Topic Modeling with Semantic Frames

Ron Fan

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master's of Science in Computer Science and Engineering

University of Washington

2019

Reading Committee:

Richard Anderson, Chair

Noah A. Smith

Program Authorized to Offer Degree:
UW CSE Combined B.S./M.S. Program

University of Washington

Abstract

Topic Modeling with Semantic Frames

Ron Fan

Chair of the Supervisory Committee:
Professor Richard Anderson
Paul G. Allen School of Computer Science and Engineering

Topic modeling encompasses a set of techniques and algorithms for extracting common themes of meaning from a collection of documents. Common topic models neglect semantics in favor of purely statistical models. Classic examples of these statistical models, such as latent Dirichlet allocation (LDA), consider documents as bags of words, largely discarding valuable data encoded in the linguistic structure of text.

Advancements in neural methods for variational parameter inference have led to new approaches for inference of latent variables. These improvements enable convenient redefinition of components in latent representation models that were previously difficult to reformulate using bounded approximations.

In this work, we augment various supervised and unsupervised topic models with semantic frames. We focus our work on adding frames as input features that do not require deeper changes to model design. Successfully improving topic models in this way would mean we can introduce improved awareness of linguistic structure or abstraction to a model by simply plugging in frame semantics alongside the raw document.

Through our experiments, we are unable to conclude with our evaluation metrics that frame semantics included in this way are able to significantly improve topic models. However, we find qualitative indications that the inclusion of frame semantics may allow models to form topics more coherently by associating topics with frames that succinctly describe the topic.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS

I wish to express my utmost respect and appreciation to my advisor Noah Smith and mentor Swabha Swayamdipta for their exceptional guidance and support. Natural language processing is absolutely fascinating, and it was an honor to be able to learn so much about it from such brilliant minds.

I am also thankful to the University of Washington for giving me the opportunity to pursue my ambitions in a world-class environment. My time here has certainly shaped my life, and for that I am eternally grateful.

Chapter 1

INTRODUCTION

Topic models (Blei et al., 2003) are statistical models that discover latent *topics* in collections of documents. Traditionally, these algorithms model documents as bags of words (BoWs), which are counters of word frequencies per document backed by a predefined vocabulary. While this approach is computationally efficient, it unrealistically ignores word ordering, and by extension, most higher-level semantics of a document (Wallach, 2006).

In the past, formulation of topic models was limited by intractable latent variable distributions (Miao et al., 2016). However, neural variational inference techniques have enabled efficient approximate posterior inference of latent variables (Kingma and Welling, 2014; Rezende et al., 2014). This has made it easier to make modifications that introduce additional complexity, such as lexical features, to topic models.

For maximum generalizability, we consider features that can be extracted from documents independently of topic model design and are amenable to being added alongside words. Semantic frames, individual lexical units of meaning, are particularly suitable for these requirements (Fillmore and Baker, 2001). Semantic frames are typically identified in sentences using frame-semantic parsing techniques (Gildea and Jurafsky, 2002). Thanks to the increased practicality of recurrent neural networks (RNNs) in recent years, neural frame-semantic parsers now support large-scale extraction of frames from documents with state-of-the-art performance (Swayamdipta et al., 2017; Roth and Lapata, 2016; Yang and Mitchell, 2017; Swayamdipta et al., 2018).

In frame semantics theory, semantic frames associate linguistic forms, such as words and phrases, with cognitive structure - frames - which determine the interpretation of those forms (Fillmore and Baker, 2009). Intuitively, topics selected from a set of documents should address the semantic content of those topics, rather than the specific words used to convey that content. However, current topic models do not utilize linguistic structure, and instead process documents only as sequences of words, often processed further to be void of even sequential order. Using frames, we can recover a deeper level of meaning from documents without requiring significant changes to topic model designs, providing the models with information that could help latent topics be determined based on meaning, rather than solely on words.

In this work, we add automatically predicted semantic frames as input features for topic models. We use the Open-SESAME[1] parser (Swayamdipta et al., 2018), which achieves state-of-the-art results on frame and argument parsing. Through the addition of frames, we intend to give topic models higher-level understanding of the documents which they model.

Our work is exploratory: we analyze the characteristics of frame-semantic parses over a large corpus, and then evaluate several of the countless possible methods for adding frames to different topic models. Based on the evaluation metrics we consider, our experiments do not indicate that the methods we consider are useful for topic models. However, we find indications that frame semantics could help topics form more coherently by associating topics with frames that summarize them. The major contributions in this work are:

- We predict semantic frames for a large dataset (IMDB movie reviews) using a state-of-the-art frame-semantic parser and characterize the results.
- We add semantic frames to a supervised topic model by (1) modeling frames as bags of frames, and (2) learning word and frame embeddings.
- We analyze the performance of a supervised topic model trained using only frames.
- We add semantic frames to an unsupervised topic model by appending frames to documents as additional vocabulary words.

[1] https://github.com/swabhs/open-sesame

Chapter 2
RELATED WORK

2.1 Topic Models

Topic models represent documents as a mixture of a set of underlying topics, shared among a set of documents. Latent Dirichlet allocation (LDA; Blei et al., 2003), the most common predecessor of modern topic models, posits, for every document in a collection of documents, the existence of some latent representation $\boldsymbol{r}$ generated from a Dirichlet distribution. LDA models topics by defining probability distributions over the word vocabulary for a specified number of topics. The latent document representation $\boldsymbol{r}$ is generated by sampling from the Dirichlet distribution, and words are subsequently generated from the latent representation by sampling once more. With this generative story as a guide, LDA and other topic models learn the latent distributions using variational Bayes or other Bayesian inference techniques.

Various modifications to LDA exist to address its weaknesses or augment it with new knowledge. Supervised LDA (sLDA), for example, models LDA with labels being generated jointly with words from the latent representation (Blei and McAuliffe, 2007). However, current supervised topic models only support labels at the document level, and are not able to express word-level labels such as semantic frames.

2.1.1 Topic Models using Neural VAEs

Recent work using neural networks has resulted in several new approaches for approximating intractable posteriors using neural models such as deep latent Gaussian models (DLGMs) or variational auto-encoders (VAEs) (Kingma and Welling, 2014; Rezende et al., 2014). Although the mathematical theories behind these models differ greatly, their implementations as neural networks follow similar patterns. By backpropagating gradients after computing some objective function, the models learn increasingly accurate parameters to an approximation of the latent posterior. The practical elegance of these models enables graphical modeling for variational inference wherein modifications can be made without significant work in deriving new formulae for approximation.

These new variational techniques have led to the development of increasingly complex topic models such as SCHOLAR, which combines key motivating ideas from sLDA and sparse additive generative models (SAGE) into a single network by using variational inference with VAEs (Card et al., 2018; Blei and McAuliffe, 2007; Eisenstein et al., 2011).

2.2 Semantic Frames and Frame-Semantic Parsing

Semantic frames are independent units of meaning based on Charles Fillmore's theory of frame semantics (Fillmore and Baker, 2001). A semantic frame is composed of a target, frame identifier, and arguments. Lexical units are lemmatized words tagged with Part-of-Speech (POS). The target is the lexical unit which evokes the frame, and is also referred to as a frame predicate. The frame identifier is a name associated with a specific frame definition. We sometimes say "frame" to refer to the frame identifier. Arguments are other nearby phrases which contribute substantially to the meaning of the frame.

In an effort to build a lexical database of these semantic frames, the FrameNet project has produced a set of over 1,200 semantic frame definitions backed by a large corpus of frame-annotated sentences (Baker et al., 1998). By training on FrameNet data, frame-semantic parsers can predict semantic frames in sentences. Approaches for this vary, although most recent work has focused on using neural network approaches, such as softmax-margin segmental RNNs (Swayamdipta et al., 2017).

Although there as has been extensive research on using FrameNet to perform benchmarking tasks like semantic role labeling (SRL), little work has been done on evaluating FrameNet's usefulness for other tasks or domains (Hartmann et al., 2017). Our goal in this work is to explore one area of the wide space of applications for semantic frames - topic modeling - and by extension, evaluate their potential for general use in diverse natural language processing and understanding tasks.

Chapter 3

A FRAME-SEMANTIC PARSING PIPELINE FOR TEXT CORPORA

3.1 Introduction

Frame-semantic parsing techniques are effective for solving shared tasks within their learned domain. However, for semantic frames to be consistently beneficial to learning document representations, it is necessary that frame-semantic parsing perform well on text corpora from different domains. We explore the performance of frame-semantic parsing on the out-of-domain IMDB informal movie review dataset (Maas et al., 2011) through quantitative and qualitative review of various characteristics of the frame-annotated results. We find that frame-semantic parsing performs well enough to produce data that could plausibly be useful as token-level syntax labels for generic tasks, but suffers various shortcomings which may negatively affect the usefulness of semantic frames in certain contexts.

We train and evaluate the Open-SESAME parser,[1] a neural frame-semantic parser based on softmax-margin segmental recurrent neural networks that achieves state-of-the-art results on frame and argument identification (Swayamdipta et al., 2017). We train Open-SESAME with its basic settings on FrameNet annotated sentences without the additional multitask syntax objective that it supports (Swayamdipta et al., 2017).

A single sentence can have multiple frames of meaning (Baker et al., 1998). So, frame-semantic parsers such as Open-SESAME output each frame separately as distinct units, thus outputting a single sentence containing multiple frames several times. Because one word can only evoke one frame, we can collapse these distinct frames into a single sentence to treat frames as per-word labels. Words can have multiple argument labels corresponding to different frames in the sentence, and we store these as variable-length lists per word.

After frame-semantic parsing on the IMDB dataset, we observe 26,657,279 words in

[1] https://github.com/swabhs/open-sesame. Although there exist other parsers for frame-semantics (Yang and Mitchell, 2017), these have not been publicly released, and are hence not adaptable to our task.

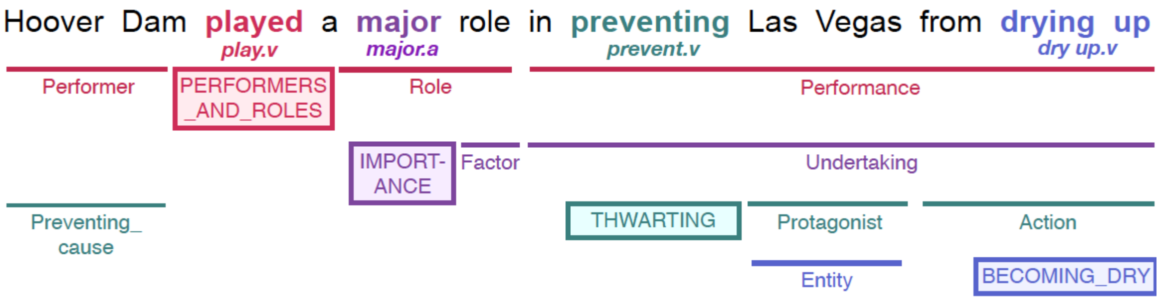

Figure 3.1: An example frame-semantic parse of a sentence with color-coded annotations of frame predicates ("targets"), frames, and arguments, from Swayamdipta et al. (2017). Frames are shown in colored blocks, and arguments are horizontal to the frame blocks.

total, with a shortest document (review) of 8 words and longest document of 2859 words. The median document length is 200 words.

3.2 Frame Predictions

In out-of-domain data, how often are words seen associated with multiple frames? The answer to this is useful because it determines feasible approaches for using frame data in models. When words are assumed to be associated with a single frame, frames can be considered simple extensions of vocabulary that naturally group semantically-similar words together. However, when words are consistently tied to different frames, we must treat them as word-level class labels.

Hereafter, we use "type" to mean a distinct word in the corpus, and "token" to mean a single instance of a type. We find that 97.88% of types across the entire corpus are associated with a single frame label, including the null frame which indicates that the type does not evoke a frame. 3.29% of types have non-null frame labels. Note that we do not enforce a vocabulary, so many types are unique and not present in FrameNet. Of these types with non-null frames, 83.16% are associated with a single frame. In total, 6,038,185 non-null frames are predicted, corresponding to 22.65% of all tokens in the corpus. This imbalance between the number of types with frames and the number of tokens with frames is expected due to the limited number of frames defined in FrameNet for uncommon types, and highlights a shortage of data that can be remedied over time.

Words with numerous frames associated with them are typically from minor lexical categories, such as conjunctions or particles (e.g. “a”, “in”, “for”), and are classified by the frame-semantic parser as parts of frame predicates. These frames often have multi-word predicates (e.g. “tend to” or “according to” where the second word is tagged as the evoker).

# Frames	**Frequency**	**Frequency (excl. NULL frames)**
1	275488 (97.88%)	7701 (75.60%)
2	4811 (1.71%)	1301 (14.05%)
3	937 (0.33%)	189 (2.04%)
4	147 (0.05%)	33 (0.36%)
≥ 5	63 (0.02%)	37 (0.40%)

Table 3.1: The frequency of certain numbers of distinct frames being associated with the same word in the IMDB dataset.

3.3 Parts of Speech

POS	**Frequency**
v	5317 (48.09%)
n	4190 (37.90%)
a	1110 (10.04%)
adv	221 (2.00%)
prep	120 (1.09%)
num	68 (0.62%)
other	30 (0.27%)

Table 3.2: The frequency of various POS tags on target lexical units in the IMDB dataset.

We count the POS tags on lexical units (LUs) identified as frame targets, finding that 86% of identified frames are verbal or nominal. Verb LUs are the most frequent predicates of frames, comprising 48% of all frames. This is promising, as verb frames generally entail a higher level of meaning through descriptions of action or state change.

3.4 Frames and Sentiment

Frame-semantic parsers perform well on shared tasks and benchmarks, but how much useful information do they accurately capture? We investigate the quality of parsed frames on the IMDB dataset with respect to a sentiment analysis task.

We evaluate the performance qualitatively by randomly sampling 20 reviews from the IMDB. We then manually go through the text of each review without looking at the label or frame and select a set of phrases which we believe are indicative of the sentiment. In situations where sentiment is not clearly attributable to one word, such as "not bad at all", we include the entire phrase. Hereafter, we refer to both single- and multi-word lexical units as phrases.

We identify 90 phrases among 20 reviews which we consider relevant for correctly determining sentiment of the document which contains them. The phrases chosen by our method include highly opinionated words such as "poor", "worst", or "amazing", as well as more nuanced phrases such as "worth checking out", "lacks any punch" or "sleepwalks through".

We search for occurrences of these phrases within the frame-annotated data. For multi-word phrases, we require that the phrase is matched word-for-word in the annotated data to consider it a repeat occurrence of the sentiment phrase we identified. We count the number of instances in which any word in a phrase is associated with a frame or argument. We calculate two sets of statistics: one including all 90 phrases, and one excluding the 63 phrases which occur more than 100 times in our observations, to better understand the ability of the frame-semantic parser to handle uncommon expressions of sentiment.

We find that approximately 63% of our selected phrases are assigned frames by the frame-semantic parser. This rate is considerably lower at 26% for low-frequency phrases, suggesting that frame-semantic parsing has trouble identifying sentiment-relevant frames for uncommon phrases. Our selected words are assigned frames considerably more frequently

Observation	**Observation Rate**
Phrase w/ frame(s)	63%
Phrase w/ arg(s)	59%
Low-freq. phrase w/ frame(s)	26%
Low-freq. phrase w/ arg(s)	60%

Table 3.3: Probabilities of different types of labels occurring in the dataset. Low-frequency phrases are observed ≤ 100 times.

than the corpus-wide average rate of 3.29% found in §3.2.

We look more closely at the frame-semantic parser output to further understand the capabilities of the parser. Table 3.4 presents 7 example instances of phrases to illustrate the types of limitations we discover. Among the examples we present, we observe the following classes of mistakes:

- A phrase is assigned an inaccurate frame, and a more accurate frame...
    - is not available (e.g. “best” is only available as Required_event in FrameNet)
    - is available (e.g. “poor” should be Desirability instead of Wealthiness)
- A phrase is not assigned a frame...
    - but could be assigned an accurate frame (e.g. “terrible” can be Desirability)
    - and is not in FrameNet at all (e.g. “worst” is not in FrameNet)
- A multi-word phrase is not accurately labeled (e.g. “great job” is assigned Being_employed instead of Desirability)

From this qualitative analysis, we tentatively conclude that while frame-semantic parsing is empirically performant, it still makes a variety of mistakes in practice. These shortcomings are attributable to both frame-semantic parsing models and FrameNet. Models make mistakes in identification, and FrameNet lacks useful frames for many phrases. In the case of FrameNet, these shortcomings have a clear remedy through the collection of more annotations, which is already underway as the project is a constant work in progress.

Phrase	**LU**	**Frame**	**Correct**	**Preferred Frame**
It is **well-made** thriller with a **talented** cast and credible situations .				
well-made	-	-	-	N/A
talented	-	-	-	N/A
Jesus , he did a **great job** in this film !				
great	-	-	-	Desirability
job	job.n	Being_employed	✓	-
This remains the **best** film made about Custer .				
best	best.v	Required_event	-	N/A
Really **terrible** and I felt like I needed a bath .				
terrible	-	-	✓	Desirability
This is truly , without exaggerating , one of the **worst** Slasher movies ever made .				
worst	-	-	-	N/A
The acting was theatrical and the sound and picture quality was extremely **poor** .				
poor	poor.a	Wealthiness	-	Desirability

Table 3.4: Examples of sentiment phrases, their contexts, and their labeled results. 'LU' is the lexical unit assigned to each word. 'Frame' is the frame predicted for the LU. 'Correct' is whether we decide that the frame is semantically accurate in the context of the sentence in which the phrase was used. 'Preferred Frame' is another frame in FrameNet associated with the same same that we consider more accurate in context. N/A indicates there is no alternative frame defined in FrameNet that is more suitable for the word.

Chapter 4

FRAMES FOR SUPERVISED TOPIC MODELING

4.1 Introduction

Supervised topic models learn latent distributions for topics and documents while optimizing to generate some document-level label. In our neural supervised topic model, we perform approximate posterior inference using a variational auto-encoder (Rezende et al., 2014; Kingma and Welling, 2014). The parameters we infer help define latent representations for the documents in the corpus. For our experiments, we only consider frames and not frame arguments.

We explore a highly generalizable approach of adding frames to models at the input level. Our methods modify inputs to topic models from bags of words into alternative forms which include frame information. We explore two categories of approaches to adding frames in this way: bags of frames, and frame embeddings.

4.2 Models

For our supervised experiments, we use SCHOLAR (Card et al., 2018) as a baseline topic model. While SCHOLAR is innovative in its support for document metadata in the form of covariates and labels, we only use labels and do not consider any covariates.

In the generative story of SCHOLAR, words in some document i are generated from a latent representation η_i. This latent representation of the distribution over vocabulary words in document i is a deterministic transformation of $\boldsymbol{r}$, the latent space of the VAE; and $\mathbf{B}$, the matrix of topics and their corresponding distributions over the vocabulary. $\boldsymbol{r}$ also generates a label after an intermediary steps applies a softmax function to produce a θ_i generating the label y_i.

We can infer an variational approximation to the posterior distribution of $\boldsymbol{r}$, $q_\Phi(r_i \mid w_i, c_i, y_i)$, where Φ consists of the parameters of the fully-connected layers used to calculate

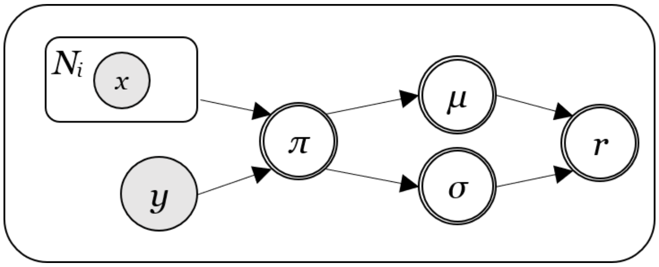

Figure 4.1: Inference model in SCHOLAR.

the variational parameters of the distribution, mean and log variance (i.e. μ_i and $\log \sigma_i^2$). Mean and log variance are computed as linear transformations of an encoded representation π_i of the document's N_i words, x_i, and sentiment label, y_i. For a multilayer perceptron encoder f_e, a BoW document representation x_i, and embedding weights W_x,

$$\pi_i = f_e([\mathbf{W}_x \boldsymbol{x}_i; \boldsymbol{y}_i]) \tag{4.1}$$

After computing a sampling-based approximation of $\boldsymbol{r}$, the SCHOLAR model generates x and y and calculates reconstruction and label loss, which are used to learn more accurate latent parameters.

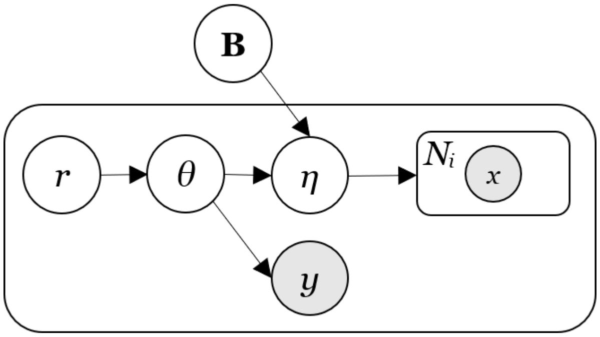

Figure 4.2: Generative story in SCHOLAR.

4.2.1 Bag of Frames

We consider modeling frames similarly to bags of words, as "bags of frames" (BoFs)—counters of the number of time each frame occurs in a document. We create a separate vocabulary for frames from the full set of frames in the FrameNet dataset. This vocabulary is used to create bags of frames from documents.

Our implementation of the BoF model learns a BoF encoder and concatenates its encodings to the encoded representation of the document's words and label. In this setup, we create a new encoder for frames with weights, $\mathbf{W}_f$. Let $\boldsymbol{F}_i$ be the BoF vector for the frames of document i. We can then calculate π_i as

$$\pi_i = f_e([\mathbf{W}_x \boldsymbol{x}_i; \mathbf{W}_f \boldsymbol{F}_i; \boldsymbol{y}_i]) \tag{4.2}$$

Backpropagation occurs through the inclusion of π_i, as a component of η_i, in the reconstruction of x and computation of KL divergence for the loss function. We call this version of the model SCHOLAR + BoF.

4.2.2 Token Embeddings

We experiment with a new approach to representing a document's words and frames for topic modeling. Instead of using bags of words and frames, we learn token-level embeddings for each vocabulary word and frame. We redefine $\mathbf{x}_i$ as the sequence of words and $\mathbf{F}_i$ as the sequence of frames in document i.

We begin by padding each document to a maximum sequence length. We replace the model's BoW inputs with sequences of word tokens. The model then learns a $V \times e$ token embedder $\mathbf{E}_x$ for word embeddings of size e. We convert the sequence of embeddings to a single vector for the MLP in π_i by summing the embeddings and appending a representation of the label, as before. We refer to this frame-less word embedding baseline as SCHOLAR$_e$ → Sum, a sum of word embeddings.

$$\pi_i = f_e\left(\left[\sum_{j=1}^{N_i}[\mathbf{E}_x(\mathbf{x}_{ij})]; y_i\right]\right) \tag{4.3}$$

Sum of Word + Frame Embeddings

We learn frame embeddings through a similar procedure, feeding in a sequence of frames and converting them to embeddings in a separate $V_f \times e$ embedder $\mathbf{E}_f$. In the inference step of the model, we redefine π_i with a sum of word embeddings concatenated with frame embeddings, and a label. We refer to this version of the model as SCHOLAR$_e$; F$_e$ → Sum, a sum of word embeddings concatenated with frame embeddings.

$$\pi_i = f_e\left(\left[\sum_{j=1}^{N_i}[\mathbf{E}_x(\mathbf{x}_{ij}); \mathbf{E}_f(\mathbf{F}_{ij})]; y_i\right]\right) \tag{4.4}$$

Fully-Connected Layer

We also experiment with using a large fully-connected layer to transform a full sequence of embeddings to a single vector. Let the maximum padded sequence length be N_i and the concatenated length of the word and frame embeddings be $2e$. We create a new $(N_i * 2e) \times e$ learned weight matrix $\mathbf{W}_e$ and flatten the sequential embedding matrix to a new $N_i \times 2e$ matrix $\mathbf{D}$ defined such that row j of $\mathbf{D}$ is defined as:

$$\mathbf{D}_j = \mathbf{E}_x(\mathbf{x}_{ij}); \mathbf{E}_f(\mathbf{F}_{ij}) \tag{4.5}$$

We then redefine π_i with a concatenation of the new document encoding and the label. We refer to this variation of the model as SCHOLAR$_e$;F_e → FC.

$$\pi_i = f_e\left([\mathbf{W}_e\mathbf{D}; \mathbf{W}_y yi]\right) \tag{4.6}$$

4.3 Evaluation

We train the above models on the IMDB movie review dataset with a training corpus of 50,000 labeled reviews. We evaluate based on generalization to held-out data, measured in perplexity, and accuracy in predicting binary sentiment labels.

4.4 Experimental Results

We configure the models with 50 topics and a 5000 word vocabulary, and use the text preprocessing techniques in the released SCHOLAR codebase[1] to better compare to previously published results from Card et al. (2018). We present our results in Table 4.1.

Word Alignment: Open-SESAME uses a different preprocessor for text tokenizing before target (predicate) identification. In order to align the two data formats, we create a heuristic frame assigner which matches SCHOLAR tokens with their respective tokens in Open-SESAME's output. Our frame assigner accounts for a variety of differences between the two formats, such as stop word removal, contraction tokenizing, and character filters.

Model	**Ppl.** ↓	**Acc.** ↑
SCHOLAR *	1857	85.69
SCHOLAR + BoF	1885	83.28
SCHOLAR$_e$ → Sum *	1859	85.43
SCHOLAR$_e$; F$_e$ → Sum	1855	84.09
SCHOLAR$_e$; F$_e$ → FC	**1829**	**85.86**

Table 4.1: Performance of various models in a supervised setting on the IMDB dataset with review sentiment as a binary label. * indicates a baseline frame-less model.

Our results do not suggest that frames provide a significant benefit to our models for supervised topic modeling. Although one of the frame models achieves the best perplexity and accuracy scores, the marginal improvement is relatively minor. Pending more rigorous testing of random seeds (we try 5 seeds), this is not a confident indicator of real improvement due to high variance in generative models resulting from random seed sampling, with accuracy varying within a range of ±2% in our experiments (Zhao et al., 2015).

Based on our qualitative analysis of frame-semantic parsing on the IMDB in §3.4, we

[1] `https://github.com/dallascard/scholar`

hypothesize that the lack of improvement in accuracy is due to a lack of sentiment-relevant words being assigned frames. We also note that perplexity of all models roughly lies within the same range, which we believe can be attributed to frames being abstract enough to have minimal influence on word-level model understanding.

4.5 Frames Only

The experiments we run in supervised topic modeling fail to show improvements based on our evaluation metrics. This raises the question of whether frames provide any form of useful information to models.

In order to investigate this concern, we briefly explore an alternative model which only trains on frame data. The model design is identical to the baseline SCHOLAR model we use. However, instead of standard word tokens, we represent documents only as frames and treat frame identifiers as tokens, ignoring individual words.

We present the accuracy of this model in Table 4.2 The model performs well above the threshold of random noise, indicating that frames provide some level of useful and actionable information that can inform topic models. The model also performs notably worse than the full model with words. This suggests that frames are not fully redundant with words and provide different information, although it is not clear if this information is a subset of information carried by words alone.

Dataset	**Acc.** ↑
Train	72.87
Test	70.59

Table 4.2: Performance of SCHOLAR with only frames and no words.

We share topics learned by the frame-only model in Table 4.3. Additionally, we provide for reference various topics learned by the baseline word-only model in Table 4.4. We refer to these as “frame topics” and “word topics”, respectively. In the context of movie reviews, topics frequently describe specific genres, tropes, or scenarios.

Criminal Trial	**Competition**	**Rebellion**
Criminal_investigation	Evaluative_comparison	Member_of_military
Trial	Retaining	Hostile_encounter
Appellations	Beat_opponent	Military
Economy	Cause_to_make_noise	Change_of_leadership
Verdict	Alliance	Irregular_combatants
Committing_crime	Finish_competition	Rebellion
Racing	**Animals**	**Art**
Vehicle	Animals	Create_physical_artwork
Team	Hunting	Social_event
Cause_impact	Natural_features	Amalgamation
Relational_natural_features	Biological_area	Fields
Ride_vehicle	Have_associated	Ineffability
Endeavor_failure	Catching_fire	Performing_arts

Table 4.3: Top 6 words from 6 selected topics from a supervised VAE trained on only frame tokens. The topic names have been manually annotated for ease of reading.

World War II	**Comedy**	**Heist**	**Martial Arts**	**Disney**
soldiers	funniest	security	martial	cried
germany	eddie	car	kung	cry
soldier	murphy	truck	arts	animals
troops	laughed	bank	hong	adorable
army	academy	chases	chan	kids
nazi	oscars	chase	jackie	animation

Table 4.4: Top 6 words from 5 selected topics from a supervised VAE trained on both words and frames. The topic names have been manually annotated for ease of reading.

We observe that word topics tend toward choosing nouns associated with each other as topics, and can often be viewed as a collection of things frequently observed in a particular scenario. These topics can contain highly-specific words that indicate the theme of the topic but are too specific to be applicable to every document associated with that theme. For example, “eddie” and “murphy” are words associated with comedy, but the actor Eddie Murphy is not in every comedic film.

Frame topics are comparatively more descriptive and less specific. Verb frames are more commonly chosen than verb words, leading to topics that are more capable of describing the actual scenario rather than identifying parts of it. Consider our examples of the “Rebellion” frame topic and “World War II” word topic which address similar themes. While “World War II” simply lists things that were part of the war, “Rebellion” tells a much more descriptive story of a military coup involving a change of leadership. Overall, this suggests that frames are capable of giving topic models access to deeper meaning in documents.

4.6 LSTM without VAE

When dealing with sequential data, the LSTM (Hochreiter and Schmidhuber, 1997) is a common choice due to its respectable performance on a variety of sequential-input tasks. We briefly explore a basic supervised sentiment classification model with no latent topic representation for a cursory evaluation of the usefulness of frames in other tasks and models.

The baseline model we compare with is a 2-layer bidirectional LSTM with learned token embeddings. To add frames, we use learned frame embeddings. We concatenate token and frame embeddings across the entire sequence before passing the sequence of concatenated embeddings to the LSTM.

Model	**Acc. ↑**
LSTM Baseline	**87.26**
LSTM w/ Frames	86.95

Table 4.5: Performance of a simple LSTM text classifier with and without frames.

We run both models on the IMDB dataset with the binary sentiment classification task, only evaluating on accuracy. The results of this experiment are presented in Table 4.5. Note that these results are not directly comparable to the ones in Table 4.1 due to differences in experimental setup with regards to preprocessing and hyperparameters. Again, we find that our model with frames is unable to outperform the baseline.

Chapter 5

FRAMES FOR UNSUPERVISED TOPIC MODELING

5.1 Introduction

Unsupervised topic models are helpful for learning useful representations of documents that can serve as features for other tasks. Given an unlabeled text corpus, we assume a latent representation $\boldsymbol{r}$ that generates the documents in the corpus. Unlike in supervised topic models, we use no labels and instead calculate loss solely from reconstruction of tokens.

5.2 Baseline Model

For our baseline model, we train a single VAE on BoW representations of documents. The VAE learns parameters μ and σ for a logistic normal prior, $p(\boldsymbol{r})$, using feed-forward networks.

The model first generates an embedding for a bag of words. It then passes this embedding to the VAE's feed-forward encoder, which creates an encoded representation π. π randomly samples a latent representation $\boldsymbol{r}$ by estimating μ and σ and setting $\boldsymbol{r} = \mu + \sigma * \epsilon$, where ϵ is random noise generated from a normal distribution.

The latent representation $\boldsymbol{r}$ is run through a decoder to reconstruct the inputs. We compute reconstruction loss and KL divergence from $\boldsymbol{r}$ and sum them to get the evidence lower bound (ELBO) shown in Equation 5.1. We call this baseline an unsupervised VAE.

$$\mathcal{L}(w_i) \approx \sum_{j=1}^{N_i} \log p(w_{ij} \mid r_i^{(s)}) - \mathrm{D}_{\mathrm{KL}}[q_\Phi(r_i \mid w_i) \| p(r_i \mid \alpha)] \tag{5.1}$$

5.3 Appending Frames

Frames in a document can be considered a high-level description of events of that document. Thus, instead of treating frames separately from words, we can assume they are the same type of input and append them to documents as additional tokens from a disjoint vocabulary.

We retain the original vocabulary used by our baseline, and extend it with the set of possible frames. To distinguish between one-word frames and tokens, we prefix frames with an `@` character. Thus, each document in the corpus is modified into a sequence of plain words followed by frame names prefixed with @ characters. We refer to this model as an unsupervised VAE with frame tokens.

5.4 Evaluation

Since these models are unsupervised, we choose to optimize based on coherence, evaluated with normalized pointwise mutual information (NPMI; Bouma, 2009). Note that the inclusion of extra frame words into vocabulary of the model renders our evaluation metrics incomparable to some extent. However, as the reference vocabulary we use to compute NPMI is constant, and our frame vocabulary is prefixed to ensure it does not overlap held-out vocabulary, there is some level of consistency in NPMI computation that allows for non-rigorous comparison between models.

5.5 Experimental Results

We present the results of our experiments in Table 5.1. We are unable to find evidence that our model of including frames into an unsupervised topic model provides measurable improvements by our metrics for unsupervised evaluation.

Model	**Ppl. ↓**	**NPMI ↑**
Unsup. VAE	**784**	**0.10**
Unsup. VAE w/ Frame Tokens	849	0.09

Table 5.1: Unsupervised VAE performance

We present selected topics from the frame model in Table 5.2 as potentially interesting results which suggest potential improvements pending further exploration of model designs. The frames chosen by these topics tend to be qualitatively good summarizations of a topic,

marriage	**prison**	**kinship**
husband	@prison	@kinship
@forming_relationships	prison	father
@personal_relationship	documentary	kids
wife	@point_of_dispute	city
marriage	issues	children

good reviews	**aesthetics**	**martial arts**
liked	work	@education_teaching
@commerce_buy	beautiful	bruce
@reading_activity	@aesthetics	martial
read	films	training
@part_whole	amazing	arts

Table 5.2: Top 5 words from 6 selected topics in the latent representation of an unsupervised VAE with frame tokens. The topic names have been manually annotated for ease of reading.

and they often encompass ideas which cannot be concisely expressed with a single topic word. Compared to the frame-only topics we present in §4.5, these topics strike a balance between specificity and broadness by selecting frames for topics when particularly appropriate. This suggests frames may help topic models model topics that are more interpretable, although the evaluation techniques we currently use do not pick up on these benefits.

Chapter 6
CONCLUSION

Do semantic frames improve topic models? Based on the experiments we conducted with in this work, we are unable to conclude that they do. Our experiments in adding frames as inputs using bags of frames, frame embeddings, and vocabulary extensions do not lead to improvements in the evaluation metrics we observe. However, we acknowledge that these metrics may not be complete in uncovering the effects of adding frames. Additionally, we find evidence supporting the hypothesis that frames provide a more meaningful class of information not captured by words.

We run a large out-of-domain corpus through a state-of-the-art frame-semantic parser and analyze its results to qualitatively evaluate the usefulness of FrameNet semantic frames for tasks where semantic labeling is not the primary goal. We find that while broad performance is quite decent, frame-semantic parsers still make several types of mistakes which negatively impact models relying on frame-semantic parses as features.

With our frame-semantic parses, we attempt to improve topic models by adding frames as input features. We first add frames to supervised topic models as bags of frames. In this formulation, we create a frame vocabulary and convert frame labels to bags of frames for each document. We then encode the BoFs and append the encoding to the BoW encoding to form the document representation.

Next, we experiment with intermediary token embeddings. We sum the embeddings in order to combine a sequence of word embeddings to a single vector. We learn separate embeddings for words and frames, and concatenate word and frame embeddings before summing them. We also experiment with combining the sequence of embeddings to a single representation using a feed-forward fully-connected neural network. In our experiments, this results in the best performance, but by a margin small enough to be considered negligible, especially in the context of the generally high variance in these models.

Using our supervised model, we train an alternate model with only frames and no words, and evaluate the latent topics it models as well as its performance on the downstream task of sentiment classification. We compare “frame topics” with “word topics” and find high-level differences in the form of content conveyed by the two types of topics.

Finally, we explore using frames in an unsupervised topic model. We try another approach, appending frames to documents as extra tokens. This changes the vocabulary, which affects comparability with our baseline. To mitigate the confounding effect, we retain the same word vocabulary and use the frame vocabulary as a strict extension to the vocabulary with a guarantee of no overlap with unseen words. We also evaluate the model’s topics, which combine frame and word vocabularies. From manual inspection of topics generated by this model, we find signs of increased topic interpretability due to selected frames adding higher scopes of meaning to topics.

The evaluation metrics we choose do not suggest that our methods of adding semantic frames to topic models are beneficial. However, although there is no significant performance improvement, adding frames to a model does not appear to degrade its performance. Evaluation of topic model performance is not a clearly solved problem, and we note that the standard evaluation metrics we use may not be entirely suitable for measuring the types of improvements we seek. While perplexity and NPMI are commonly used in literature as acceptably informative indicators, they are not definitive measurements of coherence and interpretability. Sentiment classification accuracy is similarly not necessarily a good indicator of topic modeling strength, especially given our finding that sentiment-related words are often not identified as frames. Adding frames to models worsens this challenge of precise evaluation by invalidating comparisons with baseline metrics in some designs.

With the limitations of our evaluation metrics in mind, it is possible that the models we explore in this work could be useful in different contexts with datasets and downstream tasks we have not considered. It may also be the case that other types of word-level labels on text, such as leaf nodes in constituency parse trees or semantic role labels, could be more useful than semantic frames when added in similar ways.

Our results are unfortunately largely inconclusive. While we are unable to definitively point to improvements from using frame semantics in models, we also do not find evidence that using frames strictly hinders a model. The flexibility of neural topic models suggests a variety of possible modifications which we have not explored in this work; one potential alternative is a more integrated approach in which the model generates frames from latent document representations and attempts to reconstruct frames during optimization.

Although our evaluation metrics do not indicate significant improvements, we find that models are able to incorporate frames into topics to some extent, and that frames in topics often provide descriptive interpretations. This finding, in conjunction with the variety of methods for incorporating semantic frames yet to be considered, and the room for improvement in frame-semantic parsers, which have been the focus of recent research, gives us hope that semantic frames may be used in future work to create stronger topic models capable of understanding linguistic structure.

BIBLIOGRAPHY

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *COLING-ACL*.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *NIPS*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. In *NIPS*.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference 2009*.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural models for documents with metadata. In *ACL*.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *ICML*.

Charles J. Fillmore and Collin F. Baker. 2001. Frame semantics for text understanding.

Charles J. Fillmore and Collin F. Baker. 2009. A frames approach to semantic analysis.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28:245–288.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain framenet semantic role labeling. In *EACL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *ICML*.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*.

Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. In *Proc. of ACL*, pages 1192–1202, Berlin, Germany. Association for Computational Linguistics.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *CoRR*, abs/1706.09528.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proc. of EMNLP*, pages 3772–3782.

Hanna M. Wallach. 2006. Topic modeling: beyond bag-of-words. In *ICML*.

Bishan Yang and Tom Mitchell. 2017. A joint sequential and relational model for frame-semantic parsing. In *Proc. of EMNLP*.

Weizhong Zhao, James J. Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC Bioinformatics*.