Using structured knowledge and traditional word embeddings to generate concept representations in the educational domain

Oghenemaro Anuyah, Ion Madrazo Azpiazu, Maria Soledad Pera
People and Information Research Team (PIReT)
Department of Computer Science
Boise State University
Boise, Idaho, USA
{oghenemaroanuyah,ionmadrazo}@u.boisestate.edu,solepera@boisestate.edu

ABSTRACT

To capitalize on the benefits associated with word embeddings, researchers working with data from domains such as medicine, sentiment analysis, or finance, have dedicated efforts to either taking advantage of popular, general-purpose embedding-learning strategies, such as Word2Vec, or developing new ones that explicitly consider domain knowledge in order to generate new domain-specific embeddings. In this manuscript, we instead propose a *mixed strategy* to generate enriched embeddings specifically designed for the *educational* domain. We do so by leveraging FastText embeddings pre-trained using Wikipedia, in addition to established educational standards that serve as structured knowledge sources to identify terms, topics, and subjects for each school grade. The results of an initial empirical analysis reveal that the proposed embedding-learning strategy, which infuses limited structured knowledge currently available for education into pre-trained embeddings, can better capture relationships and proximity among education-related terminology. Further, these results demonstrate the advantages of using domain-specific embeddings over general-purpose counterparts for capturing information that pertains to the educational area, along with potential applicability implications when it comes to text processing and analysis for K–12 curriculum-related tasks.

CCS CONCEPTS

• **Information systems** → **Document representation**; Information retrieval; Data encoding and canonicalization; • **Computing methodologies** → *Knowledge representation and reasoning*; Lexical semantics.

KEYWORDS

word embeddings, domain knowledge, text representation, enriched embeddings, K-12 curriculum, structured knowledge

ACM Reference Format:
Oghenemaro Anuyah, Ion Madrazo Azpiazu, Maria Soledad Pera. 2019. Using structured knowledge and traditional word embeddings to generate concept representations in the educational domain. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316583

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316583

1 INTRODUCTION

Word-embeddings have become a de-facto standard for vocabulary representation. This is due to their ability to capture word contexts and varied word relations in a numerical fashion, allowing for effective and efficient word operations via their vectors. Moreover, word embeddings has been used with great success in areas related to text-processing, given their potential to achieve good generalization [11, 29].

Among the most common strategies to generate numerical word representations we find those proposed by Mikolov et al. [21], Bojanowski et al. [2], and Pennington et al. [23]. Pre-trained embeddings created as a result of using the aforementioned strategies on well-known general-purpose corpora, including Wikipedia and GoogleNews corpus, have been widely adopted by the research community as the starting point for a broad variety of applications, from analyzing reviews or scholarly articles to enhance the recommendation process [5, 14] to examining student responses in the K–12[1] context [4] and leveraging pre-trained embeddings in order to learn bilingual word representations [1].

More often than not, embeddings pre-trained using domain-agnostic resources effectively capture word co-occurrences, as well as other semantic relationships among words. Unfortunately, relationship constraints that are inherent of domain-specific contexts may be overlooked. Khatua et al. [16] echo this observation and highlight that for domains such as biomedicine, a general-purpose corpora might not yield word embeddings that best represent the domain, no matter how large that corpora is. Instead, Khatua et al. [16], along with other authors [6, 22], argue in favor of relying on existing domain-dependent corpora, such as PubMed (for medicine-related tasks) or Geological Society and Norwegian Petroleum Directory (for oil and gas domain), in order to generate more representative embeddings, while still relying on popular modeling techniques, such as Word2Vec [21] or FastText [2], to create them. Domain-dependent corpora, however, might not be as abundant or rich in resources as general-purpose counterparts.

In order to take advantage of both large general-purpose corpora and domain-specific resources, other researchers instead bias known deep learning modeling strategies towards specific contexts by altering loss functions. Such is the case of the work introduced in [13, 24], which explicitly consider the polarity of terms in creating

[1] K–12 is a common American expression that refers to primary and secondary schooling, from Kindergarten to the 12^{th} grade.

new word embeddings—doing so is a must if these embeddings are to be used for applications where the sentiment connotation can greatly influence their outcomes, such as review analysis, recommendation systems, and natural language understanding.

Beyond health, medicine, and sentiment analysis, one domain that could benefit from the availability of domain-dependent embeddings is (primary and secondary) *education*. Aligning (web) resources to K–12 curriculum delimitations to address fundamental tasks, such as finding similar exercises [17, 25] or aiding teachers with instruction design [10, 28], requires deep understanding of domain-specific terminology and the relationship among these terms. Knowledge-bases or ontologies could serve as guidelines for identifying connections across terms particular to a domain. Yet, to the best of our understanding, such information sources are not available for our area of interest; and even if they did, jargon specific to that area would not necessarily be properly captured in the vector space.

In this manuscript, we introduce Edu2Vec, a novel *mixed strategy* to generate embeddings that are meaningful to the education domain and that can serve as the foundation for solutions to the aforementioned problems in the classroom setting. Edu2Vec leverages (i) the skip-gram model presented by Mikolov et al. [21], a well-known, state-of-the-art word modeling architecture, (ii) embeddings pre-trained using Wikipedia, (iii) Wikibooks, a rich and publicly-available corpus, and (iv) structured knowledge from educational standards.[2] Edu2Vec preserves the quality of original word embeddings while requiring minimal computation for generating representations for educational concepts, by avoiding to retrain word embeddings. Furthermore, Edu2Vec is able to generate a joint space that shares both word embeddings and representations of educational concepts, which has the potential to enable multiple educational domain applications.

We conducted initial experiments in order to demonstrate the advantages of our mixed strategy. Outcomes from these experiments, along with a preliminary qualitative assessment yielded promising results in terms of the validity of infusing minimal structured area knowledge into an existing word embeddings space for learning specific representations pertaining to the K–12 curriculum.

The remaining of this paper is organized as follows. In order to contextualize our work, we discuss background information and related literature in Section 2. Thereafter, in Section 3, we present our mixed strategy for aligning structured domain knowledge with existing word embeddings in order to better capture the connections among terms in an educational context. We discuss the results of the experiments conducted to verify the validity of our proposed strategy in Section 4. Lastly, we offer some concluding remarks and directions for future work in Section 5.

2 BACKGROUND & RELATED WORK

In this section, we briefly introduce the concept of word embeddings. Thereafter, we discuss strategies that can be used to influence embedding distributions in the space as well as literature pertaining to the generation of domain-specific word embeddings.

[2]We shared both the educational hierarchy and the embedding representations for each concept at https://bit.ly/2EGHV5O

2.1 Word Embeddings

A *word embedding* is a numerical representation of a word in a multidimensional space [21]. Each dimension represents a latent feature which partially describes the meaning of a word, facilitating the comparison, in terms of meaning, of any word pair in a numerical way. Word embeddings have been used as part of solutions to problems in multiple text-related areas of study, including Recommendation Systems [30], Information Retrieval [12, 20] and Natural Language Processing [9, 33]. Key aspects that contribute to the recent success of word embeddings are (1) their ease to both be trained and used for different languages, as they just need an unlabeled corpora for doing so, and (2) their generalization capabilities, as models that are trained on small amounts of data are able to recognize words that were never used as part of the training process. Among the most popular embedding generation strategies we find: (i) CBOW and skip-gram models presented by Mikolov et al. [21] (Word2Vec), which are based on a neural model for learning representations, (ii) the model by Bojanowski et al. [2] (FastText), which leverages sub-word information in order to improve upon Mikolov's model, and (iii) the model presented by Pennington et al. [23] (GloVe), which relies on global word-word co-occurrence statistics from a corpus in order to compute word representations.

2.2 Domain-specific Embeddings

Most of the current strategies for learning word representations rely on the distributional aspect of words among a given corpora, creating word representations that are more similar to each other the more times two words co-occur in a context window. This poses a problem for specific domains, given that co-occurrence might not always imply similarity. Sentiment analysis is an example of such domain. As mentioned by Tang et al. [27], the words *good* and *bad* tend to be located close to each other in the embedding space given that they often appear in similar contexts. Unfortunately, this spatial distribution is counterproductive for sentiment analysis, given the opposite polarity of the these two terms. To address this limitation, Tang et al. [27] explicitly encode sentiment information into the continuous representations of words. Piao and Breslin [24] focus on incorporating sentiment prediction information into the general loss function, whereas Hamilton et al. [13] combine both semantic high-quality word-embeddings with an sentiment label propagation approach.

Other domains that benefit from domain-specific embeddings are health, medicine, finance, current events, and geosciences, to name a few. In their cases, the use of domain-specific corpora of diverse sizes to learn embeddings more suitably distributed in the space is the most prominent strategy [6, 16, 18]. Alternatively, researchers have also considered leveraging knowledge resources (and in some cases multiple objective functions) in order to generate enhanced embeddings that capture additional semantic relationships, such as hyponymy and relatedness in the target domain [22, 26, 32].

Research efforts dedicated to tackle the issue of generating domain-specific representations are a must if we want to advance applications that need to process texts that incorporate specialized jargon. Hierarchical representations are a natural way to represent human understanding of domain-specific concepts, which is why we believe it is possible to leverage such structures for computing domain

dependent concept representations, with minimal computational time, as we describe in the following Section.

3 METHODOLOGY

In this section, we describe the strategy we propose for incorporating structured knowledge from the educational domain into the word embedding space. The proposed strategy consists of two steps: (i) building the educational concept hierarchy and (ii) aligning hierarchy concepts with existing, pre-trained word embeddings.

3.1 Building the Educational Hierarchy

We first build a hierarchy that captures educational knowledge. This structured knowledge, which resembles the tree structure in Figure 1(a), aligns *subjects* that pertain to the K–12 curriculum to different *grade levels*. In this structure, we also capture *topics* that correspond to the different subjects. For instance, the topics *geometry* and *chemical reactions* are aligned to the *Mathematics* and *Science* subjects, respectively. Furthermore, each of the topics in our hierarchy is connected to a number of *words*, which are terms often associated with these topics. For example, *triangles* is a term that educators expect their students to be familiar with for the *geometry* topic, whereas *molecule* is part of the vocabulary commonly affiliated with the *chemical reactions* topic. (See Figure 1(b) for another example of structured knowledge in the K–12 curriculum).

In creating our knowledge structure, we leverage several well-established *Educational Standards* that serve as guidelines to identify essential information about the K–12 curriculum, i.e., subjects, topics, and relevant vocabulary across different school grade levels:

CCSS – Common Core State Standards Established in 2009, the common core is a set of college and career-ready standards for children from Kindergarten through 12^{th} grade in English and Mathematics [7]. Some English topics include *Reading and literacy* and *Persuasive writing*, while some Mathematics topics are *Geometry*, as well as *Data Analysis, Probability, and Statistics*.

NGSS – Next Generation Science Standards The NGSS are science content standards that set expectations of what children should know in science related subjects such as Chemistry, Biology, Geography, and Solar System [8]. The NGSS enables teachers the flexibility to design classroom experiences that enhance children's interests in Science. Some NGSS topics include *Structure and Properties of Matter* and *Earth's Systems*.

ICS – Idaho Content Standards ICS details what children in Idaho public schools ought to know at the end of each grade in subjects like Sciences, Social Studies, Mathematics, and English (ICS adopts CCSS guidelines for Mathematics and English) [15]. We depend on ICS to identify topics relating to children's Social Studies, Mathematics, and English subjects.

Based on these standards, we identify four subjects: Mathematics, English, Social Studies, and Science[3]. For each of these subjects, we automatically extract educational topics using a parser that scans through the written guidelines provided by CCSS, NGSS, and ICS. A summary of the number of extracted educational topics that map to each of the subjects in our hierarchy is presented in Table 1.

Table 1: Topics and Wikibooks articles that correspond to each subject in our educational hierarchy.

Subject	Number of Topics	WikiBooks Articles
Mathematics	7	1,441
English	8	7,392
Science	40	4,500
Social Studies	27	1,645
Total	**82**	14,978[4]

In order to facilitate mapping educational concepts in the hierarchy (i.e., subjects and topics) to an existing general-purpose word embeddings (as we discuss in Section 3.2), we require the existence of *representative words* that correspond with the aforementioned educational concepts. In our case, we explore topic descriptions available on CCSS, NGSS, and ICS and identify the top–20[5] most representative words for each topic, based on TF-IDF scores. We are aware that twenty keywords may not be sufficient to capture the breath of vocabulary commonly associated with each topic. For this reason, we turn to an external, publicly available resource that can provide known relationships among educational topics and words: *Wikibooks*[6].

Wikibooks is a Wikipedia community hosted for the purpose of creating a free library of diverse textbooks. Wikibooks articles are assigned categories in order to group articles that are related to similar subjects. For instance, Wikibooks articles on *Atmospheric Layers*, *Metamorphic Rocks*, and *Effects of Air Pollution* are grouped under the category *High School Earth Science*. We argue that there is a natural connection among categories used by Wikibooks and the topics defined by Educational Standards and rely on such alignment in order to identify the articles within Wikibooks that pertain to educational concepts discussed in the K–12 curriculum. We do so using Algorithm 1, which iterates through Wikibooks articles and the 82 topics in our hierarchy and selects a Wikibooks article if its category contains one of the topics as a substring[7]. Using the resulting set of 13,543 Wikibooks articles matching the K–12 curriculum, we once again identify the top–20 most representative terms that describe each article[8] and link those with each corresponding educational topic.

Based on the grade, subject, topic, and words pairs gathered using both educational standards and Wikibooks, we use a number of relationship types to create <grade, subject>; <subject, topic>; and <topic, word> pairs. Overall, we identify 86,295 relation pairs (described below) that we use to build our educational hierarchy from domain knowledge.

[3]Note that we use Science to broadly capture subjects such as Chemistry, Physics, Geography, and Biology.

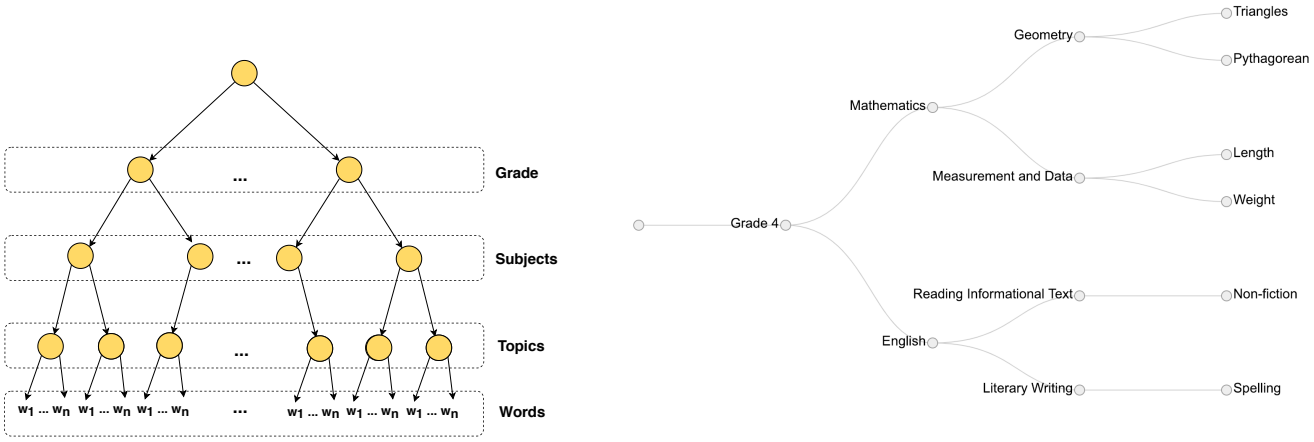
[4]This number differs from the ideal 13,543 articles being that we found a number of overlaps in Wikibooks articles for the different subjects, as some articles have multiple categories, making them align with more than one subject in some cases.

[5]We set this parameter value empirically.

[6]https://en.wikibooks.org

[7]Tokens in topics and categories may vary as a result of plurals, verb tenses, or suffixation. Hence, we first stem each of their respective tokens using the Porter stemmer algorithm [31].

[8]We identify representative terms using TD-IDF scores; with IDF calculated based on all Wikibooks articles, not just the ones matching K–12 curriculum



(a) Structured domain knowledge.

(b) Example nodes for the 4^{th} grade.

Figure 1: Education domain hierarchy, including grades, subjects, topics, and term relations.

Grade-Subject relations. We identify several relationships between grade levels and subjects by using the guidelines provided by the aforementioned educational standards. In total, our hierarchy consists of thirteen grade levels (K–12) and four subjects.

Subject-Topic relations We create relationships that capture the explicit connections among subjects and topics. This yields a hierarchy comprised of 82 unique topics for the subjects examined.

Topic-Word relations For each leaf topic in our hierarchy we match them to conceptual words, which we identify using both the educational standards and Wikibooks. Through this process, we identified a total of 15,420 unique terms for the educational topics.

Algorithm 1 Map Wikibooks articles to K–12 topics

```
Input: T, a list of K--12 Topics, WP a collection of Wikibooks articles
Output: WPT a subset of articles from WP that map to topics in T
WPT = empty set
for each t in T do
    for each w in WP do
        page_cat = extractCategory(w)
        if all(stemmed (t)) in stemmed(page_cat) then
            Add <t, w> to WPT
        end if
    end for
end for
return WPT
```

3.2 Aligning Structured Knowledge and the Word Space

In designing our mixed strategy, we aim to leverage existing general-purpose resources, i.e., pre-trained embeddings, in addition to domain-specific resources, i.e., the educational hierarchy introduced in Section 3.1. We define the latter as $H = (E, V)$, where E is the set of edges in the hierarchy and V the set of vertices. H is composed by two disjoint sub-graphs H_c and H_w that contain educational concepts (grades, subjects, topics) and words, respectively. Moreover, we define the vocabulary considered by Edu2Vec as D, which is comprised of the set of words D_w (from pre-trained word embedding vocabulary) and the set of educational concepts D_c that correspond to each vertex in H_c; $D = D_w \bigcup D_c$ and $|D_w \bigcap D_c| = 0$.

Let $X \in \mathbb{R}^{|D| \times 300}$ be a two-dimensional vector containing all the embeddings used by Edu2Vec, where the i^{th} row of X corresponds to the i^{th} term in D, $|D|$ is size of D (i.e., the number of distinct vocabulary terms) and 300 is the number of latent features used for representing each embedding. X results from stacking two vectors $X_w \in \mathbb{R}^{|D_w| \times 300}$ and $X_c \in \mathbb{R}^{|D_c| \times 300}$; the former represents the embeddings for words in D_w and the latter educational concepts in V_c. In our case, we take advantage of word embeddings pre-trained using FastText[9] in Wikipedia for initializing X_w. As for X_c, we initialize it using a random uniform distribution in range $[-1, 1]$.

In order to generate representations of educational concepts related to the K-12 curriculum that are aligned with the word representation space, we extend the model presented by Mikolov et al. [21], enabling it to consider a tree structure as input. This strategy, depicted in Figure 2, allows us to explicitly consider the domain-specific hierarchical structure H presented in Section 3.1 for learning educational concept representations X_c for H_c.

In aligning embeddings in X_c with those X_w, we maximize the following function:

[9] https://fasttext.cc/

$$\max_{X_c} \sum_{(v,v')\in V\times V} \frac{1}{d(v,v')} * P(v \mid v') \quad (1)$$

where $d(v, v')$ represents the distance between concept vertex v and v', which is used to attribute less importance to nodes that are further apart, and the probability function P is modeled as:

$$P(v \mid v') = \frac{\exp(X_v{}^\top X_{v'})}{\sum_{i=1}^{|D|} \exp(X_i{}^\top X_{v'})} \quad (2)$$

where X_v and X'_v are the embedding representations of v and v' and X_i is the embedding representation of the i^{th} token in D.

We use stochastic gradient descend [3] with a learning rate of 0.01 for minimizing the following function:

$$\min_{X_c} - \sum_{(v,v')\in V\times V} \frac{1}{d(v,v')} * P(v \mid v') \quad (3)$$

It is important to note that we only minimize in terms of X_c, excluding any update to X_w induced by the calculated gradient. This ensures that the pre-trained word embeddings are never changed, which guarantees the quality of the aforementioned embeddings remains intact, i.e., it is not affected by the minimization introduced by Edu2Vec. The outcome of the optimization procedure is a trained X_c which contains the representations of all concepts in the educational hierarchy H_c. These representations are aligned with pre-trained word representations X_w in a joint space. A joint space of both educational concepts and word embeddings is what makes possible to directly match keywords (in any text) with respect to $K-12$ curriculum terminology, which can be exploited in the development of educational text processing applications.[10]

4 EXPERIMENTAL RESULTS

We offer below insights that result from qualitative and quantitative experiments conducted to demonstrate the validity of our newly-learned educational representations.

4.1 Subject - Topic Similarity Task

To demonstrate the applicability of our domain-specific strategy, we first measure the degree of closeness for each of the 82 < *subject*, *topic* > pairs extracted from the Education Standards defined in Section 3.1. We do so by computing the cosine similarity between the keywords of each pair using the embeddings generated by Edu2Vec. To contextualize the results, we also compute the cosine similarity using general-purposed embeddings. Specifically, we use FastText word vectors for the English language trained on Wikipedia.[11] These vectors with a dimension of 300 were generated using the skip-gram model with default parameters. It is important to note that not all topics are unigrams; however, FastText's dictionary is comprised of single keywords, not n-gram phrases. For this reason, we handle topics that are n-grams phrases (for $n > 1$) by splitting the topic words into individual tokens before computing the cosine similarity between the embeddings of the < *subject*, *topic* > pairs.

We present in Figure 3 the similarity distribution computed using Edu2Vec and FastText. We see that the use of Edu2Vec leads to higher similarity scores for the majority of the pairs than when using general-purpose word embeddings. We infer from this distribution that indeed, a strategy that explicitly models structured knowledge from a given domain, i.e., education in our case, is able to better capture connections among higher-order level concepts (e.g., subjects and topics) than word-level embeddings trained with a general corpora. To present a fair picture on the different similarity distributions, we also include in Figure 3 the values computed using FastText representations only for topics that are originally unigrams. Even then, most of the yielded scores are in the [0, 0.4] range. We argue that these results serve as evidence of the benefits of integrating information from the educational domain, even if this information is limited in size, to enrich embeddings so that they can be leveraged by applications pertaining to text analysis in the corresponding domain.

4.2 Topic - Term Similarity Task

To further demonstrate the need for creating domain-specific concept representations, we also explore connections among topics and terms in the educational domain. For our analysis, we adapt the evaluation framework presented in [32], which is centered on exploring how a traditional skip-gram model and other knowledge-enriched strategies perform in identifying the topic that is the most related to a given term.

Due to the lack of existing benchmarks and datasets, we create our own using *SimpleWiki* pages[12]. We turn to SimpleWiki as it is a publicly available resource and, more importantly, it is an edition of Wikipedia that is oriented to and suitable for young audiences which uses Wikipedia's page categorization system. Much like we described in Section 3.1 for WikiBooks, we take advantage of categories assigned to each SimpleWiki page and create two datasets, which we use as ground truth. In the first dataset, DS_A, we map educational topics to high level terms associated with them. To do so, we select the subset of topics in our hierarchy that exactly match categories in SimpleWiki and extract all their respective subcategories. For example, the mathematics topic *geometry* is associated with subcategories *angles*, *conic sections*, *shapes*, and *fractals*. Using this categorization, we created 463 <*topic*, *term*, *score*> tuples, where *score* is defined as 1 being that these pairs are expected to have the optimum similarity value as subcategories for their respective SimpleWiki pages have been defined by experts. In our second dataset, DS_B, we map educational topics to more general terms commonly used within the educational domain. To create DS_B, we follow the same process outlined for DS_A but instead extracting subcategories, for each SimpleWiki category (i.e., topic) we extract the titles of pages that correspond to said category. Based on this approach, we created 2,672 <*topic*, *term*, *score*> tuples; we set *score* to 1 to capture an "ideal" similarity.

We emulate the evaluation framework in [32] and measure the degree to which general-purpose and domain-specific word representations can be effectively used to map $K-12$ vocabulary to educational topics (i.e., topics). To do so, we use error rate as our metric, which we define in Equation 4.

[10] We are aware that some topic or subject words may overlap with words in FastText. To address this limitation, we added a prefix "T_" to terms that are also topics and "S_" to terms that are subject words.

[11] https://dl.fbaipublicfiles.com/fasttext/vectors-wiki/wiki.en.zip

[12] https://simple.wikipedia.org/wiki/Category:

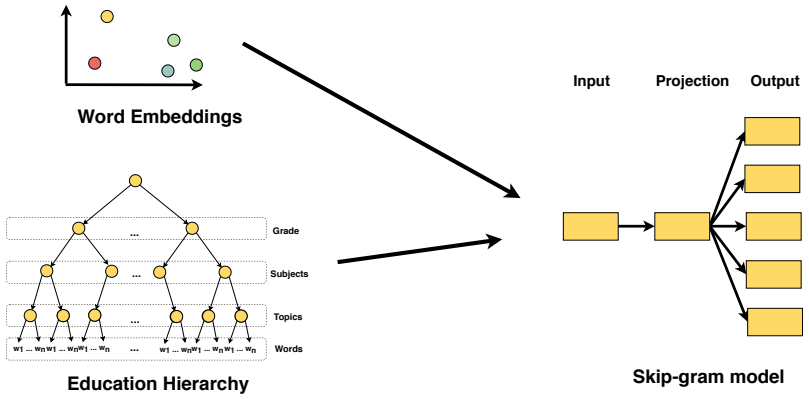


Figure 2: Proposed strategy; skip-gram model refers to version adapted based on Equations 1-3.

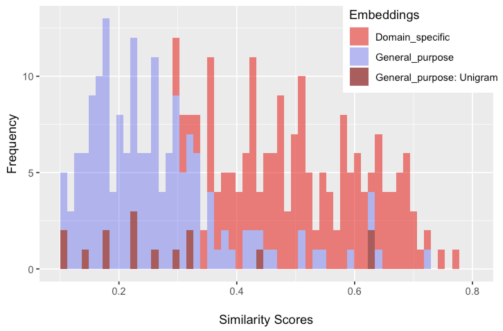


Figure 3: Similarity distribution for <subject, topic> pairs based on general-purpose (FastText) and domain-specific (Edu2Vec) strategies to create word representations.

$$Error\_rate = \frac{\sum_{i=1}^{n} Sim(topic_i, term_i) - \hat{Sim}(topic_i, term_i)}{n} \quad (4)$$

where n is the number of instances in a given dataset, $Sim(topic_i, term_i)$ is the expected similarity score for $< term, topic >$ pairs in the i^{th} instance in the dataset (1 in our case), and $\hat{Sim}(topic_i, term_i)$ is the similarity score of $< term, topic >$ predicted using Equation 5.

$$\hat{Sim}(topic_i, term_i) = \frac{emb(term_i) \cdot emb(topic_i)}{\|emb(term_i)\| \cdot \|emb(topic_i)\|} \quad (5)$$

where *emb* is used to denote the embedding generated for a given term or topic.

Task A: Mapping topics to high level terms. To conduct this analysis, we use DS_A and compute the error rate using Equation 4, based on similarity scores predicted by using general-purpose and educational embeddings, generated using FastText and Edu2Vec.

As reported in Table 2, the error rates yielded using FastText's embeddings were higher than when using Edu2Vec's embeddings. We verified this difference using a statistical significance test (T-test; $p < 0.05$). This result was not unexpected, as FastText's embeddings are based on diverse, non-education-specific information and thus are not meant to explicitly capture within the vector space connections among higher-level, conceptual terminology that naturally aligns with the K–12 curriculum.

Table 2: Error rates for the topic-term similarity task using domain-specific and domain-agnostic embeddings. "[]", indicates the variance among computed values.

Experiment	Error Rate (%) Edu2Vec	FastText
Task A	**15.1** [0.0036]	61.8 [0.0338]
Task B	**16.8** [0.0056]	69.4 [0.0316]

Task B: Mapping topics to lower level terms. For this experiment, we rely on instances in DS_B, and also measure the error rates when predicting the similarity between keywords in each <*topic*, *term*> pair, using embeddings from FastText and Edu2Vec to represent topic and term. Similar to results reported for Task A, we see that using embeddings from FastText for representing topics an terms led to higher error rates when compared to Edu2Vec's embeddings. Differences in error rates are statistically significant (T-test; $p < 0.05$). We observed a 11% increase in Edu2Vec's errors from Task A. We hypothesize this is due to the fact that it is more challenging for Edu2Vec to map more general terms to higher-level concepts, i.e., Edu2Vec is trained to explicitly account for connections among educational concepts (i.e., broader, higher level terms such as subjects and topics). For instance, it is a straightforward task for Edu2Vec to situate in close proximity the topic *shapes* and the topic *geometry*, however, mapping *shapes* to the term *vertex* is more challenging, as those two keywords are presented at different abstraction levels, one is a topic, while the other is a term.

4.3 Discussion

In the previous section, we examined the effectiveness of using Edu2Vec in order to create embeddings that can capture relationships among educational subjects, topics, and terms in the vector space. To further contextualize the applicability of our initial findings, we use a number of illustrative examples. As shown in Figure 4, topics aligned to specific subjects are in close proximity

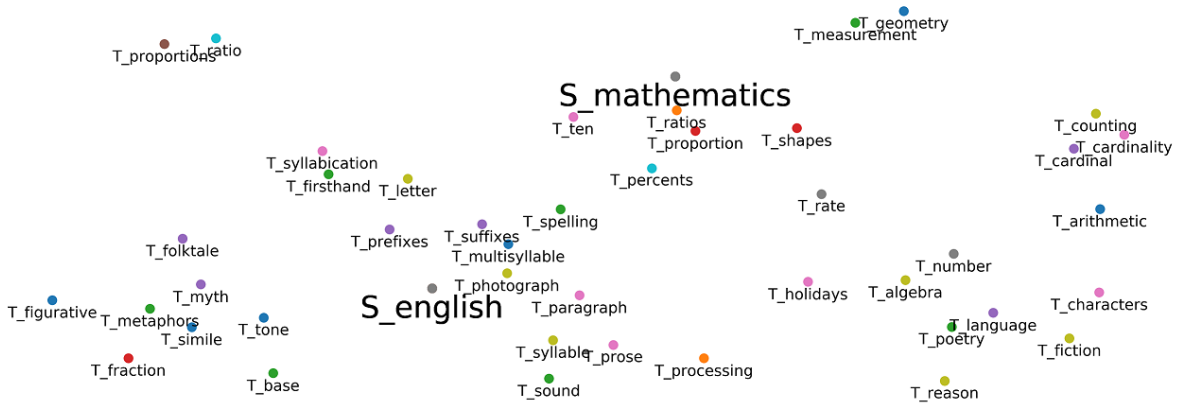


Figure 4: Space representation of subject and topic embeddings. Area around subjects *Mathematics* and *English* is shown. Dimensionality reduced using t-distributed Stochastic Neighbor Embedding (TSNE) [19] algorithm with perplexity and iteration number set to 5 and 5000 respectively.

among each other, when compared to other topics. For instance, the subject ***Mathematics*** is close to topics ***ratios***, ***proportions***, ***geometry***, and ***measurement*** but further apart from topics related to the subject ***English***, such as ***myth*** or **spelling**.

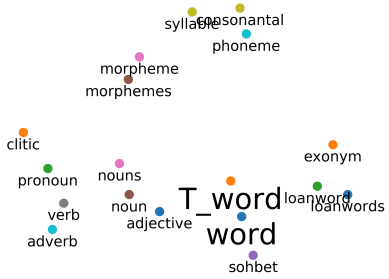


Figure 5: Closest embeddings for *word* as a topic and a term generated using Edu2Vec and FastText, respectively. Dimensionality reduced using t-distributed Stochastic Neighbor Embedding (TSNE) [19] algorithm with perplexity and iteration number set to 5 and 5000 respectively.

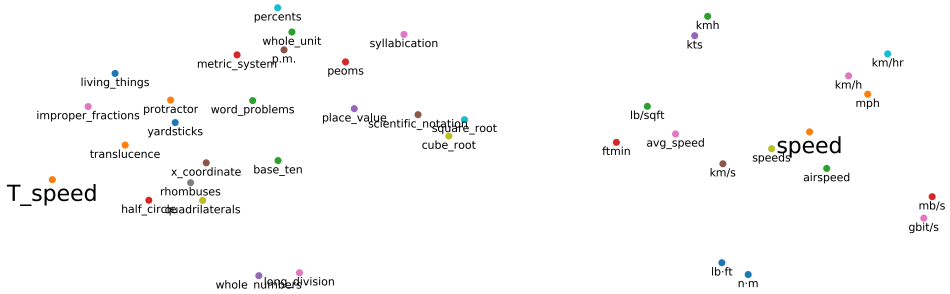
Through further manual analysis, we found that embeddings from FastText correctly capture connections among more general terms in the educational concept vector space. This is showcased in Figure 5, where we depict the embedding representation for both the educational concept ***word*** (i.e., T_word) and the term itself. In this case, there is no noticeable difference that arises by considering the word itself or the topic, in terms of neighbouring vocabulary. However, not biasing the model to favor information specific to a domain (educational in our case) may be a detriment, as then the model would not be able to identify terminology that is essential for the domain. This is indeed apparent in Figure 6, where we depict the most similar words for both the topic *T_speed* and the term *speed*. On the one hand, the term *speed* is most similar to other terms that are known to often co-occur with the keyword speed in general documents, such as *km/s* or *avg_speed*. On the other hand, the topic *T_speed* is closer to other keywords like *protractor* or *x_coordinate* that are known to appear in educational materials that focus on physical science and integrated math curriculum.

5 CONCLUSIONS & FUTURE WORK

Even if word-representation techniques are abundant, most of them are currently focused on the distributional aspect of the words. These strategies tend to generate domain-agnostic representations that can overlook important information that could greatly influence task performance in specific domains. In order to address this issue, we have introduced a novel strategy for generating representations for educational concepts. For doing so, we introduce Edu2Vec, which takes advantage of structured educational knowledge in the form of an hierarchy consisting of K–12 related information in addition to traditional pre-trained word-embeddings.

We conducted an initial assessment using labeled educational resources as well as SimpleWikipedia, demonstrating the importance of incorporating domain knowledge into the embedding-generation process, even if that knowledge is limited. These results, compounded with a preliminary qualitative analysis lead us to argue in favor of relying on embeddings generated by Edu2Vec, as opposed to traditional, general-purpose embeddings, when it comes to tasks and applications essential to the K–12 curriculum, including, but not limited to measuring the educational value of documents; filtering and ranking educational resources in response to web search tasks conducted in the classroom setting and identifying suitable documents that match pre-defined learning objectives.

In the future, we will dedicate research efforts to investigate the applicability of the proposed strategy in domains beyond Education—provided the availability of a structured domain-specific hierarchy, or the possibility to infer one from data. Furthermore, we plan on exploring strategies that can better utilize the hierarchical structure of a domain knowledge. Finally, we will consider alternative techniques for better representing educational documents incorporating not only topical information but also other aspects, such as text complexity.



(a) Closest embeddings for the ***topic*** speed (T_Speed).

(b) Closest embeddings for the ***term*** speed.

Figure 6: Closest embeddings for *speed* as a topic and a term generated using Edu2Vec and FastText, respectively. Dimensionality reduced using t-distributed Stochastic Neighbor Embedding (TSNE) [19] algorithm with perplexity and iteration number set to 5 and 5000 respectively.

ACKNOWLEDGMENTS

This work has been supported in part by the National Science Foundation Award Number 1565937.

REFERENCES

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 451–462.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[3] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 177–186.

[4] Florin Adrian Bulgarov and Rodney Nielsen. 2018. Proposition entailment in educational applications using deep neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[5] Rose Catherine and William Cohen. 2017. Transnets: Learning to transform for recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 288–296.

[6] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. 166–174.

[7] Common Core State Standards Initiative. Accessed: March 2018. Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects. http://www.corestandards.org/assets/CCSSI_ELA20Standards.pdf.

[8] National Research Council et al. 2014. *Developing assessments for the next generation science standards*. National Academies Press.

[9] Cicero dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 69–78.

[10] Michael D Ekstrand, Ion Madrazo Azpiazu, Katherine Landau Wright, and Maria Soledad Pera. 2018. Retrieving and Recommending for the Classroom. *ComplexRec 2018* 6 (2018), 14.

[11] Debasis Ganguly, Dwaipayan Roy, Mandar Mitra, and Gareth JF Jones. 2015. Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. ACM, 795–798.

[12] Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2018. Neural vector spaces for unsupervised information retrieval. *ACM Transactions on Information Systems (TOIS)* 36, 4 (2018), 38.

[13] William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*, Vol. 2016. NIH Public Access, 595.

[14] Hebatallah A Mohamed Hassan, Giuseppe Sansonetti, Fabio Gasparetti, and Alessandro Micarelli. 2018. Semantic-based tag recommendation in scientific bookmarking systems. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 465–469.

[15] Idaho State Department of Education. Accessed: March 2018. Idaho Content Standards. https://www.sde.idaho.gov/academic/standards/.

[16] Aparup Khatua, Apalak Khatua, and Erik Cambria. 2019. A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks. *Information Processing & Management* 56, 1 (2019), 247–257.

[17] Qi Liu, Zai Huang, Zhenya Huang, Chuanren Liu, Enhong Chen, Yu Su, and Guoping Hu. 2018. Finding similar exercises in online education systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1821–1830.

[18] Yue Liu, Tao Ge, Kusum S Mathews, Heng Ji, and Deborah L McGuinness. 2018. Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. *arXiv preprint arXiv:1804.04225* (2018).

[19] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[20] Ion Madrazo Azpiazu, Nevena Dragovic, Oghenemaro Anuyah, and Maria Soledad Pera. 2018. Looking for the Movie Seven or Sven from the Movie Frozen?: A Multi-perspective Strategy for Recommending Queries for Children. In *Proceedings of the 2018 Conference on Human Information Interaction& Retrieval*. ACM, 92–101.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[22] Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. Evaluation of Domain-specific Word Embeddings using Knowledge Resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

[23] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[24] Guangyuan Piao and John G Breslin. 2018. Financial aspect and sentiment predictions with deep neural networks: an ensemble approach. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 1973–1977.

[25] Jiří Rihák and Radek Pelánek. 2017. Measuring Similarity of Educational Items Using Data on Learners' Performance. *Educational Data Mining* (2017).

[26] Isabel Segura-Bedmar, Víctor Suárez-Paniagua, and Paloma Martínez. 2015. Exploring word embedding for drug name recognition. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*. 64–72.

[27] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1555–1565.

[28] Khoi-Nguyen Tran, Jey Han, Danish Contractor, Utkarsh Gupta, Bikram Sengupta, Christopher J Butler, and Mukesh Mohania. 2018. Document Chunking and Learning Objective Generation for Instruction Design. *arXiv preprint arXiv:1806.01351* (2018).

[29] Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. ACM, 363–372.

[30] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1235–1244.

[31] Peter Willett. 2006. The Porter stemming algorithm: then and now. *Program* 40, 3 (2006), 219–223.

[32] Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM international conference on*

conference on information and knowledge management. ACM, 1219–1228.

[33] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 13, 3 (2018), 55–75.