Catalog: An educational content tagging system

Saad M. Khan
FineTune Learning
saad@finetunelearning.com

Joshua Rosaler
FineTune Learning
josh@finetunelearning.com

Jesse Hamer
FineTune Learning
jesse@finetunelearning.com

Tiago Almeida
FineTune Learning
tiago@finetunelearning.com

ABSTRACT

We present Catalog, an educational content classification and alignment system that tags learning and assessment content in a semantically meaningful and accurate manner. Unlike other approaches that rely on keywords or search terms and crosswalks between knowledge taxonomies, Catalog utilizes powerful NLP, specifically language models based on the Transformer architecture, to encode content in a context attentive fashion. This allows us to capture deep conceptual and contextual relations in content to classify it against a wide variety of educational standards and taxonomies. We present results from empirical studies demonstrating efficacy of our approach in classifying learning content to the Next Generation Science Standards (NGSS).

Keywords

Content tagging/classification, NLP, Transformer Networks

1. INTRODUCTION

Tagging educational content with the most relevant learning and assessment standards and education search terms is one of the most critical elements in creating highly efficacious content. This enables the tracking of student skill gaps, recommendation of remediative learning content and mastery of discipline topics, skills and cross cutting capabilities. With the ever growing volume of digital learning content and educational standards [12] the demands on tagging content are not being met by current solutions.

Current processes to tag content typically starts with raw untagged content that has to be manually reviewed, understood and analyzed by subject matter experts (SMEs) and then classified against a particular education standard e.g. the NGSS [10] resulting in the first set of foundational standards tags. Typically, these standards are hierarchical and utilize a taxonomic knowledge representation to capture the knowledge structure including core disciplinary knowledge, skills and/or cross cutting capabilities. Given the foundational tags one can transfer onto any number of desired taxonomies, for instance the Common Core State Standards [11], using taxonomy crosswalks [13]. Crosswalks are essentially mappings from one standard's taxonomy to another that have for the most part been developed by SMEs and are many times proprietary limiting their applicability.

While in theory this process seems to offer a relatively scalable solution to the content tagging problem, in practice it is inefficient and has significant limitations. Firstly, the initial step of creating the foundational tags is manually executed and highly subjective, making it expensive and error prone. But even when that is done well the taxonomy crosswalks do not offer a perfect solution, because these crosswalks are not one-to-one mappings between the tags of one taxonomy and the other. Due to the hierarchical nature of the standards taxonomies and how they are designed and crafted by SMEs, oftentimes there are vast differences in the levels of knowledge abstraction, resulting in many-to-many mappings for the crosswalks connecting them. The end result is that for a given unit of content even when there is a foundational tag available and using an associated crosswalk, SMEs still have to make the final adjudication of the most appropriate tag in the target standard's taxonomy.

To address these challenges we have developed Catalog, an automated content classification system that leverages recent advances in NLP, specifically the Transformer architecture. This allows us to analyze educational content with richer context-aware text embeddings and pre-trained language models. We have evaluated the accuracy of our approach with promising results on an OpenStax Biology textbook [14] with ground truth NGSS tags (human experts labeled). We believe Catalog can significantly help streamline and accelerate manual workflows around content tagging and curation. These are applicable for both existing or new content, enriching existing content tags for more targeted search, discovery and recommendation as well as maintaining content alignments as educational standards evolve.

2. TECHNICAL APPROACH AND SYSTEM DETAILS

2.1 Transformers and Text Embeddings

At the core of Catalog's content classification tagging system is the Transformer architecture, first proposed by Vaswani et al in 2017 [2]. Catalog utilizes a series of pre-trained Transformer models [5, 6] to encode text-based content in vectorized features which are then further used to analyze the probability that the content is related to a textual description of the target taxonomy. Further details of this approach are presented in the following subsections. Here we present a brief overview of the Transformer architecture.

By eschewing the sequentially-processed nature of previous deep-learning NLP architectures (like LSTMs [3] and GRUs [4]) in favor of *multi-head attention*, the Transformer architecture is highly parallelizable and scalable, allowing for richer context-aware text embeddings and a substantial pre-training capacity which allows for a transfer learning approach to NLP tasks. Since its inception, research into the Transformer architecture has exploded, with variants such as Google's BERT

[5] and OpenAI's GPT series [6, 7, 8] topping several NLP benchmarks, such as the multitask GLUE suite [9].

As indicated above, Transformers are a deep-learning architecture based on the *attention mechanism*. The original formulation of the Transformer architecture used a variant known as *scaled dot-product attention*, defined as

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})\, V,$$

where the matrices Q and K are called the *queries* and *keys*, respectively, and each have column-dimension d, while the matrix V is called the *values* and has column-dimension d'. When the queries, keys, and values are all equal to some matrix X, the resulting operation is called *self-attention*. The rows of this matrix X correspond to context-independent feature vectors of the tokens of the input text, each with a small *positional encoding vector* added so that the model is aware of each token's position within the input sequence. Self-attention can be thought of as an operation which recomputes each token as a linear combination of the other tokens, where the weights of the linear combination correspond a scaled dot-product similarity score (the $\frac{QK^T}{\sqrt{d}}$ term).

In this way, (potentially long-range) interactions between tokens are captured. To allow the Transformer to learn different patterns of interaction, several matrices of learnable weights are used to compute *multi-head self-attention*:

$$\text{MultiHead}(X, X, X) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O,$$

where

$$\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V),$$

and all of the W matrices consist of learnable weights. After multi-head self-attention is computed, the resulting feature vector is fed to a single-hidden-layer feedforward neural network for aggregation and resizing. These two consecutive operations, multi-head self-attention followed by the feedforward neural network, constitute the core of a *Transformer block*. A *Transformer model*, then, is built by chaining several Transformer blocks together, each potentially with their own set of weight matrices.

2.2 How Catalog Works

Catalog's AI-powered content tagging system utilizes a Transformer-based semantic matching engine to rank taxonomic categories by their semantic similarity to given educational content. The semantic matching algorithm works as follows. We are given a collection of textual descriptions of taxonomic categories (e.g. NGSS [10]), which we refer to as "documents," and the raw text of educational content, referred to as the "query" that needs to be classified. For each document, we produce a string of input text by combining it with the query along with a small amount of connective text. Using a Transformer model pre-trained for next token prediction, we then process the input string to convert the query tokens into feature vectors. These feature vectors are then further processed to produce probabilities for each query token, conditioned on the document text. Additionally, we process the query text by itself in order to determine unconditioned probabilities for the query tokens. Finally, a match score is produced for each document by comparing the conditioned vs. unconditioned query token probabilities and then aggregating these into a single real-valued score. Documents are then ranked according to these scores, with a higher score indicating a higher match similarity.

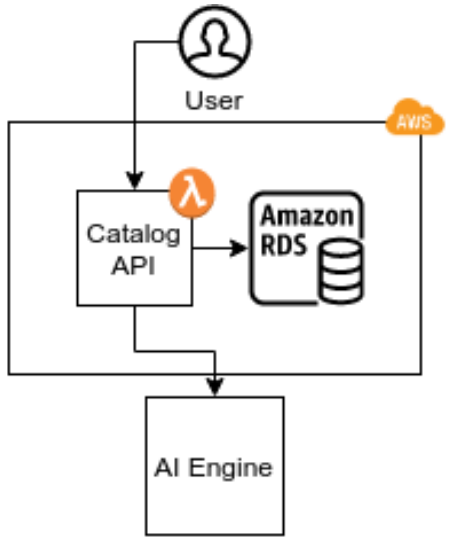
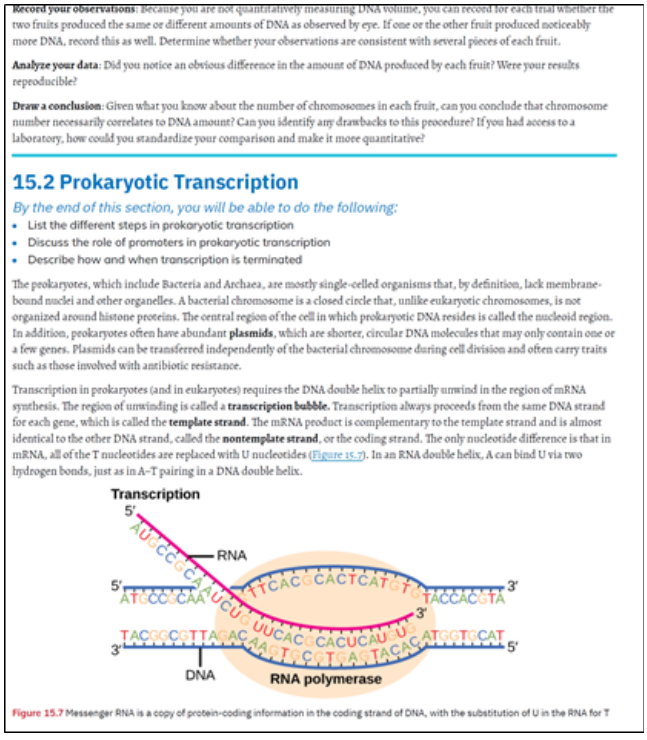


Figure 1: Catalog system architecture is designed to be modular with disturbed services hosted on AWS.


Record your observations: Because you are not quantitatively measuring DNA volume, you can record for each trial whether the two fruits produced the same or different amounts of DNA as observed by eye. If one or the other fruit produced noticeably more DNA, record this as well. Determine whether your observations are consistent with several pieces of each fruit.

Analyze your data: Did you notice an obvious difference in the amount of DNA produced by each fruit? Were your results reproducible?

Draw a conclusion: Given what you know about the number of chromosomes in each fruit, can you conclude that chromosome number necessarily correlates to DNA amount? Can you identify any drawbacks to this procedure? If you had access to a laboratory, how could you standardize your comparison and make it more quantitative?

15.2 Prokaryotic Transcription

By the end of this section, you will be able to do the following:

- List the different steps in prokaryotic transcription
- Discuss the role of promoters in prokaryotic transcription
- Describe how and when transcription is terminated

The prokaryotes, which include Bacteria and Archaea, are mostly single-celled organisms that, by definition, lack membrane-bound nuclei and other organelles. A bacterial chromosome is a closed circle that, unlike eukaryotic chromosomes, is not organized around histone proteins. The central region of the cell in which prokaryotic DNA resides is called the nucleoid region. In addition, prokaryotes often have abundant **plasmids**, which are shorter, circular DNA molecules that may only contain one or a few genes. Plasmids can be transferred independently of the bacterial chromosome during cell division and often carry traits such as those involved with antibiotic resistance.

Transcription in prokaryotes (and in eukaryotes) requires the DNA double helix to partially unwind in the region of mRNA synthesis. The region of unwinding is called a **transcription bubble.** Transcription always proceeds from the same DNA strand for each gene, which is called the **template strand**. The mRNA product is complementary to the template strand and is almost identical to the other DNA strand, called the **nontemplate strand**, or the coding strand. The only nucleotide difference is that in mRNA, all of the T nucleotides are replaced with U nucleotides (Figure 15.7). In an RNA double helix, A can bind U via two hydrogen bonds, just as in A–T pairing in a DNA double helix.

Transcription

RNA

DNA

RNA polymerase

Figure 15.7 Messenger RNA is a copy of protein-coding information in the coding strand of DNA, with the substitution of U in the RNA for T

Figure 2: Sample page from the OpenStax Biology 2e textbook used in our experiments.

System architecture and implementation wise, Catalog has two core architectural components as shown in figure 1. The first is a Lambda API Endpoint that leverages the serverless architecture and serves mainly as an interface between the user and the Transformer-based semantic query process, the "AI Engine". It authenticates users' requests, manages requests and accesses the system database. The second major component, "AI Engine"

manages content processing and returns the match scores from classification.

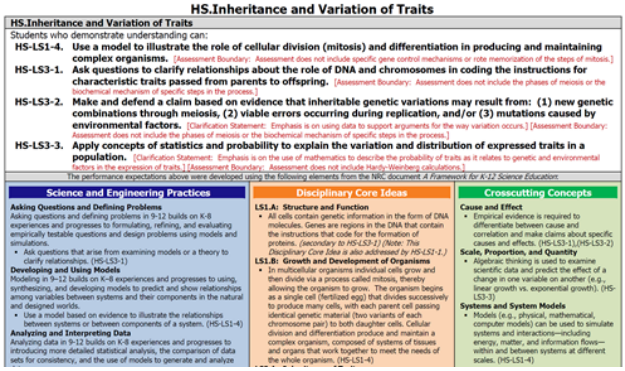


Figure 3: Sample from NGSS High School Life Science Biology performance expectation (PE) standards. Image shows 4 of the 24 unique PE standards used in our experiment.

2.3 Experimental Results

We tested the accuracy and performance of our approach on a learning content dataset extracted from the OpenStax Biology 2e high school textbook [14]. The dataset consists of approximately 500 pages of content spanning 98 chapter/subchapter sections that ranged from 410 to 545 words each. Each of the book's 98 sections is annotated with NGSS High School Life Sciences (HS-LS) performance expectation (PE) tags [10], also provided by OpenStax [14] and served as the ground truth labels in our experiment. There are a total of 24 unique Biology NGSS PE standards applicable to our dataset, essentially rendering this a 24 class classification problem. Figure 2 shows a sample from the OpenStax textbook and in figure 3 we include sample PEs from the 24 NGSS standards used in our experiment. We note that these ground truth labels are not necessarily unique: each section is associated with one to three NGSS tags.

Topic documents for the 24 PE standards were assembled from the Topic Arrangements of the NGSS that includes descriptions of PEs, Science and Engineering Practices, Disciplinary Core Ideas, Crosscutting Concepts. Because our model predicts NGSS tags for a given OpenStax section by ranking them, we assess performance by computing the *top-n overall accuracy*, that is, the proportion of predictions which have at least one ground truth label in their top-*n* ranked predictions (note that for $n = 1$, this is just the traditional overall accuracy measure). For comparison, we had an SME perform this classification exercise manually i.e. provide up to three suggested NGSS PE tags for each of the 98 book sections in our dataset. This SME is a high-school science teacher in a New York city school district and is highly experienced with the NGSS standards.

Before examining the results of this experiment, we note that one NGSS standard, HS-LS1-2, was severely overrepresented in our dataset, accounting for nearly 42% of all ground truth tags, more than 5 times the next-most-represented tag. To account for this in our accuracy computations, we decided to take 1000 random subsamples of this class, and then average the top-*n* accuracy over these subsamples. Figure 4 shows the resulting NGSS tag distribution of such a subsample.

Figure 5 shows the top-*n* accuracy averaged over the 1000 subsamples as a function of *n*. When compared to ground truth, the semantic query model achieved 51%, 73%, and 77% top-1, top-2, and top-3 overall accuracy, respectively, among the 24 NGSS PE standards. In contrast, the SME achieved 48%, 68%,

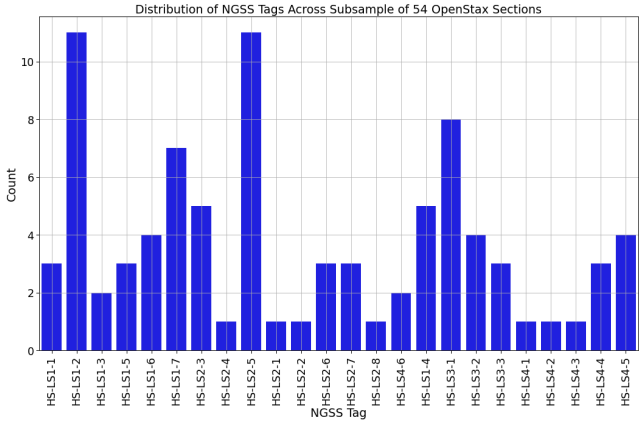


Figure 4: Distribution of NGSS tags across subsample of data. In all, 55 OpenStax sections are associated with tag HS-LS1-2, whereas each subsample randomly selects only 11 of these to be commensurate with the next most represented tag, HS-LS2-5.

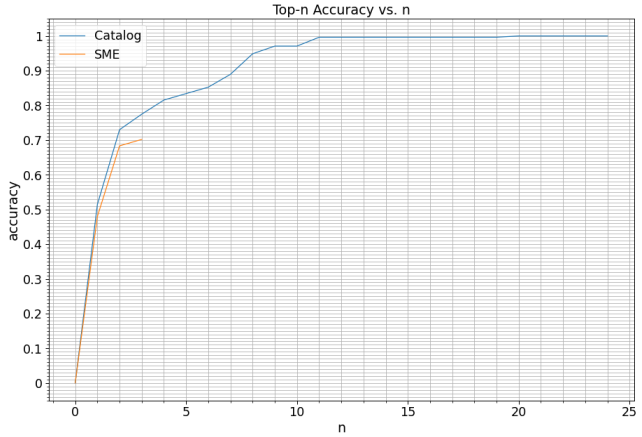


Figure 5: Top-*n* Accuracy vs. *n* for 98 items of section text from the OpenStax Biology 2e textbook, tagged against the NGSS High School Life Sciences performance expectation standards (as above, *n* is the number of top predictions within which at least one ground truth label must fall for the prediction to be counted as correct).

and 70% top-1, top-2, and top-3 overall accuracy, respectively. It should also be noted that it took the SME 520 minutes to complete the manual classification of the dataset, whereas our system completed processing in only approximately 2 minutes.

3. CONCLUSION

In this paper we have introduced Catalog, a NLP based content classification system that utilizes recent advances in transfer learning approaches to deeply and accurately tag educational content against popularly used learning standards. Unlike other approaches that rely on keywords or search terms and crosswalks between knowledge taxonomies, Catalog is built on a language modeling architecture that understands the deep semantic structure and relationship between concepts, topics, learning objectives and other attributes of content. We have presented early results from empirical studies demonstrating efficacy of our

approach in classifying learning content to the Next Generation Science Standards (NGSS).

4. ACKNOWLEDGEMENTS

We would like to thank Sara Vispoel for reviewing our results and insightful discussions and comments on the NGSS standards and taxonomic representations used in K-12 life science education and high school biology in particular.

5. REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 1993), 795-825. DOI= http://doi.acm.org/10.1145/161468.16147.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 5998-6008.

[3] Hochreiter, S. and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9, 8, 1735-1780.

[4] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on EMNLP* (Oct. 2014), 1724-1734. DOI=10.3115/v1/D14-1179

[5] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT (1)*, 4171-4186. DOI=10.18653/v1/N19-1423.

[6] Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. 2018. Improving language understanding by generative pre-training. *Technical report.* *https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf*.

[7] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. 2019. Language models are unsupervised multitask learners. *Technical report* *https://openai.com/blog/better-language-models/*.

[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems.*

[9] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S. 2018a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353-355.

[10] NGSS Lead States. 2013. Next Generation Science Standards: For states, by states. *Washington, DC: The National Academic Press*.

[11] National Governors Association Center for Best Practices, Council of Chief State School Officers 2010. Common Core State Standards. *National Governors Association Center for Best Practices, Council of Chief State School*.

[12] Scott-Little, C., Lesko, J., Martella, J., & Milburn, P. (2007). Early Learning Standards: Results from a National Survey to Document Trends in State-Level Policies and Practices. Early Childhood Research & Practice, 9(1), n1.

[13] Conley, D. T. (2011). Crosswalk Analysis of Deeper Learning Skills to Common Core State Standards. Educational Policy Improvement Center (NJ1).

[14] https://openstax.org/details/books/biology-2e