TopicEq: A Joint Topic and Mathematical Equation Model for Scientific Texts

Michihiro Yasunaga, John D. Lafferty
Yale University
michihiro.yasunaga@yale.edu

Abstract

Scientific documents rely on both mathematics and text to communicate ideas. Inspired by the topical correspondence between mathematical equations and word contexts observed in scientific texts, we propose a novel topic model that jointly generates mathematical equations and their surrounding text (*TopicEq*). Using an extension of the correlated topic model, the context is generated from a mixture of latent topics, and the equation is generated by an RNN that depends on the latent topic activations. To experiment with this model, we create a corpus of 400K equation-context pairs extracted from a range of scientific articles from arXiv, and fit the model using a variational autoencoder approach. Experimental results show that this joint model significantly outperforms existing topic models and equation models for scientific texts. Moreover, we qualitatively show that the model effectively captures the relationship between topics and mathematics, enabling novel applications such as topic-aware equation generation, equation topic inference, and topic-aware alignment of mathematical symbols and words.

Introduction

Technical scientific articles, such as those from physics and computer science, rely on both mathematics and text to communicate ideas. Most existing work in natural language processing (NLP) and machine learning studies these two components separately. For instance, text-based topic models have been used widely on scientific articles to uncover their semantic structure (Blei, Ng, and Jordan 2003; Blei and Lafferty 2006; Newman et al. 2010a). For mathematics, recent work (Lan et al. 2015; Zanibbi et al. 2016; Deng et al. 2017) has studied methods to model and generate mathematical equations, for example using RNNs. However, ultimately these two components should be processed together in a seamless manner. Algorithms for automated understanding of scientific documents should extract the information encoded by not only words but also mathematical equations. At the same time, equations should ideally be modeled with the help of the surrounding text, as the meaning of an equation depends not only on its constituent symbols and syntax, but also on the context in which it appears (Wang et al. 2015; Krstovski and Blei 2018).

To this end, this paper proposes a topic-equation model that *jointly* generates equations and their surrounding text in scientific documents (*TopicEq*), and demonstrates that the model can effectively achieve the aforementioned two goals. The intuition behind the model is illustrated in the sample passages in Figure 1, which shows how the topic of the word context is often indicative of the distinctive types of equations used, and vice versa. For instance, equations appearing in the topic of relativity (with context words like "back hole", "Einstein") tend to involve a series of tensors like $G_{\mu\nu}$ and $T_{\mu\nu}$, while equations used in the topic of optimization (context words "gradient", "optimal") may use norms, the $\min$ operator, and often their combinations. Ideally, the strings of mathematical symbols in the equations should aid the training of topic models, and the context words should aid the modeling and understanding of the equations.

Our model formalizes this intuition for scientific texts by generating each equation and its context passage using a shared latent topic. Specifically, we apply a topic model to the context passage, and use the same latent topic proportion vector in a recurrent neural network (RNN) to generate the equation as a sequence of symbols. To develop and experiment with this model, we construct a large corpus of context-equation pairs, extracted from the LaTeX source of arXiv articles across a range of scientific domains (*ContextEq-400K*). We fit the model on this corpus using approximate inference

Black holes in Einstein gravity. As a warm-up exercise, in this section, we will briefly review the observation made by Padmanabhan [14] by generalizing his discussion to a more general spherically symmetric case. In Einstein's general relativity, the gravitational field equations are

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = 8\pi G T_{\mu\nu}$$

where $G_{\mu\nu}$ is Einstein tensor and $T_{\mu\nu}$ is the energy-momentum tensor of matter field. On the other hand, for a general static, spherically symmetric spacetime, its metric can be written down as .....

(snippet from Cai and Ohta (2010))

We give the derivation for the primal-dual subgradient update, as composite mirror-descent is entirely similar. We need to solve update (3), which amounts to

$$\min_x \eta\langle \bar{g}_t, x\rangle + \frac{1}{2t}\delta\|x\|_2^2 + \frac{1}{2t}\langle x, \mathrm{diag}(s_t)x\rangle + \eta\lambda\|x\|_1$$

Let $\hat{x}$ denote the optimal solution of the above optimization problem. Standard subgradient calculus implies that when $|\bar{g}_{t,i}| \leq \lambda$ the solution is $\hat{x} = 0$. Similarly, when $\bar{g}_{t,i} \leq -\lambda$, then $\hat{x} > 0$, the objective is differentiable, and the solution is obtained by setting the gradient to zero. .....

(snippet from Duchi et al. (2011))

Figure 1: The words in a given technical context often characterize the distinctive types of equations used, and vice versa. **Top** topic: Relativity; **bottom** topic: Optimization.

based on a variational autoencoder approach.

Our evaluation shows that this joint model significantly outperforms alternative topic models and RNN equation models for scientific texts. We further show that the model enables novel applications that bridge topics and mathematical equations. Concretely, the paper makes the following contributions.

- The first study of jointly modeling topics and mathematics in scientific texts.
- Better topic models for scientific texts: Joint training with the RNN equation model boosts the quality of topic modeling. This greatly outperforms the topic model that includes equations simply as bags of tokens, suggesting that equations' syntax-level information captured by the RNN is useful for topic modeling.
- Better equation models: Joint topic modeling provides the narrative context for equation prediction, and improves the quality/grammaticality of the RNN equation model.
- Our model successfully captures the relationship between mathematical equations and topics (words), enabling interpretable handling of equations. For instance, we illustrate that the model enables topic-aware equation generation and equation topic inference. We also present a variant of this model that learns topic-aware associations between mathematical symbols and words.
- The model is unsupervised, and enables the aforementioned tasks and applications without manual labels.

Related Work

Our work is connected to a wide range of recent research, from topic models to mathematical equation processing.

Topic models. Topic models provide a powerful tool to extract the semantic structure of texts in the form of the latent topics—usually multinomial distributions over words. Starting from LDA (Blei, Ng, and Jordan 2003), topic models have been studied extensively (Teh et al. 2005; Blei and Lafferty 2006; 2007; Hall, Jurafsky, and Manning 2008), especially for scientific articles. However, while mathematical equations play an essential role in scientific documents, topic models capable of processing equations besides word texts are yet to be studied. This work shows that incorporating joint modeling of equations via an RNN boosts the performance of topic modeling for scientific texts.

Recent work (Cao et al. 2015; Larochelle and Lauly 2012) has proposed neural topic models, leveraging the flexibility and representation power of neural networks. In particular, (Miao, Yu, and Blunsom 2016; Miao, Grefenstette, and Blunsom 2017; Srivastava and Sutton 2017) employ neural variational inference to train topic models; we will apply their technique to fit our model.

Language models & equation models. Language modeling aims to learn a probability distribution over a sequence of words. It is a fundamental task in NLP, with a plethora of applications including text generation. RNN-based language models are shown effective for sequences with long-term dependencies (Mikolov et al. 2010; Jozefowicz et al. 2016).

Similar to language models, equation models are useful for various tasks involving equation generation, such as semantic parsing (Roy, Upadhyay, and Roth 2016) and handwriting / optical character recognition (Deng et al. 2017). The use of RNNs to model LaTeX was illustrated by (Karpathy 2015) for an algebraic geometry text. This work also employs an RNN to model each equation as a sequence of LaTeX tokens (or "symbols," interchangeably).

Neural topic-language models. Our model architecture is motivated by joint topic-language models. Such models typically extract latent topics of a given document via a topic model, and utilize the topic knowledge to improve an RNN language model. Mikolov and Zweig (2012) incorporate the topic vector of a pre-trained LDA model into an RNN language model; recent work (Dieng et al. 2017; Lau, Baldwin, and Cohn 2017; Wang et al. 2018) trains neural topic and language models jointly, as we will do here.

Key distinctions can be made between our work and these models. First, while previous work uses topic models to improve language modeling on the same word text, our task models two different modalities: word text and equations. In this sense, our work is related to (Blei and Jordan 2003), which extends LDA to model image-text pairs. Moreover, taking advantage of these two modalities, we also present a variant of the TopicEq model that learns topic-aware association between mathematical symbols and words.

The second difference lies in the RNN equation model we propose. While (Dieng et al. 2017; Ahn et al. 2016; Lau, Baldwin, and Cohn 2017) integrate the topic knowledge into either the output layer of the LSTM or the word predictions of the language model, we embed the topic proportion vector inside the LSTM, to enable the topic knowledge to have deeper influence on equation generation. Experimental results show that this method of incorporating topic information is more effective than the existing methods for improving the quality of equation modeling.

Mathematical equation processing. Some work has processed equations as bags of math symbols to extract their features for searching (Sojka and Líška 2011) and clustering (Lan et al. 2015). Zanibbi et al. (2016) introduce tree-based representations for equations for mathematical information retrieval tasks. Most recently, Deng et al. (2017) propose RNN-based models to generate equations. We will show that RNN-based equation processing can capture syntactic features of equations, and provides more effective help for topic modeling than bag of token-based equation processing does.

Finally, our work of modeling equations with contexts is related to (Krstovski and Blei 2018), which fits equation embeddings using surrounding words. While they limit the equation domains (i.e., ML, AI), this work aims to uncover topics for texts and equations from a range of scientific domains. This work also models each equation itself as a sequence of symbols, which is not studied in their work.

The TopicEq Model

Our starting point is the correlated topic model (Blei and Lafferty 2007), which models the topic proportion vector through a latent Gaussian vector. We extend this model to

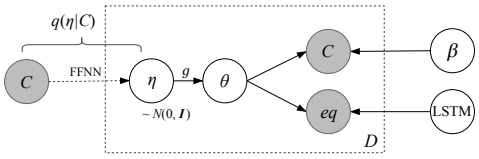

Figure 2: Graphical structure underlying the TopicEq model.

the setting where each "document" consists of a displayed equation *eq* and its surrounding text $C = \{w_n^{(c)}\}_{n=1}^N$, which we call the equation's *context*. Our joint model assumes that each equation and its context are generated from the same latent topic vector θ; see Figure 2. Concretely, the generative process for a given $D = (C, eq)$ is

$$\eta \sim \mathcal{N}(0, I), \quad \theta = g(\eta) \tag{1}$$

$$w_n^{(c)} \mid \theta \sim \text{Mult}(\theta^T \beta) \tag{2}$$

$$eq \mid \theta \sim \text{LSTM}(\theta) \tag{3}$$

where $g(\eta) = \text{softmax}(W_g \eta + b_g)$. Note that this is equivalent to placing a logistic normal distribution on θ where the latent Gaussian has mean b_g and covariance $W_g W_g^T$. The parameters W_g, b_g, the topics β, and the weights in the LSTM are to be estimated from data. Expressing the model as shown in Figure 2 emphasizes the connection with neural topic models such as (Miao, Grefenstette, and Blunsom 2017); we will apply their model training technique.

Both the words and the equation are generated in a way that depends on the topic proportion vector θ. The topics $\beta^T = (\beta_1, \ldots, \beta_K)$ are distributions over a word vocabulary with V words; the context words $w_n^{(c)}$ are then drawn from the mixture $\theta^T \beta$, similar to (Wang et al. 2018). We employ an RNN to generate *eq* as a sequence of mathematical tokens, where the vocabulary is extracted from the set of LaTeX tokens. Specifically, to generate an equation conditioned on the latent topic proportion vector θ (equivalently η), we consider a *Topic-Embedded LSTM* (TE-LSTM), an extension of the LSTM (Hochreiter and Schmidhuber 1997) where the t-th update is

$$\begin{aligned} i_t &= \sigma\left(W_i[x_t; h_{t-1}; \theta] + b_i\right) \\ f_t &= \sigma\left(W_f[x_t; h_{t-1}; \theta] + b_f\right) \\ \tilde{c}_t &= \tanh\left(W_c[x_t; h_{t-1}; \theta] + b_c\right) \\ o_t &= \sigma\left(W_o[x_t; h_{t-1}; \theta] + b_o\right) \end{aligned}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \quad h_t = o_t \odot \tanh\left(c_t\right).$$

Here $[x_t; h_{t-1}; \theta]$ denotes the concatenation of the current input, previous state and topic proportion vector; σ is the sigmoid function and $\odot$ denotes the Hadamard product. The probability of the next token in the equation is $p(y_t \mid y_{1:t-1}) = \text{softmax}\left(W_y h_t + b_y\right)$. Thus, the TE-LSTM embeds θ inside the LSTM cell to reflect the topic knowledge for equation generation. As a joint topic-equation model, it is similar to the topic-language model of (Wang et al. 2018).

Writing the equation as a sequence of tokens $eq = y_{1:T}$, the training objective is the marginal likelihood of C and *eq*

$$p(C, y_{1:T}) = \int_\eta p(\eta) p(C|\eta) \prod_{t=1}^{T} p(y_t | y_{1:t-1}, \eta) d\eta \tag{4}$$

Since its direct optimization is intractable, we employ variational inference (Jordan et al. 1999). Denoting the variational distribution by $q(\eta)$, we maximize the variational lower bound (ELBO) for the log-likelihood, $\log p(C, y_{1:T})$:

$$\begin{aligned} \mathcal{L} = \mathbb{E}_{q(\eta)}\big[\log p(C|\eta)\big] - D_{\text{KL}}\big(q(\eta) \,\|\, p(\eta)\big) \\ + \mathbb{E}_{q(\eta)}\left[\sum_{t=1}^{T} \log p(y_t \mid y_{1:t-1}, \eta)\right] \end{aligned} \tag{5}$$

Following recent approaches to neural topic-language models (Miao, Grefenstette, and Blunsom 2017; Dieng et al. 2017; Wang et al. 2018), we compute $q(\eta)$ as a function of the context C using the variational autoencoder technique (Kingma and Welling 2014). Specifically, we use a feed-forward neural network (FFNN) as an inference network to parameterize the mean and variance vectors of the (diagonal) Gaussian variational distribution $q(\eta \mid C)$. We then use samples from q to optimize Eq 5. The parameters of the inference network, the topic model, and the equation model are jointly trained by stochastic gradient descent.

We also include a topic diversity regularization term to Eq 5, following (Xie, Deng, and Xing 2015). We observed that this technique prevents learning generic, redundant topics.

Experiments

We study the performance of the proposed model on a corpus of context-equation pairs constructed from arXiv articles. We quantitatively show that our joint topic-equation model provides superior fits than alternative topic models and equation models. We further demonstrate its efficacy through qualitative analyses and novel applications, such as equation generation and equation topic inference.

Dataset Construction (*ContextEq-400K*)

To obtain a dataset of context-equation pairs, we used scientific articles published on arXiv.org. We sampled 100k articles from all domains in the past 5 years, and split them into train, validation and test sets (80%, 10%, 10%). For each article, we parsed its LaTeX source and extracted single-line display equations that have five consecutive sentences both before and after the equation, which are used to define the word context. Following (Deng et al. 2017), we further tokenized each equation into a sequence of LaTeX tokens (e.g., `\sigma`, `^`, `{`, `2`, `}`) and kept those of length 20–150, yielding the final corpus of 400K equation-context pairs. An equation has 63 tokens on average. The context size of 10 sentences is similar to the document size used in recent work of topic-language models (Dieng et al. 2017; Wang et al. 2018).

Experimental Setup

We fit the TopicEq model end-to-end on the train set and evaluate its performance on the test set.

Preprocessing. For the topic modeling of context passages, we first removed all the inline math expressions in the text. We then followed the preprocessing steps in (Wang et al. 2018) to tokenize and lowercase all words, exclude stopwords and words appearing in fewer than 100 documents;

Topic Model	50	100 (# Topics)
LDA (context only)	.085	.083
Ours (context only)	.085	.084
Ours (context + Eq BOW)	.087	.086
Ours (context + Eq LSTM)	**.097**	**.094**
Ours (context + Eq LSTM shuffled)	.086	.085

Table 1: Topic coherence of different topic models, evaluated on the held-out arXiv data. Our full TopicEq model is shown as "Ours (context + Eq LSTM)."

Quantum physics	spin energy field electron magnetic state states hamiltonian
Particle physics	higgs neutrino coupling decay scale masses mixing quark
Astrophysics	mass gas star stellar galaxies disk halo radius luminosity
Relativity	black metric hole schwarzschild gravity holes einstein
Number theory	prime integer numbers conjecture integers degree modulo
Graph theory	graph vertex vertices edges node edge number set tree
Linear algebra	matrix matrices vector basis vectors diagonal rank linear
Optimization	problem optimization algorithm function solution gradient
Probability	random probability distribution process measure time
Machine learning	layer word image feature sentence model cnn lstm training

Table 2: Topics learned by the TopicEq model. Left: topic name (summarized by us). Right: top words in topic.

this resulted in a vocabulary size of 8,660. For equations, we use the 1,000 most frequent LaTeX tokens as our vocabulary.

Model setting. For the inference network $q(\eta|C)$, we use a 2-layer FFNN with 300 units, similar to (Miao, Yu, and Blunsom 2016; Miao, Grefenstette, and Blunsom 2017). The equation TE-LSTM architecture has two layers and state size 500, with dropout rate 0.5 applied to each layer (Srivastava et al. 2014). The parameters of the TopicEq model are jointly optimized by Adam (Kingma and Ba 2015), with batch size 200, learning rate 0.002, and gradient clipping 1.0 (Pascanu, Mikolov, and Bengio 2012).

Topic Model Evaluation

We first study the topic modeling performance of TopicEq, by evaluating the coherence of the learned topics β (Chang et al. 2009; Newman et al. 2010b; Mimno et al. 2011). Specifically, following (Lau, Newman, and Baldwin 2014), we compute the normalized PMI metric on the held-out test set. As our TopicEq model incorporates joint, RNN-based equation model, to analyze its effect, we compare the full TopicEq model with the following baseline topic models:

- LDA (context only): we apply LDA to the word text
- Ours (context only): TopicEq without the equation model
- Ours (context + Eq BOW): TopicEq's joint LSTM equation model (Eq 3) is replaced by a baseline bag-of-tokens model similar to that for context words.

The evaluation results are summarized in Table 1. The full TopicEq model is shown as "Ours (context + Eq LSTM)" in the table. We observe that TopicEq's topic model component (2nd row) performs on a par with LDA (1st row), but it achieves a significant boost (+0.01) when trained together with the LSTM equation model (4th row). Adding equations as bag of tokens (3rd row) does improve topic models marginally (+0.002), but the improvement made by using joint LSTM equation model is 5 times greater. These results show that a joint RNN equation model provides significant information to aid topic modeling of scientific texts.

Equation Model	Perplexity 50	Perplexity 100	Error (%) 100
No joint training			
LSTM (no topic)	5.81	5.81	15.3
LSTM + LDA	5.54	5.52	13.4
Joint training with topic model			
TD-LSTM (Lau et al. 2017)	5.44	5.41	12.5
TE-LSTM (**Ours**)	**5.36**	**5.34**	**11.7**

Table 3: Performance of different equation models, evaluated on held-out arXiv data. We report the perplexity metric (for # topics 50, 100 if topic info is used), and the syntax error rate of generated LaTeX equations (for # topics 100).

Why is the RNN helpful? We hypothesize that one reason why the joint RNN equation model is more helpful than the bag-of-tokens equation model is that the RNN also captures syntax-level information in equations. But one might argue that the introduction of the RNN itself was useful for topic modeling (e.g. as a form of regularization). To study our hypothesis, we re-trained TopicEq with each equation's token order randomly shuffled in the training data—thus corrupting the syntactic information of each equation. The result is shown in Table 1 as "Ours (context + Eq LSTM shuffled)." This time, the topic model performance degrades severely and falls to the level of the baseline topic model, "Ours (context only)". This result supports the claim that the original TopicEq's joint RNN actually captured syntactic features of equations, providing more effective help for topic modeling than a bag-of-token equation model does.

This idea also makes intuitive sense. Mathematical equations use a much smaller vocabulary (symbols / variables) than word texts, and thus often need phrase or syntax-level information to aid topic modeling. For example, in the equations in Figure 1, phrases like $T_{\mu\nu}$ (use of super/sub-scripts for a tensor) and $\lambda\|x\|_1$ (regularization term) provide rich information to identify the topics (relativity and optimization), while the corresponding bags of tokens $\{\mu, \nu, T\}$ and $\{1, \lambda, x, |\}$ themselves do not provide as much help.

Learned topics. To visualize the topic modeling performance, we sampled 10 topics learned by TopicEq (Table 2). They intuitively reflect the scientific topics of arXiv articles.

Equation Model Evaluation

Next, we evaluate the equation model component of TopicEq by measuring the test set perplexity. Additionally, as the grammaticality of equations can be measured using the LaTeX compiler, we also evaluate the syntax error rate of generated equations. We compare our TE-LSTM with

- a generic LSTM (no topic knowledge)
- LSTM + LDA: the topic vector θ obtained from a pre-trained LDA is concatenated to the output of LSTM

and a recent topic-dependent LSTM applied to our task

Topic	Generated Equations
Quantum physics	• $E = \hbar\frac{\partial^2 S}{\partial t^2}(\frac{\partial \varphi}{\partial c}) - \frac{k}{\hbar^2}\frac{\partial B}{\partial t}\,(t + \partial_t \delta)$. • $\Psi_{\text{pr}} = \sum_{\mathbf{l}}(\psi_{\mathbf{r}+\uparrow} - \psi^{\dagger}_{\mathbf{r}\downarrow}) + \sum_{\mathbf{r}'}(\psi^{\dagger}_{\mathbf{r}\downarrow,\uparrow} - \psi_{\mathbf{r}\downarrow}\sigma^{\dagger})$.
Particle physics	• $\mathcal{H} = \frac{2}{4}(\partial_\mu \phi)^2 + 2m\phi_\nu(\phi) + \frac{1}{2}m^2(\phi)(1-\phi^2)^2$. • $m_{\text{eff}}(M) = 1.4 \cdot 10^{-13}\,\text{GeV}$.
Relativity	• $\mathcal{M} = \frac{2}{2}g^{\mu\nu}(f_{\mu\nu,\mu} - g_{\mu\nu,\nu} + g_{\nu\nu,b}f_{\mu,\nu}) + \frac{1}{2}g^{\mu\nu}$. • $T_{\mu\nu} = \int_0^{\infty} ds_{\mu\nu}ds^2 + a_\mu^2 dr^2 + r^2 d\Omega^2$.
Number theory	• $(2^k)^k + (1^n + 1)(1 + p^k) = 1$.
Linear algebra	• $\text{tr}(E_\varepsilon X^*) = U^\top(\text{tr}(V_\varepsilon X))$. • $\phi_h(\theta, y) = \left\{X \in \text{Span}\left(P_c(\mathbf{T}[x, \text{x}])\right)\right\}$.
Optimization	• $\min_p p(x)$ subject to $\|p^x - y\|_2 \leq m_p$. • $w^+ = w_t + g_t\|u_t - \nabla \mathbf{u}^*\|_2^2$.
Probability	• $\mathbb{P}(r_\tau < t) = \mathbb{E}_{\tau_{\text{wist}}}(N_\tau)$. • $T^*(t) = \lim_{t\to\infty}\mathbb{E}[\|N(t) + \mathbb{E}[\varphi_t(x)]^*\|]$

Table 4: The TopicEq model generates equations that reflect the characteristics of given topics. Left: topic (picked from Table 2). Right: equations generated by the model conditioned on the given topic (one-hot topic vector θ).

- TD-LSTM (Lau, Baldwin, and Cohn 2017): θ is added to the output of LSTM via a dense layer.

TD-LSTM and our TE-LSTM are jointly trained with our topic model component. As Table 3 shows, all the topic-dependent LSTMs are superior to the vanilla LSTM in both the perplexity metric and syntax error metric. Moreover, our TE-LSTM outperforms TD-LSTM, suggesting that the model better incorporates topic knowledge by embedding θ inside the LSTM. We also find that compared to (Wang et al. 2018)'s Mixture-of-Expert LSTM, our model achieves similar performance in this task while requiring fewer parameters and much less training time (40% reduction). In total, compared to the generic LSTM, our TE-LSTM equation model reduces test perplexity by 8% (relative) and syntax error rate by 3.5% (absolute). This result suggests that incorporating context/topic information can improve the quality and grammaticality of equation modeling.

Qualitative Analysis & Applications

Topic-aware Equation Generation

The TopicEq model can generate meaningful equations from specified topics, using Eq 3 (TE-LSTM). For example, given a topic k, we let θ be the one-hot vector representing the topic; conditioned on θ, and starting from `<START>` token, we keep sampling the next LaTeX token until the `<END>` token is generated. Table 4 shows several topics picked from Table 2 (left), and equations generated from each of these topics (right). We see that the artificial equations generated by the model clearly reflect the distinctive characteristics of the given topics. For instance, derivatives, and number + units are generally used for physics; electron configuration $\uparrow, \downarrow$ for quantum physics; series of tensors like $T_{\mu\nu}$ for relativity; prime number p for number theory; $\mathbb{E}$, $\mathbb{P}$ clauses for probability. We also note that the equations generated by our TE-LSTM use not only topic-specific symbols but also topic-specific phrases and syntax (e.g., a set definition is used for linear algebra; "$\min$ subject to" clause for optimization). These qualitative results support that TopicEq is capable of fully incorporating topic information for equation modeling.

Topic Gradation	Generated Equation (Greedily decoded)
Astrophysics (100%)	$G_{\text{eff}} = \frac{1}{2}\left(\frac{M_{\text{eff}}}{M_\odot}\right)^{-1}\left(\frac{M_{\text{eff}}}{M_\odot}\right)^{-1}$
⋮	$G_{\text{eff}} = \frac{1}{2}\left(\frac{M_{\text{eff}}}{M_\odot}\right)^{-1}$
⋮	$G_{\text{eff}} = \frac{1}{2}\left(\frac{1}{2} + \frac{1}{2}\right)$
50% — 50%	$G_{\text{s}} = \frac{1}{2}(1 - \frac{1}{2})$
⋮	$G_{\text{s}}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{x} + \mathbf{x}^T\mathbf{x}$
⋮	$G_{\text{s}}(\mathcal{C}) = \mathcal{C}(\mathcal{C}) + \mathcal{C}(\mathcal{C})$.
Graph theory (100%)	$G_i = \{(x, y) \in \mathbb{R}^n : x_i = x_i\}$
Optimization (100%)	$L = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2$
⋮	$L = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2$
⋮	$L = \frac{1}{2}\sum_{i=1}^n\sum_{i=1}^n(x_i - x_i)^2 + \sum_{i=1}^n(x_i - x_i)^2$
50% — 50%	$L = \frac{1}{2}\sum_{i=1}^n\sum_{i=1}^n(x_i - x_i)^2 + \sum_{i=1}^n x_i^2$
⋮	$L = \frac{1}{N}\sum_{i=1}^N\sum_{i=1}^N(x_i - x_i)^2$
⋮	$L = \frac{1}{N}\sum_{i=1}^N\mathbb{E}[\mathbf{x}_i^T\mathbf{x}_i]$,
Statistics (100%)	$L = \frac{1}{N}\sum_{i=1}^N\mathbb{E}[\mathbf{x}_i^T\mathbf{x}_i]$

Table 5: We let the TopicEq model greedily generate equations while smoothly changing θ between two topics (via linear interpolation). Left: given topic pair and its interpolation. Right: generated equation (for the first topic pair, we let the model generate from G; for the second pair, from $L =$).

Context words	Inferred Topics	Generated Equations
star gravity einstein mass galaxies	58% Astrophysics 36% Relativity 3% Quantum physics	• $\left(\frac{m_b}{M_\odot}\right)^{b_\nu} \ll g\sqrt{\frac{\tilde{\Phi}_0}{\eta_{\text{eff}}}}$. • $G(r) = \int_{r_0}^{r_0} dr\sqrt{\log(r)(\overline{r} + r_0(w))}$.
data training likelihood model gradient	62% Machine learning 21% Statistics 15% Optimization	• $L = -\frac{1}{N}\sum_{i=1}^N\mathcal{R}_{R_i}(\mathbf{r}_i) + V^r(\mathbf{r}_i)$. • $\text{argmax}_{\mathcal{U}}\,\mathbb{E}_{W\sim\Psi}\log\exp\left[\Lambda(\widetilde{W}) - H\right]$.

Table 6: Given a set of context words picked from an article abstract (1st column), we let TopicEq infer its topic proportions (2nd col) and generate equations (3rd col).

Mixtures of topics. The model can also generate equations from a mixture of topics by setting θ accordingly. To qualitatively analyze the space of the topic vector θ in terms of equation generation, we let the model generate equations while smoothly changing θ between two topics (i.e., one-hot vectors θ_1 and θ_2) via linear interpolation: $\theta(t) = (1-t)\theta_1 + t\theta_2$ for $t \in [0, 1]$. In Table 5, for two examples we show the given topic pair and its interpolation (left), and the equation greedily decoded from each $\theta(t)$ (right). We let the model start all equations from G in the first example (astrophysics and graph theory), and from $L =$ in the second example (optmization and statistics). In both cases we observe that the generated equations make a smooth transition from one topic to the other — e.g., for the first example, from using $M_{\text{eff}}/M_\odot$ (astrophysics) to using linear algebraic term $\mathbf{x}^T\mathbf{x}$, and finally a set notation (graph theory). In the second example, where the two topics optimization and statistics are closely related, the generated equations make a very intuitive transition: from an optimization objective with norms and

Given Equation ⟦ ⟧ shows the correct formula name for readers	Inferred Topic (showing top 5 words)	
	by our TopicEq	by bag-of-token baseline
#1 $i\hbar\frac{\partial}{\partial t}\|\Psi(\mathbf{r},t)\rangle = \hat{H}\|\Psi(\mathbf{r},t)\rangle$ ⟦Schrödinger Equation⟧	hamiltonian, spin, particle, interaction, wave ✓	time, operator, space, hamiltonian, system ✓
#2 $F = \frac{d(mv)}{dt}$ ⟦Newton's 2nd Law of Motion⟧	velocity, particle, pressure, motion, force ✓	time, velocity, particle, diffusion, force ✓
#3 $W + \Delta U = \int f \cdot dx - mgh$ ⟦Potential energy & Work⟧	direction, force, surface, strain, stress ?	method, order, solution, numerical, problem ✗ *(vague)*
#4 $f_m = \sigma(W_f h_{m-1} + U_f x_m + b_f)$ ⟦LSTM⟧	layer, word, image, feature, network ✓	function, section, problem, condition, solution ✗ *(vague)*
#5 $P(X\|Y) = \frac{P(Y\|X)P(X)}{P(Y)}$ ⟦Bayes' Theorem⟧	random, variable, probability, distribution, entropy ✓	probability, random, theorem variable, distribution ✓
#6 $\lim_{n\to\infty} P\left(\sqrt{n}(S_n - \mu) \le z\right) = \Phi\left(\frac{z}{\sigma}\right)$ ⟦Central Limit Theorem⟧	measure, random, process, gaussian, convergence ✓	probability, random, theorem variable, distribution ✓
#7 $f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$ ⟦Taylor Expansion⟧	coefficients, series, expansion fourier, polynomial ✓	polynomial, series, function, convergence, order ✓
#7′ $h(b) = h(a) + \frac{h'(b)}{1!}(b-a) + \frac{h''(b)}{2!}(b-a)^2 + \cdots$ ⟦Taylor Expansion⟧	coefficients, series, expansion fourier, polynomial ✓	function, integral, equation point, solution ✗ *(fooled)*

Table 7: The TopicEq model can infer the appropriate topic for equations from various domains, with better precision and consistency than bag-of-token baseline. Left: given equation. Right: topic inferred by our model and the baseline. ✓ indicates that the inferred topic is correct; ✗ not good. We verified that the *exact* same equations did not appear in the training data.

regularization terms (top), to using summation terms (middle) and finally expectations (bottom; statistics topic). These observations support that TopicEq learns smooth representations for the latent topic vector θ (especially for a mixture of closely related topics), regarding equation generation.

Finally, we illustrate that the model can generate equations from a given set of context words. Specifically, we let the model infer the topic proportion θ of the context words via the inference network $q(\eta|C)$, and then generate equations from θ via Eq 3 (TE-LSTM). As Table 6 shows, the model is able to infer the right topic mixture (2nd column) and generate equations that reflect those topics (e.g., solar mass $M_\odot$ and radius r are used for the top example; loss function L, $\arg\max$, and $\mathbb{E}$ for the bottom example).

Equation Topic Inference

Identifying the topic of equations is an important task that allows readers to obtain semantic descriptions for equations unfamiliar to them. However, while some work (Schubotz et al. 2016; Stathopoulos et al. 2018) has studied the task of identifying the meaning of individual mathematical symbols, no prior work has succeeded in providing descriptions to entire equations from various domains.

Our TopicEq model can be utilized to identify the topic of given equations. Specifically, with a trained TopicEq model, for a given equation *eq*, we find the topic $k \in [K]$ (so θ is a one-hot vector) that maximizes the likelihood $p(eq \mid \theta)$ in Eq 3, which is parametrized by our topic-dependent LSTM. Table 7 shows examples of equations across different domains (1st column), and the most likely topic inferred by our model for each equation (2nd column). We used $K = 100$ topics in this task. We observe that the TopicEq model correctly identifies the domains or even finer topics (e.g., note the distinction between **#5** and **#6**) for most of the given equations.

Is an RNN necessary for this task? We repeated this experiment using a bag of tokens model for equations in Eq 3 (instead of LSTM), to analyze whether the RNN equation model provides an advantage over the bag of tokens-based approach in this task. As can be seen in Table 7, 3rd column, this bag-of-tokens baseline performs as well in **#1** and **#2**, which have topic-specific variables like $\hbar$, ψ, v, but fails in **#3** and **#4**, which consist of a relatively generic set of symbols $\{f, h, m, U, W, x\}$ and require recognizing phrases like $\int f \cdot dx$ (work) and $\sigma(Wh + b)$ (neural network layer) to identify the correct topic. Indeed, the topics predicted for **#3** and **#4** are very generic and similar. Similarly, the bag-of-tokens baseline fails to distinguish **#5** and **#6**, most likely because it does not recognize the phase and syntax-level differences between these two equations. Finally, for **#7** (Taylor Expansion), we also experimented with **#7'**, where we just changed some variable names without altering the equation's meaning and syntax. While our TopicEq still recognizes this to be the same topic as **#7**, the bag-of-tokens baseline is fooled by the changed variable names and predicts a wrong topic. These observations suggest that the RNN equation model can capture phrase and syntax-level information, and can consistently infer the correct topics for equations from various domains. The TopicEq model could be used to help readers interpret equations unfamiliar to them.

Extension: Topic-aware alignment between mathematical tokens and words

Mathematical symbols (including variables) carry different meanings in different contexts or topics. Prior work (Pagael and Schubotz 2014; Schubotz et al. 2016; Stathopoulos et al. 2018) has studied the task of identifying meanings of math variables using surrounding words, but its topic dependence has not been modeled explicitly. Here we present a variant of the TopicEq model that captures *topic-dependent* alignment between mathematical tokens and words from scientific document data. Specifically, we aim to learn the most probable

Math symbol	Topics			
	No Topic	Probability	Quntum physics	Graph theory
E	energy, expectation, elliptic curve	expectation, expected value	electric field, energy	edge, spectral sequence
M	mass, matrix	martingale, maximum	magnetic moment, mass	matroid, matching
p	polynomials, momentum, probability	probability, poisson, distribution	momentum, proton, pressure	path, perimeter, probability
T	temperature, transpose, transfer matrix	stopping time, test statistic	temperature, thermal conductivity	tree, trees, triangulation
V	potential, voltage, visibility, volume	variance, volatility	voltage, potential energy	vertex, volume, SVD
σ	conductivity, variance, normal distribution	standard deviation, normal distribution	conductivity, pauli matrices	permutation, simplex
$\|$	norm, distance, conditional	conditional probability	absolute value	triangle inequality, cardinality

Table 8: Top word phrases **predicted by our topic-aware alignment model** for each math symbol. We show the prediction results for three of the learned topics (3rd-5th column), as well as the non-topic baseline (2nd column).

descriptions (word phrases) w associated with a given math symbol s, under a given topic or topic mixture θ: $p(w \mid s, \theta)$.

Baseline alignment model. We use the equations and context texts from our *ContextEq* corpus. Similar to (Pagael and Schubotz 2014), we consider that the descriptions of math symbols often appear in the sentence immediately before or after the given equation (*immediate context*). We then consider a simple alignment model between symbols s in the equation and phrases w in the immediate context, such that

$$w \sim \text{Mult}(\text{softmax}(A\boldsymbol{s})) \tag{6}$$

Here vector $\boldsymbol{s} \in \mathbb{R}^L$ is the bag-of-tokens representation of the equation. $A \in \mathbb{R}^{M \times L}$ is the alignment matrix we estimate from the data, by maximizing the likelihood $p(w \mid \boldsymbol{s})$. L, M are the vocab sizes of symbols and word descriptions. For the vocabulary of word descriptions, we collect the titles of Wikipedia pages that contain mathematical equations. We then use the top 2,000 phrases that appear in our arXiv dataset. For math symbols, we use the top $L = 200$.

To predict w given a single symbol s, we set $\boldsymbol{s}$ to be the one-hot vector representing s, as a surrogate.

Topic-aware alignment model. To model $p(w \mid s, \theta)$, we want the alignment matrix A to depend on θ. Motivated by the tensor factorization method in (Song, Gan, and Carin 2016), we let

$$A(\theta) = W_a \cdot \text{diag}(W_b\theta) \cdot W_c \tag{7}$$

where $W_a \in \mathbb{R}^{M \times F}$, $W_b \in \mathbb{R}^{F \times K}$, $W_c \in \mathbb{R}^{F \times L}$ are parameters to estimate. F is the number of factors, which we set to be equal to the number of topics K. To jointly perform topic modeling and alignment learning, we consider a variant of TopicEq, where we just replace Eq 3 by this topic-dependent alignment model. We train it on the *ContextEq* corpus.

Results and Discussion

Table 9 shows the perplexity of the baseline / topic-aware alignment models evaluated on the held-out test set. We observe that the topic information significantly improves the alignment between math symbols and word descriptions, reducing the perplexity by more than 33% (relative).

Qualitative results. Table 8 shows the actual top phrases predicted by the alignment models for several math symbols that are used in a wide range of domains. The proposed TopicEq variant indeed learns the topic-dependent alignment between symbols and words. For instance, it associates E with "expectation" for the probability topic, "electric field" for quantum physics, and "edge" for graph theory, which makes intuitive sense. On the other hand, the baseline (no topic) model associates E with "energy", which is simply the description that appears most frequently across all articles. This is another example where the TopicEq framework can be used to capture the relation of topics and mathematics.

Alignment Model	50	100 (# Topics)
Baseline (no topic)	602	602
Topic-Aware	**406**	**387**

Table 9: Test perplexity for phrase prediction.

Topic Model	50	100 (# Topics)
Context Only	.085	.084
with joint Alignment Model	**.088**	**.087**

Table 10: Topic coherence evaluation for each topic model.

Utility. We also note that our topic-aware alignment model can be conditioned on a mixture of topics by setting θ accordingly. Given a context text and equation, this model can infer the topic proportion by the topic model component, and then use the topic-aware alignment component to infer the most probable meaning of each variable in the given equation. This could aid readers to comprehend scientific documents containing mathematics unfamiliar to them.

Effect on topic modeling. In Table 10, we compare our baseline topic model (top) and this TopicEq variant with the alignment component (bottom). The joint alignment model provides moderate improvements for topic modeling quality.

Conclusion

Motivated by the topical correspondence between text and mathematical equations observed in scientific documents, we proposed *TopicEq*, a joint topic-equation model that generates the text by a topic model and the equations by a topic-dependent RNN. This joint model outperforms existing topic models and equation models for scientific texts. We also qualitatively analyzed TopicEq, and showed its applications and extensions, such as equation topic inference and topic-aware alignment of mathematical symbols and words.

References

Ahn, S.; Choi, H.; Pärnamaa, T.; and Bengio, Y. 2016. A neural knowledge language model. *arXiv:1608.00318*.

Blei, D. M., and Jordan, M. I. 2003. Modeling annotated data. In *SIGIR*.

Blei, D. M., and Lafferty, J. D. 2006. Dynamic topic models. In *ICML*.

Blei, D. M., and Lafferty, J. D. 2007. A correlated topic model of science. *The Annals of Applied Statistics* 17–35.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR*.

Cao, Z.; Li, S.; Liu, Y.; Li, W.; and Ji, H. 2015. A novel neural topic model and its supervised extension. In *AAAI*.

Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J. L.; and Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*.

Deng, Y.; Kanervisto, A.; Ling, J.; and Rush, A. M. 2017. Image-to-markup generation with coarse-to-fine attention. In *ICML*.

Dieng, A. B.; Wang, C.; Gao, J.; and Paisley, J. 2017. Topicrnn: A recurrent neural network with long-range semantic dependency. In *ICLR*.

Hall, D.; Jurafsky, D.; and Manning, C. D. 2008. Studying the history of ideas using topic models. In *EMNLP*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1999. An introduction to variational methods for graphical models. *Machine learning* 37(2):183–233.

Jozefowicz, R.; Vinyals, O.; Schuster, M.; Shazeer, N.; and Wu, Y. 2016. Exploring the limits of language modeling. *arXiv:1602.02410*.

Karpathy, A. 2015. The unreasonable effectiveness of recurrent neural networks. Blog posting, May 21.

Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *ICLR*.

Krstovski, K., and Blei, D. M. 2018. Equation embeddings. *arXiv:1803.09123*.

Lan, A. S.; Vats, D.; Waters, A. E.; and Baraniuk, R. G. 2015. Mathematical language processing: Automatic grading and feedback for open response mathematical questions. In *ACM Conference on Learning@ Scale*.

Larochelle, H., and Lauly, S. 2012. A neural autoregressive topic model. In *NIPS*.

Lau, J. H.; Baldwin, T.; and Cohn, T. 2017. Topically driven neural language model. In *ACL*.

Lau, J. H.; Newman, D.; and Baldwin, T. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*.

Miao, Y.; Grefenstette, E.; and Blunsom, P. 2017. Discovering discrete latent topics with neural variational inference. In *ICML*.

Miao, Y.; Yu, L.; and Blunsom, P. 2016. Neural variational inference for text processing. In *ICML*.

Mikolov, T., and Zweig, G. 2012. Context dependent recurrent neural network language model. *SLT* 12:234–239.

Mikolov, T.; Karafiát, M.; Burget, L.; Černockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *Interspeech*.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *EMNLP*.

Newman, D.; Baldwin, T.; Cavedon, L.; Huang, E.; Karimi, S.; Martinez, D.; Scholer, F.; and Zobel, J. 2010a. Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web* 8(2-3):169–175.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010b. Automatic evaluation of topic coherence. In *NAACL*.

Pagael, R., and Schubotz, M. 2014. Mathematical language processing project. In *CICM*.

Pascanu, R.; Mikolov, T.; and Bengio, Y. 2012. On the difficulty of training recurrent neural networks. *arXiv:1211.5063*.

Roy, S.; Upadhyay, S.; and Roth, D. 2016. Equation parsing: Mapping sentences to grounded equations. In *EMNLP*.

Schubotz, M.; Grigorev, A.; Leich, M.; Cohl, H. S.; Meuschke, N.; Gipp, B.; Youssef, A. S.; and Markl, V. 2016. Semantification of identifiers in mathematics for better math information retrieval. In *SIGIR*.

Sojka, P., and Líška, M. 2011. Indexing and searching mathematics in digital libraries. In *CICM*.

Song, J.; Gan, Z.; and Carin, L. 2016. Factored temporal sigmoid belief networks for sequence learning. In *ICML*.

Srivastava, A., and Sutton, C. 2017. Autoencoding variational inference for topic models. In *ICLR*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*.

Stathopoulos, Y.; Baker, S.; Rei, M.; and Teufel, S. 2018. Variable typing: Assigning meaning to variables in mathematical text. In *NAACL*.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2005. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS*.

Wang, Y.; Gao, L.; Wang, S.; Tang, Z.; Liu, X.; and Yuan, K. 2015. Wikimirs 3.0: a hybrid mir system based on the context, structure and importance of formulae in a document. In *JCDL*.

Wang, W.; Gan, Z.; Wang, W.; Shen, D.; Huang, J.; Ping, W.; Satheesh, S.; and Carin, L. 2018. Topic compositional neural language model. In *AISTATS*.

Xie, P.; Deng, Y.; and Xing, E. 2015. Diversifying restricted boltzmann machine for document modeling. In *KDD*.

Zanibbi, R.; Davila, K.; Kane, A.; and Tompa, F. W. 2016. Multi-stage math formula search: Using appearance-based similarity metrics at scale. In *SIGIR*.