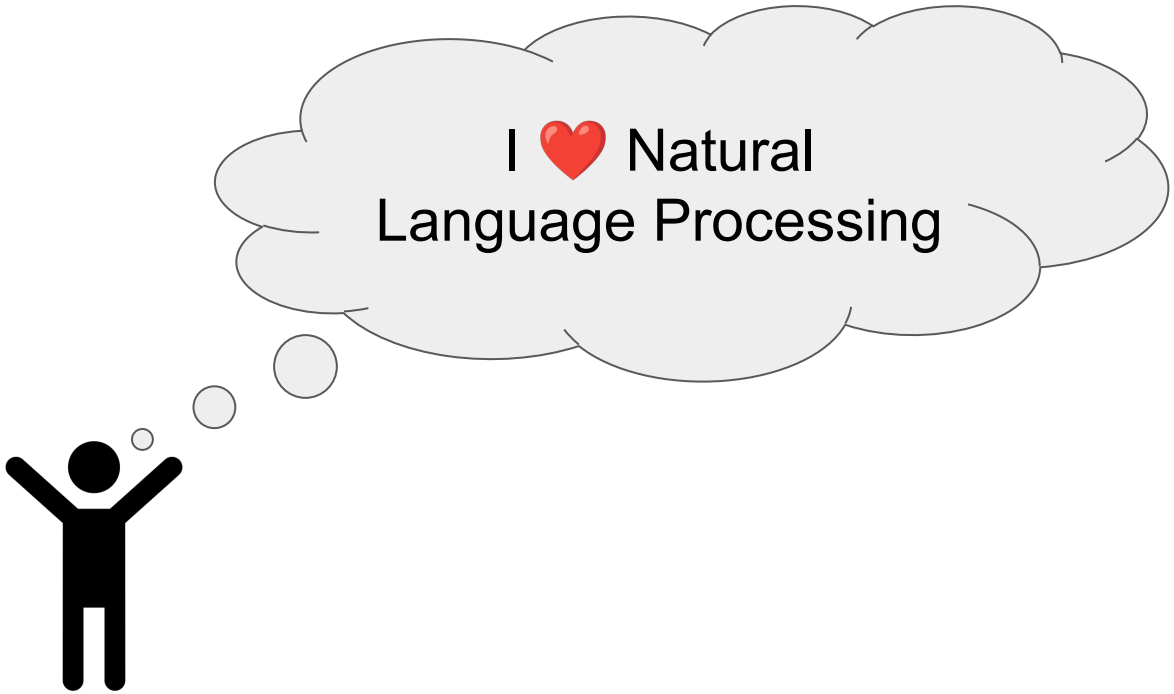
Constructing Knowledge Graphs from Web Data

Aidan San

Motivation

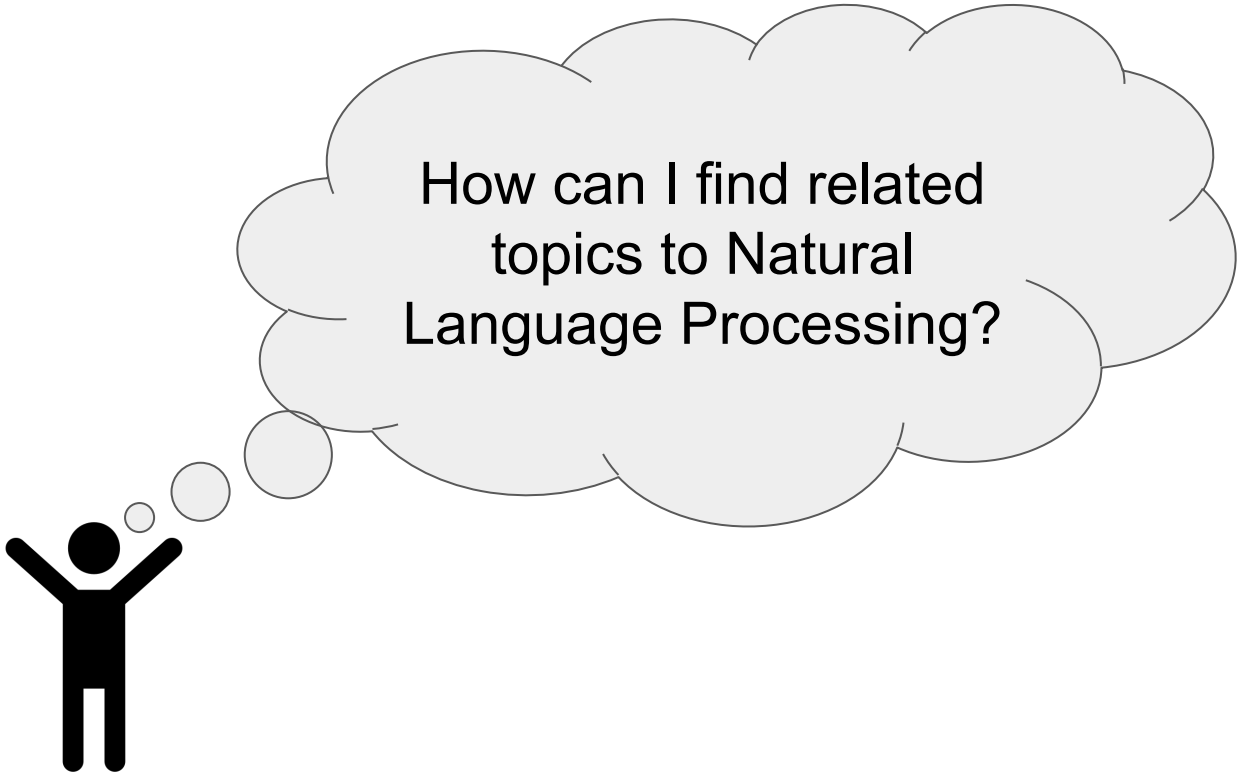
A student is trying to build their own online learning curriculum

Motivation

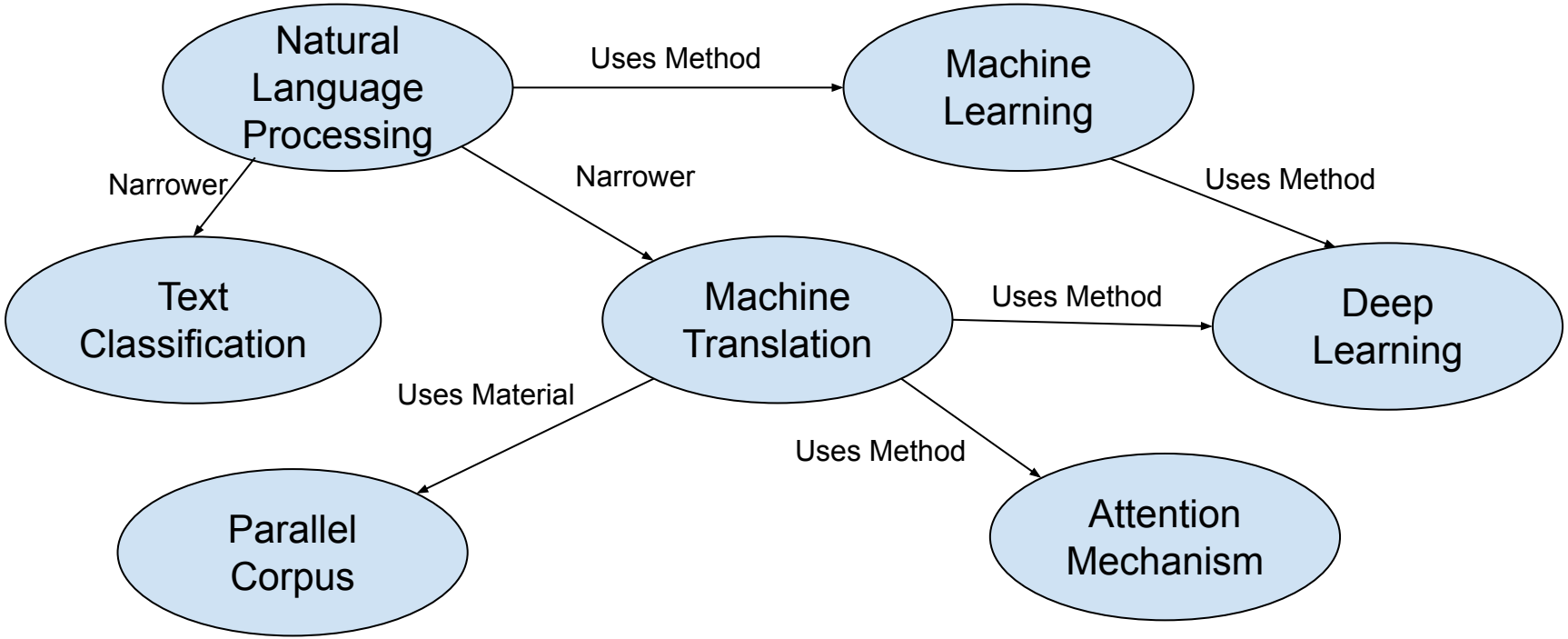


Motivation

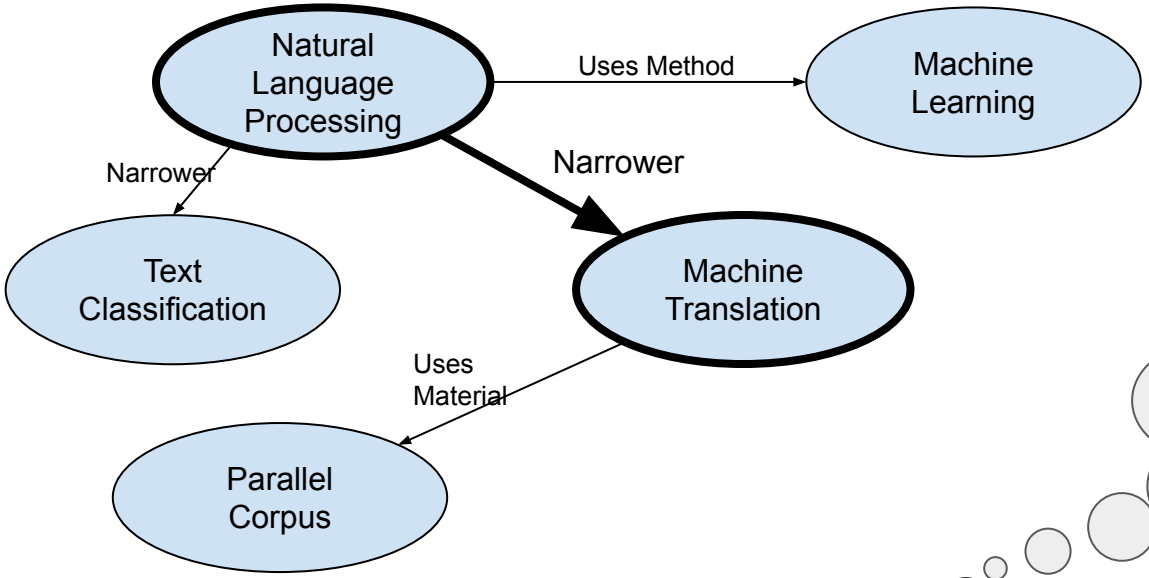
Motivation

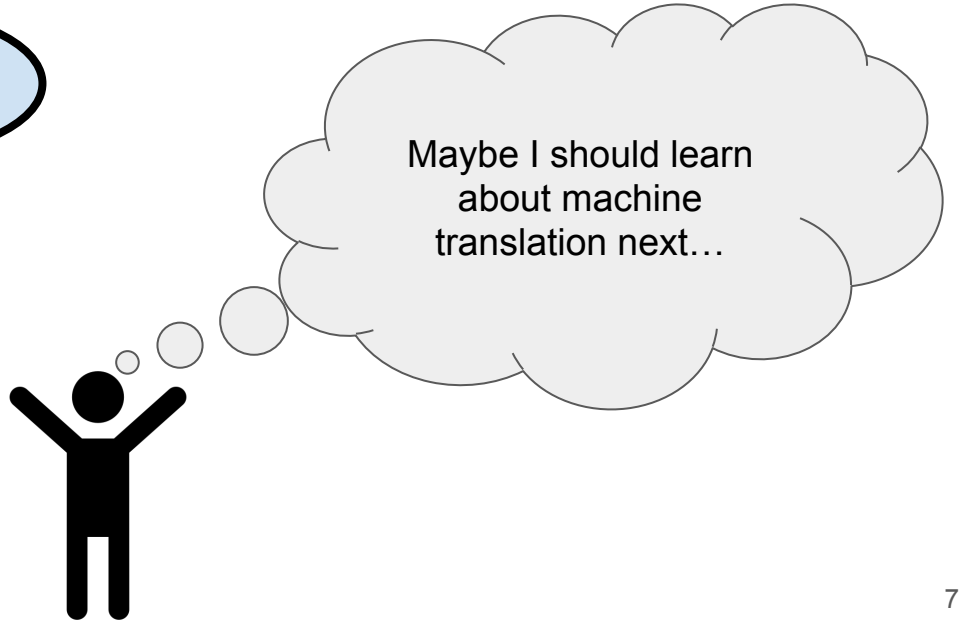

How can I find related topics to Natural Language Processing?

Knowledge Graphs



Motivation


Natural Language Processing
Uses Method
Machine Learning
Narrower
Narrower
Text Classification
Machine Translation
Uses Material
Parallel Corpus

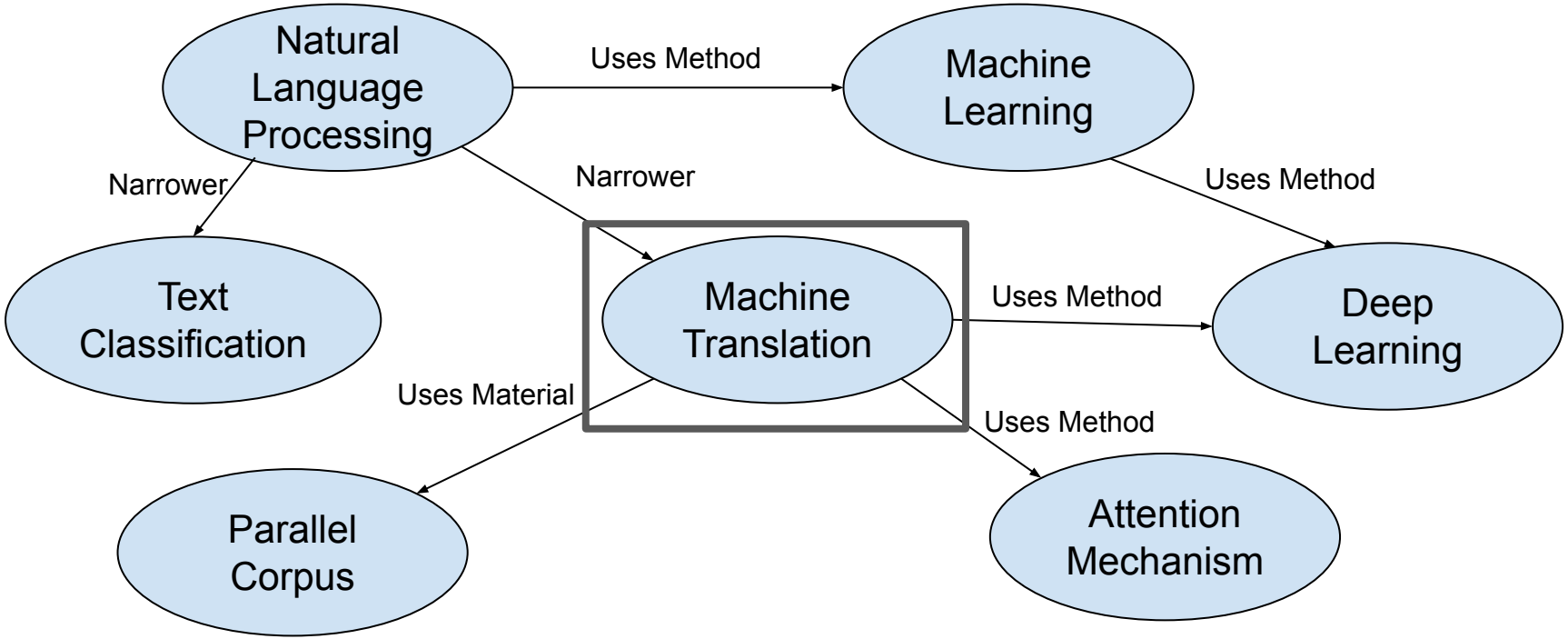

Maybe I should learn about machine translation next…

Outline

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

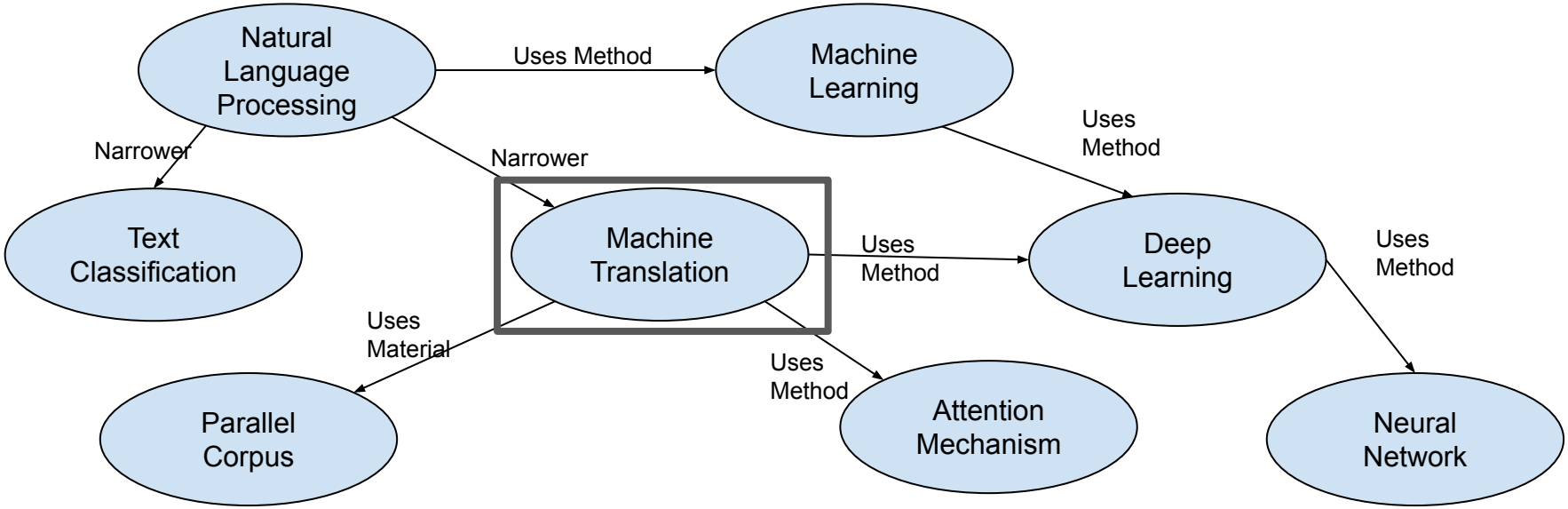
What are Knowledge Graphs?

What are Knowledge Graphs? - Entities

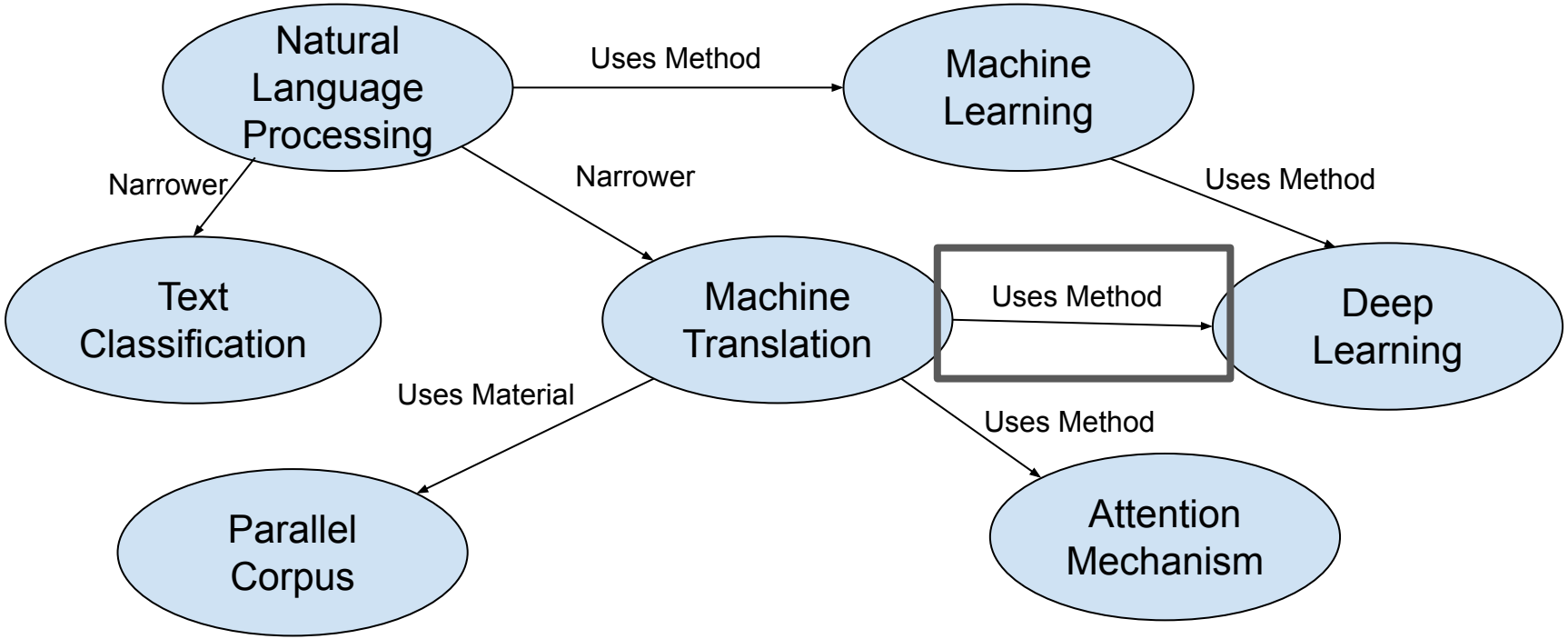


What are Knowledge Graphs? - Entities

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

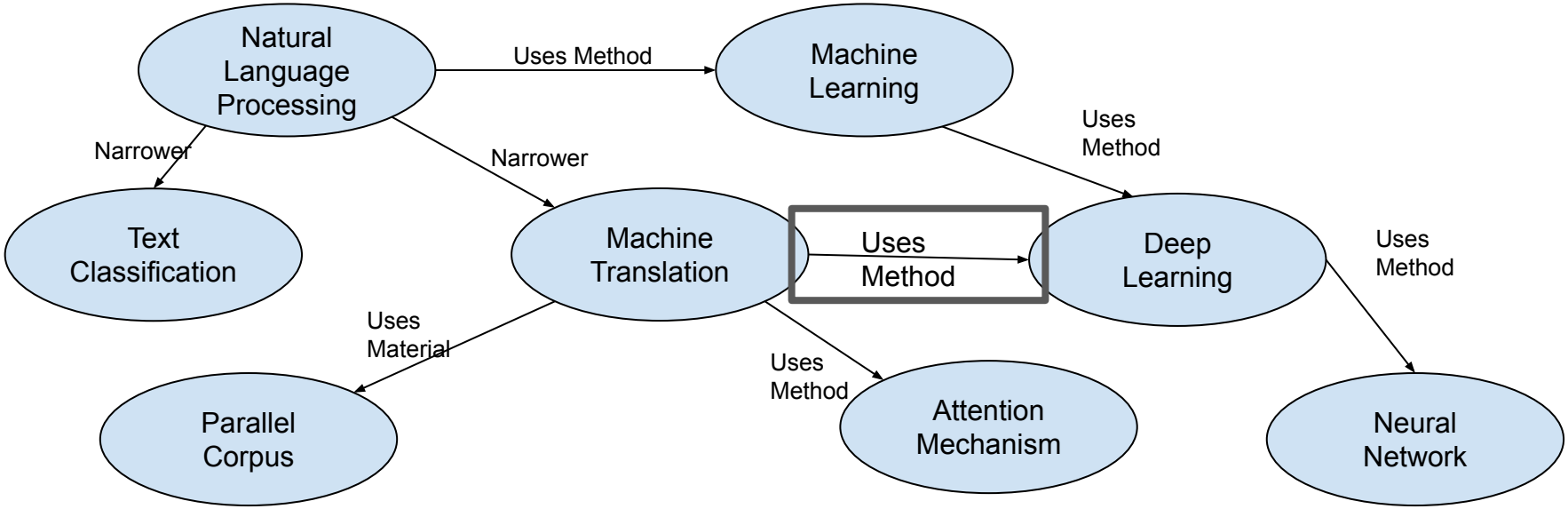


What are Knowledge Graphs? - Relations



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

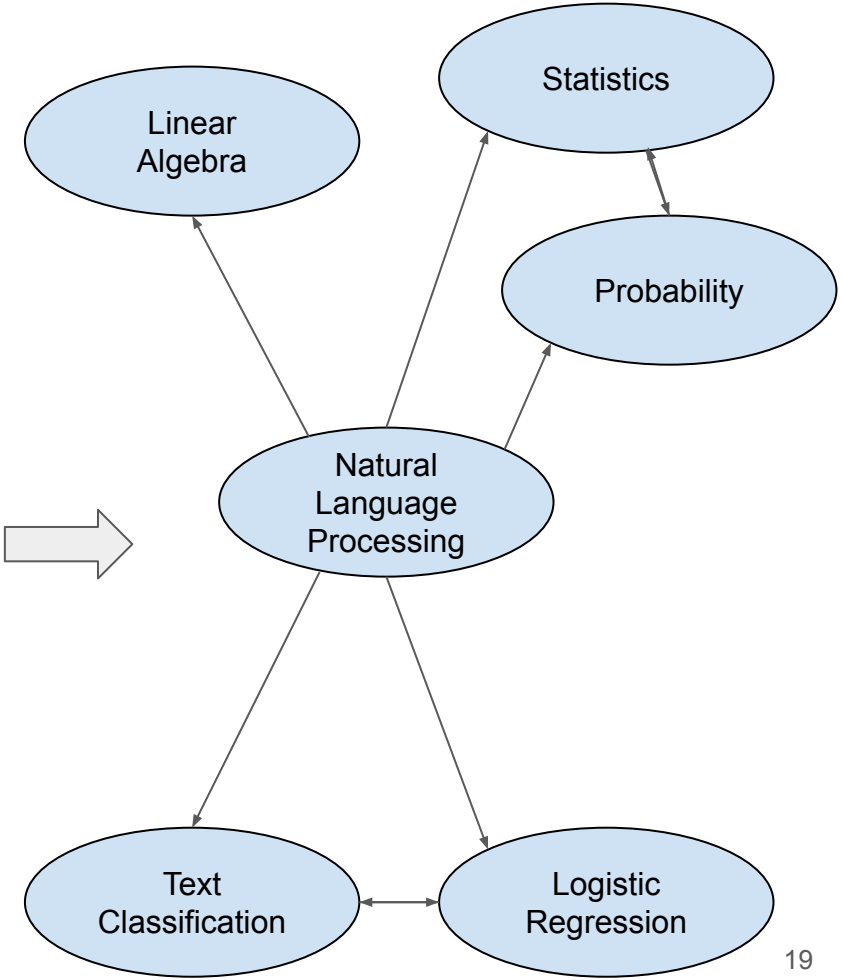
- **Week 1**: Aug. 25
  - Topics: Course information, review of linear algebra
  - Course Materials: Slides
- **Week 2**: Aug. 30, Sept. 1
  - Topics: Review of probability and statistics
- **Week 3**: Sept. 6, 8
  - Topics: Text classification, logistic regression
  - Readings: JE 02
  - Course Materials: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

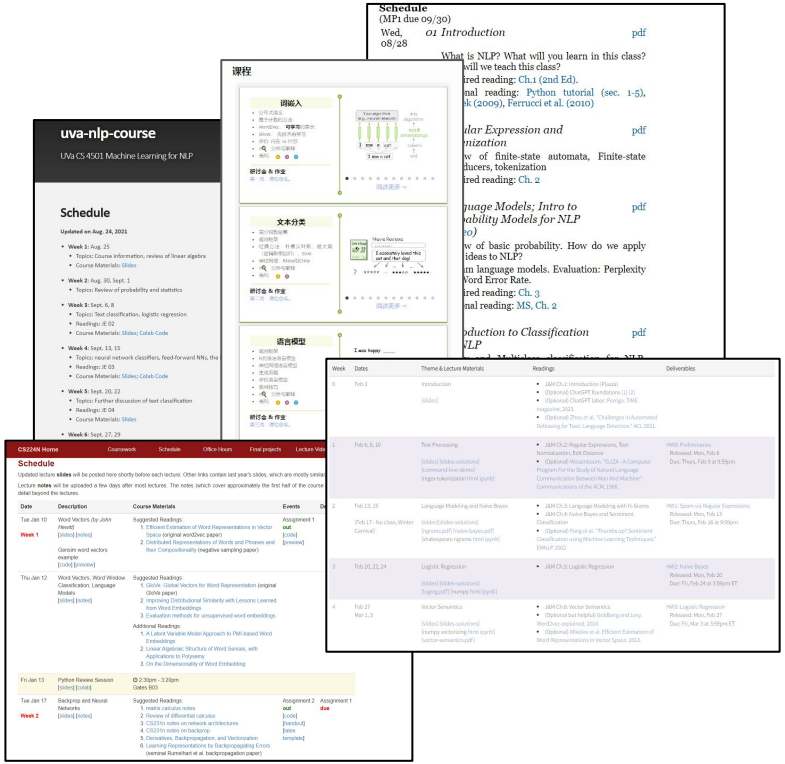
Schedule

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) giv
detail beyond the lectures.

Date	Description	Course Materials	Events	Deadlines
Tue Jan 10 **Week 1**	Word Vectors *(by John Hewitt)* [slides] [notes] Gensim word vectors example: [code] [preview]	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]	
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



Web Extraction Model

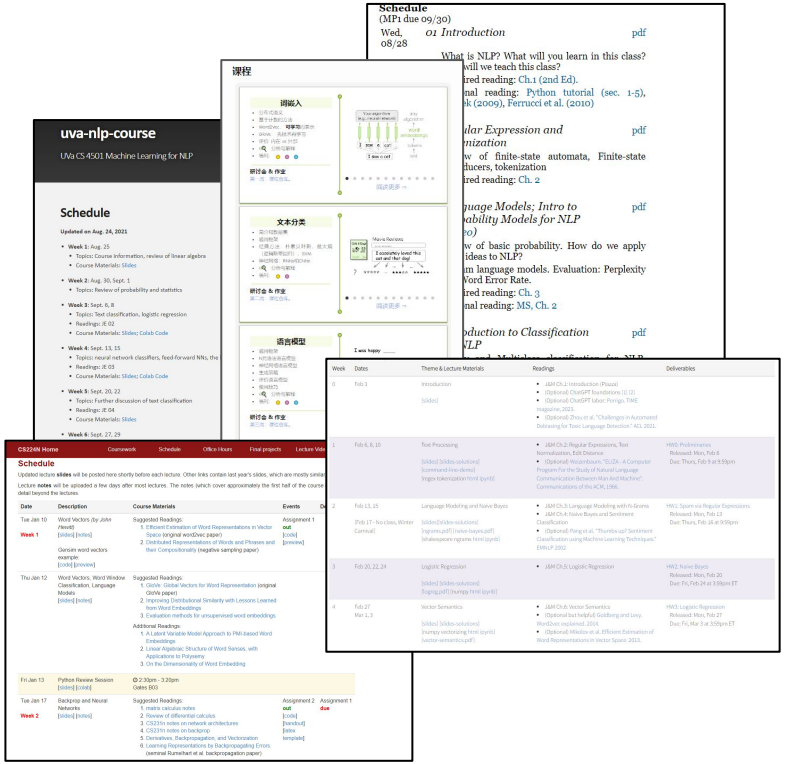
Extracted Concepts

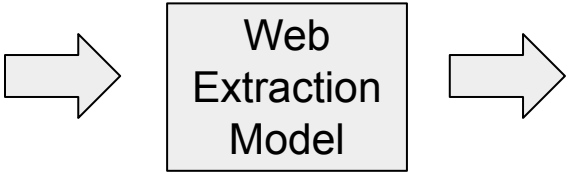
“Language Models”

“Tokenization”

“Logistic Regression”

We want a model to extract these concepts





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

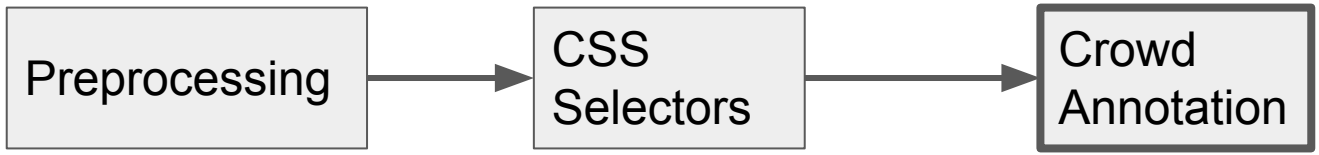


PLAtE - Crowd Annotation

- 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

amazon mechanical turk

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

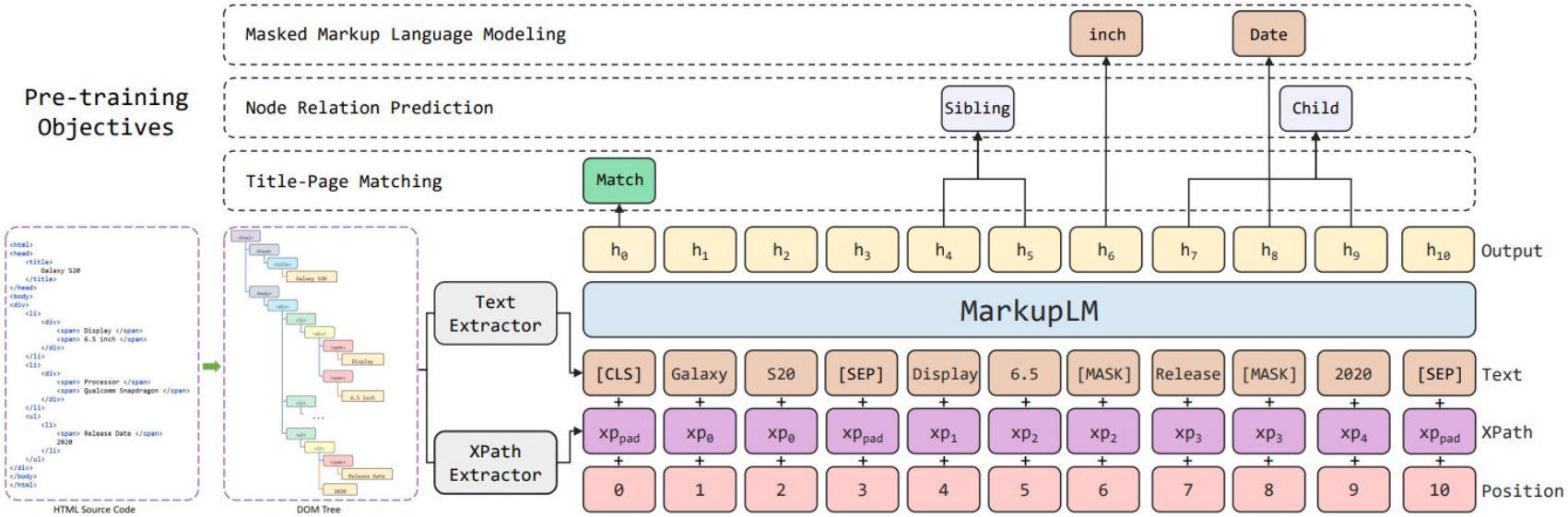
Are there any pieces of text which belong to the **Product Name** and appear **BELOW** the the Target Product

PLAtE - Modeling

MarkupLM [Li, Xu, and Wei 2022]

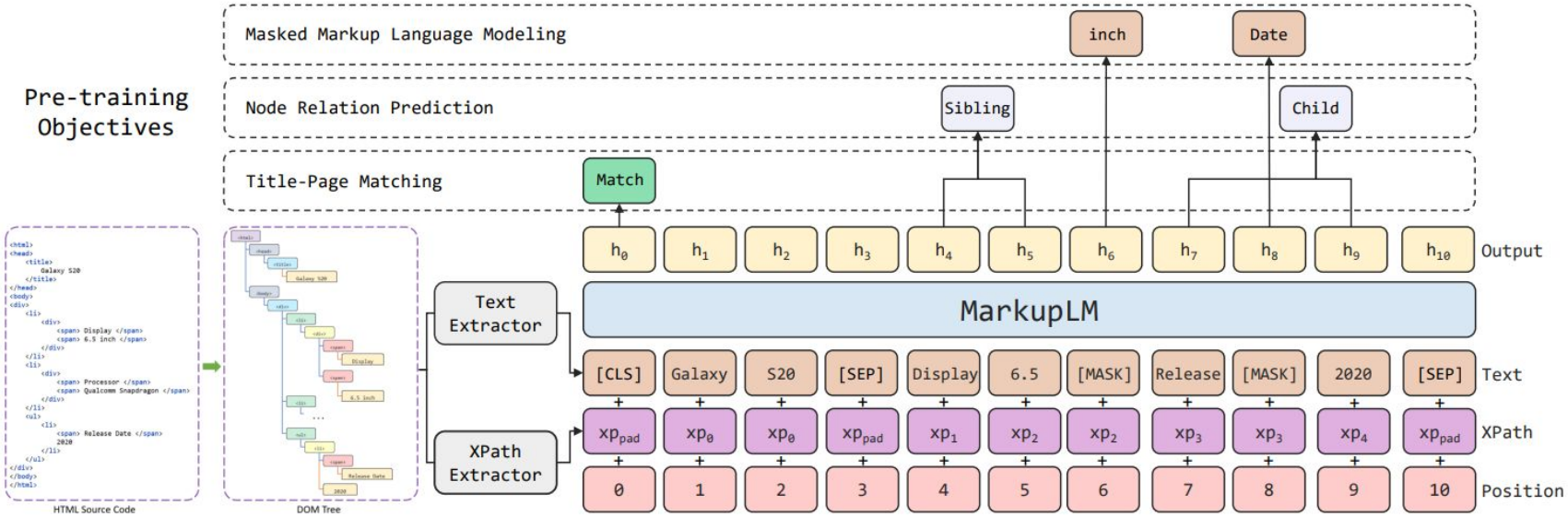
We also evaluated two other models

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



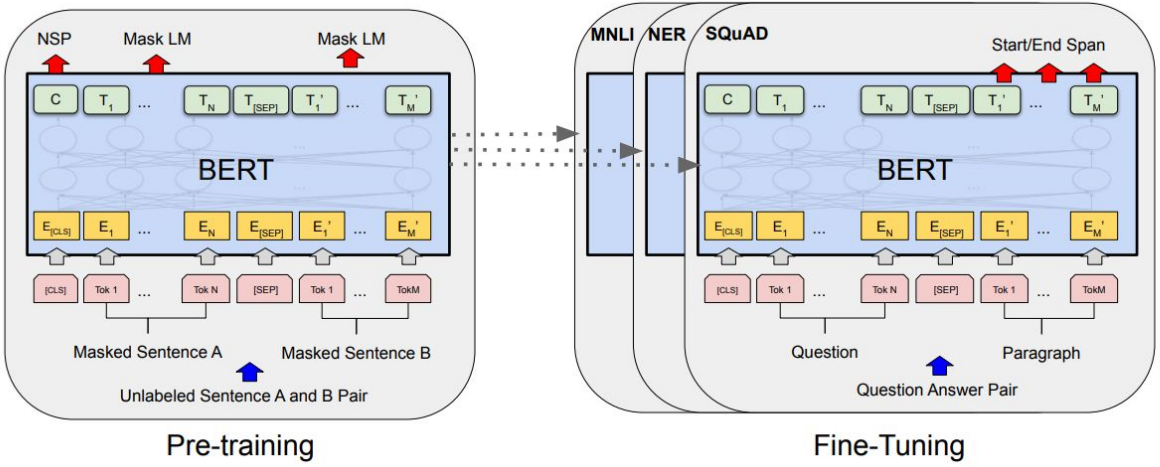
MarkupLM

- Transformer Architecture
  - Same architecture as LaMDA and GPT



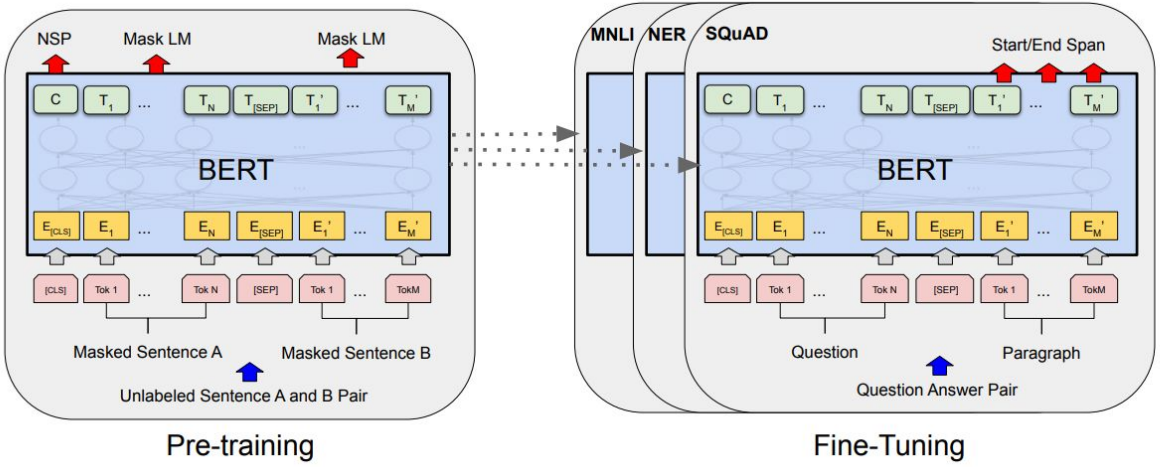
KG Construction Background - Pre-training

- 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 <u>teach</u> 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”

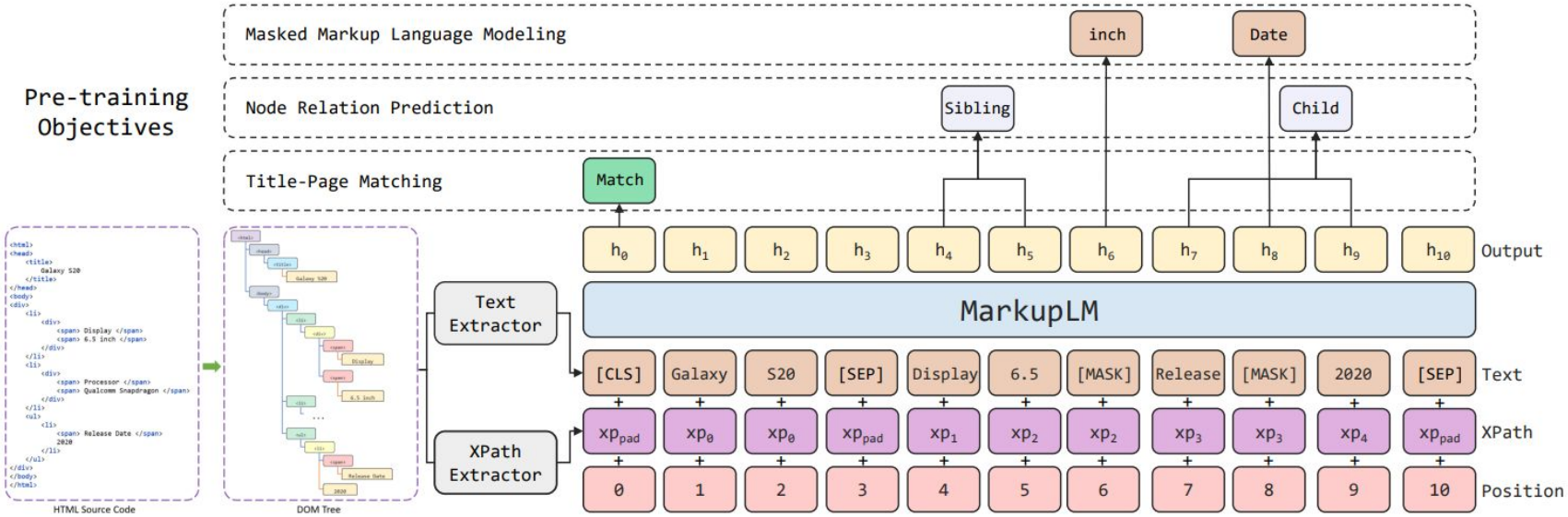
I want to _____ CS1110

Model predicts “teach” ✓

Model predicts a different word ✗

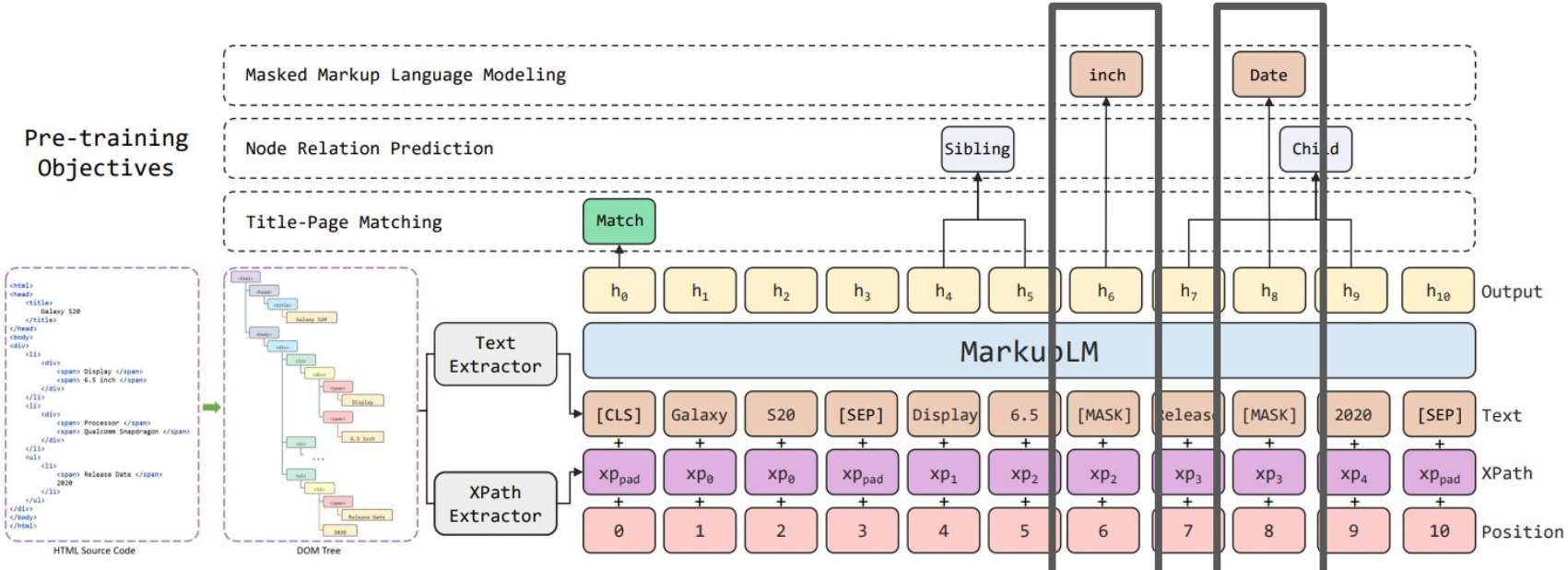
MarkupLM

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



MarkupLM

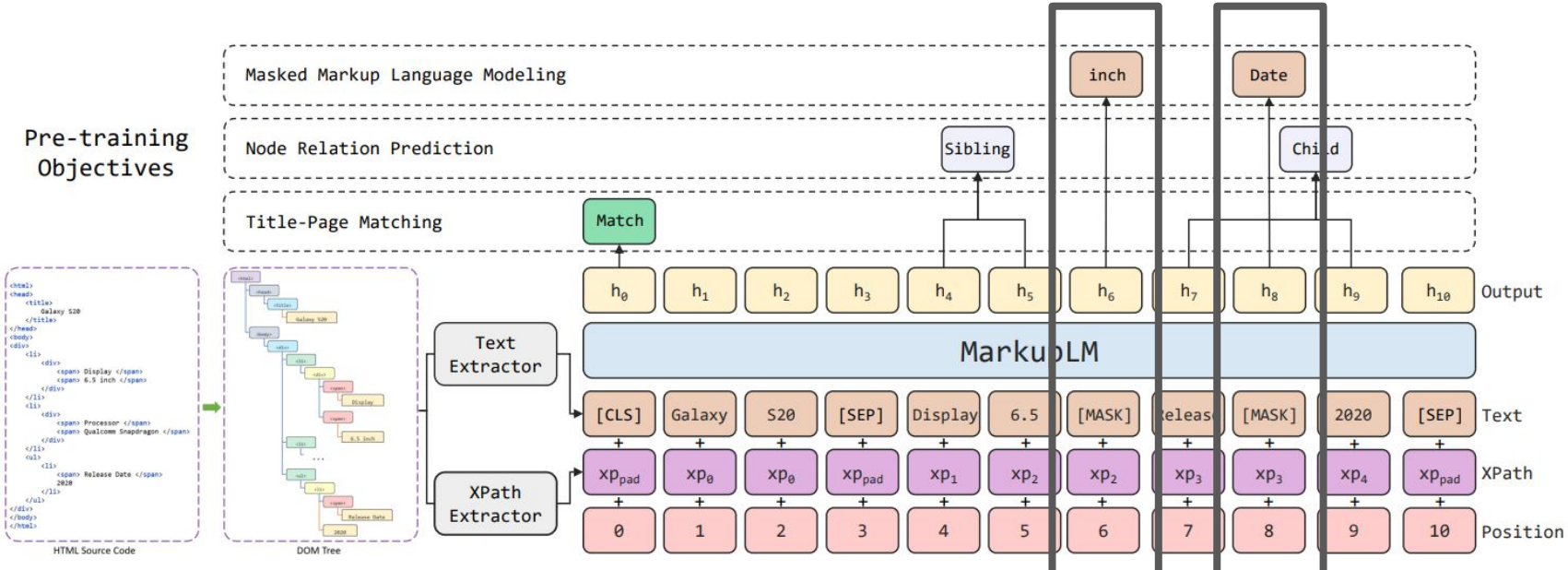
- Masked Markup Language Modeling



MarkupLM

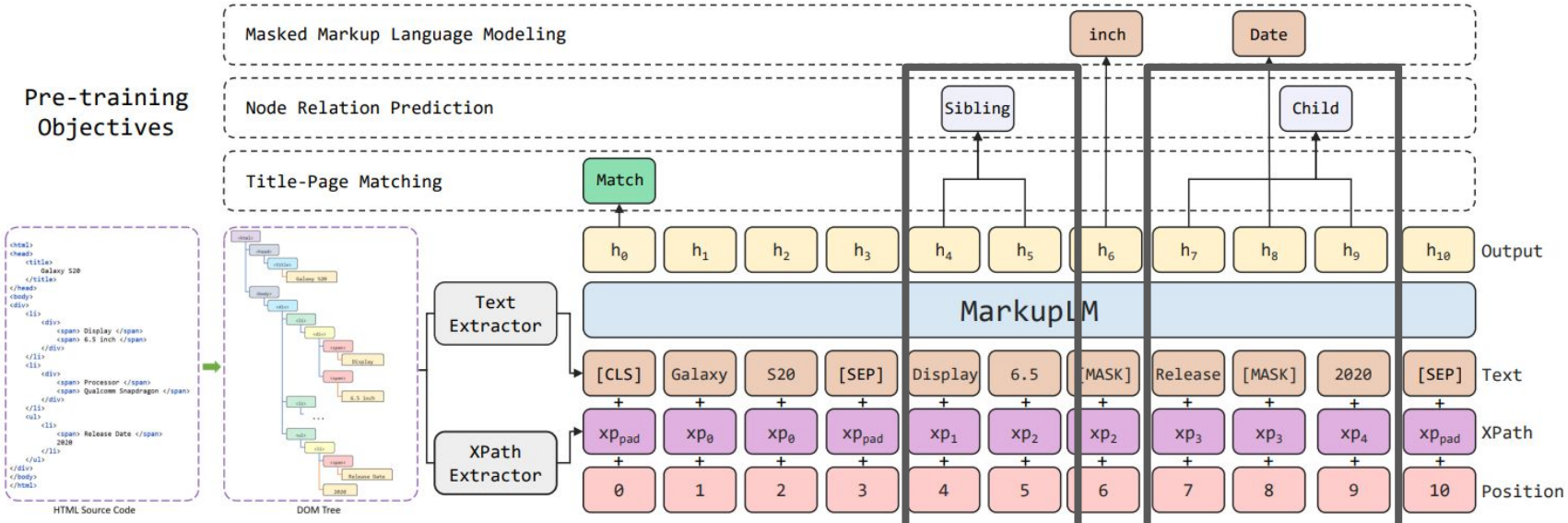
- Masked Markup Language Modeling

What should the [MASK] be?



MarkupLM

- Node Relation Prediction

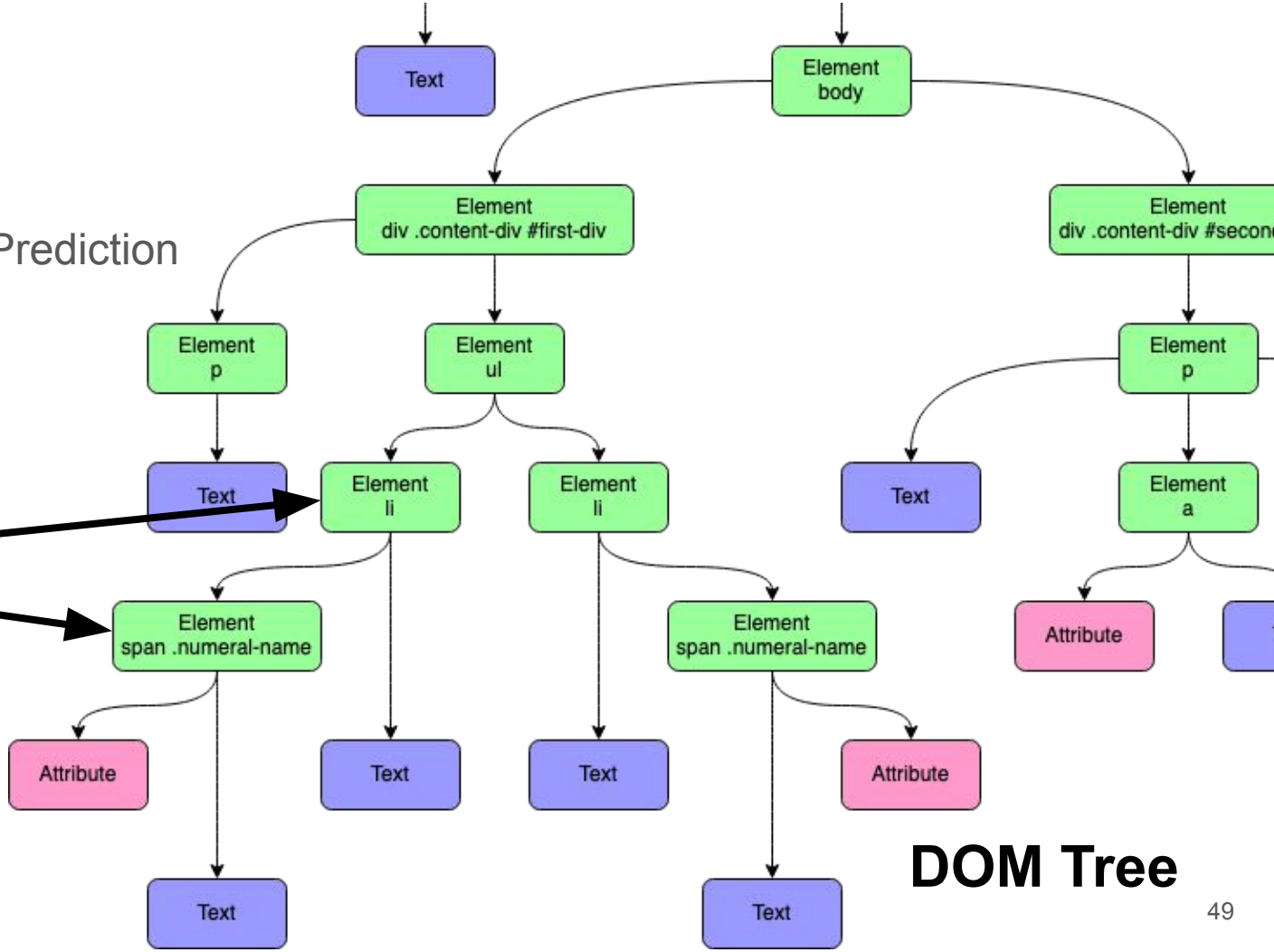


MarkupLM

- Node Relation Prediction

What is the relation?

Child



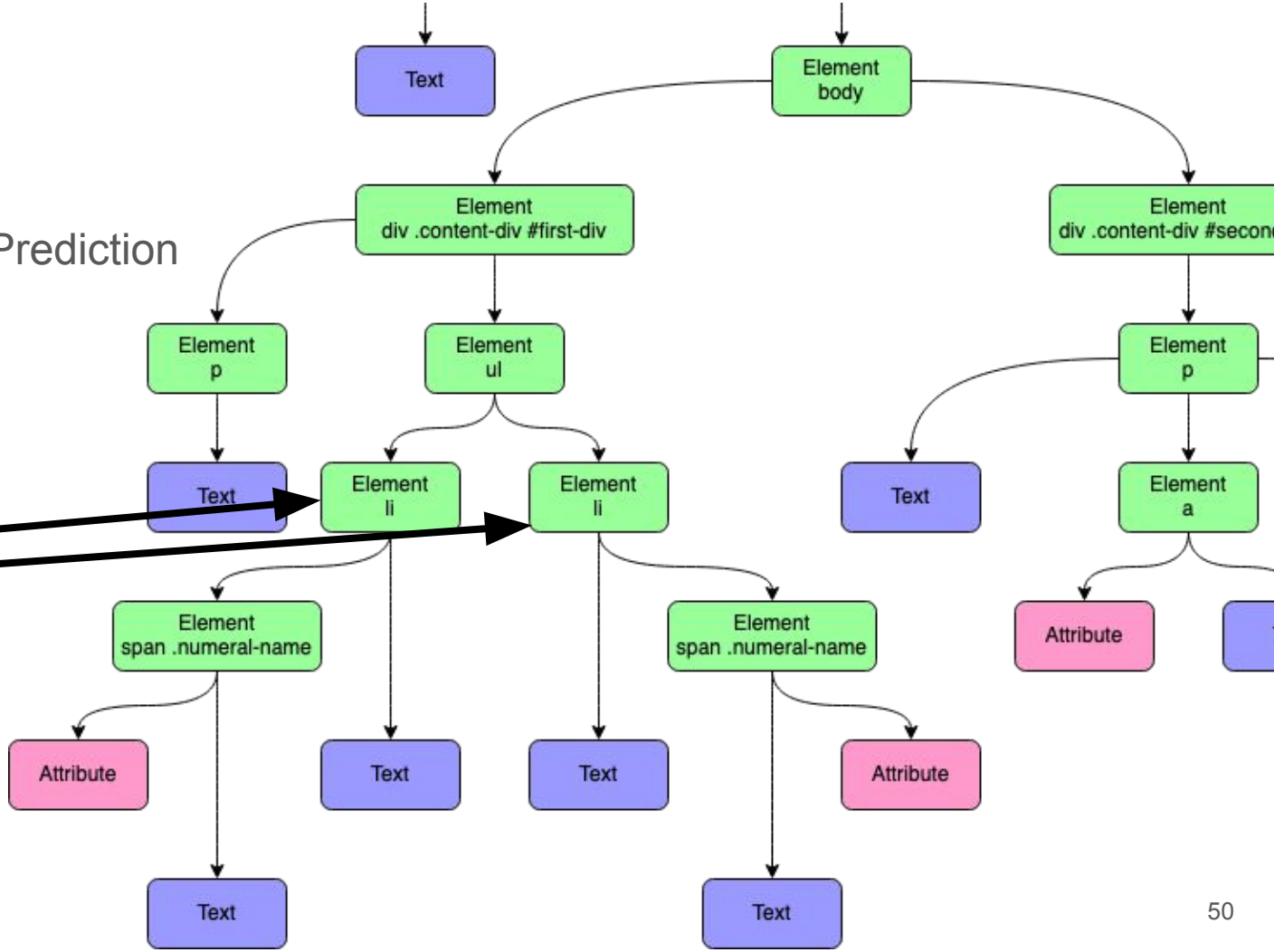
DOM Tree

MarkupLM

- Node Relation Prediction

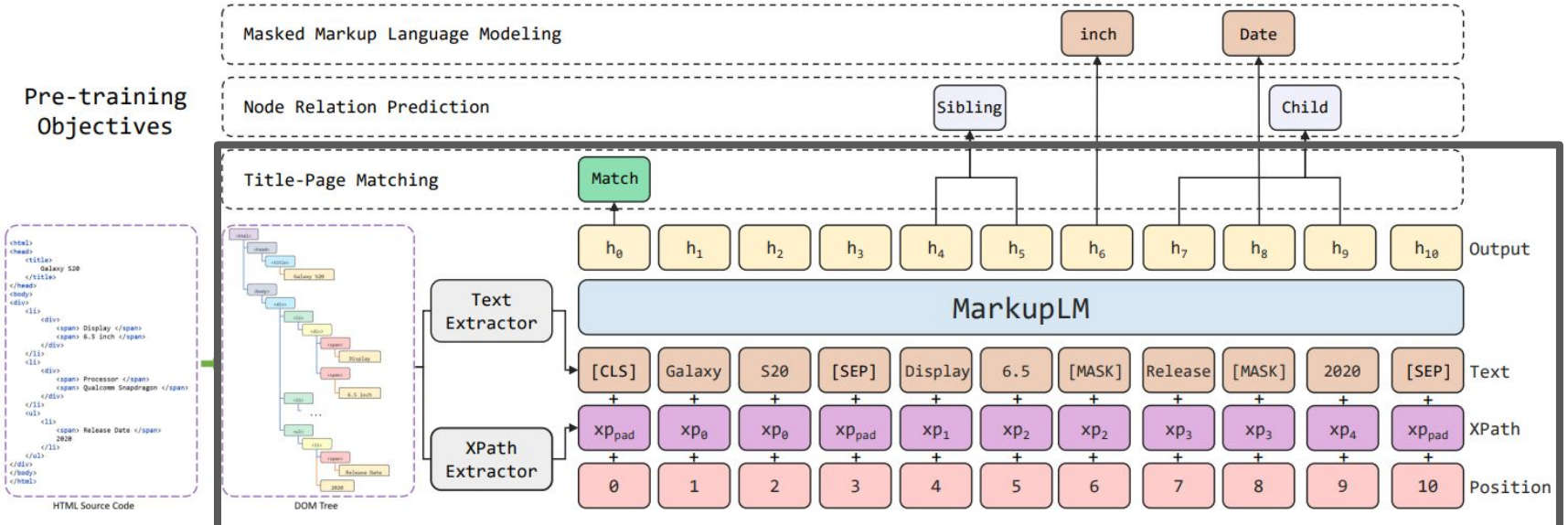
What is the relation?

Sibling



MarkupLM

- Title-Page Matching



MarkupLM

- Title-Page Matching

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

```
<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   How do we represent and capture who does what to whom?   Required reading: "
    <a href="https://web.stanford.edu/~jurafsky/slp3/18.pdf">Ch. 18</a>
    "   Optional reading: "
    <a href="http://aclweb.org/anthology/J/J05/J05-1004.pdf">Palmer et al. (2005)</a>
    ", "
```

MarkupLM

- Title-Page Matching

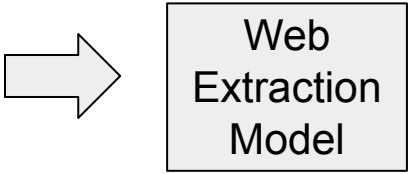
Does the <title> match the rest of the page?

```
<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   How do we represent and capture who does what to whom?   Required reading: "
    <a href="https://web.stanford.edu/~jurafsky/slp3/18.pdf">Ch. 18</a>
    "   Optional reading: "
    <a href="http://aclweb.org/anthology/J/J05/J05-1004.pdf">Palmer et al. (2005)</a>
    ", "
```

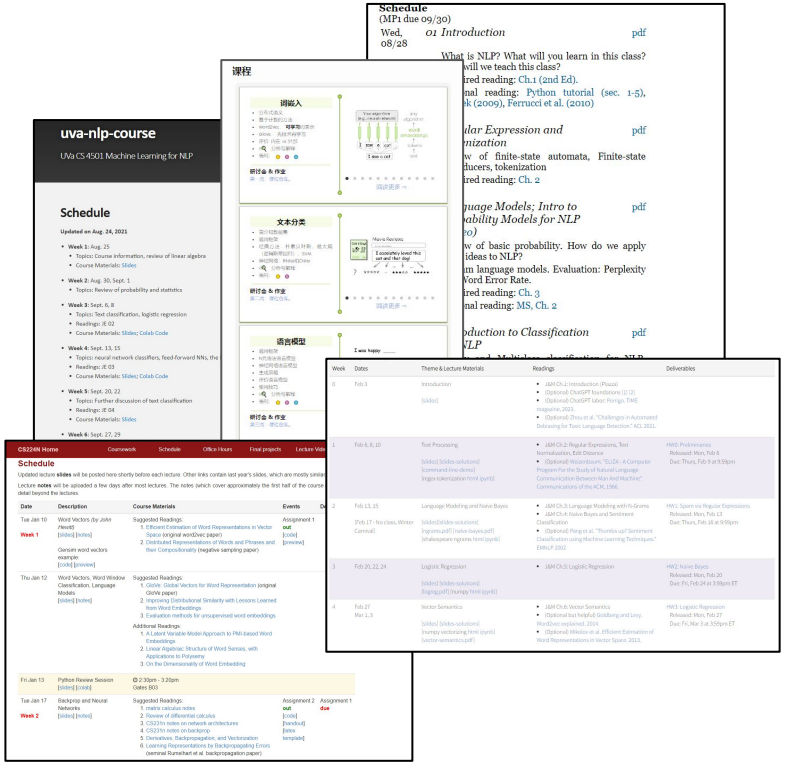
Applying our Dataset

Train on PLAtE

PLAtE Dataset



Perform extraction on webpages of interest



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

Artificial Intelligence Is Booming—So Is Its Carbon Footprint

Greater transparency on emissions could also bring more scrutiny

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

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

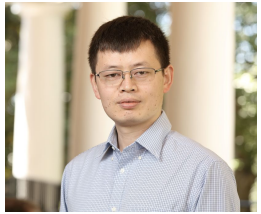
With the rise of the popular new chatbot ChatGPT, colleges are restructuring some courses and taking preventive measures.

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 Ciemiewicz


ChengXiang Zhai


Assma Boughoula


Sanxing Chen


Xiaodong Liu


Renqin Cai


Hongning Wang

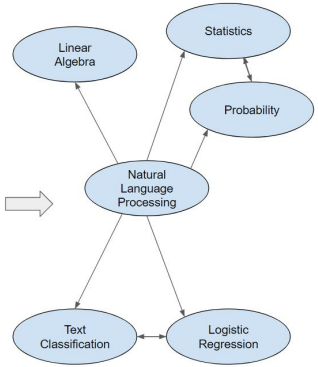

Jibang Wu

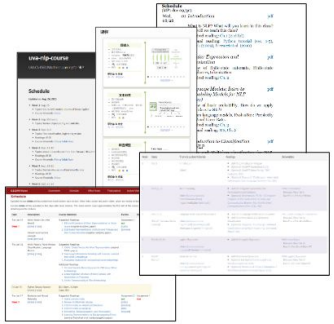

Chong Wang

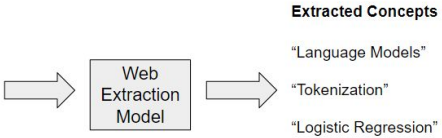
Schedule

Updated on Aug. 24, 2021

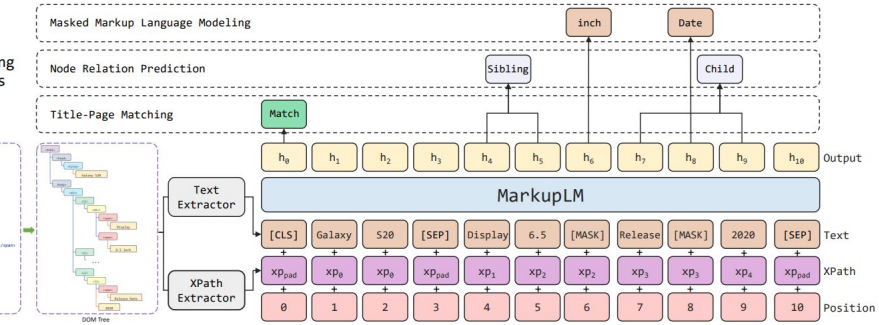
- **Week 1**: Aug. 25
  - Topics: Course information, review of linear algebra
  - Course Materials: Slides
- **Week 2**: Aug. 30, Sept. 1
  - Topics: Review of probability and statistics
- **Week 3**: Sept. 6, 8
  - Topics: Text classification, logistic regression
  - Readings: JE 02
  - Course Materials: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







Thank you!



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

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

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

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://mlnlp-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*.

Dessí, 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). Lamda: 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	0.843	0.652	0.735
DOMLM	0.871	0.815	0.655	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	0.861	0.718	0.728	0.722	0.716	0.764
MarkupLM	0.861	0.769	0.805	0.787	0.771	0.870