ASTM: An Attentional Segmentation based Topic Model for Short Texts

Jiamiao Wang*, Ling Chen†, Lu Qin† and Xindong Wu‡
*School of Computer Science and Information Engineering, Hefei University of Technology, China
†Centre for Artificial Intelligence, University of Technology Sydney, Australia
‡MiningLamp Software Systems, Haidian District, Beijing 100084, China
Email: wjmzjx@163.com, Ling.Chen@uts.edu.au, Lu.Qin@uts.edu.au, wuxindong@mininglamp.com

***Abstract*—To address the data sparsity problem in short text understanding, various alternative topic models leveraging word embeddings as background knowledge have been developed recently. However, existing models combine auxiliary information and topic modeling in a straightforward way without considering human reading habits. In contrast, extensive studies have proven that it is full of potential in textual analysis by taking into account human attention. Therefore, we propose a novel model, Attentional Segmentation based Topic Model (ASTM), to integrate both word embeddings as supplementary information and an attention mechanism that segments short text documents into fragments of adjacent words receiving similar attention. Each segment is assigned to a topic and each document can have multiple topics. We evaluate the performance of our model on three real-world short text datasets. The experimental results demonstrate that our model outperforms the state-of-the-art in terms of both topic coherence and text classification.**

***Index Terms*—topic model, word embedding, topic embedding, short texts , human attention**

I. Introduction

Topic modeling as an unsupervised learning approach represents documents via latent topics. It has become a prevalent technique for various tasks on unstructured texts, such as document classification, information retrieval and content-based recommendation. However, conventional topic models, such as LDA [1], are not competent enough in handling short texts because of the data sparsity in terms of word co-occurrences. Inspired by the observation that human beings understand short texts well by bringing background knowledge, a number of models using distributed representation technologies (i.e., [2]–[8]) have been proposed recently.

Existing topic models of distributed representations can be generally grouped into two categories: those with background knowledge learned internally [7], [8] and those using externally trained background information [2]–[6]. The former jointly learns topics and word embeddings in a collaborative manner. Joint learning allows to acquire word embeddings that are strongly related with latent topics. However, methods of this category attempt to extract more information from the given data through different angles (e.g., topics representing the global context and word vectors representing the local context). No external or new knowledge is introduced to facilitate topic modelling from sparse short texts. The latter uses the pre-trained word embeddings as background knowledge to enrich semantics in topic modeling. The obvious advantage is that the supplementary information comes from a big extra corpus so that the knowledge could be more comprehensive. Nevertheless, the prevalent word embedding methods, like Word2vec [9] and GloVe [10], use only one vector to represent each word. The resulted embeddings cannot address the issues of homonymy and polysemy. Therefore, the mechanism that combines topic modeling and word embedding is particularly important because the word embeddings trained on external corpus may not represent the exact semantics of the words in the given data. The straightforward combination mechanisms of existing models fail to enrich information without including noises.

Recently, motivated by the observation that people usually assemble text into segments assigned with different attention levels while reading documents, approaches introducing attention mechanisms into text analysis have been proposed and demonstrated to be promising. For example, Recurrent Attentional Topic Model (RATM) [11] studies the influence of *reading order* at sentence level, and employs a sentence-based attention mechanism to acquire the influence from successive sentences in topic modeling. Wang et al. [12] take *reading time* as the attention weights, where the reading time is measured by POS (part of speech) and the carried information. However, existing models cannot be applied directly to short texts. A document in short text datasets may have only one sentence, which renders RATM incapable because RATM models sentence level attention. The model [12] is also not applicable because it is designed for sentence representation, rather than topic modeling.

Considering the limitations of existing topic models using distributed representation technologies and models integrating attention mechanisms, we propose in this paper an Attentional Segmentation based Topic Model (ASTM) for short texts by taking into account human reading habits. When people read documents, especially short texts, they usually unconsciously do two things: one is to attach background knowledge and the other is to segment a document into relatively coherent fragments based on semantics and positions. For example, the short text "*japan radiation fear sparks south korea diaper rash*" may be segmented into the following fragments when people try to comprehend the meaning, *japan radiation fear*, *south korea*, and *diaper rash*. Clearly, different parts have different importance in expressing different topics of

the document. Because of the different importance, the parts will receive different human attention. We thus develop an attention mechanism to segment a document into different parts according to attentional signals received by individual words. Since the attentional signals represent the importance of words in respective topics, we obtain them from word embeddings and topic embeddings. Generally, ASTM models topics from short texts through the following two steps. Firstly, we use an *attentional LDA* to learn initial topic embeddings to avoid cold start, where pre-trained word embeddings are integrated. Secondly, ASTM puts words receiving similar attentional signals into same segmentations and assigns a topic to all words in each segmentation in order to obtain more coherent topics.

The main contributions of our work are summarized as follows.

- To the best of our knowledge, ASTM is the first effort that aims to integrate both human reading habits, adding background knowledge and paying different attention to segmentations, to discover topics from short texts.
- By deriving topic embeddings and also word attentional signals using both externally pre-trained word embeddings and internally modelled word topic distributions, ASTM is able to enjoy the benefits brought by the auxiliary information and capture the exact semantics of the words in the given dataset at the same time.
- By assuming each segmentation belongs to only one topic, the data sparsity of short texts can be relieved. Meanwhile, each document can have multiple topics, determined by its segmentations.
- Our extensive experiments on real world data sets show that ASTM outperforms the state-of-the-art in terms of both topic coherence and document classification accuracy.

The remainder of this paper is organized as follows. We discuss related works in Section II to provide background of this research. Section III presents the technical details of ASTM. Section IV shows the results of comparative experiments that we have done to evaluate the performance of ASTM. Finally, we summarize our work in SectionV.

II. Related Work

A. Topic Models for Short Texts

Topic modeling from short texts is handicapped by a poverty of co-occurrence information. Hence, early works on short texts focus on making full use of word co-occurrences to solve the problem of data sparsity. For instance, the Dirichlet Multinomial Mixture model (DMM) [13] assumes that a short text is related to only one topic. That is, all words in each document are generated from the same topic distribution. The Biterm Topic Model (BTM) [14] aggregates document-level co-occurrences and extracts topics by modeling the word pairs in the whole corpus. The Self-Aggregation based Topic Model (SATM) [15] clusters short texts into long pseudo-texts based on semantic similarity before applying the topic model. Hesam et al. [16] focus on topical coherence and propose a LDA-based topic model by introducing segmentation. In their model, the topic assigned to a word in a segmentation is extracted from document specific topic distributions or segmentation specific topic distributions.

B. Topic Models with Word Embeddings

The technology of word representation [9], [10], [17] has become increasingly mature. It gives rise to an emerging demand to improve the quality of topic detection with word embeddings. As discussed in the previous section, existing approaches can be divided into two categories: 1) those based on externally pre-trained embeddings, and 2) those based on jointly learned embeddings.

Models in the first category adopt a two-step setting, where word embeddings are pre-trained from external corpora, and a combination mechanism is used to integrate the topic model with word embeddings. For example, LFTM [2] combines the two techniques by generating words from a two-component mixture of a topic-word component and a latent feature component. GPUDMM [3] promotes the co-occurrence of words with related semantics under the same topic based on the generalized Pólya urn (GPU) model, where semantic relevance between words is obtained by computing similarity of word embeddings. Xun et al. [6] treat each document as a Gaussian topic over word embeddings, where the topic is expressed as a multivariate Gaussian distribution instead of a multinomial distribution. Embedding-based Topic Model (ETM) [5] uses word embeddings to aggregate short texts into long pseudo-texts, and then discovers topics from pseudo-texts by a Markov Random Field regularized model.

Methods in the second category simultaneously learn topics and word embeddings to achieve mutual reinforcement. For example, Collaborative Language Model (CLM) [8] assumes that the weight of a word under a topic is proportional to the inner product of the word embedding and the topic embedding, and learns topics and word embeddings by matrix factorization at the same time. Skip-gram Topical word Embedding (STE) [7] predicts surrounding words for a given word by taking into account the topic distribution in order to learn topic-specific word embeddings, and learns topics based on the topic assignments of surrounding words.

We have discussed the strengths and weaknesses for models of the two categories respectively in the previous section. Basically, models using externally pre-trained word embeddings are able to enrich semantics of short texts by adding supplementary information. However, noise may be introduced at the same time if the combination mechanism is designed carefully. Models relying on internally and jointly learned word embeddings may acquire more relevant embeddings. Nevertheless, the potential of the models exerting to extract more knowledge from the given dataset may be limited.

C. Attention-based Topic Models

A wide range of applications have engaged attention mechanisms in textual analysis, including machine translation [18]–

TABLE I: A list of variable notations

Notation	Meaning
K	the number topics
V	the number of words
D	the total number of documents in a dataset
S	the total number of segmentations in a dataset
Seg	all segmentations in a dataset
N_s	the number of words in segmentation s
α, β	the two hyperparameters for φ, θ
φ_k	the topic-word distribution of topic k
θ_d	the document-topic distribution of document d
a_w^k	the attention weight of word w under topic k
ω_w	the pre-trained word embedding of word w
τ_k	the kth topic embedding
$\mathbb{S}_{out,w}$	a binary indicator that determines whether a word w is similar to a word w_{out}
V_i	the K-length attentional signal vector of the ith word
n_k	the number of segmentations assigned to topic k
$n_{w\|k}$	the number of word w assigned to topic k
$n_{*\|k}$	the total number of words assigned to topic k
$n^{\neg s}, n^{\neg d}$	document d or segmentation s is excluded

[20], document summarization [21]–[23], textual entailment [24], where tremendous gains have been achieved. However, very limited research has been done to introduce attention mechanisms to topic modeling. Recurrent Attentional Topic Model (RATM) [11] uses the attention mechanism to capture the relevance among successive sentences during Bayesian process. Wang al et. [12] derive attentional signals from human reading time and the amount of information delivered by words and POS tags. The two models are not applicable to topic modeling from short texts.

III. Attentional Segmentation Topic Model

In this section, we present the technical details of ASTM, from model description to estimation and parameter inference. Notations used in this paper are listed in Table I.

A. Model Description & Generative Process

There are two phases in our method. In the first phase, an attentional LDA is used to learn initial topic embeddings $\boldsymbol{\tau}$ to avoid cold start. In the second phase, we compute attentional signals received by individual words using both topic embeddings and word embeddings. The derived attentional signals of each word thus represent the importance of the word in the corresponding topic. Next, we group words receiving similar attentional signals into same segmentations. Finally, topics are assigned to segmentations.

Phase 1: Fig. 1 shows the attentional LDA used to learn initial topic embeddings $\boldsymbol{\tau}$. The generative process is as follows:

1. For each topic $k \in \{1, 2, \cdots, K\}$, draw $\varphi_k \sim Dir(\beta)$;
2. For each document $d \in \{1, 2, \cdots, D\}$:
   a) Draw a topic distribution $\theta_d \sim Dir(\alpha)$;
   b) For each word w in document d:
      i) Draw a topic $z_n \sim Multi(\theta_d)$;
      ii) Calculate the attentional signal a_{z_n};
      iii) Draw a word $w_n \sim Multi(a_{z_n} \cdot \varphi_{z_n})$.

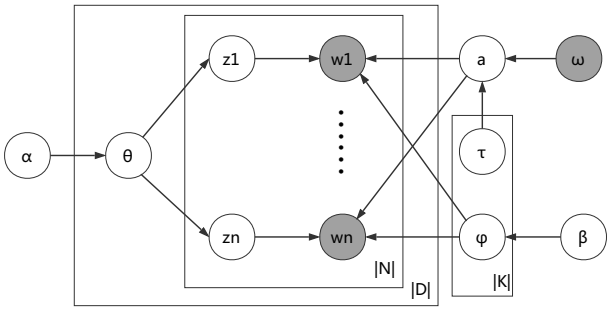


Fig. 1: Phase 1: Attentional LDA.

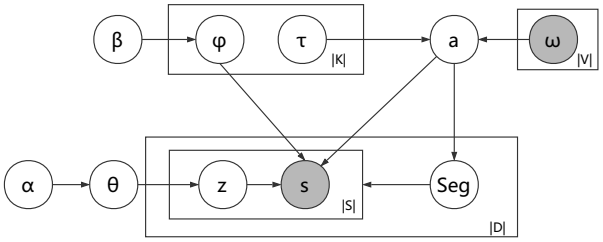


Fig. 2: Phase 2: Attentional Segmentation based Topic Model.

The algorithmic details of Phase 1 will be discussed in the next subsection. Basically, when sampling a topic for a word, the attentional signals representing the importance of the word in corresponding topics will be taken into account. The output of the phase is the initialized topic embeddings.

Phase 2: Fig. 2 shows the graphical model of ASTM.

1. Draw a topic distribution $\theta \sim Dir(\alpha)$ for the whole document collection;
2. For each topic $k \in \{1, 2, \cdots, K\}$, draw $\varphi_k \sim Dir(\beta)$;
3. For each document $d \in \{1, 2, \cdots, D\}$:
   a) Calculate segmentations Seg according to attentional signals a;
   b) For each segmentation $s \in Seg$:
      i) Draw a topic $z_s \sim Multi(\theta)$ for all words in the segmentation s;
      ii) For each position n in s: draw a word $w \sim Multi(\varphi_{z_s})$;

Phase 2 models topics for segmentations, which are iteratively refined based on updated attentional signals received by words.

B. Algorithms

The pseudo codes of ASTM are presented in Algorithm 1.

At first, a vanilla LDA is used to initialize the topic assignments (Line 1). In Lines 2-3, we design a simple strategy to deal with words without word embeddings. Existing models directly remove those words. However, it may deteriorate performance on a dataset with a lot number of such words (e.g.,

a social media data set). For instance, we have found there are 105 words without word embeddings in the Oscars dataset [1], which are mainly hashtags and neologisms (e.g.,"oscarsfail" and "envelopgate"). Such words are invented all the time, leaving the pre-trained word embeddings easily outdated. To address this issue, given a word without pre-trained word embedding w_{out}, we take the average word embeddings of words similar to w_{out} in the topic space as an approximate embedding of w_{out}. A nonparametric probabilistic sampling strategy is proposed as follows:

$$\mathbb{S}_{out,w} \sim Bernoulli\Big(CosSim(p_w, p_{out})\Big) \tag{1}$$

$$p(z=k|w) = \frac{p(w|z=k)p(z=k)}{\sum_{q=1}^{K} p(w|z=q)p(z=q)} \tag{2}$$

$$\omega_{out} = \frac{\sum_{w=1}^{V} \mathbb{S}_{out,w} \cdot \omega_w}{\sum_{w=1}^{V} \mathbb{S}_{out,w}} \tag{3}$$

In Eq. 1, a binary variable $S_{out,w}$ is sampled from a Bernoulli distribution to indicate whether the embedding of a word w will be used to compute the word embedding of w_{out}, where $CosSim(p_w, p_{out})$ is the cosine similarity between two words in terms of a *K*-length vector (i.e. p_w and p_{out}). Each element of the vector p_w is the conditional topic probability z for w, which is defined in Eq. 2. Therefore, according to Eq. 1 and 2, if a word w is highly relevant to the word w_{out} in terms of the topic distribution similarity, its word embedding ω_w is more likely to participate the composition of word embedding ω_{out} as in Eq. 3.

Topic embeddings are then learned via the attentional LDA (Lines 4-9). In Lines 5-6, we update the topic embedding of τ_k using the negative log likelihood as the cost function. The objective function with L_2 regularization is defined as:

$$L_k = -\sum_{i=1}^{V} \varphi_{k,i} \cdot \log(\frac{exp(\omega_w \cdot \tau_k)}{\sum_{i=1}^{V} exp(\omega_{w_i} \cdot \tau_k)}) + \epsilon \parallel \tau_k \parallel_2^2 \tag{4}$$

The partial derivative of the objective function L_k with respect to the *jth* dimension of topic embedding τ_k is:

$$\frac{\partial L_k}{\partial \tau_{k,j}} = -\sum_{i=1}^{V} \varphi_{k,i}(1 - \frac{exp(\omega_w \cdot \tau_k)}{\sum_{i=1}^{V} exp(\omega_{w_i} \cdot \tau_k)}) + 2\epsilon\tau_{k,j} \tag{5}$$

The design of the objective function will be discussed in next subsection.

Lines 7-9 sample topics for words. As discussed before, attentional LDA samples topics for words by taking into account the attention signals of words, which represent the importance of the words in respective topics. Thus, we compute the attention weight a_w^k of word w under topic k as follows.

$$a_w^k = \frac{exp(\omega_w \cdot \tau_k)}{\sum_{i=1}^{V} exp(\omega_{w_i} \cdot \tau_k)} \tag{6}$$

Then, we assign a topic k to the ith word in document d from:

[1]https://www.kaggle.com/madhurinani/oscars-2017-tweets

Algorithm 1: attentional segmentation topic model.

Input: corpus D, word embeddings $\boldsymbol{\omega}$, Threshold t
Output: Segmentations and posterior topic-word distribution

1 Initialize topic assignments using vanilla LDA ;
2 **for** *words* w_{out} *without word embeddings* **do**
3 sample the word embeddings ;
4 **for** *initial iteration* $initIter$ **do**
5 **for** *topic* $k = 1, 2, \ldots, K$ **do**
6 compute τ_k ; /* Eq. 4 & 5 */
7 **for** *document* $d \in \{1, 2, \ldots, D\}$ **do**
8 **for** $i = 1$ *to* N_d **do**
9 sample topic z_{d_i} from $P(z_{d_i} = k|\boldsymbol{z}_{\neg \boldsymbol{d_i}}, \boldsymbol{\tau}, \boldsymbol{\omega})$;
10 Seg = SampleSegmentations($D, \boldsymbol{\tau}, \boldsymbol{\omega}, t$);
11 Initialize topic assignments for segmentations Seg;
12 **for** *iteration* $iter$ **do**
13 **for** $s \in Seg$ **do**
14 sample topic z_s from $P(z_s = k|\boldsymbol{z}_{\neg \boldsymbol{s}}, \boldsymbol{\tau}, \boldsymbol{\omega})$;
15 **if** $iter\%100==0$ **then**
16 **for** *topic* $k = 1, 2, \ldots, K$ **do**
17 compute τ_k ; /* Eq. 4 & 5 */
18 Seg = SampleSegmentations($D, \boldsymbol{\tau}, \boldsymbol{\omega}, t$);

$$P(z_{d_i} = k|\boldsymbol{z}_{\neg \boldsymbol{d_i}}, \boldsymbol{\tau}, \boldsymbol{\omega}) \propto a_{w_{d_i}}^k (n_{*|k}^{\neg d_i} + \alpha)(\frac{n_{w_{d_i}|k}^{\neg d_i} + \beta}{n_{*|k}^{\neg d_i} + V\beta}) \tag{7}$$

After phase 1, the learned topic embeddings will be used to obtain the initial set of segmentations Seg using *SampleSegmentation* (Line 10). Algorithm 2 illustrates the main idea of *SampleSegmentation*. Basically, we cut a document into segmentations of consecutive words. Hence, we calculate the similarity for adjacent words only, based on their attention weights $V_i = \{a_{w_i}^1, a_{w_i}^2, \ldots, a_{w_i}^K\}$, where $a_{w_i}^k$ is the attention weight received by w_i under topic k using the word embedding of w_i and the topic embedding of τ_k as in Eq. 6. Therefore, V_i models the human attention when reading this word, combining with background knowledge (word embedding ω_{w_i}) and the semantic information contained in the current dataset (topic embedding τ_k). We group neighbouring words into the same segmentation if they receive similar attention.

Next, we initialize topic assignments for segmentations by adopting the topic of the first word in each segmentation (Line 11). During the iterative process of sampling topics for Seg (Lines 12-18), we sample topics for segmentations instead of words as follows. We extra a topic z_s to all words in a segmentation s:

$$P(z_s = k|\boldsymbol{z}_{\neg \boldsymbol{s}}, \boldsymbol{\tau}, \boldsymbol{\omega}) \propto \tag{8}$$
$$\frac{(n_k^{\neg s} + \alpha)}{|S| - 1 + K\alpha} \frac{\prod_{w \in s}(n_{w|k}^{\neg s} + \beta) \cdot a_w^k}{\prod_{p=1}^{N_s}(n_{*|k}^{\neg s} + V\beta + p - 1)})$$

Algorithm 2: SampleSegmentations($D, \boldsymbol{\tau}, \boldsymbol{\omega}, t$)

Input: Documents D, word embeddings $\boldsymbol{\omega}$, topic embeddings $\boldsymbol{\tau}$, and Threshold t
Output: segmentations

1 **for** *document* $d \in \{1, 2, \ldots, D\}$ **do**
2 $segment.add(w_1)$;
3 **for** $i = 1$ *to* $N_d - 1$ **do**
4 $V_i = \{a^1_{w_i}, a^2_{w_i}, \ldots, a^K_{w_i}\}$;
5 $V_{i+1} = \{a^1_{w_{i+1}}, a^2_{w_{i+1}}, \ldots, a^K_{w_{i+1}}\}$;
6 **if** $CosSim(V_i, V_{i+1}) > t$ **then**
7 $segment.add(w_{i+1})$;
8 **else**
9 $segmentations.add(segment)$;
10 $segment = null$;
11 $segment.add(w_{i+1})$;
12 $segmentations.add(segment)$;
13 return $segmentations$;

Here, $n^{\neg s}_k$ is the number of segmentations assigned to topic k, $n^{\neg s}_{w|k}$ is the number of word w assigned to topic k, and $n^{\neg s}_{*|k}$ is the total number of words assigned to topic k. The symbol $\neg$ indicates that document d or segmentation s is excluded.

Moreover, the frequency of topic sampling is different to that of updating topic embeddings τ and segmentations Seg. For more efficient learning, we update τ and Seg every 100 iterations (Lines 15-18).

C. Learning Topic Embeddings

We now discuss the design of the objective function (Eq. 4) which is used to update topic embeddings. In our work, topics can be represented in two ways: 1) as a distribution over words (e.g., $\varphi_{k,i}$) and 2) as the real valued vectors in the same space as word embeddings (e.g., τ). Therefore, in Eq. 4, the cross-entropy loss function captures the relationship between topics and words in the two spaces to learn topic embeddings. The relationship in the topic-word distribution space is straightforward. Therefore, we focus on explaining capturing the relationship between topics and words in the embedding space, $\frac{exp(\omega_w \cdot \tau_k)}{\sum_{i=1}^{V} exp(\omega_{w_i} \cdot \tau_k)}$, as follows.

Consider two words w_i, w_j and a topic k. Let $relevance(w, k)$ be the variable representing the relationship between a word w and a topic k, so that $relevance(w, k) \to 1$ if w and k are related, and $relevance(w, k) \to 0$ if otherwise. We now consider how to capture the same relationship through their embeddings. Let ω_i and ω_j be the word embeddings of w_i and w_j respectively, and τ_k be the topic embedding of the topic k. If there is a function $F(\omega_i, \omega_j, \tau_k)$ defined on the three arguments, it would be preferable to have $F(\omega_i, \omega_j, \tau_k) = \frac{relevance(w_i,k)}{relevance(w_j,k)}$ so that the $F(\omega_i, \omega_j, \tau_k)$ encodes the relationship between the two words and a topic by having the properties summarized in Table II.

We further analyze the structure of $F(\omega_i, \omega_j, \tau_k)$ as follows.

TABLE II: $F(\omega_i, \omega_j, \tau_k)$ satisfies the following conditions

$F(\omega_i, \omega_j, \tau_k)$	$relevance(w_i, k) \to 1$	$relevance(w_i, k) \to 0$
$relevance(w_j, k) \to 1$	$F(\omega_i, \omega_j, \tau_k) \to 1$	$F(\omega_i, \omega_j, \tau_k) \to 0$
$relevance(w_j, k) \to 0$	$F(\omega_i, \omega_j, \tau_k) \gg 1$	$F(\omega_i, \omega_j, \tau_k) \to 1$

TABLE III: Data Statistics. #label: number of categories, #docs: number of documents, V: size of the vocabulary, AvgLength: average number of words in a document

Dataset	#label	#docs	V	AvgLength
Oscars	-	5,966	609	7.609
Snippets	8	1,707	2,904	7.607
Title	7	10,193	5,352	6.543

1. As word embeddings are inherently linear structures, it is natural to use the difference between ω_i and ω_j to represent their relationship. Therefore, $F(\omega_i, \omega_j, \tau_k)$ can be simplified as $F(\omega_i - \omega_j, \tau_k)$.

2. Given the equation $F(\omega_i, \omega_j, \tau_k) = \frac{relevance(w_i,k)}{relevance(w_j,k)}$, the right hand side is a scalar. Therefore, we can use inner product to convert the argument on the left hand side into a scalar for consistency. Then, $F(\omega_i - \omega_j, \tau_k)$ can be replaced by $F\Big((\omega_i - \omega_j)^T \tau_k\Big)$.

Based on the analysis, it is ideal to have

$$F\Big((\omega_i - \omega_j)^T \tau_k\Big) = \frac{relevance(w_i, k)}{relevance(w_j, k)} \tag{9}$$

Suppose that $F(\mathcal{X}) = exp(\mathcal{X})$. Then, it can be derived from Eq. 9 to have :

$$F\Big((\omega_i - \omega_j)^T \tau_k\Big) = \frac{exp(\omega_i \cdot \tau_k)}{exp(\omega_j \cdot \tau_k)} = \frac{relevance(w_i, k)}{relevance(w_j, k)} \tag{10}$$

Thus, we have $exp(\omega_i \cdot \tau_k) = relevance(w_i, k)$. Since $Softmax(\mathcal{X}) \propto exp(\mathcal{X})$, we capture the relationship between topics and words in the word embedding space using $\frac{exp(\omega_w \cdot \tau_k)}{\sum_{i=1}^{V} exp(\omega_{w_i} \cdot \tau_k)}$ in the objective function (Eq. 4) because it is easy to obtain derivative for $Softmax(\mathcal{X})$. We use L-BFGS[2] [25] to learn topic embeddings $\boldsymbol{\tau}$.

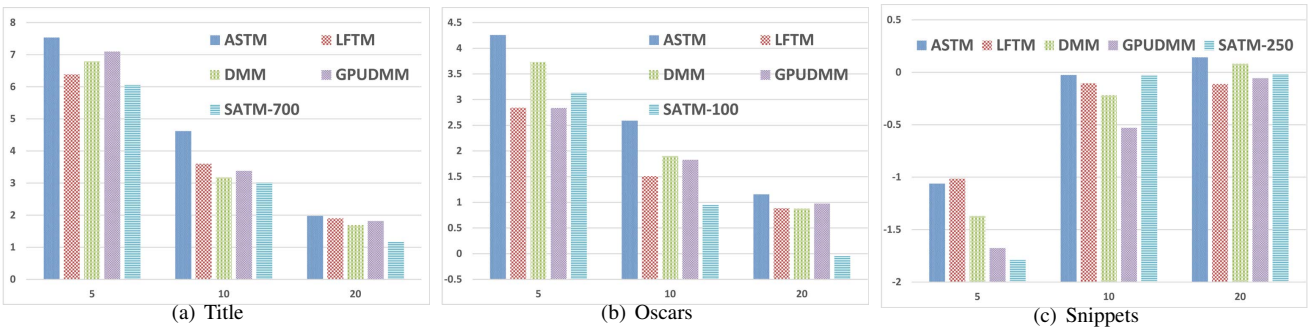
IV. Experimental Results

A. Datasets Description and Setup

Data sets. We study the empirical performance of ASTM on three publicly available short text datasets, including tweets about the 2017 Academy Awards (Oscars for short)[2], Web Snippets (Snippets for short) [26] [3], and TagMyNewsTitle (Title for short) [27] [3]. Statistics of the three datasets after preprocessing are summarized in Table III.

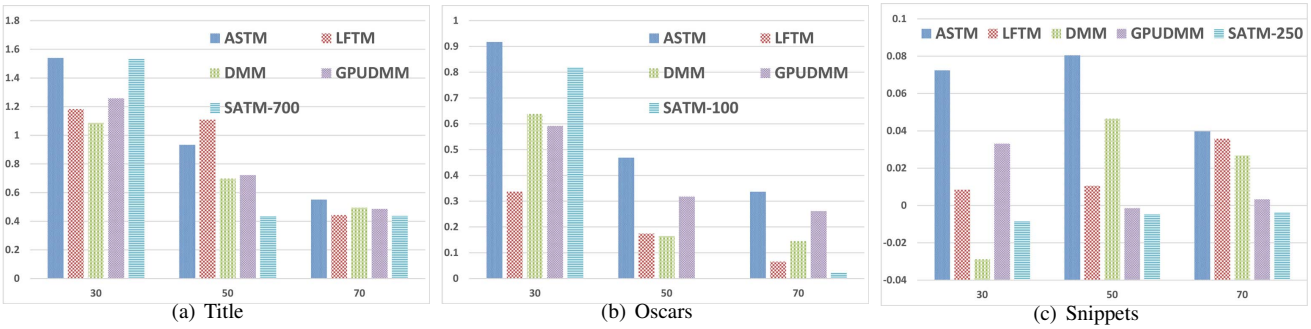
Word Embeddings. In this paper, we use the word embeddings pre-trained by Stanford GloVe as the auxiliary information. According to [2], Word2vec [4] and GloVe produce similar results.

[2]https://www.kaggle.com/madhurinani/oscars-2017-tweets
[3]http://acube.di.unipi.it/tmn-dataset/
[4]https://code.google.com/p/word2vec/



(a) Title (b) Oscars (c) Snippets

Fig. 3: NPMI scores on three datasets, with K ranging from 5 to 20



(a) Title (b) Oscars (c) Snippets

Fig. 4: NPMI scores on three datasets, with K ranging from 30 to 70

Baselines. We compare the performance of ASTM [5] with the following methods. The DMM and SATM are classical topic models designed for short texts. The LFTM and GPUDMM are methods integrating word embeddings into topic models.

- DMM [13] is a classical short text topic model adopting a straightforward assumption that each document is related to only one topic.
- SATM [15] learns topics from long pseudo-documents consisting of short texts. In this model, the number of pseudo-documents need to be set in advance, which affects topic qualities. We set it to 100, 250 and 700, respectively, on the three datasets to obtain the best classification results, as denoted in Fig 3 and Fig 4.
- LFTM [2] generates words from a two-component mixture of a topic-word component and a latent feature component. We select glove-DMM from the four variations of LFTM since it is the closest to our model. In particular, glove-DMM integrates into DMM the word embeddings pre-trained by GloVe.
- GPUDMM [3] uses the generalized Pólya urn (GPU) model to promote the co-occurrence of similar words under the same topic, where the similarity is computed by word embeddings.

Parameter Settings. For ASTM, we set the default parameter values as $\alpha = 0.1, \beta = 0.01, t = 0.5$. For baseline approaches, we follow the parameter settings in their respective papers.

[5]Our implementation is available at https://github.com/wjmzjx/ASTM

B. Topic Coherence Evaluation

NPMI score. We use the Normalized Pointwise Mutual Information (NPMI) [2] score to measure the semantic coherence of discovered topic distributions. For each topic, we consider the top 15 most probable words and use the English Wikipedia of 20837 articles [6] and 30607 news[7] as the external corpora to calculate the following,

$$NPMI(k) = \sum_{1 \le i < j \le N} \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \tag{11}$$

where $P(w_i, w_j)$ is the probability that two words w_i and w_j co-occur in a 5-word sliding window. A higher NPMI score indicates better topic coherence.

We compute the average NPMI score over all topics, discovered by the five comparative models, on the three datasets with the number of topics K ranging from 5 to 70. The results have been plotted in Fig. 3 and Fig. 4. We observe that, firstly, ASTM achieves the best performance in all settings on the Oscars dataset. ASTM achieves the best performance on the other two datasets except for one setting (i.e. $K = 50$ on Title and $K = 5$ on Snippets). The reason why ASTM clearly outperforms the other models on the Oscars dataset is probably because the data set has more words without word embeddings compared to the other two datasets. While the other models simply ignore such words, ASTM uses a

[6]https://einstein.ai/research/the-wikitext-long-term-dependency-language-modeling-dataset

[7]http://acube.di.unipi.it/datasets/

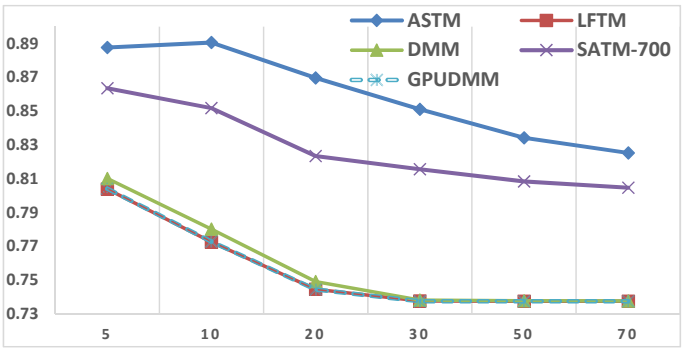


Fig. 5: Classification accuracy on Title dataset.

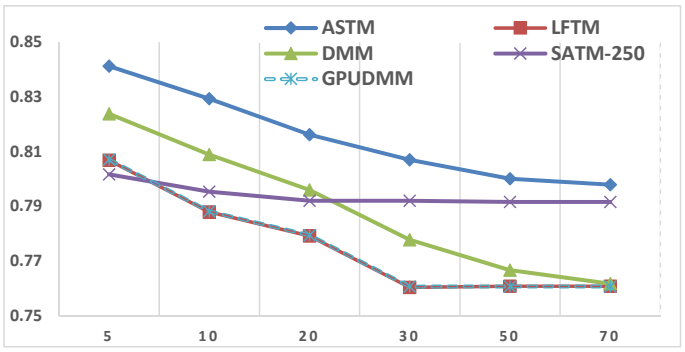


Fig. 6: Classification accuracy on Snippets dataset.

sampling strategy to approximate their word embeddings to retrieve more background information.

Secondly, comparing the four baseline models, we find that the two models integrating word embeddings (i.e. LFTM and GPUDMM) do not always perform better than the two models without introducing background information (i.e. DMM and SATM). It indicates that the simple combination mechanisms used by existing topic models to integrate word embeddings may introduce noises instead of useful information. Take the Oscars dataset as an example. The data set contains tweets about the 2017 Academy Awards. The word embeddings of some words may be too over-generalized to capture the right semantic meaning of the word in the data sets. For instance, the word "moonlight" refers to a movie in the dataset, while its embedding is too general – the five words closest to it in the embedding space are: "moon", "blue", "velvet", "night", "sunny", none of which can help to realize the word refers to a movie. In such cases, the over-generalized word embeddings bring noise which makes the word embedding integrated models worse than the vanilla DMM.

C. Short Text Classification Evaluation

We evaluate the performance of the five topic models in terms of short text classification. We employ Naive Bayes classifier [8] in this paper because it has a strong probabilistic foundation for Expectation-Maximization [13], [28], [29]. Fig. 5 and Fig. 6 show the average classification accuracy of the five models on the two labeled datasets (i.e. Title and Snippets), with respect to different numbers of topics. The following observations are made.

Firstly, ASTM remarkably outperforms all the other models on both datasets. The significant improvement achieved by ASTM compared to the DMM-based models (i.e. DMM, GPUDMM, LFTM (glove-DMM)) validates that it is effective to assign a topic to a segmentation induced from human attention instead of assigning a topic to a short text document. In other words, the assumption that a short text document having only one topic may enhance the topic coherence but negatively affects the accuracy of document classification.

[8]Java-ML is employed with its parameters chosen by five-fold cross-validation on the training set, which can be download from http://java-ml.sourceforge.net/.

Secondly, comparing the three DMM-based models (i.e. DMM, GPUDMM, LFTM (glove-DMM)), we find that the DMM model is slightly better than the other two models on the Title data set and clearly better on the Snippets data set, even though DMM does not incorporate word embeddings. It again suggests that the combination mechanism should be carefully designed to integrate word embeddings into topic models to avoid detrimental effects. For example, in the Title dataset, the group of two words, "white" and "sox", denotes an American baseball team. Intuitively, there should be a high possibility for the two words to co-occur in the same topic so that we can correctly classify the documents containing the group into the category of "sports". However, the word embeddings of the two words are not close at all since the distributed representation is trained on an external large corpus. The ways used by existing models to incorporate word embeddings are not sophisticated enough to address this issue. For example, GPUDMM directly calculates the vector similarity between word embeddings. In contrast, our ASTM model learns attention signals, corresponding to the importance of words under respective topics, using both word embeddings trained from external corpora and topic embeddings learned from the given data set. The resulted performance ascertain that the combination mechanism is critical to better performance.

D. Examples of Segmentations

We further examine whether the segmentations generated by ASTM are consistent with human reading habits. Four examples, two from the Oscars dataset and two from the Title dataset, are shown in Table IV and Table V with annotations illustrating ASTM results. In particular, a bracket denotes a segmentation of words. All words in a segmentation are assigned to the same topic. Topics are marked in different colors. We reserve stop words in the examples to facilitate understanding and mark them gray. We analyze the results on the four examples in details as follows.

Two tweets from the Oscars dataset shown in Table IV. Firstly, ASTM accurately groups the full names of the three persons (i.e., Demi Lovato, Meryl Streep and Justin Timberlake) into same segmentations. Secondly, the three names/segmentations are assigned to different topics mapping the characteristics of the corresponding persons. For example,

TABLE IV: The topics in Oscars dataset

(demi lovato) was (stunning) at the (vanity fair #oscars party)	
(meryl streep) and (justin timberlake) (backstage) at the 89th (annual academy awards) #oscars	
Topic1	emma stone backstage brie larson crying congratulates heytheredaiiiah starts accountants worked final center fiasco song dream mintmovie adele ryan gosling
Topic5	ryan gosling years incredible lin manuel moana auli cravalho princess voice taking vaiieta united states american stage enews gt performance
Topic7	party fair vanity wore thefashioncourt gown andrew garfield fall back couture seriesbrasil evans nina chris gold demi blue dress lovato
Topic18	year start mix city movie resist funder trumprussia russiagate york streep political election ratings low london hit california movies day
Topic28	academy awards red carpet dakota matt damon interview ben video stars jamie looked hq star stunning added dallas jamiedornan fashion

TABLE V: The topics in Title dataset

(twins 1), (white sox 0): (twins' liriano) (throws) (no-hitter) against (white sox)	
(japan radiation fear) (sparks) (south korea) diaper (rush)	
Topic1	red sox white baseball blue san puts wings moves rays sweep indians boston force bay black angels grand hot closer
Topic6	japan nuclear crisis quake plant power plans energy safety workers fears wal mart panel radiation recovery japanese water tsunami control
Topic11	back video makes pay start school offer lede alabama free tech good lose money vows private cash schools give leave
Topic23	mets yankees home hits phillies giants jays dodgers twins braves road injury tigers marlins cubs pitch straight reds pirates miss
Topic24	south libya north war talks forces nato libyan rebels government army africa town korea military troops sudan carolina held gaddafi

Demi Lovato is more relevant to the Vanity Fair Party (Topic7) instead of the Academy Awards Ceremony because she does not have an award nomination. The topic of Meryl Streep (Topic18) is related to politics because of her speech at the Golden Globe Awards. As for the opening guest Justin Timberlake (Topic5), Twitter users are keen to discuss the reunion between him and Ryan Gosling since they were both child stars from the Mickey Mouse Club.

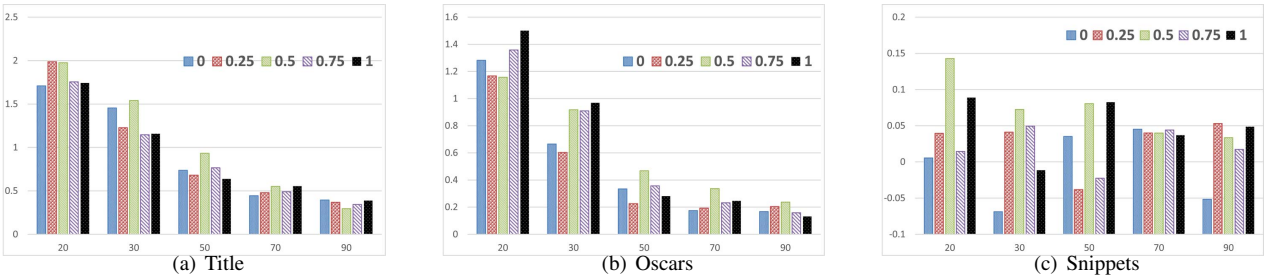
Two news titles from the Title dataset shown in TableV. The first one is labeled as "sports". As we can see from the title, two baseball teams (i.e., Minesota Twins and Chicago White Sox) are successfully identified by ASTM. One notable thing is that the two teams are assigned to different topics (i.e., topic23 and topic1 respectively). By taking a closer look at the content of the two topics, we realize that while they are both related to baseball teams, the two topics are generated because of externally trained word embeddings. More specifically, the teams involved in the Topic1 have the same characteristic that their team names contain words describing colors, such as *Boston Red Sox*, *Chicago White Sox* and *Toronto Blue Jays*. Meanwhile, the characteristic shared by teams involved in the Topic23 is that their team names contain words of animals, such as *Detroit Tigers*, *Chicago Cubs* and *Florida Marlins*. This observation proves again that externally trained word embeddings need to be used cautiously.

The second title is labeled as "health", and it summarizes a buying frenzy from South Korean mothers since the food bans sparked by the risk of radiation contamination from Japan's damaged nuclear power stations. ASTM segments this title into four segmentations and assigns four different topics. Firstly, we notice that the segmenting is in line with human reading habits and correctly determines the subject ("japan radiation fear"), predicate ("sparks"), and object ("south korea"). Secondly, we observe that the title does relate to multiple topics, such as health (Topic6), and world (Topic24), which justifies our assumption to assign multiple topics rather than one topic to a short text document. Moreover, we find that ASTM is able to group "japan radiation fear" together, although the word embedding of "japan" may suggest a low similarity with "radiation". By modelling attentional signals of words from externally trained word embeddings and internally learned topic embeddings, ASTM addresses the noise issue confronted by existing distributed representation based topic models to some degree.

E. Discussion of Threshold

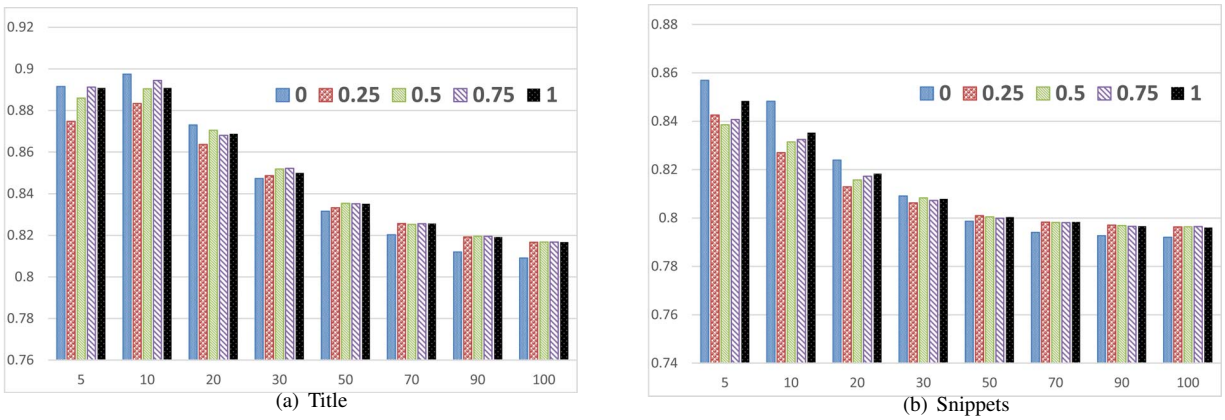
Recall that ASTM generates segmentations by computing the cosine similarity between pairs of consecutive words in terms of their attentional weights. Two words will be grouped into the same fragment if their cosine similarity is greater than some pre-specified threshold t. We conduct experiments to evaluate the influence of the threshold t on the performance of ASTM. In particular, we vary t from 0 to 1 with an interval of 0.25. Note that, although the cosine similarity between two vectors of attentional weights varies from -1 to 1, negative values indicate the attentional signals received by two words are dissimilar, which is contrary to our assumption that only neighboring words receiving similar attention will be grouped. Therefore, it is meaningful only to consider positive values for t.

Fig. 7 shows the NPMI scores with respect to the different settings of t on the three datasets respectively. It can be observed that the topic coherence is sensitive to the value t, especially on the Snippets data set. We take a closer look at the data set and discover that Snippets is different from the other two datasets in the way that consecutive words are not necessarily semantically related. According to [26], the Snippets data set consists of a set of search snippets which are "noisy and less topic-focused". We notice a situation in Snippets where there are many synonyms in the same document such as "wikipedia" and "wiki", and they are separated from each other with some less relevant words in between. Since ASTM considers only the similarity between neighboring words, whether such synonyms can be grouped into the same segmentations depends on the selection of the threshold t as well as the relevance of the words between such



(a) Title (b) Oscars (c) Snippets

Fig. 7: NPMI scores on three datasets, with K ranging from 20 to 90 and threshold t from 0 to 1



(a) Title (b) Snippets

Fig. 8: Classification accuracy on two datasets, with K ranging from 5 to 100 and threshold t from 0 to 1

synonyms. Incorrectly putting synonyms into different segmentations is equivalent to deleting their relevance, which has a great impact on topic coherence. On the other two datasets where documents are more topic-focused, the influence of t on topic coherence is much smaller, especially when the number of topics is large. We observe that when the number of topics is sufficient, similar words tend to have high similarity values (e.g., CosSim$>$0.9). Consequently, varying the threshold from 0 to 0.75 does not affect the grouping of such words. That explains why topic coherence becomes less sensitive to the threshold t when the number of topic K increases on Title and Oscars.

The influence of threshold on classification accuracy is shown in Fig. 8. It can be observed that the influence of the threshold decreases on the two labeled datasets as the number of topics K increases. When K is greater than 30, different thresholds hardly affect the classification accuracy any more. As discussed before, this is because the similarity value of words receiving similar attention tend to be high when the number of topics is great, so that lower threshold values do not cause much changes to the topic-word distributions, which are used to train Naive Bayes classifier. Interestingly, although the topic coherence fluctuates with respect to the variation of the threshold t on the Snippets data set, the classification accuracy on the same dataset is less sensitive to the threshold.

By noticing that $t = 0.5$ gives the best NPMI score at most K, and considering the impact of threshold on classification performance, we set t as 0.5 in other experiments to ensure that both tasks achieve satisfactory performance.

V. Conclusion

Inspired by human reading habits, we proposed a topic model named ASTM for short texts in this paper. In this model, we take word embeddings as background knowledge to relieve the sparsity of short texts and use attentional signals to segment individual documents into segmentations, where a segmentation is a coherent sequence of neighboring words. Unlike the simple combination mechanisms of existing models that improve the topic sampling by directly using word embeddings, we consider word embeddings and topic embeddings jointly to compute attention signals. On one hand, our strategy ensures the information brought by word embeddings will be used only if they are consistent with the given dataset. On the other hand, the strategy models the attention received by words as the importance of the words under respective topics, which conforms with human habits in understanding texts. The experiments on three publicly available short text datasets demonstrate the effectiveness of ASTM in terms of both topic coherence and document classification performance. The experimental results, coupled with conclusions from existing studies [3], [11], [12], reveal that taking into account of human attention and extra background knowledge can be useful for enhancing textual analysis.

As an ongoing work, we consider improving our model by using some self-adaptive sliding window based nonparametric

sampling strategy to generate segmentations to avoid the influence of the similarity threshold. We are also interested in applying ASTM on review datasets to analyze the characteristics of user attention to build personality-based recommendation systems.

VI. Acknowledgments

This research has been supported by the National Key Research and Development Program of China under grant 2016YFB1000901, the National Natural Science Foundation of China (NSFC) under awards 91746209, 61503114 and 61503112, and partially supported by ARC Discovery Grant DP180100966.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[2] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," *TACL*, vol. 3, pp. 299–313, 2015.

[3] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic modeling for short texts with auxiliary word embeddings," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, 2016, pp. 165–174.

[4] J. He, Z. Hu, T. Berg-Kirkpatrick, Y. Huang, and E. P. Xing, "Efficient correlated topic modeling with topic embedding," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, 2017, pp. 225–233.

[5] J. Qiang, P. Chen, T. Wang, and X. Wu, "Topic modeling over short texts by incorporating word embeddings," in *Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II*, 2017, pp. 363–374.

[6] G. Xun, V. Gopalakrishnan, F. Ma, Y. Li, J. Gao, and A. Zhang, "Topic discovery for short texts using word embeddings," in *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, 2016, pp. 1299–1304.

[7] B. Shi, W. Lam, S. Jameel, S. Schockaert, and K. P. Lai, "Jointly learning word embeddings and latent topics," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, 2017, pp. 375–384.

[8] G. Xun, Y. Li, J. Gao, and A. Zhang, "Collaboratively improving topic discovery and word embeddings by coordinating global and local contexts," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, 2017, pp. 535–543.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, 2013, pp. 3111–3119.

[10] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1532–1543.

[11] S. Li, Y. Zhang, R. Pan, M. Mao, and Y. Yang, "Recurrent attentional topic model," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, 2017, pp. 3223–3229.

[12] S. Wang, J. Zhang, and C. Zong, "Learning sentence representation with guidance of human attention," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 2017, pp. 4137–4143.

[13] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 2014, pp. 233–242.

[14] X. Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: topic modeling over short texts," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 12, pp. 2928–2941, 2014.

[15] X. Quan, C. Kit, Y. Ge, and S. J. Pan, "Short and sparse text topic modeling via self-aggregation," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, 2015, pp. 2270–2276.

[16] H. Amoualian, W. Lu, É. Gaussier, G. Balikas, M. Amini, and M. Clausel, "Topical coherence in lda-based models through induced segmentation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 2017, pp. 1799–1809.

[17] Rumelhart, E. David, Hinton, E. Geoffrey, Williams, and J. Ronald, "Learning representations by back-propagating errors," vol. 323, no. 6088, pp. 399–421, 1986.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 6000–6010.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[20] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, 2015, pp. 1412–1421.

[21] P. Ren, Z. Chen, Z. Ren, F. Wei, J. Ma, and M. de Rijke, "Leveraging contextual sentence relations for extractive summarization using a neural attention model," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, 2017, pp. 95–104.

[22] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, 2015, pp. 379–389.

[23] P. Ren, Z. Chen, Z. Ren, F. Wei, J. Ma, and M. de Rijke, "Leveraging contextual sentence relations for extractive summarization using a neural attention model," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, 2017, pp. 95–104.

[24] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kociský, and P. Blunsom, "Reasoning about entailment with neural attention," *CoRR*, vol. abs/1509.06664, 2015.

[25] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 1-3, pp. 503–528.

[26] X. H. Phan, M. L. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, 2008, pp. 91–100.

[27] D. Vitale, P. Ferragina, and U. Scaiella, "Classification of short texts by deploying topical annotations," in *Advances in Information Retrieval - 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings*, 2012, pp. 376–387.

[28] K. Nigam, A. McCallum, S. Thrun, and T. M. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, no. 2/3, pp. 103–134, 2000.

[29] A. Fang, I. Ounis, P. Habel, C. Macdonald, and N. Limsopatham, "Topic-centric classification of twitter user's political orientation," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, 2015, pp. 791–794.