Finding Educationally Supportive Contexts for Vocabulary Learning with Attention-Based Models

Sungjin Nam[1], **Kevyn Collins-Thompson**[2], **David Jurgens**[2], **Xin Tong**[2]
[1]ACT, Inc., 500 Act Dr, Iowa City, IA 52243
[2]University of Michigan, 105 S State St, Ann Arbor, MI 48109
sungjin.nam@act.org
{kevynct,jurgens,xstong}@umich.edu

Abstract

When learning new vocabulary, both humans and machines acquire critical information about the meaning of an unfamiliar word through contextual information in a sentence or passage. However, not all contexts are equally helpful for learning an unfamiliar 'target' word. Some contexts provide a rich set of semantic clues to the target word's meaning, while others are less supportive. We explore the task of finding educationally supportive contexts with respect to a given target word for vocabulary learning scenarios, particularly for improving student literacy skills. Because of their inherent context-based nature, attention-based deep learning methods provide an ideal starting point. We evaluate attention-based approaches for predicting the amount of educational support from contexts, ranging from a simple custom model using pre-trained embeddings with an additional attention layer, to a commercial Large Language Model (LLM). Using an existing major benchmark dataset for educational context support prediction, we found that a sophisticated but generic LLM had poor performance, while a simpler model using a custom attention-based approach achieved the best-known performance to date on this dataset.

Keywords: contextual vocabulary learning, language acquisition curriculum, attention-based model

1. Introduction

We learn the vast majority of our new vocabulary with significant help from context. Humans acquire the meanings of unknown words partially and incrementally by repeated exposure to clues in the surrounding text or conversation (Frishkoff et al., 2008). As part of literacy training, contextual word learning methods can help students by teaching them different techniques for inferring the meaning of unknown words by recognizing and exploiting semantic cues such as synonyms and cause-effect relationships (Heilman et al., 2010). However, not all contexts are equally supportive of learning a word's meaning. As Figure 1 shows, there can be wide variation in the amount and type of information about a 'target' word to be learned, via semantic constraints implied by the context. Humans are very good at 'few-shot learning' of new vocabulary from such examples, but the instructional *quality* of initial encounters with a new word is critical. Identifying the degree and nature of supportive contexts in authentic learning materials is an important problem to solve for designing effective curricula for contextual word learning (Webb, 2008).

Predicting and characterizing educationally supportive contexts for learning differs from other context-based prediction tasks, such as n-gram prediction or cloze completion. For example, some contexts are better than others for learning because they provide more effective support for inferring the meaning of the target word. Generic natural language processing models may not capture this target-specific educational supportiveness, which is critical in determining instructive quality in contextual word learning applications.

1) My friends, family, and I all really like *tesgüino*.
2) There is a bottle of *tesgüino* on the table.
3) Brewers will ferment corn kernels to make *tesgüino.*

Figure 1: These sentences have the same length but provide very different contextual information about the meaning of the target word, *tesgüino.* We explore computational models that can quantify the degree and nature of this target-specific *educationally supportive contexts*.

In this study, we introduce examples of predicting the degree and understanding the nature of the *educationally supportive contexts* with respect to the meaning of a target word to be learned. We show that the application of deep learning using a model based on BERT (Devlin et al., 2019), combined with an attention layer, gives the best-known accuracy to date on an existing key benchmark multi-sentence context dataset (Kapelner et al., 2018). We also compared our custom models for prediction with the generic use of a sophisticated recent LLM and showed identifying educationally supportive contexts was a challenging task for this LLM. We believe our results are applicable not only to developing educational curricula for vocabulary instruction, but also to NLP tasks like few-shot machine learning of new words or concepts from text.

2. Educationally Supportive Contexts

Both humans and language models use context words to infer word meaning. In education, contextual word learning is an instructional method that teaches students how to infer the meaning of unknown words by recognizing and utilizing semantic cues (Frishkoff et al., 2008; Heilman et al., 2010). Beck et al. (1983) note that not all contexts are equal by characterizing the supportiveness of contexts for learning new words, distinguishing between pedagogical vs. natural. Both high- *and* low-supportive contexts play important roles in optimizing long-term retention of new vocabulary, as they invoke different but complementary learning mechanisms (van den Broek et al., 2018). Exposing a reader to a carefully chosen contextual curriculum can lead to significantly better long-term retention of new words (Frishkoff et al., 2016).

Multiple studies in NLP have focused on related tasks such as predicting a missing word (Zweig and Burges, 2011; Shaoul et al., 2014), predicting the meaning, substitutes, or properties of a particular word (McCarthy and Navigli, 2007; Kremer et al., 2014; Wang et al., 2017; Pavlick and Pasca, 2017; Pilehvar and Camacho-Collados, 2019), or measuring surprise at seeing a particular word in the context (Peyrard, 2018). These complementary tasks focus on the word itself, rather than the context, to learn word meaning and models of language (Warren, 2012). Indeed, the missing-word task forms the basis for most pre-training of large language models (e.g., masked word prediction Peters et al., 2018; Devlin et al., 2019). A few previous studies investigated how contextual information can be used to infer the meaning of synthetically generated target words (Lazaridou et al., 2017; Herbelot and Baroni, 2017). However, they also relied on the assumption that the provided context contains enough information to make an inference, by manually selecting the training sentences for synthetic words. In general, work in NLP typically assumes informative contexts are given, and does not predict or characterize the varying degrees of contextual supportiveness with respect to a target word, especially for educational scenarios.

A handful of NLP studies have computationally approached characterizing the supportive contexts of curriculum materials for vocabulary learning. The most related work to our study, Kapelner et al. (2018), achieved their best prediction performance using a random forest model with over 600 different hand-specified text features on their own dataset of multi-sentence text passages. Unlike our study, however, they did not explore the use of deep learning frameworks for learning feature representations automatically for enhanced prediction. The REAP project (Collins-Thompson and Callan, 2004) used NLP methods to identify appropriate contexts for vocabulary learning, but focused on filtering entire web pages by tagging sentences with specific criteria, not individual prediction of supportive contexts. Similarly, Hassan and Mihalcea (2011) used feature engineering and a supervised classifier to label entire documents as “learning objects” for concepts (e.g., computer science), but did not focus on quantifying or characterizing the supportiveness of context passages for specific target words. Our work bridges distinct fields, exploring the first use of deep learning for the educational supportiveness prediction task, not only avoiding the need for extensive feature engineering, but also providing a mechanism for interpretability, to characterize how different cues in a particular context contribute to information about a given target word.

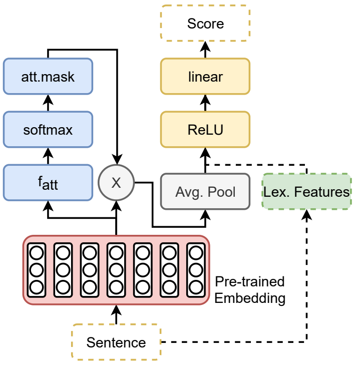

Figure 2: Our model consists of a pre-trained embedding (red) with a masked attention block (blue) to create attention-weighted context vectors, and a regression block (yellow) to predict the numeric educational supportive context score. We also tested lexical features from Kapelner et al. (2018) (green) as complementary additional model input.

3. Attention-Based Models

Our model is inspired by a prior attention-based model that classified customer sentiment towards particular product aspects by capturing the relationship between context words and a target word (Liu et al., 2018). In Liu et al. (2018), the attention mechanism was used to capture the different sentiment polarities of context words with respect to the target concept/aspect. They showed this approach can be effective for sentences with multiple aspects or complex structures. In our work, we used the attention mechanism to capture the different amounts of contextual support of words with respect to the target word to learn.

Our model predicts contextual supportiveness by identifying which aspects contribute to identifying a word’s meaning. We used pre-trained components (red block in Figure 2) to retrieve vector representations of contexts and the target word. For each model, we treated the target word as an unknown (e.g., `<UNK>` for ELMo or `[MASK]` for BERT models) token so that the model must use contextual information to infer the meaning of the 'unknown' target word.

The target and context word vectors are passed to the attention layers (blue blocks). Using an attention mechanism (Luong et al.), we calculated the weights between the masked target word and context words (f_{att}) and get the attention-weighted average of context vectors. We also explored complementing our model's representation with lexical features added from Kapelner et al. (2018) (green block) (see Appendix C for details) [1].

It is important to understand why our application of attention mechanisms for this educational prediction task differs from generic attention-based models (e.g., transformer-based language models). When children learn to read, literacy strategies can teach them to look for clues in the surrounding context to infer the meaning of an unknown word. This includes exercises to find synonyms, antonyms, cause-effects, and other relationships. Since not all relationships of nearby context words are equally helpful or easy for a child to recognize, our model, together with the extensive labeled dataset we use for training, focuses on capturing the nuances of the types of word-context relationships that are specifically helpful for language learners. In contrast, the attention weights from generic self-attention layers may capture a variety of relationships between the words in an input text sequence, but these are not necessarily *educationally supportive* relationships that are most accessible and helpful for learning.

4. Dataset

In contextual word learning, the meaning of a target word can be determined from information in surrounding sentences. To test our models in the multi-sentence scenario, we used an existing dataset from the only previous study, to our knowledge, on educational supportiveness of contexts (Kapelner et al., 2018). Those authors selected 933 words for advanced exams such as the ACT, SAT, and GRE. Based on these target words, they collected 67,833 contexts from the DictionarySquared database. On average, each target word had 72.7 (σ^2=20.7) contexts. They categorized target words into ten difficulty levels, and these levels were not correlated with annotated supportiveness scores.

This multi-sentence dataset contains over 67k passages selected from the existing database ($\mu = 81$ words, $\sigma^2 = 42$). Second, each context contains one of 933 unique target words, which were selected to range across difficulty levels. Third, crowdworkers for the dataset annotated the educational supportiveness of context passages (with target word included) using a four-point Likert scale (roughly corresponding to the four categories in (Beck et al., 1983)). Table 1 provides example multi-passage items derived from that dataset.

High Supportive: As with ginger, turmeric has *salubrious* properties. It is an antiseptic, applied as a paste to cuts and abrasions, and is taken with food to aid digestion. *Target Word:* salubrious *Avg. Score:* 3.22
Low Supportive: With the increase in the number of *clandestine* laboratory seizures throughout the country, there has been a corresponding escalation of problems confronting state and local agencies that are called to the scene of these laboratories. *Target Word:* clandestine *Avg. Score:* 1.89

Table 1: The multi-sentence context dataset from Kapelner et al. (2018) consists of passages collected from the DictionarySquared database, with crowd-sourced supportiveness ratings.

5. Experiment: Finding Educationally Supportive Contexts

We evaluated how effectively a range of models predicted the educational supportiveness of contexts. We computed both RMSE of human vs. machine prediction (min-max scaled), and the ROCAUC for three binary prediction problems: predicting the passages in the lowest 20%, median 50%, and highest 20% of annotated supportiveness scores. The latter high-precision setting corresponds to the goal of the prior study (Kapelner et al., 2018). All reported results are based on 10-fold cross-validation. Each fold was randomly selected based on the target word to ensure the model did not see sentences with the same target word during training.

5.1. Baseline Models

We used multiple baseline models for the analysis. For simple baselines, we used a dummy model (`Base:Avg`) that always predicts the average supportiveness score from a fold; a linear regression model (`Base:Length`) based on sentence length; and a ridge regression model (`Base:BoW`) using the co-occurrence information of context words – The last method could be similar to mutual information score, as it is based on word-count features.

We included the random forest model (`Base:RF_Lex`) from Kapelner et al. (2018). Thanks to data provided by those authors, our implementation of the random forest model was

[1] Codes can be found at https://github.com/sungjinnam/contextual_informativeness

	RMSE(↓)	Lower 20%(↑)	50:50(↑)	Upper 20%(↑)
Base: Avg.	0.173 (0.170, 0.176)	0.500 (0.500, 0.500)	0.500 (0.500, 0.500)	0.500 (0.500, 0.500)
Base: Length	0.173 (0.170, 0.176)	0.511 (0.505, 0.517)	0.507 (0.500, 0.514)	0.502 (0.495, 0.509)
Base: BoW	0.201 (0.199, 0.204)	0.643 (0.630, 0.656)	0.599 (0.588, 0.610)	0.585 (0.575, 0.595)
Base: RF_Lex	0.157 (0.154, 0.159)	0.736 (0.729, 0.743)	0.698 (0.691, 0.705)	0.680 (0.669, 0.692)
Base: GPT-3.5	0.290 (0.288, 0.291)	0.541 (0.536, 0.546)	0.540, (0.537, 0.544)	0.546, (0.542, 0.551)
Base: ELMo	0.152 (0.146, 0.159)	0.768 (0.757, 0.779)	0.729 (0.721, 0.737)	0.705 (0.696, 0.715)
Ours: ELMo+Att	0.153 (0.149, 0.156)	0.770 (0.760, 0.780)	0.727 (0.720, 0.734)	0.701 (0.689, 0.713)
Ours: ELMo+Att+Lex	0.152 (0.149, 0.155)	0.789 (0.779, 0.799)	0.746 (0.739, 0.754)	0.725 (0.719, 0.731)
Base: BERT	0.139 (0.136, 0.142)	0.807 (0.797, 0.817)	0.764 (0.757, 0.772)	0.751 (0.739, 0.763)
Ours: BERT+Att	**0.138 (0.136, 0.140)**	0.816 (0.806, 0.825)	0.777 (0.770, 0.785)	0.768 (0.757, 0.778)
Ours: BERT+Att+Lex	0.145 (0.142, 0.149)	**0.822 (0.814, 0.831)**	**0.782 (0.775, 0.788)**	**0.773 (0.765, 0.781)**

Table 2: Average RMSE (lower is better(↓)) and binary classification results (ROCAUC, higher is better(↑)) with the multi-sentence context dataset (Kapelner et al., 2018). Adding the attention block (`+Att`) to the BERT-based model performed significantly better than the baseline and ELMo-based models. Adding lexical features from the original paper (Kapelner et al., 2018) (`+Lex`) further increased prediction performance. Numbers in parentheses are the 95% confidence interval. Numbers in bold indicate the best-performing model for each evaluation criterion.

able to replicate the reported R^2 scores (e.g., 0.179 vs. 0.177), using lexical features of contexts. The model used lexical features, including 600+ hand-specified features, such as n-gram frequencies from Google API, Coh-Metrix (McNamara et al., 2014), and sentiment analysis results (Crossley et al., 2017), psycholinguistic (Crossley et al., 2016). More details about the model and lexical features can be found in Appendix A.

We also used pre-trained language models. We used ELMo (`Base:ELMo`) (Peters et al., 2018) and BERT (`Base:BERT`) (Devlin et al., 2019) to predict the educational supportiveness scores without the additional attention block (i.e., without the blue blocks from Figure 2). These baselines are expected to perform better than the simple baselines.

We included a commercial LLM GPT-3.5 (OpenAI, 2023) baseline model (`GPT-3.5`) for comparison. Our input prompt carefully used the same description language for each rating level given to human raters in (Kapelner et al., 2018). It also included few-shot demonstration examples showing sample passages for each possible supportiveness rating, as well as a secondary task of evaluating whether expert knowledge might be required to fully understand the passage, to properly calibrate for typical non-expert users of the system (see Appendix D for details).

5.2. Results

Overall prediction results are shown in Table 2. The sentence-length baseline (`Base:Length`) showed near-random classification performance, since the contexts were long enough and less correlated to the number of words. The co-occurrence baseline model (`Base:BoW`) showed significantly better performance than `Base:Length`, but not better than the pre-trained baselines. The previous best-known prediction model on this dataset (`RF_Lex`) did significantly better than the simple baselines, but not as well as the deep learning-based models.

Despite significant work on prompt engineering and the use of few-shot examples, we obtained weaker results than expected using the commercial LLM (`GPT-3.5`). It is possible that further prompt refinements or the use of more powerful models might result in much better accuracy. However, the results provide evidence that obtaining good educational supportiveness prediction results with this dataset is a non-trivial challenge for even a sophisticated general-purpose LLM.

Based on 95% confidence intervals, all ELMo- and BERT-based models performed significantly better than the other baseline models. For our custom model, we found that adding the attention block alone (`BERT+Att`) provided marginal gains over the BERT baseline. The complementary model (`BERT+Att+Lex`) that concatenated attention-weighted context vectors with lexical features, performed significantly better than all baseline models, giving the best overall performance on all three binary prediction tasks.

6. Conclusion

Our custom models tuned for predicting the educational supportiveness of contexts performed significantly better than the off-the-shelf commercial LLM, and our deep learning frameworks significantly improved over the prior best prediction model from (Kapelner et al., 2018) that used large numbers of engineered features. Beyond education scenarios, we believe our results also motivate further research on educational supportiveness prediction as an important additional benchmark task for general LLM evaluation suites.

7. Discussion and Limitations

First, a more complete model for predicting educationally supportive contexts would include personalized and user group-oriented components such as users' background knowledge of the target concept or L1 vs. L2 learner profile. However, such models can be challenging to learn and evaluate. For example, individual differences in word knowledge (Borovsky et al., 2010) or language proficiency (Elgort et al., 2018) may result in different levels of processing linguistic information.

Second, our attention-based model still has room to improve. Like many ML models, our model tends to make centralized predictions, over-predicting the low-scored contexts and under-predicting the high-scored contexts. This might be related to the encoder models' limited vocabulary size compared to distinctive vocabulary, semantic cue usages in low— or high-scored contexts, or the location of the target word that would affect the amount of nearby context. In future work, we will systematically investigate characteristics of the less accurate predictions, including the role of pre-trained NLP models' limited vocabulary size, difficulty processing grammatically complex or incorrect sentences, or variability between different semantic clue types.

Third, we focused on developing a more capable model for predicting the amount of educationally supportive contexts in vocabulary learning. In a future study, qualitative or quantitative evaluations of the model, including 1) analyzing attention weights of contextual words with anecdotal examples and 2) comparing different contextual word types' attention weights in scale with synthetic sentences with similar sentence structures, would provide an in-depth analysis of model behaviors.

Lastly, we believe our approach would have broader NLP implications. For example, the attention weights for the context words can be useful by providing more detailed explanations of specific semantic clues for language learners. Further evaluations on the quality of the attention weights will be beneficial to designing interactive applications for literacy. Our model could also be used in applications like curriculum learning (e.g., use our model to decide more/less informative training materials to design the training curriculum for NLP models to learn about new concepts), reinforcement learning (e.g., use our model as a reward model for tuning the larger generative model to provide more context-rich output), and information retrieval (e.g., use our model to rank more informative context on top for the query).

Ethics Statement

This study only investigated datasets and pre-trained language models developed for the English language. Since our model is based on pre-trained models like ELMo, BERT, and GPT-3.5, our results may also inherit different biases, including gender, race, culture, or word frequency, from these base models. Further studies using non-English datasets, accounting for more inflected languages' complex grammatical rules, or semantic biases associated with different cultures, would be a valuable future extension of this research.

Acknowledgements

We thank Gwen Frishkoff, Rada Mihalcea, and Ryan Baker for their feedback on an early version of this work. This research was sponsored in part by a grant from the Michigan Institute for Data Science (MIDAS), with additional support from the University of Michigan School of Information.

References

Isabel L Beck, Margaret G McKeown, and Ellen S McCaslin. 1983. Vocabulary development: All contexts are not created equal. *The Elementary School Journal*, 83(3):177–181.

Arielle Borovsky, Marta Kutas, and Jeff Elman. 2010. Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition*, 116(2):289–296.

Kevyn Collins-Thompson and Jamie Callan. 2004. Information retrieval for language tutoring: an overview of the REAP project. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 544–545. ACM.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4):1227–1237.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2017. Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, 49(3):803–821.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Irina Elgort, Marc Brysbaert, Michaël Stevens, and Eva Van Assche. 2018. Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, 40(2):341–366.

Gwen Frishkoff, Kevyn Collins-Thompson, and Sungjin Nam. 2016. Dynamic support of contextual vocabulary acquisition for reading: An intelligent tutoring system for contextual word learning. In Scott A Crossley and Danielle S McNamara, editors, *Adaptive Educational Technologies for Literacy Instruction.*, chapter 5, pages 69–81. Taylor & Francis, Routledge, New York, NY, USA.

Gwen Frishkoff, Kevyn Collins-Thompson, Charles Perfetti, and Jamie Callan. 2008. Measuring incremental changes in word knowledge: Experimental validation and implications for learning and assessment. *Behavior Research Methods*, 40(4):907–925.

S. Hassan and R. Mihalcea. 2011. Learning to identify educational materials. *ACM Trans. Speech Language Process.*, 8(2).

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, Maxine Eskenazi, Alan Juffs, and Lois Wilson. 2010. Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 20(1):73–98.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: Acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309.

Adam Kapelner, Jeanine Soterwood, Shalev Nessaiver, and Suzanne Adlof. 2018. Predicting contextual informativeness for vocabulary learning. *IEEE Transactions on Learning Technologies*, 11(1):13–26.

Peter Kolb. 2008. Disco: A multilingual database of distributionally similar words. In *Proceedings of KONVENS*.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us - analysis of an “all-words” lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549.

Angeliki Lazaridou, Marco Marelli, and Marco Baroni. 2017. Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, 41:677–705.

Qiao Liu, Haibin Zhang, Yifu Zeng, Ziqi Huang, and Zufeng Wu. 2018. Content attention model for aspect based sentiment analysis. In *Proceedings of the 2018 World Wide Web Conference*, pages 1023–1032. International World Wide Web Conferences Steering Committee.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

OpenAI. 2023. Gpt 3.5 (version 0613).

Ellie Pavlick and Marius Pasca. 2017. Identifying 1950s American jazz musicians: Fine-grained IsA extraction via modifier composition. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2099–2109.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Maxime Peyrard. 2018. A formal definition of importance for summarization. *arXiv preprint arXiv:1801.08991*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273.

Cyrus Shaoul, Harald Baayen, and Chris Westbury. 2014. N-gram probability effects in a cloze task. *The Mental Lexicon*, 9(3):437–472.

Gesa SE van den Broek, Atsuko Takashima, Eliane Segers, and Ludo Verhoeven. 2018. Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning*, 68(2):546–585.

Su Wang, Stephen Roller, and Katrin Erk. 2017. Distributional modeling on a diet: One-shot word learning from text only. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 204–213.

Paul Warren. 2012. *Introducing Psycholinguistics*. Cambridge University Press.

Stuart Webb. 2008. The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, 20(2):232–245.

Geoffrey Zweig and Christopher JC Burges. 2011. The Microsoft Research Sentence Completion Challenge. *Microsoft Research Technical Report MSR-TR-2011–129*.

A. Hyperparameters and Lexical Features

The ridge regression baseline model (`Base:BoW`) was trained with `scikit-learn`'s default alpha value for the baseline model using bag-of-words co-occurrence information. The co-occurrence matrix was built for words that appeared more than five times in the training data. The score thresholds used for the low, median, and high-supportiveness binary prediction tasks were 2.13, 2.6, and 3.0, respectively.

We successfully replicated the random forest model (`Base:RF_Lex`) used in (Kapelner et al., 2018) by following the original paper's settings. We used features and settings provided by the authors, setting the number of estimators as 500 and bootstrapping sample size as 10000. Additional lexical sophistication features included the top ten words with synonymous words from the target. These top ten context words frequently collocate with the target words, the frequency of the target word, context words' politeness, age of acquisition, and meaningfulness of context words (Crossley et al., 2016; Kolb, 2008).

During the training of the ELMo- and BERT-based models, we fine-tuned the pre-trained models. Because of the differences in the number of trainable parameters of each pre-trained model, we used different learning rates for each ELMo-based ($1e{-}3$) and BERT-based ($3e{-}5$) model. Other hyper-parameters remained constant across models (batch size: 16, iteration: 3). The dimension of the ReLu layer was 256. The dimensions for the attention block layers were the same as those of the pre-trained embeddings (ELMo: 1024, BERT: 768).

B. Computing Resource for Training

For this study, we used a single NVIDIA 2080 TI GPU with Intel i7 CPU. For the multi-sentence context dataset, it took approximately 30 minutes per fold. We used pre-trained versions of ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) from TensorFlow Hub `https://tfhub.dev/`. Our ELMo-based model with attention block had about 426k trainable parameters, while the BERT-based counterpart had about 7.3M trainable parameters.

C. Model Structure

The input for the attention layers (blue blocks in Fig. 2) is the target word and context tokens vector. Using a multiplicative attention mechanism (Luong et al.), we calculated the relationship between the token that replaced the target word and context words (f_{att}). We used $softmax$ to normalize the output of the attention layer. The output of the softmax layer masked non-context tokens as zero, to eliminate the weights for padding and the target word ($att.mask$). The masked attention output was then multiplied with the contextual vectors from the pre-trained model to generate attention-weighted context vectors. We also explored complementing our model's representation with lexical features from Kapelner et al. (2018) (green block) by concatenating their features with our attention-weighted context vectors (`BERT+Att+Lex` in Table 2). The regression layers (yellow blocks) used an average pooling result of attention-weighted context vectors. The layers comprised a ReLU layer and a fully connected linear layer that estimated the score of educational supportiveness on a continuous scale. We used root mean square error (RMSE) as a loss function. We compared the use of pre-trained versions of ELMo (Peters et al., 2018) and BERT-Base (12 layers, 768 dimensions) (Devlin et al., 2019) models from TensorFlow Hub in developing our model.

To avoid overfitting, we selectively updated the pre-trained models' parameters: for ELMo-based models, we updated parameters that determine the aggregating weights of LSTM and word embedding layers; for BERT-based models, we updated the parameters for the last layer.

D. GPT Parameters and Prompt

We used the OpenAI `gpt-3.5-turbo-0613` model with the following parameters:

```
temperature=0, max_tokens=2500,
```

top_p=1, frequency_penalty=0, presence_penalty=0.

The prompt contains few-shot learning examples for the main educational supportiveness prediction task as well as instructions for two ancillary tasks: (1) a binary ’accessibility’ prediction task of labeling passages that likely require subject-specific expertise or not, to be accessible to student readers, and (2) an explanation of what context words contributed clues to the target word meaning, with relative weights. The prompt also included batch format instructions so that multiple passages (for this study, ten) could be evaluated with a single API call.

You are an assistant who is a linguistic expert, but is not a subject matter expert for any other subject. I will give you 10 target word-passage pairs. Each target word-passage pair is in the format below:

target word ||| passage

Below is an example:

abrogate ||| But the fight against parental notification is really only one example of many attempts to water down traditional values and even abrogate the original terms of American democracy. Freedom prospers when religion is vibrant and the rule of law under God is acknowledged. When our Founding Fathers passed the First Amendment, they sought to protect churches from government interference. ...

For each target word-passage pair, quantify how useful the passage is for learning the complete and correct definition of the target word contained in the passage.

Output your prediction as a real number on the following scale:

+4 Very Helpful. After reading the passage, a student will have a complete and correct understanding of what the target word means.

+3 Somewhat Helpful. After reading the passage the student will have some idea of the meaning of the target word.

+2 Neutral. After reading the passage it neither helps nor hinders a student’s understanding of the word’s meaning.

+1 Bad. This passage is misleading about the target word meaning, too difficult, or otherwise inappropriate.

After the prediction score, output "Expertise required" if the language in the passage contains vocabulary or acronyms that require expert knowledge, or "Accessible" if no specialized domain language is required to understand the passage.

After outputting the prediction, output the list of at most 5 context words, not including the target word, that give a student the most significant clues about the meaning of the target word. For each context word, assign a score from 0 to 100 that indicates how strongly it gives evidence about the meaning of the target word.

Do not include the reasonings behind your choice of context words, just output the context words.

Below are four examples of an input with its corresponding output:

User - "bifurcate ||| ... The AAA International Arbitration Rules are even more explicit, stating that the tribunal "may in its discretion direct the order of proof, bifurcate proceedings, exclude cumulative or irrelevant testimony or other evidence, and direct to the parties to focus their presentations on issues the decision of which could dispose of all or part of the case." ..."

Assistant - "target_word: bifurcate
Score: +1
Accessibility: Accessible
Context words: AAA International Arbitration Rules (20), tribunal (10), order of proof (10), proceedings (5), case (5)"

User - "bifurcate ||| The parties will invariably want to bifurcate the issues of costs and attorneys fees (if applicable) and the amount of any punitive damages from the primary arbitration. They may also wish to consider bifurcating liability and damages, though such a request, if granted, will probably create a substantial hiatus between arbitration phases. ..."

Assistant - "target_word: bifurcate
Score: +2
Accessibility: Accessible
Context words: parties (10), issues (10), costs (10), attorneys fees (10)"

User - "bifurcate ||| ... In birds, the auditory nerve contacts nucleus magnocellularis (NM). Axons emanating from neurons in NM bifurcate and innervate nucleus laminaris (NL) on both sides of the brain. NL neurons have symmetric dorsal and ventral dendrites. The dorsal dendrites receive input from ipsilateral NM while the ventral dendrites receive input from contralateral NM. ..."

Assistant - "target_word: bifurcate
Score: +3
Accessibility: Expertise required
Context words: auditory nerve(20), nucleus magnocellularis(30), neurons(15), nucleus laminaris(25)"

User - "bifurcate ||| First, the pastor needs to integrate their sermon preparation and personal Bible study. In the past, it has been argued by some that a pastor should bifurcate their personal Bible study from their sermon preparation. I wholeheartedly disagree with this counsel because a preacher needs to preach from their life in Scripture. ..."

Assistant - "target_word: bifurcate
Score: +4
Accessibility: Accessible
Context words: pastor(10), sermon preparation(15), personal Bible study(15), preacher(10)"

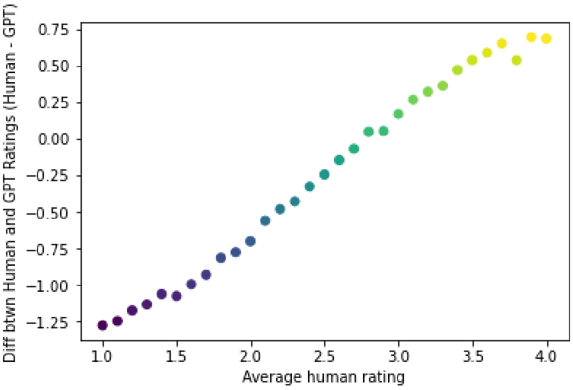

Figure 3: Residual of LLM predicted scores as a function of average human rating for multi-sentence contexts.

Your response should be a list of 10 items corresponding to the list of 10 word passage pairs given in the input.

The format of each item should follow the example response below:

target_word: bifurcate
Score: +3
Accessibility: Expertise required
Context words: auditory nerve (20), nucleus magnocellularis(30), neurons(15), nucleus laminaris(25)

When assuming expertise in subject matter for the agent that rates contextual informativeness, the prompt changed the role of the system to the following: Assistant is a linguistic expert that analyzes the informativeness of a passage in helping readers learn the complete and correct definition of the target word in the passage

E. Additional GPT Results

Figure 3 shows the residual plot of RMSE as a function of human rating. The GPT model tended to have overly conservative predictions, overpredicting supportiveness for passages rated low-quality by humans, and underpredicting supportiveness for high-quality passages.

We also examined how prediction errors were connected with the need for subject area expertise, as marked by the 'accessibility: expertise required' label assigned by the LLM. Overall, 1.176% of passages were marked as 'Expertise required' with highly technical language. For passages where the LLM underpredicted supportiveness, this fraction was 1.103%, while for passages where the LLM overestimated supportiveness, the fraction was 1.970%. This discrepancy suggests the need for further refinement of the prompt design to improve the LLM's calibration of what contexts require subject expertise, and how expertise should be factored into a realistic prediction for the target user population (typically non-expert students).