Example-Driven Course Slides on Natural Language Processing Concepts

Natalie Parde
Natural Language Processing Laboratory
Department of Computer Science
University of Illinois Chicago
parde@uic.edu

Abstract

Natural language processing (NLP) is a fast-paced field and a popular course topic in many undergraduate and graduate programs. This paper presents a comprehensive suite of example-driven course slides covering NLP concepts, ranging from fundamental building blocks to modern state-of-the-art approaches. In contributing these slides, I hope to alleviate burden for those starting out as faculty or in need of course material updates. The slides are publicly available for external use and are updated regularly to incorporate new advancements.

1 Introduction

Natural language processing is advancing rapidly, making it exciting and difficult to teach. In designing NLP courses, one must build a delicate pedagogical balance between more classical concepts and newer, trendier topics, connecting algorithms to linguistic fundamentals and emphasizing important overlaps with contemporary deep learning. Some material may require frequent updates, and some may grow or shrink in importance over time. Here, I present a comprehensive suite of slides covering classical and modern NLP concepts, designed and iteratively revised over the course of five years at a large, public, minority-serving institution. These slides are primarily based on material from Jurafsky and Martin (2023), and supplemented by material originally published by Pustejovsky and Stubbs (2012), Huyen (2023), and Liu et al. (2023).

2 Intended Use

2.1 Course Structure

These slides were developed for two NLP courses at the University of Illinois Chicago (UIC): *CS 421: Natural Language Processing*, and *CS 521: Statistical Natural Language Processing*. Both are assignment- and project-based courses housed in UIC's Department of Computer Science. Most courses at UIC, including CS 421 and CS 521, are on the semester cycle with 16-week semesters (15 weeks of instructional material and one week reserved only for final exams). Material is delivered via lectures throughout most of CS 421, whereas only five weeks of lectures are included in CS 521 (with the remainder of the course seminar-based). The slides presented within this paper are from the CS 421 and CS 521 lectures during the Fall 2023 (CS 421) and Spring 2024 (CS 521) semesters.

2.2 Student Population

UIC is a large, public university located in Chicago in the United States of America. It is a Minority Serving Institution (MSI), Hispanic Serving Institution (HSI), and Asian American and Native American Pacific Islander Serving Institution (AANAPISI). Fifty percent of undergraduates at UIC receive Federal Pell Grants,[1] and nearly half of first-year students are first-generation college students. UIC's Department of Computer Science includes 2,135 undergraduates, 271 masters students, and 170 PhD students. Nearly 90% of the graduate students in the Department of Computer Science are international students.

CS 421 is classified as a technical elective for the Bachelor of Science (BS) in Computer Science, and it is an option or technical elective for several other CS programs and concentrations. It is required for the BS in CS + Linguistics program (itself housed in the Department of Linguistics). As a result, undergraduates enter the course with broad-ranging technical backgrounds. At the graduate level, CS 421 and CS 521 both count towards the breadth requirement for the CS PhD qualifier exam, meaning that while they are particularly popular for students interested in artificial intelligence research, they are also regularly taken by students from other

[1]These grants are awarded to U.S. undergraduates with exceptional financial need: https://studentaid.gov/understand-aid/types/grants/pell.

Proceedings of the Sixth Workshop on Teaching NLP, pages 4–6

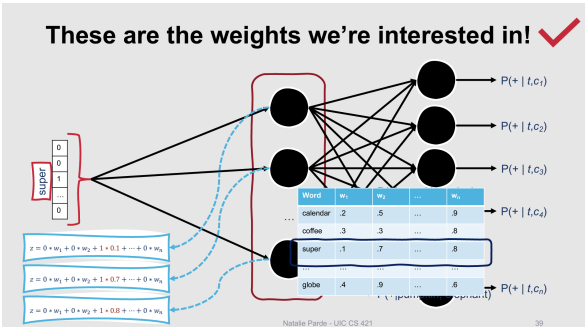

Figure 1: An example CS 421 slide demonstrating how weights are learned using the Word2Vec algorithm.

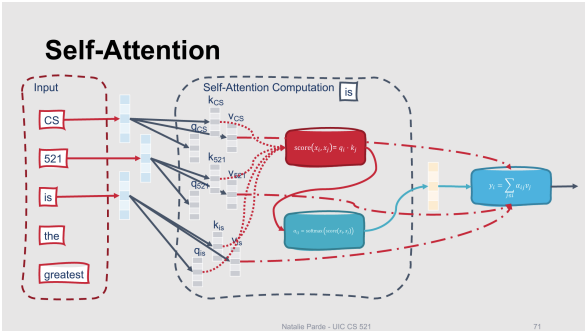

Figure 2: An example CS 521 slide demonstrating how self-attention is computed.

CS subfields. While CS 421 is open to both undergraduate and graduate students (usually slightly more undergraduates than graduate students are enrolled), CS 521 is open only to graduate students and requires CS 421 as a prerequisite. In recent years, CS 421 and CS 521 have had enrollments of approximately 90 and 35 students, respectively. CS 521 is purposely capped at that enrollment size to foster an atmosphere conducive to discussion. CS 421 enrollment caps fluctuate depending on departmental needs and instructional bandwidth.

3 Description of Materials

The slides include many worked-out examples, often drawing from the source textbook(s) and paper(s) but extending or updating them. Videos covering earlier versions of some slides, broken into short segments, are publicly available.[2] Topics covered in the included slides are listed in §3.1. Readings for CS 421 are drawn nearly exclusively from the third edition draft of *Speech and Language Processing* by Jurafsky and Martin (2023). Readings for CS 521 are from *Speech and Language Processing* as well as from *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications* by Pustejovsky and Stubbs (2012), various blogs, and research papers.

3.1 Included Slides

CS 421. Slide decks include: (1) Dialogue Systems and Chatbots; (2) Text Preprocessing and Edit Distance; (3) N-Gram Language Models, Naive Bayes, and Evaluating Text Classifiers; (4) Logistic Regression and Vector Semantics; (5) Word Embeddings and Feedforward Neural Networks; (6) Overview of Deep Learning; (7) Hidden Markov Models and Part-of-Speech Tagging; (8) Constituency Grammars and Constituency Parsing; (9) Dependency Parsing and Logical Representations of Sentence Meaning; (10) Relation and Event Extraction and Temporal Reasoning; (11) Word Senses and WordNet and Semantic Role Labeling; (12) Affective Lexicons and Linguistic Background for Coreference Resolution; and (13) Coreference Resolution and Discourse Coherence. An example slide from the lecture on word embeddings, illustrating how weights are learned using Word2Vec (Mikolov et al., 2013), is provided in Figure 1.

CS 521. Slide decks include: (1) Data Collection; (2) Deep Learning Architectures for Sequence Processing; (3) Machine Translation, Question Answering, and Encoder-Decoder Models; (4) Transfer Learning with Pretrained Language Models and Contextual Embeddings; and (5) Generative AI. An example slide from the lecture on encoder-decoder models, illustrating how self-attention (Vaswani et al., 2017) is computed for language tasks, is provided in Figure 2.

3.2 Slide Organization

Slides in CS 421 can broadly be separated into (1) building blocks of contemporary NLP models, and (2) specific language tasks. The first slide deck falls into neither group but instead was designed to present an exciting application of NLP to students during the first week of class. Building blocks introduce necessary information for completing deliverables throughout the semester, and specific language tasks are presented afterward to enhance understanding of NLP's strong connection to language and ensure familiarity with common terminology and tasks referenced in more advanced NLP classes. Slides in CS 521 build upon one another sequentially. Given the fast-paced nature of NLP research, these are updated more frequently;

[2] https://www.youtube.com/@NatalieParde_NLP

after topics become entrenched in NLP practice, they are either shifted to the CS 421 slide sequence or adapted for more introductory CS 421 use.

3.3 Slide Use and Reuse

The slides discussed here and made available with the *Sixth Workshop on Teaching NLP* are also publicly available on my personal website, which requires no special access privileges to download materials.[3] The slides are made available as PDF files for direct reuse. Source files in Microsoft Powerpoint (.pptx) format are available upon request via email. Slides are shared under a CC BY-NC-SA 4.0 license,[4] which allows non-commercial uses of the work with attribution; any adaptations of the work must be shared under the same licensing terms. All images used in the slides were acquired in one of the following ways: screen captures, images or icons provided via Microsoft Powerpoint or Canva, graphics created by me using shapes and/or other drawing tools in those software applications, or downloads from public domain sources (e.g., https://commons.wikimedia.org).

3.4 Slide Updates

Slides will be regularly updated each time that I teach CS 421 or CS 521, and the updated versions will be made available at the same link on my website[3] (older versions of the slides are also available at that link). Given the dynamic nature of the field, updates can be substantial between course iterations. For instance, as large language models (LLMs) loom larger in the public consciousness, they are also likely to be a larger driver for enrollment for CS 421; previously, the more advanced machine learning concepts powering contemporary LLMs were less appropriate and of less interest to individuals fitting the target CS 421 profile. Course topics in CS 421 may correspondingly shift to include more focus on accessible introductory generative AI concepts and more discussion of tradeoffs between general-purpose LLMs and specially designed tools (e.g., syntactic parsers) for the tasks covered. Another topic included in CS 521 that could be ripe for adaptation to CS 421 would be practical guidelines for data development and use, for example through coverage of data sheets, open versus closed data and models, data anonymization, fair use, and data and model release.

[3] https://www.natalieparde.com/teaching.html

[4] https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode.en

4 Conclusion

This paper described comprehensive, example-driven slides designed and iteratively revised over the span of five years at a large public university to scaffold students' understanding of NLP. The slides are publicly available, and they are regularly updated whenever I teach CS 421 or CS 521; these updates can be accessed via my website.[3] I hope that these materials are useful for new NLP faculty or current faculty looking to update their course content in this fast-paced field.

Acknowledgements

Thank you to the anonymous reviewers, who had excellent ideas for future course lectures; I plan to incorporate these in upcoming semesters. Time spent developing these materials was sponsored in part by a faculty startup grant from the University of Illinois Chicago. CS 421 and CS 521 were both originally conceived by Barbara Di Eugenio, and her syllabi were used as the basis for the first time I taught each of these courses. My own material naturally evolved over time from that starting point. This paper describes only the material included in CS 421 and CS 521 when I teach it personally; content included in course sections taught by other faculty members may vary considerably.

References

Chip Huyen. 2023. Rlhf: Reinforcement learning from human feedback.

Daniel Jurafsky and James H. Martin. 2023. *Speech and Language Processing (3rd Edition)*. Draft.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.