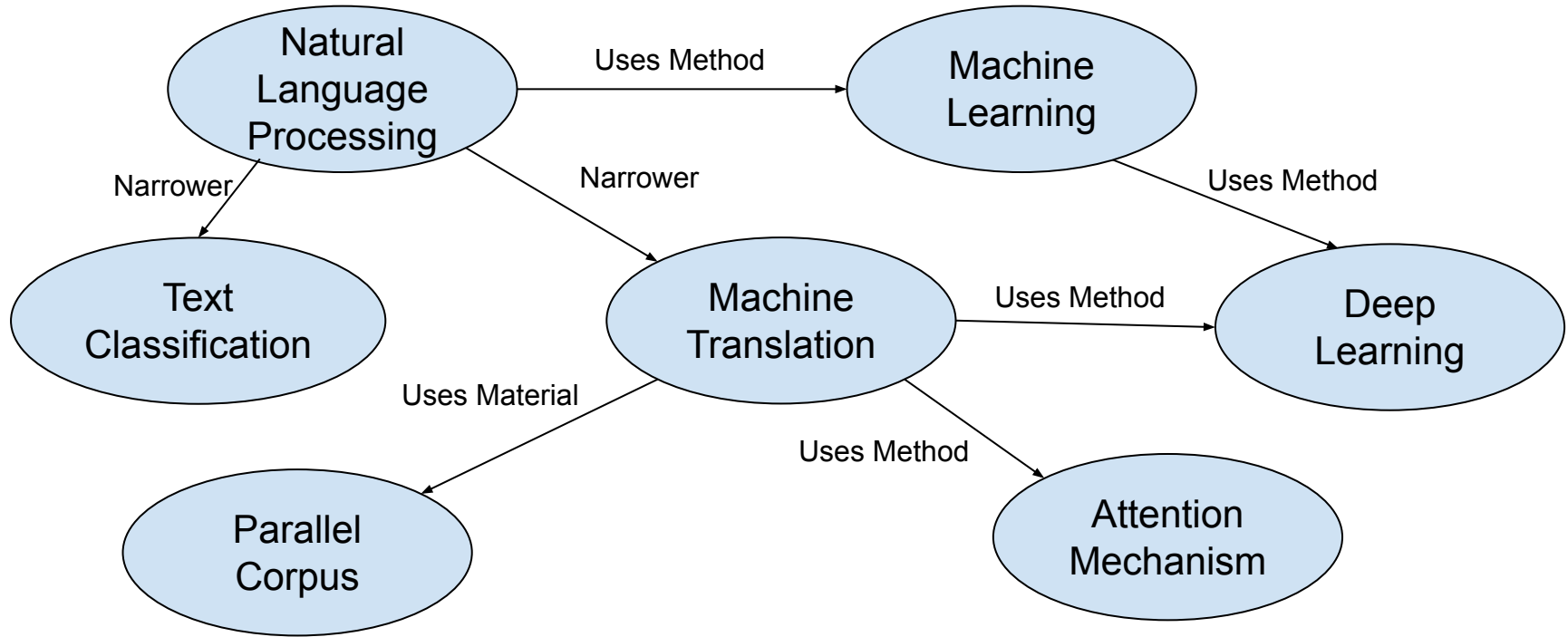
Motivation



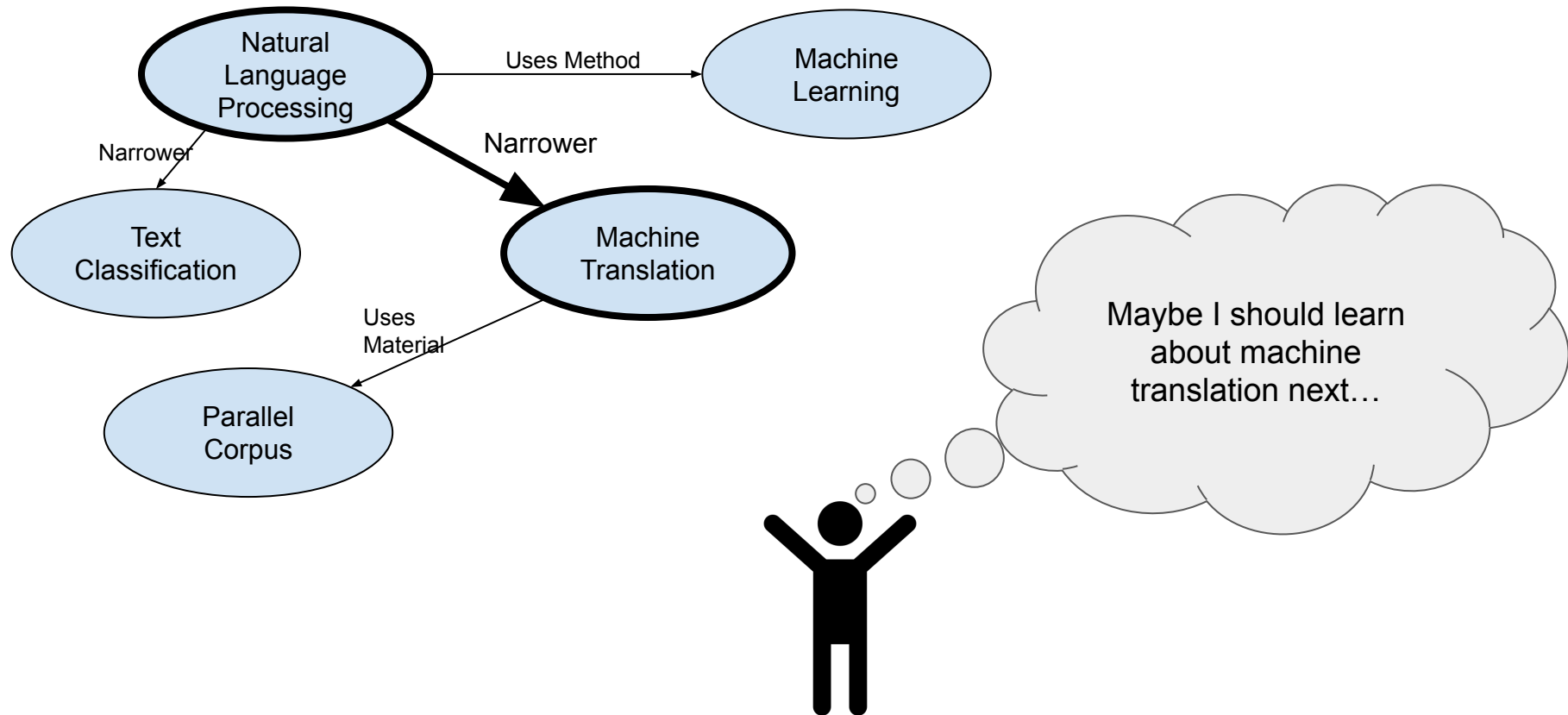
Motivation



Knowledge Graphs



Motivation

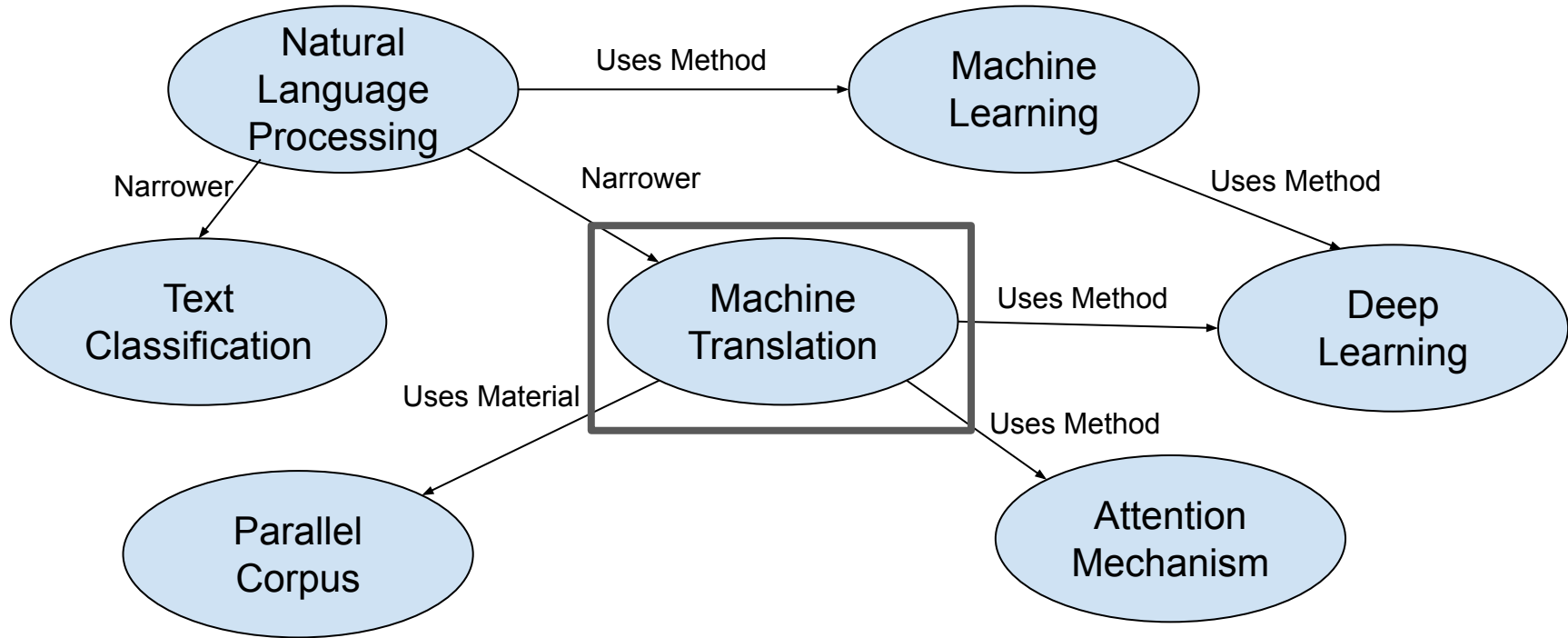


Outline

1. What are Knowledge Graphs?
2. How are Knowledge Graphs constructed?
3. What are the ethical implications?

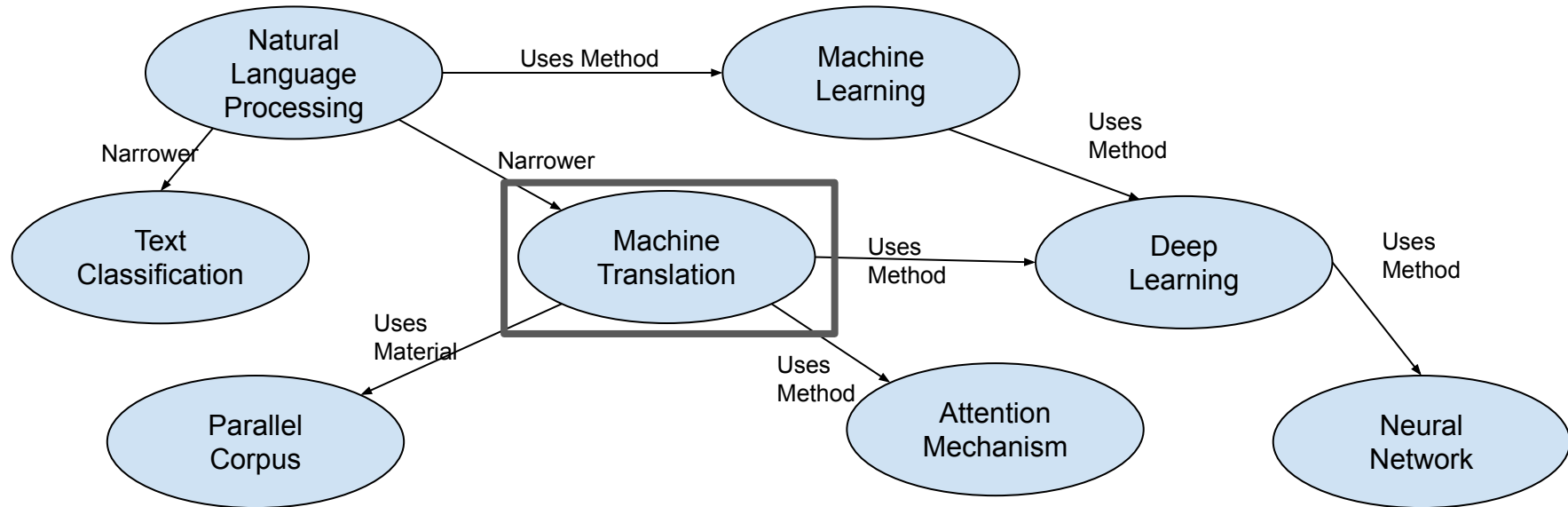
What are Knowledge Graphs?

What are Knowledge Graphs? - Entities

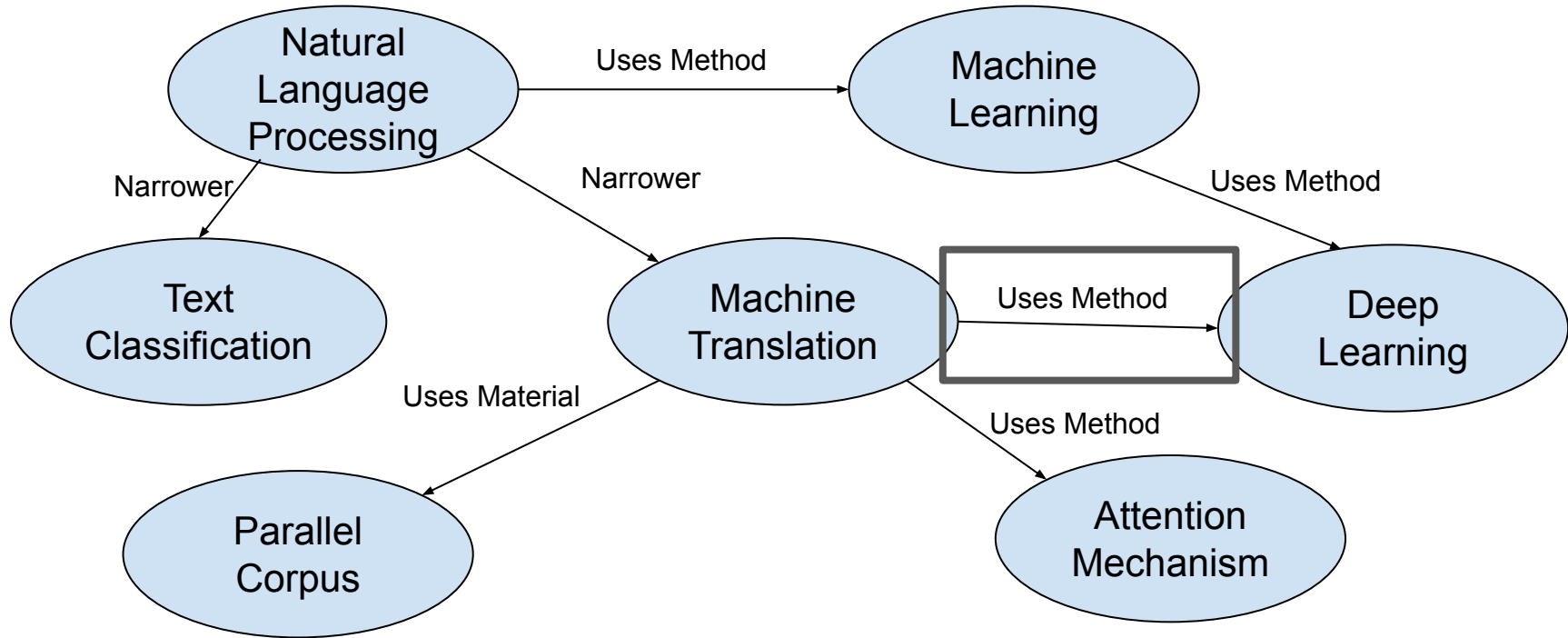


What are Knowledge Graphs? - Entities

- A concept
- Example: *Machine translation*
- Vertices in the graph

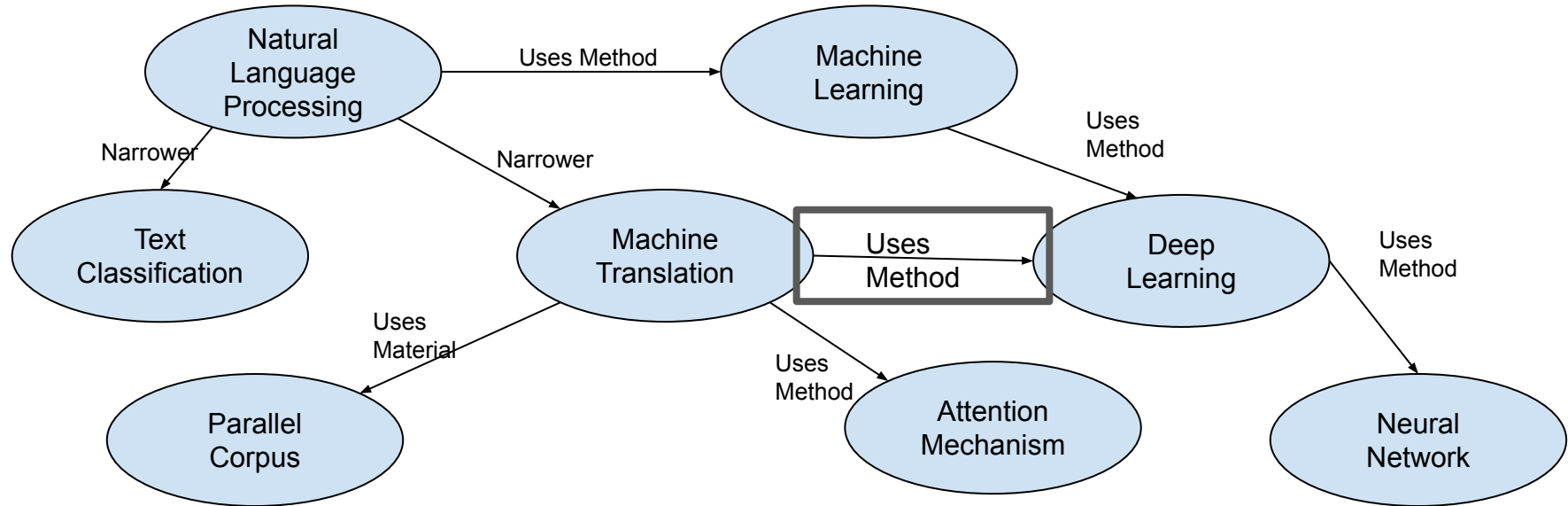


What are Knowledge Graphs? - Relations



What are Knowledge Graphs? - Relations

- Connects two entities in a graph
- Example: *Uses Method*
- Edges in the graph



How can we construct Knowledge Graphs?

Can we build KGs from course websites?

Schedule

Updated on Aug. 24, 2021

- **Week 1:** Aug. 25
 - Topics: Course information, review of linear algebra
 - Course Materials: [Slides](#)
- **Week 2:** Aug. 30, Sept. 1
 - Topics: Review of probability and statistics
- **Week 3:** Sept. 6, 8
 - Topics: Text classification, logistic regression
 - Readings: JE 02
 - Course Materials: [Slides](#); [Colab Code](#)

Can we build KGs from course websites?

Schedule

Updated on Aug. 24, 2021

- **Week 1:** Aug. 25
 - Topics: Course information, review of linear algebra
 - Course Materials: [Slides](#)
- **Week 2:** Aug. 30, Sept. 1
 - Topics: Review of probability and statistics
- **Week 3:** Sept. 6, 8
 - Topics: Text classification, logistic regression
 - Readings: JE 02
 - Course Materials: [Slides](#); [Colab Code](#)

We can utilize the page structure information

Segment Page

Schedule

Updated on Aug. 24, 2021

- **Week 1:** Aug. 25
 - Topics: Course information, review of linear algebra
 - Course Materials: [Slides](#)
- **Week 2:** Aug. 30, Sept. 1
 - Topics: Review of probability and statistics
- **Week 3:** Sept. 6, 8
 - Topics: Text classification, logistic regression
 - Readings: JE 02
 - Course Materials: [Slides](#); [Colab Code](#)

Extract Fields

Schedule

Updated on Aug. 24, 2021

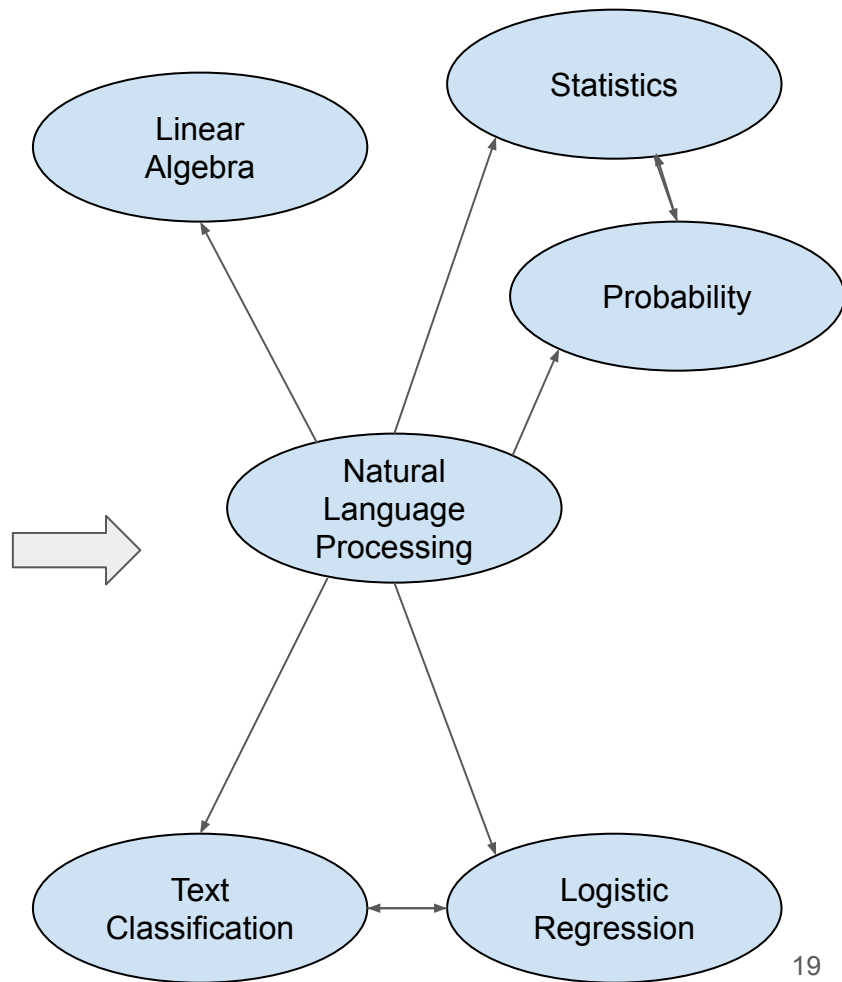
- **Week 1:** Aug. 25
 - Topics: Course information, review of linear algebra
 - Course Materials: [Slides](#)
- **Week 2:** Aug. 30, Sept. 1
 - Topics: Review of probability and statistics
- **Week 3:** Sept. 6, 8
 - Topics: Text classification, logistic regression
 - Readings: JE 02
 - Course Materials: [Slides](#); [Colab Code](#)

Build Graph

Schedule

Updated on Aug. 24, 2021

- **Week 1:** Aug. 25
 - Topics: Course information, review of linear algebra
 - Course Materials: [Slides](#)
- **Week 2:** Aug. 30, Sept. 1
 - Topics: Review of probability and statistics
- **Week 3:** Sept. 6, 8
 - Topics: Text classification, logistic regression
 - Readings: JE 02
 - Course Materials: [Slides](#); [Colab Code](#)



Extract Fields

Schedule

Updated on Aug. 24, 2021

- **Week 1:** Aug. 25
 - Topics: Course information, review of linear algebra
 - Course Materials: [Slides](#)
- **Week 2:** Aug. 30, Sept. 1
 - Topics: Review of probability and statistics
- **Week 3:** Sept. 6, 8
 - Topics: Text classification, logistic regression
 - Readings: JE 02
 - Course Materials: [Slides](#); [Colab Code](#)

Can't we write rules to do the extraction?

e.g.

Take the text after “Topics:”

Different course pages have different formats

uva-nlp-course

UvA CS 4501 Machine Learning for NLP

Schedule

Updated on Aug. 24, 2021

- **Week 1:** Aug. 25
 - Topics: Course information, review of linear algebra
 - Course Materials: [Slides](#)
- **Week 2:** Aug. 30, Sept. 1
 - Topics: Review of probability and statistics
- **Week 3:** Sept. 6, 8
 - Topics: Text classification, logistic regression
 - Readings: JE 02
 - Course Materials: [Slides](#); [Colab Code](#)
- **Week 4:** Sept. 13, 15
 - Topics: neural network classifiers, feed-forward NNs, the back-propagation algorithm
 - Readings: JE 03
 - Course Materials: [Slides](#); [Colab Code](#)

CS224N Home

[Coursework](#)[Schedule](#)[Office Hours](#)[Final projects](#)[Lecture Videos](#)[E](#)

Schedule

Updated lecture **slides** will be posted here shortly before each lecture. Other links contain last year's slides, which are mostly similar.

Lecture **notes** will be uploaded a few days after most lectures. The notes (which cover approximately the first half of the course content) give detail beyond the lectures.

Date	Description	Course Materials	Events	Deadlines
Tue Jan 10	Word Vectors (<i>by John Hewitt</i>) [slides] [notes]	Suggested Readings: 1. Efficient Estimation of Word Representations in Vector Space (original word2vec paper) 2. Distributed Representations of Words and Phrases and their Compositionality (negative sampling paper)	Assignment 1 out [code] [preview]	
	Gensim word vectors example: [code] [preview]			
Thu Jan 12	Word Vectors, Word Window Classification, Language Models [slides] [notes]	Suggested Readings: 1. GloVe: Global Vectors for Word Representation (original GloVe paper) 2. Improving Distributional Similarity with Lessons Learned from Word Embeddings 3. Evaluation methods for unsupervised word embeddings Additional Readings: 1. A Latent Variable Model Approach to PMI-based Word Embeddings 2. Linear Algebraic Structure of Word Senses, with Applications to Polysemy 3. On the Dimensionality of Word Embedding		
Fri Jan 13	Python Review Session [slides] [colab]	⌚ 2:30pm - 3:20pm Gates B03		

We want a model to extract these concepts

uva-nlp-course

Uva CS-4501: Machine Learning for NLP

Schedule

Updated on Aug. 24, 2021

- Week 1: Aug. 25
 - Topics: Course Information, review of linear algebra
 - Course Materials: slides
- Week 2: Aug. 30, Sept. 1
 - Topics: Review of probability and statistics
- Week 3: Sept. 6, 8
 - Topics: Text classification, Logistic regression
 - Readings: 8.02
 - Course Materials: slides, Code Code
- Week 4: Sept. 13, 15
 - Topics: neural network classifiers, feed forward NNs, the
 - Readings: 8.03
 - Course Materials: slides, Code Code
- Week 5: Sept. 20, 22
 - Topics: Further discussion of text classification
 - Readings: 8.04
 - Course Materials: slides
- Week 6: Sept. 27, 29

课程

What is NLP? What will you learn in this class?
What will we teach this class?
What is NLP? What will you learn in this class?
What will we teach this class?
What is NLP? What will you learn in this class?
What will we teach this class?

What is NLP? What will you learn in this class?
What will we teach this class?
What is NLP? What will you learn in this class?
What will we teach this class?
What is NLP? What will you learn in this class?
What will we teach this class?

Schedule

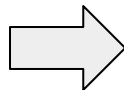
(MP) due 09/30
Wed, 08/25

01 Introduction pdf

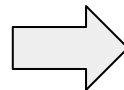
What is NLP? What will you learn in this class?
What will we teach this class?
What is NLP? What will you learn in this class?
What will we teach this class?
What is NLP? What will you learn in this class?
What will we teach this class?

What is NLP? What will you learn in this class?
What will we teach this class?
What is NLP? What will you learn in this class?
What will we teach this class?
What is NLP? What will you learn in this class?
What will we teach this class?

Week	Dates	Theme/Lecture Materials	Readings	Deliverables
0	Fri 1	Introduction	<ul style="list-style-type: none">Jarvis et al. Introduction (Phase)Optional: ChatGPT Foundations (1/1)Optional: ChatGPT Lab: Prompt, Time, Response (1/1)Optional: ChatGPT at "Challenges in Natural Language Processing" (1/1)	
1	Fri 6, 8, 10	Text Processing	<ul style="list-style-type: none">Jarvis et al. Regular Expressions, Text Normalization, Edit DistanceOptional: Resources: "SLD: A Computer Representation for the Study of Natural Language Communication Between Man and Machine" Communications of the ACM, 1964	<ul style="list-style-type: none">HW1: Introduction, Released Mon, Feb 8 Due: Thurs, Feb 9 at 9:00pm
2	Fri 13, 15	Language Modeling and Naïve Bayes	<ul style="list-style-type: none">Jarvis et al. Language Modeling with Naïve BayesJarvis et al. Naïve Bayes and Sentiment ClassificationOptional: Peng et al. "Towards a Sentiment Classification using Machine Learning Techniques" EMNLP 2002	<ul style="list-style-type: none">HW2: Spam vs. Not-Spam Classification, Released Mon, Feb 13 Due: Thurs, Feb 14 at 9:00pm
3	Fri 20, 22, 24	Logistic Regression	<ul style="list-style-type: none">Jarvis et al. Logistic Regression	<ul style="list-style-type: none">HW3: Naïve Bayes, Released Mon, Feb 20 Due: Fri, Feb 21 at 9:00pm ET
4	Fri 27, Sat 1, 3	Vector Semantics	<ul style="list-style-type: none">Jarvis et al. Vector SemanticsOptional: Wang et al. "Efficient Estimation of Word Representations in Vector Space" 2013	<ul style="list-style-type: none">HW4: Logistic Regression, Released Mon, Feb 27 Due: Fri, Mar 1 at 9:00pm ET



Web
Extraction
Model



Extracted Concepts

“Language Models”

“Tokenization”

“Logistic Regression”

We want a model to extract these concepts

uva-nlp-course

Uva CS-4501: Machine Learning for NLP

Schedule

Updated on Aug. 24, 2021

- Week 1: Aug. 25
 - Topic: Course Information, review of linear algebra
 - Course Materials: slides
- Week 2: Aug. 30, Sept. 1
 - Topic: Review of probability and statistics
 - Readings: 8.02
 - Course Materials: slides, CodeCade
- Week 3: Sept. 6, 8
 - Topic: neural network classifiers, feed forward NNs, the
 - Readings: 8.03
 - Course Materials: slides, CodeCade
- Week 4: Sept. 20, 22
 - Topic: Further discussion of text classification
 - Readings: 8.04
 - Course Materials: slides
- Week 5: Sept. 27, 29

课程

课程介绍

- 课程目标
- 课程大纲
- 课程资源
- 课程评价

文本分类

- 课程目标
- 课程大纲
- 课程资源
- 课程评价

语言模型

- 课程目标
- 课程大纲
- 课程资源
- 课程评价

Schedule

(MP) due 09/30
Wed, 08/25

01 Introduction

What is NLP? What will you learn in this class?
What will we teach this class?
Read reading: Ch.1 (and E4).
Read reading: Python tutorial (sec. 1-5).
Read reading: MS, Ch. 2

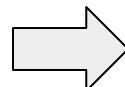
2. word

What is NLP? What will you learn in this class?
What will we teach this class?
Read reading: Ch.1 (and E4).
Read reading: Python tutorial (sec. 1-5).
Read reading: MS, Ch. 2

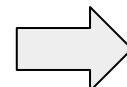
3. word

What is NLP? What will you learn in this class?
What will we teach this class?
Read reading: Ch.1 (and E4).
Read reading: Python tutorial (sec. 1-5).
Read reading: MS, Ch. 2

Week	Dates	Theme & Lecture Materials	Readings	Deliverables
0	Fri 1	Introduction	<ul style="list-style-type: none">JARPA's Introduction (Phase)Optional: ChatGPT Foundation (1/1)Optional: ChatGPT Lab: Prompt, Time, Response (1/1)	
1	Fri 6, Su 8, 10	Text Processing	<ul style="list-style-type: none">JARPA's Regular Expressions, Text Normalization, Edit DistanceOptional: Introduction: "SLD: A Computer Interpreter for the Study of Natural Language Communication Between Man and Machine"Communication of the ACM, 1964	<ul style="list-style-type: none">HW1: IntroductionReleased: Mon, Feb 8Due: Thurs, Feb 9 at 9:00pm
2	Fri 13, Su 15	Language Modeling and Hidden Markov Models	<ul style="list-style-type: none">JARPA's Language Modeling with Markov ModelsJARPA's Hidden Markov Models and Sentiment ClassificationOptional: Peng et al. "Formalizing Sentiment Classification using Deep Learning Techniques"EMNLP 2002	<ul style="list-style-type: none">HW2: Spoken vs. Written SentimentReleased: Mon, Feb 13Due: Thurs, Feb 14 at 9:00pm
3	Fri 20, Su 22, 24	Logistic Regression	<ul style="list-style-type: none">JARPA's Logistic Regression	<ul style="list-style-type: none">HW3: Hidden Markov ModelsReleased: Mon, Feb 20Due: Fri, Feb 24 at 9:00pm
4	Fri 27, Sun 29	Vector Semantics	<ul style="list-style-type: none">JARPA's Vector SemanticsOptional: Word Embedding and Word SimilarityWordNet, 2004Optional: Mikolov et al. "Efficient Estimation of Word Representations in Vector Space, 2013"	<ul style="list-style-type: none">HW4: Logistic RegressionReleased: Mon, Feb 27Due: Fri, Mar 3 at 9:00pm



Web
Extraction
Model



Extracted Concepts

“Language Models”

“Tokenization”

“Logistic Regression”

But we need training
data for the model!

PLAtE - Outline

Can we build a dataset to train web extraction models?

PLAtE - Outline

Can we build a dataset to train web extraction models?

1. Introduction
2. Dataset Construction
3. Modeling

Aidan San, Yuan Zhuang, Jan Bakus, Colin Lockard, David Ciemiewicz, Sandeep Atluri, Kevin Small, Yangfeng Ji, Heba Elfardy. 2023. PLAtE: A Large-scale Dataset for List Page Web Extraction. ***Under Review at the 2023 Conference of the Association for Computational Linguistics: Industry Track***

PLAtE - Introduction

- Want to extract data from list page (multiple item) websites
- Build dataset to train models to perform web extraction
- Extracted data can be used to build knowledge graphs
- **Shopping domain**

Schedule

Updated on Aug. 24, 2021

- **Week 1:** Aug. 25
 - Topics: Course information, review of linear algebra
 - Course Materials: [Slides](#)

- **Week 2:** Aug. 30, Sept. 1

- Topics: Review of probability and statistics

PLAtE - Dataset Construction

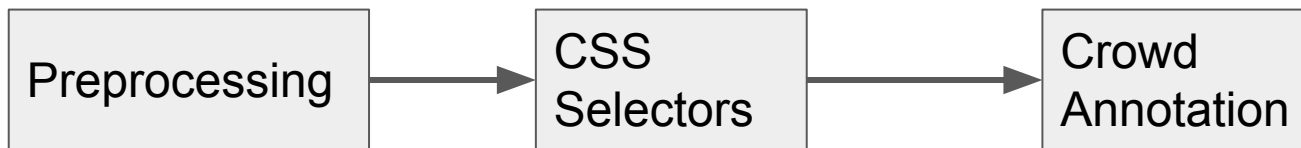


PLAtE - Dataset Construction

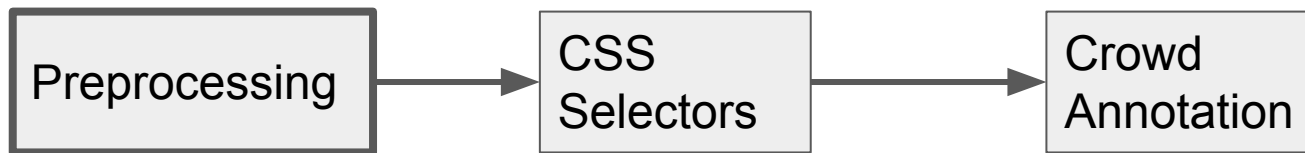


PLAtE - Preprocessing

1. Collect from Common Crawl (270 million pages)
2. Filter using List-page Classifier
3. Filter using List-page Heuristic
4. Remove duplicate URLs
5. Remove Non-English pages
6. Filter by Popular Websites
7. Filter out inappropriate content
8. Select 43 Domains with most pages (6694 Pages)

PLAtE - Preprocessing

1. Collect from Common Crawl (270 million pages)
2. Filter using List-page Classifier
3. Filter using List-page Heuristic
4. Remove duplicate URLs
5. Remove Non-English pages
6. Filter by Popular Websites
7. Filter out inappropriate content
8. Select 43 Domains with most pages (6694 Pages)

**~ 40 thousand
times smaller!**

PLAtE - Dataset Construction

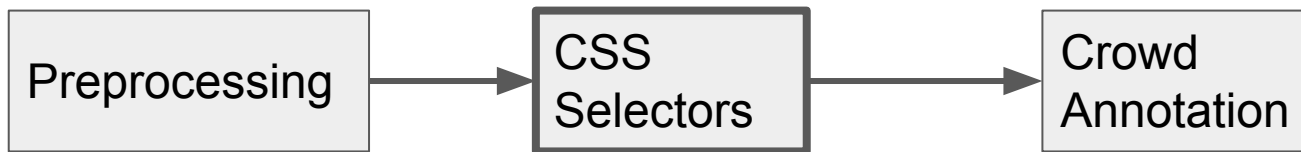


PLAtE - CSS Selectors

```
<tr>
  <td class="date">Fri, 08/30 </td>
  <td class="lecture">02 </td>
  <td class="topic">Regular Expression and Tokenization</td>
  <td class="slides"><a class="future" href="./Slides/Lecture02.pdf">pdf</a></td>
</tr>
<tr>
  <td> </td><td> </td>
  <td class="blurb" colspan="2">Review of finite-state automata, Finite-state transducers, to
</tr>
<tr>
  <td> </td><td> </td>
  <td class="blurb" colspan="2">Required reading: <a href="https://web.stanford.edu/~jurafsky
</tr>
<tr class="blankrow">
  <td colspan="5"> </td>
</tr>
<!------->
```


PLAtE - CSS Selectors

td.topic

```
<tr>
  <td class="date">Fri, 08/30 </td>
  <td class="lecture">02 </td>
  <td class="topic">Regular Expression and Tokenization</td>
  <td class="slides"><a class="future" href="./Slides/Lecture02.pdf">pdf</a></td>
</tr>
<tr>
  <td> </td><td> </td>
  <td class="blurb" colspan="2">Review of finite-state automata, Finite-state transducers, to
</tr>
<tr>
  <td> </td><td> </td>
  <td class="blurb" colspan="2">Required reading: <a href="https://web.stanford.edu/~jurafsky
</tr>
<tr class="blankrow">
  <td colspan="5"> </td>
</tr>
<!------->
```

PLAtE - Dataset Construction

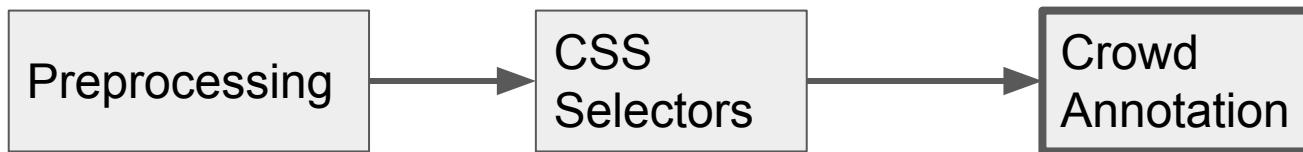


PLAtE - Crowd Annotation

- Crowd working site (Mechanical Turk) for validation and cleaning
- Require perfect performance on custom qualification task
 - 20% qualified
 - 77 annotators
- Block spammers that spend less than 20 seconds
- 1 annotator for train set
- 3 annotators for development and test set



PLAtE - Crowd Annotation

[Best bang for the Buck](#)

START OF TARGET PRODUCT SECTION

Honeywell
42 Pt. Indoor Portable Evaporative Air Cooler
[Check Price](#)
[Bottom Line](#)

[It's technically an air cooler, as it doesn't use a compressor or refrigerant. Works best in dry climates.](#)

[Editor's Notes](#)

[It's much cheaper to run than a traditional air conditioner, but it does require a constant supply of ice and water.](#)

END OF TARGET PRODUCT SECTION

BLACK+DECKER
BPACT08WT 8,000 BTU Portable Air Conditioner
[Check Price](#)
[Bottom Line](#)

[An energy-efficient choice that comes with a built-in air filter. Perfect for smaller rooms.](#)

[Editor's Notes](#)

Portable Air Conditioner FAQ

Q. Why should I choose a portable air conditioner over a window unit or central air conditioning?

A. Buying a portable air conditioner is much cheaper than installing central air conditioning, and it can be used in homes where it isn't possible or practical to fit a window unit.

Does the **Target Product** section in the Highlighted Webpage Panel contain information about at least one product? (If you answer **No** to this question please disregard all other questions and then submit)

☐ Yes ☐ No

Does the **Target Product** section in the Highlighted Webpage Panel contain information about more than one product? (If you answer **Yes** to this question please disregard all other questions and then submit)

☐ Yes ☐ No

Product Name Subtask

[View Product Name Selection](#)

Q1: Please checkmark **ALL** of the pieces of text (if any) in A1 which belong to the **Product Name**

A1

☐ Honeywell 42 Pt. Indoor Portable Evaporative Air Cooler

Are there any pieces of text which belong to the **Product Name** and **DO NOT** appear in A1, but **DO** appear **IN** the the Target Product Section? (e.g. is any text for the **Product Name** missing in A1)?

☐ Yes ☐ No

Are there any pieces of text which belong to the **Product Name** and appear **ABOVE** the the Target Product Section?

☐ Yes ☐ No

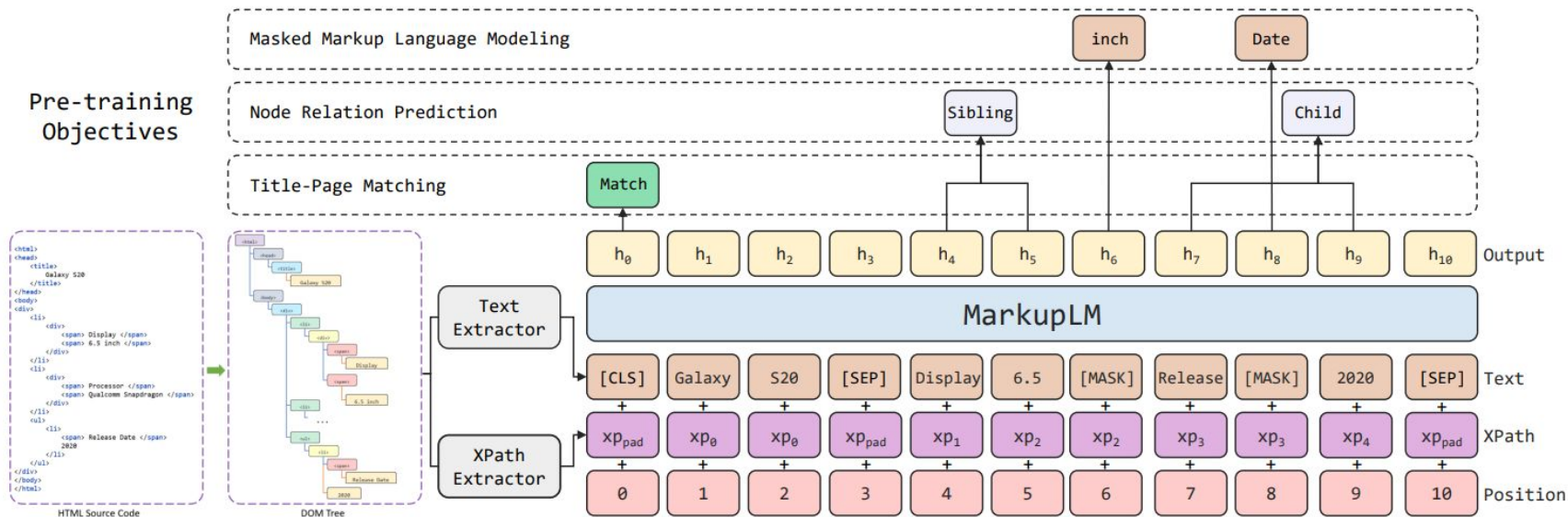
Are there any pieces of text which belong to the **Product Name** and appear **BELOW** the the Target Product

PLAtE - Modeling

MarkupLM [Li, Xu, and Wei 2022]

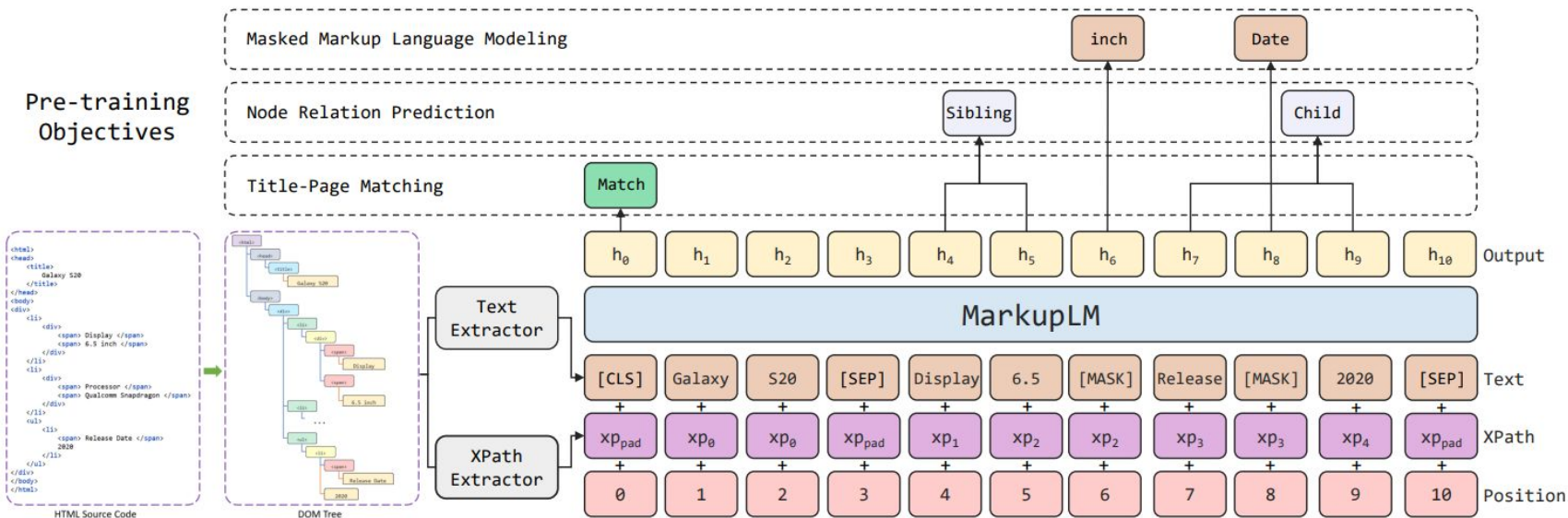
We also evaluated two other models

- RoBERTa [Liu et al 2019]
- DOM-LM [Deng et al 2022]



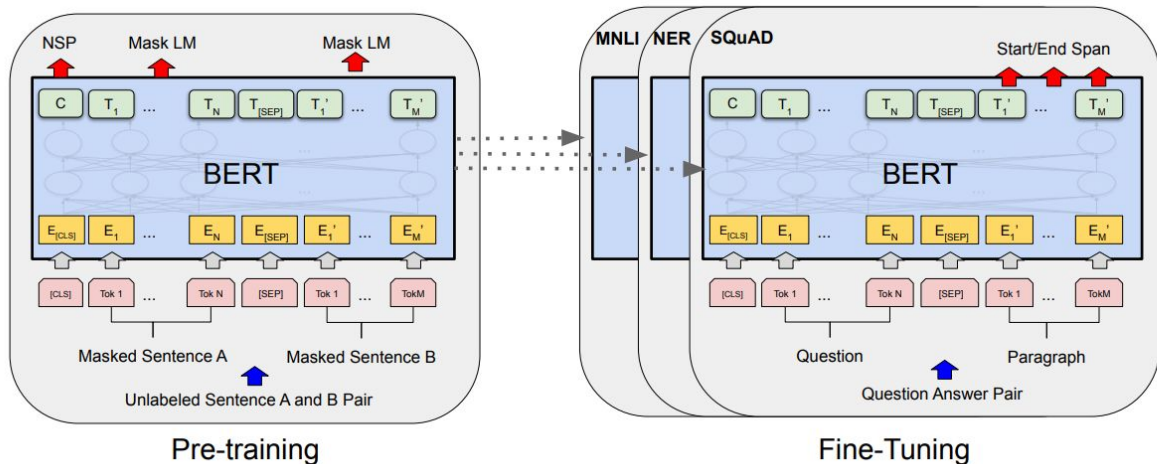
MarkupLM

- Transformer Architecture
 - Same architecture as LaMDA and GPT



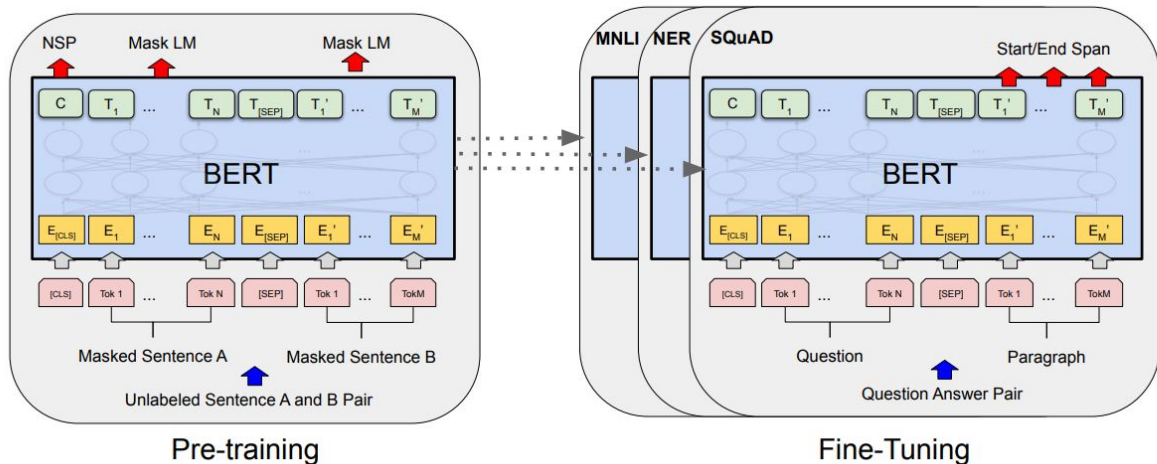
KG Construction Background - Pre-training

- Try to train the model about language before training it for a specific task



KG Construction Background - Pre-training

- **Masked Language Modeling**
- Pass the model lots of text and train the model to learn to “fill in the blanks”



KG Construction Background - Pre-training

- **Masked Language Modeling**
- Pass the model lots of text and train the model to learn to “fill in the blanks”

I want to teach CS1110

KG Construction Background - Pre-training

- **Masked Language Modeling**
- Pass the model lots of text and train the model to learn to “fill in the blanks”

I want to _____ CS1110

KG Construction Background - Pre-training

- **Masked Language Modeling**
- Pass the model lots of text and train the model to learn to “fill in the blanks”

I want to _____ CS1110

Train model to predict what should fill in the blank

KG Construction Background - Pre-training

- **Masked Language Modeling**
- Pass the model lots of text and train the model to learn to “fill in the blanks”

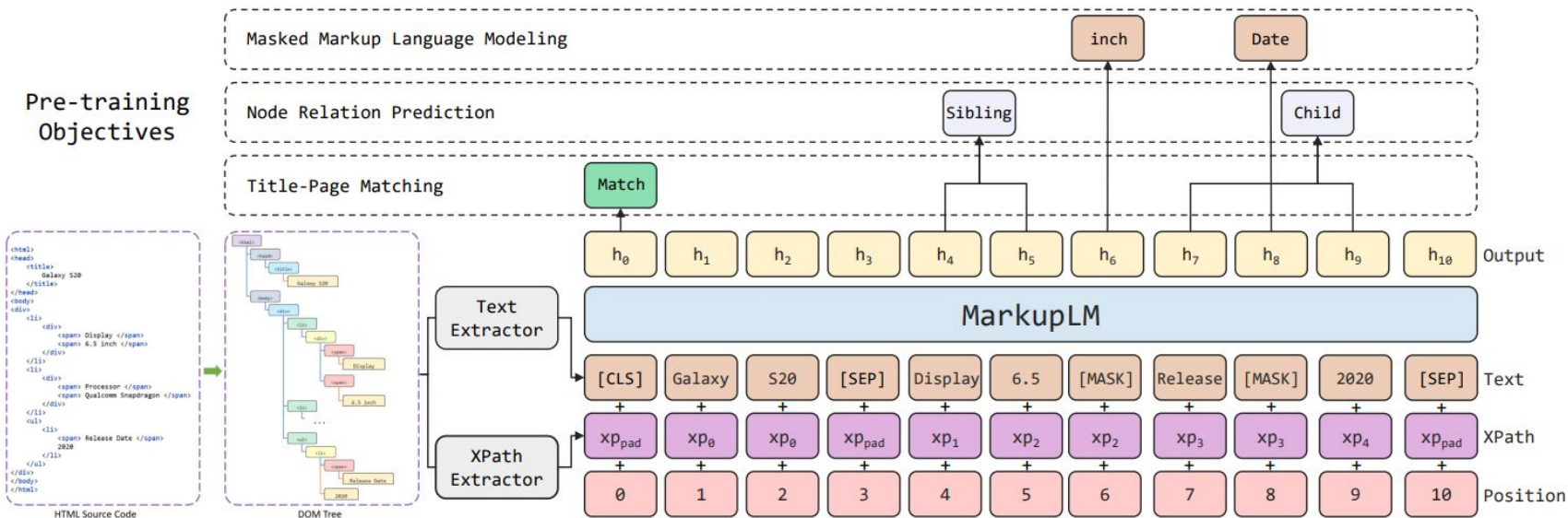
I want to _____ CS1110

Model predicts “teach” 

Model predicts a different word 

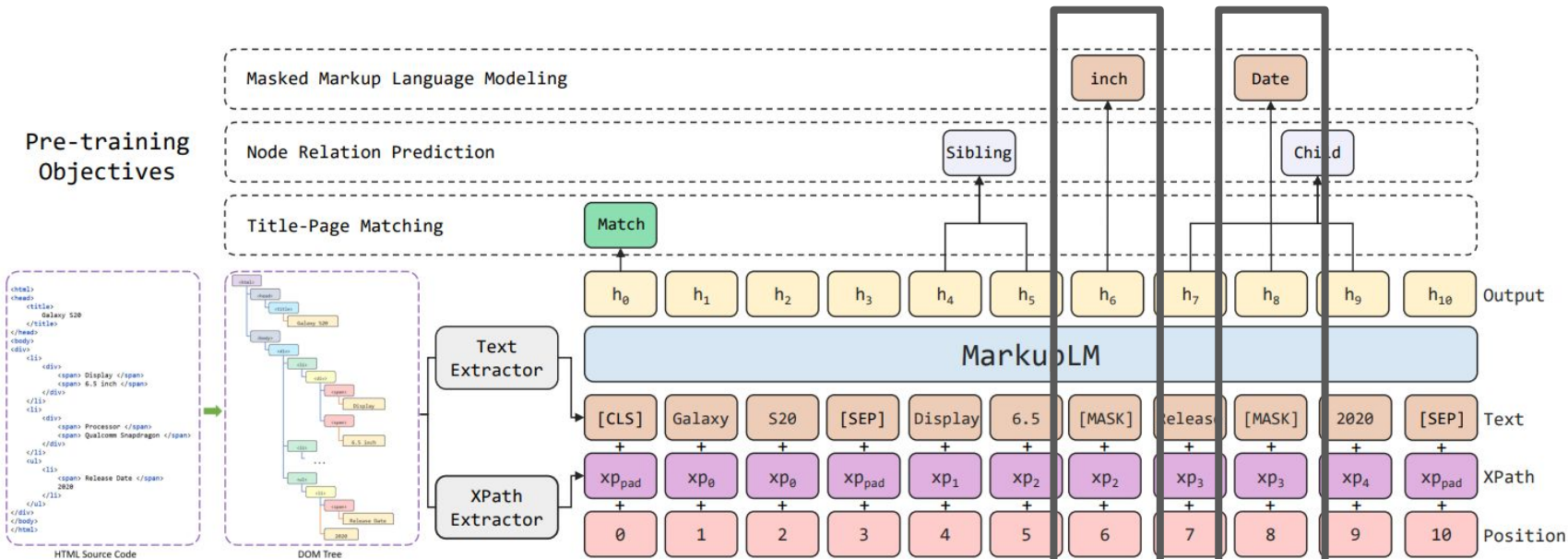
MarkupLM

- Pretrained on 3 Tasks:
 - Masked Markup Language Modeling
 - Node Relation Prediction
 - Title-Page Matching



MarkupLM

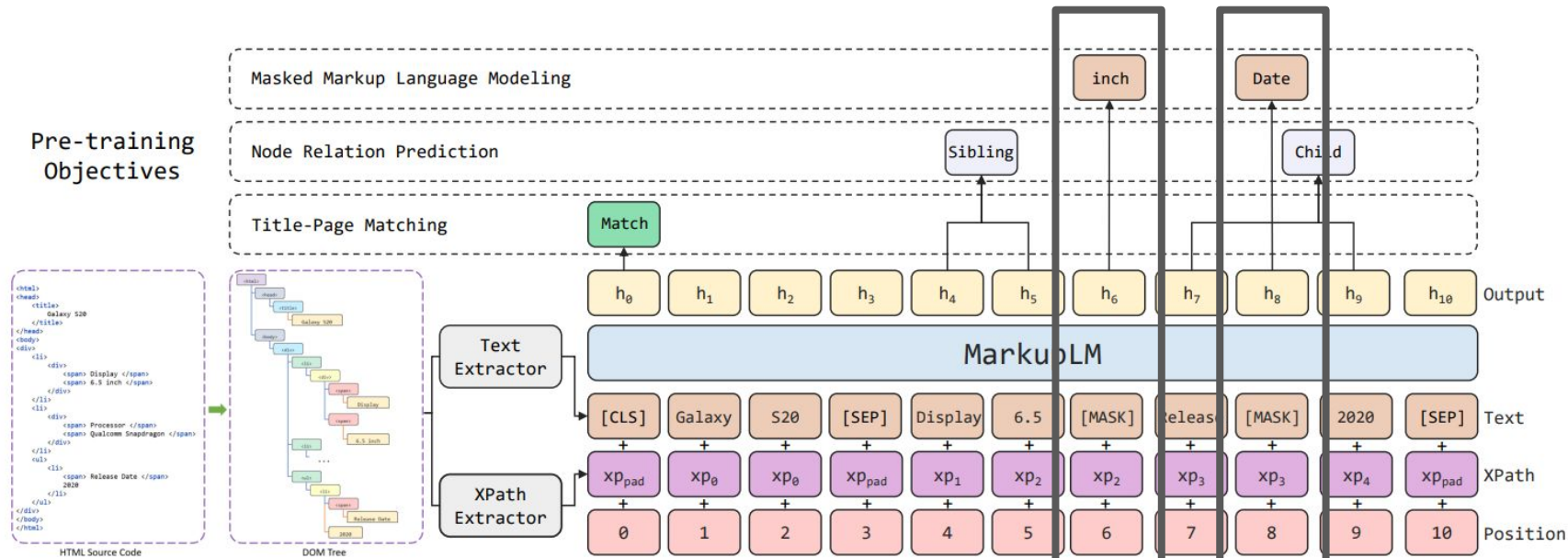
- Masked Markup Language Modeling



MarkupLM

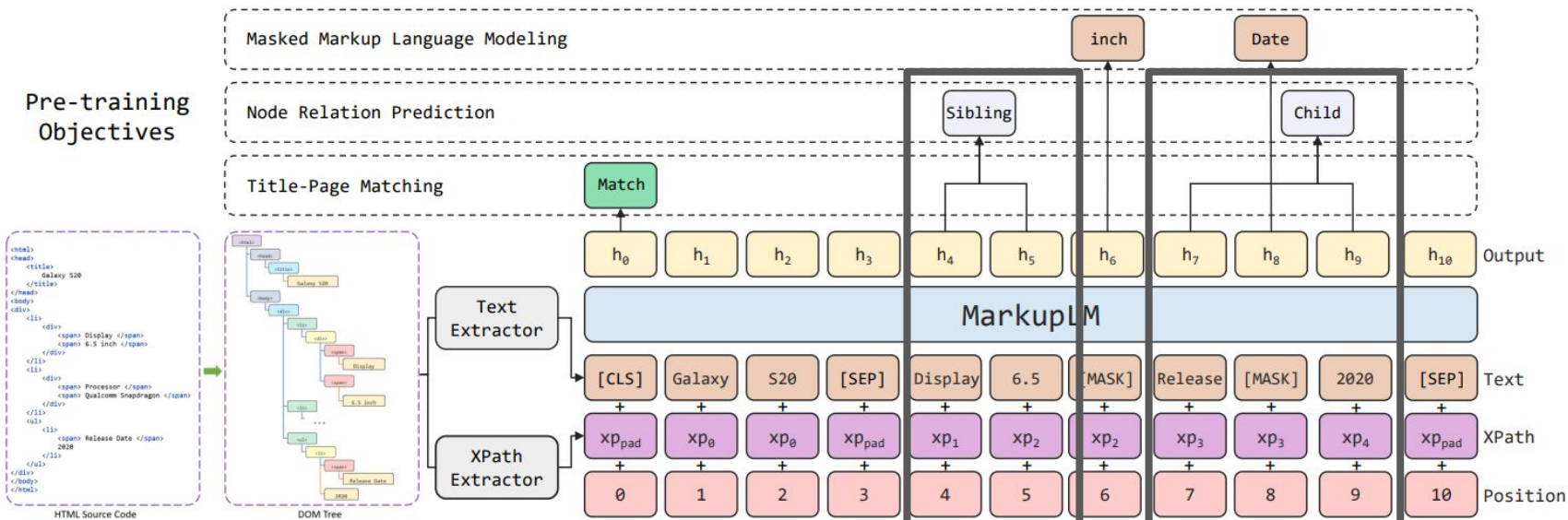
- Masked Markup Language Modeling

What should the [MASK] be?



MarkupLM

- Node Relation Prediction

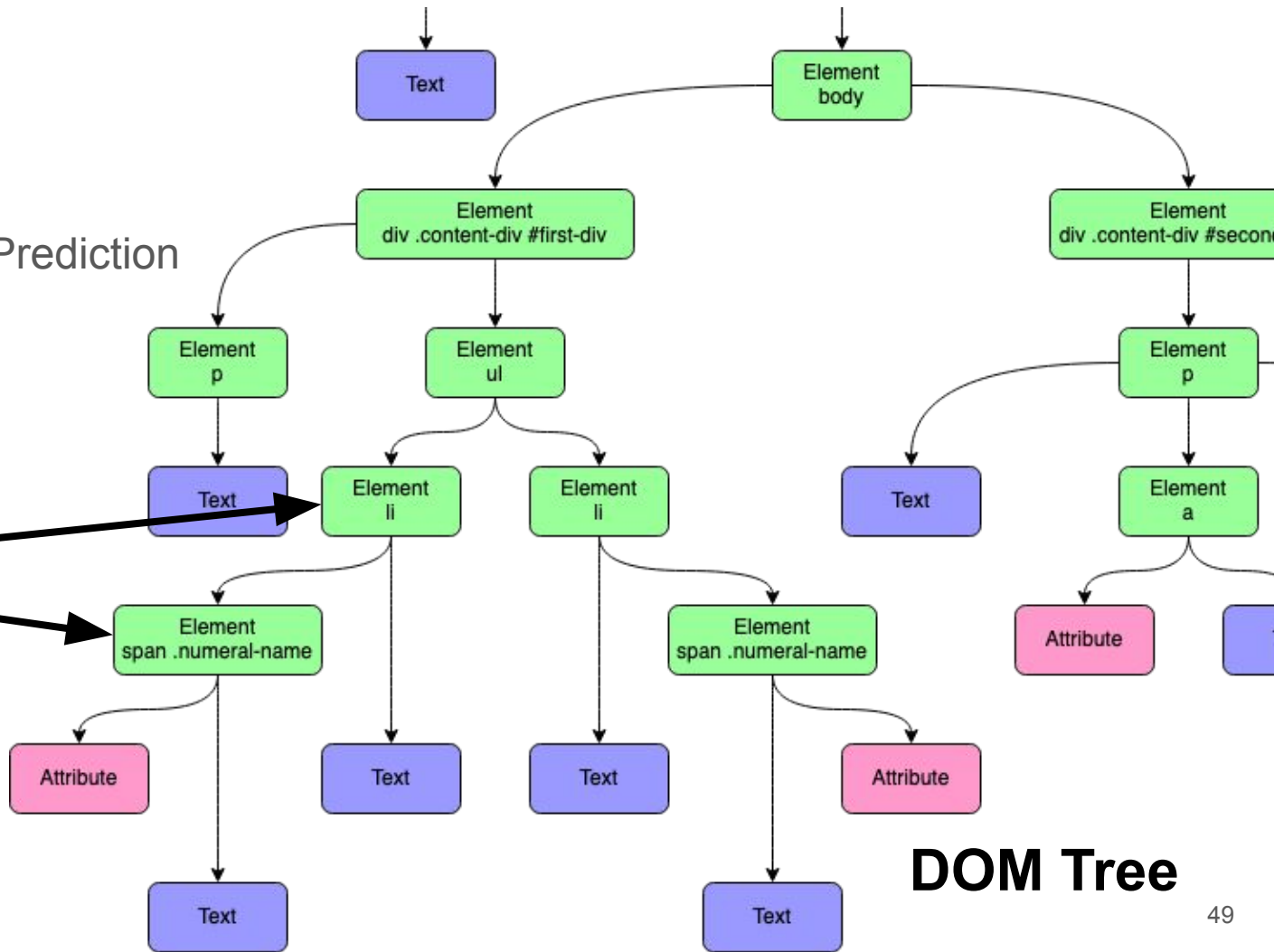
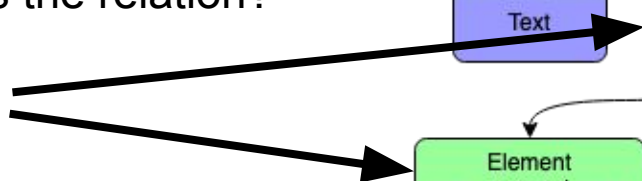


MarkupLM

- Node Relation Prediction

What is the relation?

Child



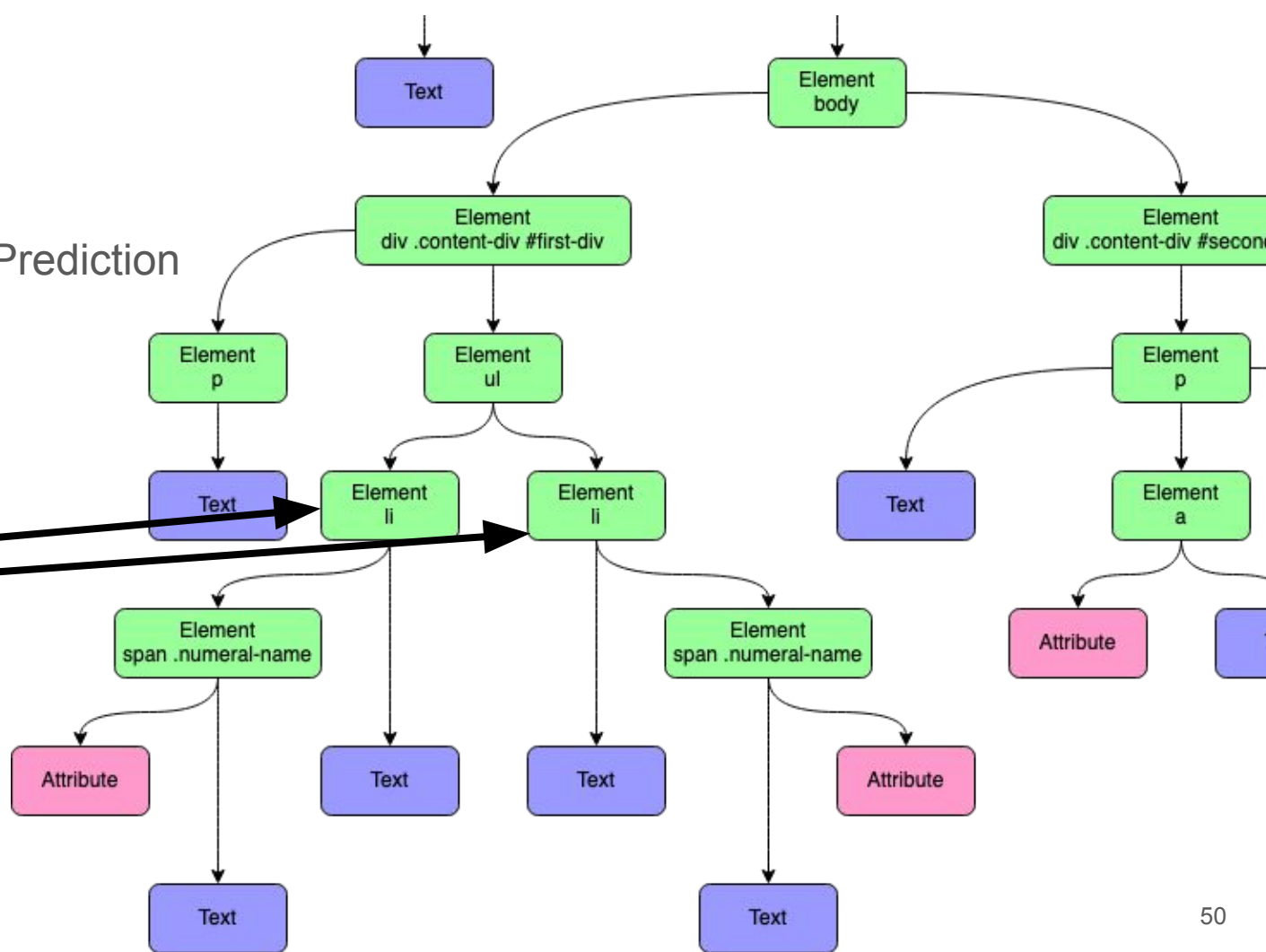
DOM Tree

MarkupLM

- Node Relation Prediction

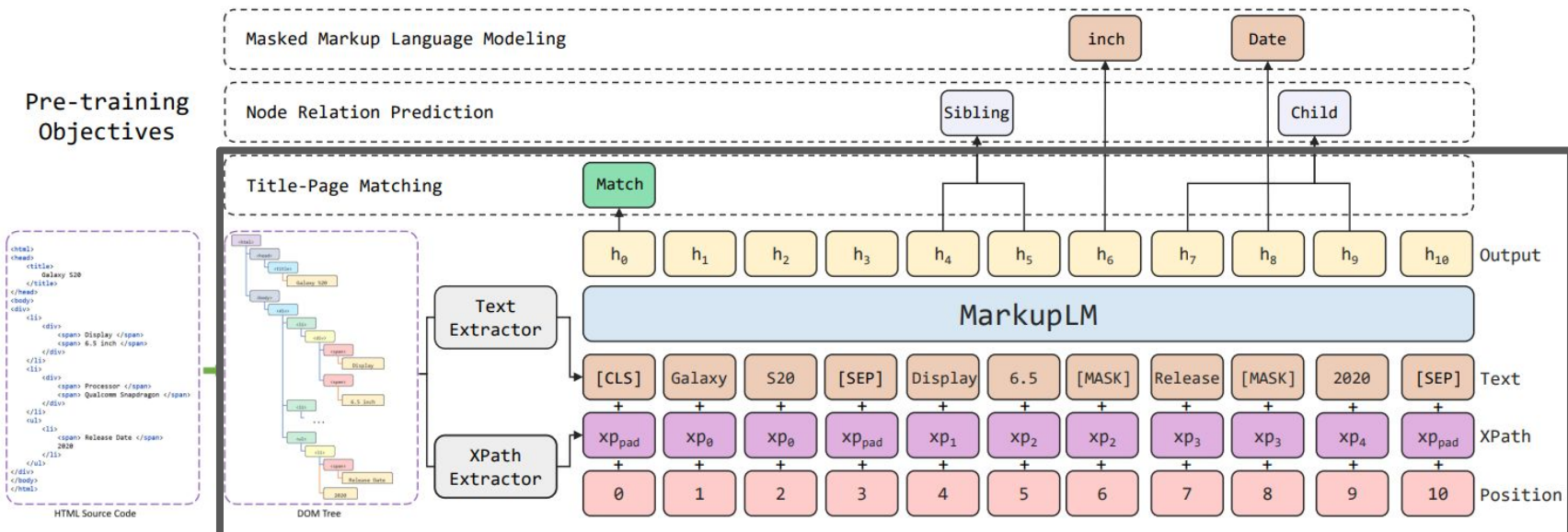
What is the relation?

Sibling



MarkupLM

- Title-Page Matching



MarkupLM

Does the <title> match the rest of the page?

- Title-Page Matching

```
<html>
  <head>
    <meta http-equiv="content-type" content="text/html; charset=iso-8859-1">
    <title>CS 447: Natural Language Processing (Fall 2019)</title>
    <link href="cs447.css" rel="stylesheet" type="text/css">
  </head>
  <body>
    <div id="header">...</div>
    <div id="menubar">...</div>
    <div id="text">...</div>
    " Fri, 11/08 22 Sentence Semantics II: Semantic Role Labeling "
    <a class="future" href="./Slides/Lecture22.pdf">pdf</a>
    " MP3 due. "
    <a href=".">MP4</a>
    " out &nbsp; How do we represent and capture who does what to whom? &nbsp; Required reading: "
    <a href="https://web.stanford.edu/~jurafsky/slp3/18.pdf">Ch. 18</a>
    " &nbsp; Optional reading: "
    <a href="http://aclweb.org/anthology/J/J05/J05-1004.pdf">Palmer et al. (2005)</a>
    " , "
```

MarkupLM

Does the <title> match the rest of the page?

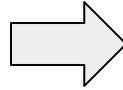
- Title-Page Matching

```
<html>
  <head>
    <meta http-equiv="content-type" content="text/html; charset=iso-8859-1">
    <title> CS446/ECE449 Spring 2023 | Machine Learning </title>
    <link href="cs447.css" rel="stylesheet" type="text/css">
  </head>
  <body>
    <div id="header">...</div>
    <div id="menubar">...</div>
    <div id="text">...</div>
    " Fri, 11/08 22 Sentence Semantics II: Semantic Role Labeling "
    <a class="future" href="./Slides/Lecture22.pdf">pdf</a>
    " MP3 due. "
    <a href=".">MP4</a>
    " out &nbsp; How do we represent and capture who does what to whom? &nbsp; Required reading: "
    <a href="https://web.stanford.edu/~jurafsky/slp3/18.pdf">Ch. 18</a>
    " &nbsp; Optional reading: "
    <a href="http://aclweb.org/anthology/J/J05/J05-1004.pdf">Palmer et al. (2005)</a>
    ", "
```

Applying our Dataset

Train on PLAtE

PLAtE
Dataset



Web
Extraction
Model

Perform extraction on webpages of interest

uva-nlp-course
UvA CS-4501: Machine Learning for NLP

Schedule
Updated on Aug. 24, 2021

- Week 1:** Aug. 25
 - Topics: Course Information, review of linear algebra
 - Course Materials: slides
- Week 2:** Aug. 30, Sept. 1
 - Topics: Review of probability and statistics
- Week 3:** Sept. 6, 8
 - Topics: Text classification, Logistic regression
 - Readings: 8.02
 - Course Materials: slides, Code, Code
- Week 4:** Sept. 13, 15
 - Topics: neural network classifiers, feed forward NNs, the
 - Readings: 8.03
 - Course Materials: slides, Code, Code
- Week 5:** Sept. 20, 22
 - Topics: Further discussion of text classification
 - Readings: 8.04
 - Course Materials: slides
- Week 6:** Sept. 27, 29

Schedule
Updated on Sep. 24, 2021

Updated course details will be posted here shortly before each lecture. Other links contain last year's slides, which are mostly correct. **Course materials will be updated a few days after each lecture.** The index (which covers approximately the first half of the course) is also beyond the lecture.

Date	Description	Course Materials	Events
Tue Jan 10	Week 1: Intro to NLP	Suggested Readings: 1. Efficient Estimation of Word Representations in Vector Spaces (original and revised paper) 2. Distributed Representations of Words and Phrases and their Compositionality (negative sampling paper) [link] [link]	Assignment 1 due
Thu Jan 12	Week 2: Word Vectors, Word Window	Suggested Readings: 1. GloVe: Global Vectors for Word Representation (paper) 2. Improving Distributional Similarity with Lessons Learned from Word Embeddings 3. Evaluating methods for unsupervised word embeddings Additional Readings: 4. A Latent Semantic Model Approach to PhN-based Word Embeddings 5. Linear Algebra: Structure of Word Spaces, with Applications to Polygraphs 6. On the Dimensionality of Word Embedding	
Fri Jan 13	Python Review Session	0 2:30pm - 3:20pm Code: 100	
Tue Jan 17	Week 3: Text Classification	Suggested Readings: 1. Naïve Bayes Classifier 2. Review of different classifiers 3. CS231n notes on neural architectures 4. CS231n notes on image classification 5. Convolution, Backpropagation, and Visualization 6. Convolutional Neural Networks for Visual Question Answering (arXiv:1608.08220v1)	Assignment 2 due

课程

课程介绍

- 本课程是自然语言处理领域的入门课程，旨在帮助学生了解该领域的最新进展，并掌握相关的理论和实践技能。
- 课程内容包括：自然语言处理的基本概念、文本分类、机器翻译、问答系统等。
- 课程将采用讲授、实验和作业相结合的方式进行。

文本分类

- 本课程将介绍文本分类的基本概念和理论，包括朴素贝叶斯分类器、支持向量机等。
- 课程将介绍文本分类的常见应用，如垃圾邮件过滤、情感分析等。
- 课程将介绍文本分类的常见数据集和评估指标。

语言模型

- 本课程将介绍语言模型的基本概念和理论，包括马尔可夫链、隐马尔可夫模型等。
- 课程将介绍语言模型的常见应用，如机器翻译、文本生成等。
- 课程将介绍语言模型的常见数据集和评估指标。

Schedule
(MP) due 09/30
Wed, 08/25

01 Introduction pdf

What is NLP? What will you learn in this class?
What will we teach this class?
Read reading: Ch.1 (and E4).
Read reading: Python tutorial (sec. 1-5).
Read reading: Ch.2 (and E4).
Read reading: Ch.3 (and E4).
Read reading: Ch.4 (and E4).
Read reading: Ch.5 (and E4).
Read reading: Ch.6 (and E4).
Read reading: Ch.7 (and E4).
Read reading: Ch.8 (and E4).
Read reading: Ch.9 (and E4).
Read reading: Ch.10 (and E4).
Read reading: Ch.11 (and E4).
Read reading: Ch.12 (and E4).
Read reading: Ch.13 (and E4).
Read reading: Ch.14 (and E4).
Read reading: Ch.15 (and E4).
Read reading: Ch.16 (and E4).
Read reading: Ch.17 (and E4).
Read reading: Ch.18 (and E4).
Read reading: Ch.19 (and E4).
Read reading: Ch.20 (and E4).
Read reading: Ch.21 (and E4).
Read reading: Ch.22 (and E4).
Read reading: Ch.23 (and E4).
Read reading: Ch.24 (and E4).
Read reading: Ch.25 (and E4).
Read reading: Ch.26 (and E4).
Read reading: Ch.27 (and E4).
Read reading: Ch.28 (and E4).
Read reading: Ch.29 (and E4).
Read reading: Ch.30 (and E4).
Read reading: Ch.31 (and E4).
Read reading: Ch.32 (and E4).
Read reading: Ch.33 (and E4).
Read reading: Ch.34 (and E4).
Read reading: Ch.35 (and E4).
Read reading: Ch.36 (and E4).
Read reading: Ch.37 (and E4).
Read reading: Ch.38 (and E4).
Read reading: Ch.39 (and E4).
Read reading: Ch.40 (and E4).
Read reading: Ch.41 (and E4).
Read reading: Ch.42 (and E4).
Read reading: Ch.43 (and E4).
Read reading: Ch.44 (and E4).
Read reading: Ch.45 (and E4).
Read reading: Ch.46 (and E4).
Read reading: Ch.47 (and E4).
Read reading: Ch.48 (and E4).
Read reading: Ch.49 (and E4).
Read reading: Ch.50 (and E4).
Read reading: Ch.51 (and E4).
Read reading: Ch.52 (and E4).
Read reading: Ch.53 (and E4).
Read reading: Ch.54 (and E4).
Read reading: Ch.55 (and E4).
Read reading: Ch.56 (and E4).
Read reading: Ch.57 (and E4).
Read reading: Ch.58 (and E4).
Read reading: Ch.59 (and E4).
Read reading: Ch.60 (and E4).
Read reading: Ch.61 (and E4).
Read reading: Ch.62 (and E4).
Read reading: Ch.63 (and E4).
Read reading: Ch.64 (and E4).
Read reading: Ch.65 (and E4).
Read reading: Ch.66 (and E4).
Read reading: Ch.67 (and E4).
Read reading: Ch.68 (and E4).
Read reading: Ch.69 (and E4).
Read reading: Ch.70 (and E4).
Read reading: Ch.71 (and E4).
Read reading: Ch.72 (and E4).
Read reading: Ch.73 (and E4).
Read reading: Ch.74 (and E4).
Read reading: Ch.75 (and E4).
Read reading: Ch.76 (and E4).
Read reading: Ch.77 (and E4).
Read reading: Ch.78 (and E4).
Read reading: Ch.79 (and E4).
Read reading: Ch.80 (and E4).
Read reading: Ch.81 (and E4).
Read reading: Ch.82 (and E4).
Read reading: Ch.83 (and E4).
Read reading: Ch.84 (and E4).
Read reading: Ch.85 (and E4).
Read reading: Ch.86 (and E4).
Read reading: Ch.87 (and E4).
Read reading: Ch.88 (and E4).
Read reading: Ch.89 (and E4).
Read reading: Ch.90 (and E4).
Read reading: Ch.91 (and E4).
Read reading: Ch.92 (and E4).
Read reading: Ch.93 (and E4).
Read reading: Ch.94 (and E4).
Read reading: Ch.95 (and E4).
Read reading: Ch.96 (and E4).
Read reading: Ch.97 (and E4).
Read reading: Ch.98 (and E4).
Read reading: Ch.99 (and E4).
Read reading: Ch.100 (and E4).

ular Expression and pdf

nization

w of finite-state automata, Finite-state
ducers, tokenization
red reading: Ch. 2

uage Models; Intro to pdf

ability Models for NLP

(o)

w of basic probability. How do we apply
ideas to NLP?
in language models. Evaluation: Perplexity
Word Error Rate
red reading: Ch. 3
red reading: MS, Ch. 2

duction to Classification pdf

Web Extraction Model

Extracted Concepts

“Language Models”

“Tokenization”

“Logistic Regression”

Connecting Back to Education...

- Build knowledge graphs out of textbook and lecture transcripts
- Education domain

Assma Boughoula, [Aidan San](#), and ChengXiang Zhai. 2020. Leveraging Book Indexes for Automatic Extraction of Concepts in MOOCs. In Proceedings of the Seventh ACM Conference on **Learning @ Scale** (L@S '20). Association for Computing Machinery, New York, NY, USA, 381–384.

We'll also introduce two important ideas that are often used with these masked language models. The first is the idea of **fine-tuning**. Fine-tuning is the process of taking the network learned by these pretrained models, and further training the model, often via an added **neural net** classifier that takes the top layer of the network as input, to perform some downstream task like named entity tagging or question answering or coreference. The intuition is that the pretraining phase learns a language model that instantiates a rich representations of word meaning, that thus enables the model to more easily learn ('be fine-tuned to') the requirements of a downstream

What are the ethical implications?

Ethical Implications

- Language Models
 - Bias towards demographic groups
 - Can be used to generate misinformation (phishing)
 - ChatGPT trained with underpaid outsourced labor
 - Massive energy costs to training LLMs

60 MINUTES OVERTIME >

ChatGPT and large language model bias

Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic

Artificial Intelligence Is Booming—So Is Its Carbon Footprint

Greater transparency on emissions could also bring more scrutiny

Disinformation Researchers Raise Alarms About A.I. Chatbots

Researchers used ChatGPT to produce clean, convincing text that repeated conspiracy theories and misleading narratives.

Ethical Implications


- Education
 - Language models enable cheating
 - Data diversity
 - Subjects
 - Institutions
 - Languages

Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach

With the rise of the popular new chatbot ChatGPT, colleges are restructuring some courses and taking preventive measures.

Future Work

Future Work

- Build a knowledge graph from extracted data
- User studies
 - Can students effectively use these graphs to build self-taught curricula?
- Improve methods for building KGs for education
 - Build larger web extraction datasets
 - Domain adaptation
 - Shopping -> Education  Good project for undergrads
- More advanced modeling methods
 - Transformer models
 - Large language models

Acknowledgments



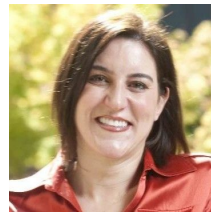
Yangfeng Ji



Heba Elfardy



Kevin Small



Tanya Roosta



Yuan Zhuang



Jan Bakus



Sandeep Atluri



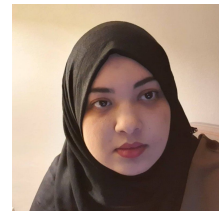
Colin Lockard



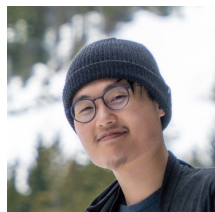
David Cierniewicz



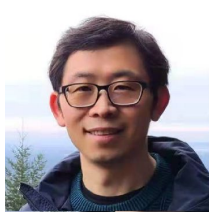
ChengXiang Zhai



Assma Boughoula



Sanxing Chen



Xiaodong Liu



Renqin Cai



Hongning Wang



Jibang Wu

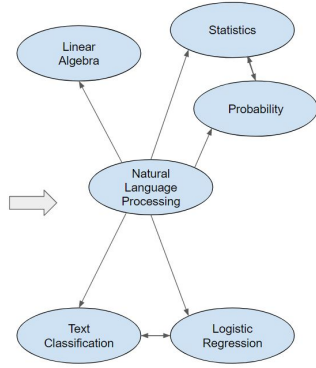


Chong Wang

Schedule

Updated on Aug. 24, 2021

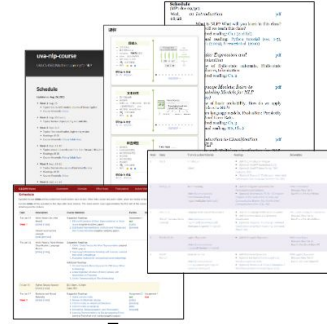
- **Week 1:** Aug. 25
 - Topics: **Course information, review of linear algebra**
 - Course Materials: **Slides**
- **Week 2:** Aug. 30, Sept. 1
 - Topics: **Review of probability and statistics**
- **Week 3:** Sept. 6, 8
 - Topics: **Text classification, logistic regression**
 - Readings: JE 02
 - Course Materials: **Slides; Colab Code**



Preprocessing

CSS
Selectors

Crowd
Annotation



Web
Extraction
Model

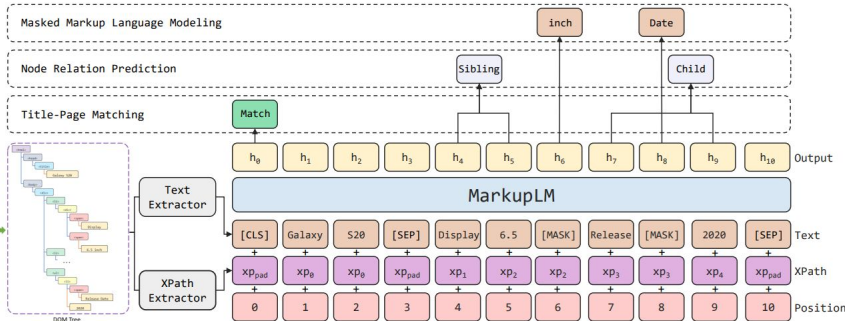
Extracted Concepts

"Language Models"

"Tokenization"

"Logistic Regression"

Thank you!



60 MINUTES OVERTIME
ChatGPT and large language models

Artificial Intelligence Is Booming—So Is Its Carbon Footprint

Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less **Disinformation Researchers Raise Alarms About A.I. Chatbots**

Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach

ChatGPT to produce clean, convincing text that racy theories and misleading narratives.

With the rise of the popular new chatbot ChatGPT, colleges are restructuring some courses and taking preventive measures.

References

Person Icon: https://www.flaticon.com/free-icon/man_9567040?term=person&page=1&position=82&origin=search&related_id=9567040

Dom Tree Image: <https://www.linode.com/docs/guides/traversing-the-dom/>

MTurk Image: <https://twitter.com/amazonmturk/photo>

Roberta Image: <https://paperswithcode.com/model/roberta-large-sst>

Course Webpages: <https://web.stanford.edu/class/cs224n/index.html#schedule>, <http://yangfengji.net/uva-nlp-course/schedule.html>,
<https://courses.grainger.illinois.edu/cs447/fa2019/syllabus.html>, <https://mlnp-world.github.io/NLP-Course-Chinese/>,
<https://www.cs.williams.edu/~kkeith/teaching/s23/cs375/>

Textbook: <https://web.stanford.edu/~jurafsky/slp3/11.pdf>

News articles: <https://www.cbsnews.com/news/chatgpt-large-language-model-bias-60-minutes-2023-03-05/>,
<https://time.com/6247678/openai-chatgpt-kenya-workers/>,
<https://www.bloomberg.com/news/articles/2023-03-09/how-much-energy-do-ai-and-chatgpt-use-no-one-knows-for-sure>,
<https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html>,
<https://www.nytimes.com/2023/01/16/technology/chatgpt-artificial-intelligence-universities.html>

https://web.stanford.edu/~vinayc/kg/notes/What_is_a_Knowledge_Graph.html

References

- San, A., Bakus, J., Lockard, C., Ciemiewicz, D., Ji, Y., Atluri, S., ... & Elfardy, H. (2022). PLATe: A Large-scale Dataset for List Page Web Extraction. *arXiv preprint arXiv:2205.12386*.
- Boughoula, A., San, A., & Zhai, C. (2020, August). Leveraging Book Indexes for Automatic Extraction of Concepts in MOOCs. In *Proceedings of the Seventh ACM Conference on Learning@ Scale* (pp. 381-384).
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.
- Deng, X., Shiralkar, P., Lockard, C., Huang, B., & Sun, H. (2022). DOM-LM: Learning Generalizable Representations for HTML Documents. *arXiv preprint arXiv:2201.10608*.
- Dessi, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., & Motta, E. (2022, October). CS-KG: A large-scale knowledge graph of research entities and claims in computer science. In *The Semantic Web—ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings* (pp. 678-696). Cham: Springer International Publishing.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Janowicz, K., Yan, B., Regalia, B., Zhu, R., & Mai, G. (2018, October). Debiasing Knowledge Graphs: Why Female Presidents are not like Female Popes. In *ISWC (P&D/Industry/BlueSky)*.
- Li, J., Xu, Y., Cui, L., & Wei, F. (2021). Markuplm: Pre-training of text and markup language for visually-rich document understanding. *arXiv preprint arXiv:2110.08518*.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., ... & Le, Q. (2022). Lambda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

PLAtE - Results

Model	Attribute Extraction			
	Dev F1	Test P	Test R	Test F1
RoBERTa	0.851	<u>0.843</u>	0.652	<u>0.735</u>
DOMLM	<u>0.871</u>	0.815	<u>0.655</u>	0.726
MarkupLM	0.866	0.839	0.620	0.711

PLAtE - Results

Model	Segmentation					
	Dev F1	Test P	Test R	Test F1	Test ARI	Test NMI
RoBERTa	0.839	0.692	0.665	0.678	0.693	0.744
DOMLM	<u>0.861</u>	0.718	0.728	0.722	0.716	0.764
MarkupLM	<u>0.861</u>	<u>0.769</u>	<u>0.805</u>	<u>0.787</u>	<u>0.771</u>	<u>0.870</u>