Concept Extraction Dataset from Presentation Slides of CS Courses

See the poster + references

Raja Krishnaswamy[1*], Jessica Hsin[1], Titus Kurian[1], Owen Finucan[1], Timothy Mooney[1], Jonah Smith[1], Naveena Nagaraju[1], Yixiao Li[2], Jacob Hoffman[1], Nathan Ong[3], Michael Miller Yoder[1], Daniel Mosse[1]

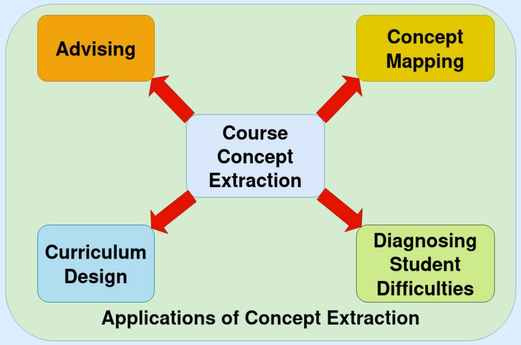
[1]University of Pittsburgh [2]Penn State University [3]Duquesne University

University of Pittsburgh
Learning Research and Development Center

Overview

Problem: creating explicit lists of course concepts takes significant time and effort

.



Solution: Course Concept Extraction (CCE), i.e. automatically extracting the concepts of a course from course materials.

- Focus on **presentation slides:**
  - Abundant course-specific data
  - Well-suited for extracting the concepts for a specific course.
  - N**o ground-truth dataset exists** for CCE on slides.

Contribution:

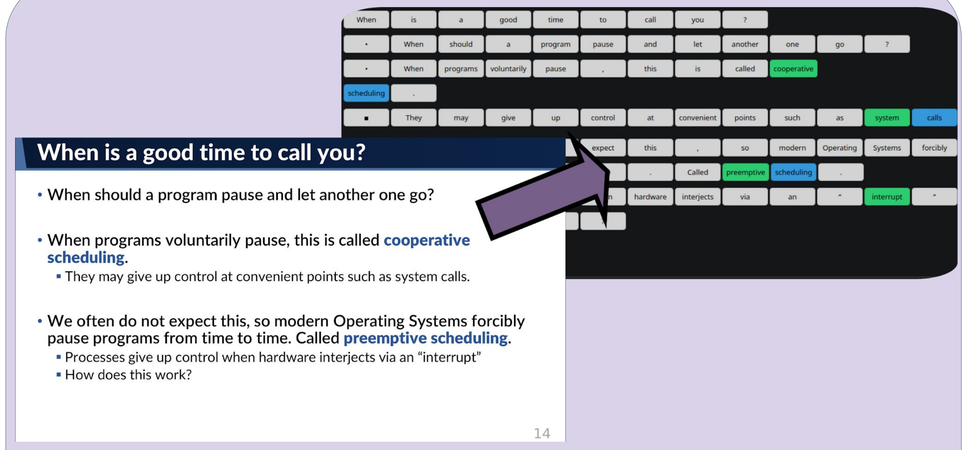
- **Assembled novel dataset** of CS presentation slides with concept ground truths
- **Evaluated** five **SotA models**, demonstrating the task's difficulty.

Background

Concept extraction has been pursued in various domains:

- **medicine** (Abbas et al. 2021; Yang et al. 2020)
- **law** (Modi et al., 2023).
- **education corpora**
  - **code submissions for assignments** (Moraes, 2021)
  - **textbooks** (Li et al., 2022; Wang et al. 2023)
  - **MOOC video lecture transcripts** (Albahr et al., 2021)
  - **combinations** of them (Boughoula et al., 2020).

Rely on **abundant text**; but course presentations are **succinct** and **visual**



An Example of Concept Extraction from Presentation Slides

Dataset

- Eight undergraduate CS courses from the University of Pittsburgh
- Three different versions/formats:
  a. originally delivered PowerPoint/PDF slides,
  b. the text extracted
  c. concept-annotated file.

Table 1. Dataset Characteristics: Means.

	Slide	Slide Deck	Course
Words	63.7	1883	41488
Tagged Concepts	3.136	88.6	1872
Unique Concepts	0.494	20.5	231.5
Slides	1	28.6	596.9
Slide Decks	-	1	20.9

The mean is read as row per column.

Total number of elements in the Dataset

Element	Total
Words	314388
Tagged Concepts	14976
Unique Concepts	2360
Slides	4775
Slide Decks	167
Courses	8

Annotation

- Extracted typed text in the lecture slides by using pdftotext.
- Developed an annotation software
  - Annotation on a token-by-token basis
- Annotated by six experts
  - followed an iteratively developed codebook
  - each annotated 1 full course and parts of others
  - Fleiss's Kappa interannotator agreement was Moderate (κ = 0.50)
    - Landis-Koch scale (Landis & Koch, 1977).

Evaluation Results

We evaluate a SotA LLM, **GPT 6 Sol (OpenAI, 2026)**; a highly optimized classifier leveraging long-range graph and semantic features, **GEST (Martínez-Cruz et al., 2025)**; a SotA keyword extractor, **TNT-KID (Martinc et al., 2022)**; and two encoder baselines, **RoBERTa (Y. Liu et al., 2019)** and **XLNet (Z. Yang et al., 2019)**.

Error Analysis

- GEST extracts important phrases that are not key takeaways.
- GPT 6 is less precise and extracts very long concepts.
- XLNet and RoBERTa output short concepts and partial words
- All models struggle on generic terms, symbols/code, and prior knowledge.

Quantitative Results

Table 4. Model Evaluation on Course Concept Extraction

Model	Prec.	Recall	F1 Score
GEST	**0.47**	0.64	**0.54**
GPT 6 Sol	0.43	**0.70**	0.52
RoBERTa	0.40	0.61	0.48
TNT-KID (Base)	0.34	0.10	0.15
TNT-KID (Chunk)	0.29	0.23	0.25
XLNet	0.38	0.57	0.45

GEST, XLNet, and RoBERTa were finetuned and evaluated using a leave-one-course-out strategy, while the others were directly evaluated without finetuning.

Conclusion

- Created a novel dataset comprised of lecture slides from eight CS courses and their corresponding concept annotations
- Evaluated SOTA models on the CCE task for the dataset, and none were able to get a F1-score above 0.54.
- Further research must be done to obtain a well-performing automatic course concept extractor.