It All Starts with Entities: A Salient Entity Topic Model

Wu, C.; Kanoulas, E.; de Rijke, M.

DOI
10.1017/S1351324919000585
Publication date
2020
Document Version
Final published version
Published in
Natural Language Engineering

Natural Language Engineering (2020), **26**, pp. 531–549
doi:10.1017/S1351324919000585

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

It all starts with entities: A Salient entity topic model

Chuan Wu[1,2,*], Evangelos Kanoulas[2] and Maarten de Rijke[2]

[1] School of Information Management, Wuhan University, Wuhan, China and [2]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
*Corresponding author. Email: wuchuan114@gmail.com

(Received 25 October 2018; revised 26 July 2019; accepted 10 September 2019; first published online 22 November 2019)

Abstract

Entities play an essential role in understanding textual documents, regardless of whether the documents are short, such as tweets, or long, such as news articles. In short textual documents, all entities mentioned are usually considered equally important because of the limited amount of information. In long textual documents, however, not all entities are equally important: some are salient and others are not. Traditional entity topic models (ETMs) focus on ways to incorporate entity information into topic models to better explain the generative process of documents. However, entities are usually treated equally, without considering whether they are salient or not. In this work, we propose a novel ETM, Salient Entity Topic Model, to take salient entities into consideration in the document generation process. In particular, we model salient entities as a source of topics used to generate words in documents, in addition to the topic distribution of documents used in traditional topic models. Qualitative and quantitative analysis is performed on the proposed model. Application to entity salience detection demonstrates the effectiveness of our model compared to the state-of-the-art topic model baselines.

Keywords: Entity Salience; Entity Topic Model

1. Introduction

The importance of entities has been well recognized in domains as diverse as data mining (Shen *et al.* 2013), knowledge representation (Xie *et al.* 2016), language technology (Levit *et al.* 2014), and information retrieval (Balog 2018). Downstream applications in the aforementioned domains have benefited from modeling entities as vital sources of information in the generative process of documents. This has led to the development of a range of entity topic models (ETMs), with entities either treated as external labels of documents (Rosen-Zvi *et al.* 2004) or observed variables (Erosheva, Fienberg, and Lafferty 2004). For example, the Author Topic Model (ATM) (Rosen-Zvi *et al.* 2004) assumes a topic distribution for each author, representing the research interest of authors. To generate a word in a document, an author is selected and a topic is sampled from the topic distribution of the author, before sampling a word from the topic distribution of the selected topic. In contrast, entities can be viewed as observed variables different from words in documents. For example, Link-LDA (LLDA) (Erosheva *et al.* 2004) models references of papers as observed variables to model the generation of academic articles.

One limitation of existing ETMs is that none of them takes the salience of entities into account. Entity salience reflects the importance of an entity for a particular document. Entity salience can be characterized by local scoping and invariable perception (Gamon *et al.* 2013). Research about entity salience in web pages shows that fewer than 5% of the entities mentioned on a web page are salient for the page (Gamon *et al.* 2013). Intuitively, salient entities should play a more important role in the process of generating documents than non-salient entities.

In this work, we propose a novel topic model, *Salient Entity Topic Model* (SETM), that models salient entities in the generative process of documents. We model the generative process as a three-step procedure: (1) sample a topic distribution for a document from a Dirichlet prior; (2) sample salient entities using the topic distribution of the document; and (3) sample words from the joint topic distribution combined from document topic distribution and salient entity topic distributions. The advantage of Salient Entity Topic Model (SETM) is that it models the mutual reinforcement between topics and entity salience. For example, if an entity is likely to be salient under a given topic, it will not only have higher probability to show up in documents around this topic, but also have higher probability to be generated as a salient entity. Another advantage of SETM is that if an entity e_a is salient in document d_a and document d_b is semantically similar to d_a, then an entity e_b (in d_b) that is similar to e_a is likely to be salient in d_b.

The assumption behind our model is that stories are built upon a story line (topic) and a set of main characters in the story (salient entities). Imagine that a news reporter is writing a news article about a specific story. The primary thing under consideration is what the document is about, and which is modeled by the topic distribution of the document. The second thing is which entities are salient entities in the story described in the document. And finally, other words and entities are added to the document to complete the story.

Experiments on a publicly available dataset show that our model better explains and models the generative process of documents, outperforming state-of-the-art methods. Both a qualitative and quantitative analysis is performed to demonstrate that by taking salient entities of documents into consideration, our model is better than similar models. The code of our work is published at Github.[a]

The main contributions of this work are as follows:

1. We propose a novel Salient Entity Topic Model (SETM) to model the generation of documents.
2. We derive a Gibbs sampling algorithm for parameter estimation.
3. We demonstrate the effectiveness of SETM to model text by performing both a qualitative and a quantitative analysis.

2. Related work

Topic models have been widely used in text analysis. Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) models each document as a mixture of topics and automatically generates summaries of topics in terms of a multinomial distribution over words. The original LDA has been extended in a wide variety of directions. Recent topic model extensions are either designed for specific tasks, such as multi-label classification (Li, Ouyang, and Zhou 2015a,b) and opinion mining (Wang, Chen, and Liu 2016), or particular kinds of texts, such as short texts (Zhang, Mao, and Zeng 2016; Bicalho *et al.* 2017; Qiu and Shen 2017; Li *et al.* 2018).

On the other hand, the notion of entity salience is attracting more attention (Gamon *et al.* 2013; Tran *et al.* 2015; Escoter *et al.* 2017; Xiong *et al.* 2018). Gamon *et al.* (2013) propose the task of identifying salient entities on web pages. Tran *et al.* (2015) take entity salience into consideration in ranking entities for summarization of high-impact events. Escoter *et al.* (2017) group business news stories based on the salience of named entities. Xiong *et al.* (2018) propose a Kernel Entity Salience Model to better estimate entity salience in documents so as to improve text understanding and retrieval.

In this work we extend it by considering salient entities in modeling the generative process of documents. In this context, there are three branches of closely related work, such as entities as sources, entities as observed variables, and entities as entity topics. Below, we first summarize

[a] https://github.com/setm2nle/salient-entity-topic-model.

these three types of related work and clarify the differences between our model and the previous work. Then, we discuss related work on topic labeling and clarify their difference with our work.

2.1 Entities as source of information

In some scenarios, entities represent an external source of information that generates documents. For example, the Author Model (AM) (McCallum 1999) models document content and its authors' interests, where each author (that corresponds to an external entity) corresponds to one topic. To generate a word, an author z is sampled uniformly from a set of authors of the document, and then a word w is generated by sampling from an author-word multinomial distribution. The ATM (Rosen-Zvi *et al*. 2004) extends AM by introducing a topical layer between authors and words. An author's interests are modeled with a mixture of topics. Each document is associated with a set of observed authors. To generate a word, an author x is chosen uniformly from this set, then a topic is selected from the topic distribution of author x, and then a word is generated by sampling from a topic-specific multinomial distribution over words. The Author Recipient Topic Model (ARTM) (McCallum, Corrada-Emmanuel, and Wang 2005) takes the recipient of messages into account. In ARTM, recipients of a message are also considered as authors of the message, and contribute to the generation of a particular message.

In all the previous models, authors/entities are external labels, such as senders or recipients of messages. Similarly, we also consider salient entities as a source of information to generate documents. The distinguishing feature of our model is that we use entities that are both *observed* and *salient* in documents to model the sources of information. This distinction is important because unlike authors as external labels of documents, salient entities can serve as external labels and as representations of the content of documents at the same time. Hence, we hypothesize that salient entities capture more of the available information.

2.2 Entities as observed variables

Entities are different semantic units from words, and hence they should be modeled as a special kind of observed variable. LLDA (Erosheva *et al.* 2004) models the generation of academic articles. In academic articles, references of papers can be viewed as entities. In the document generation process of LLDA, a topic distribution is sampled from a Dirichlet prior in the same way as in LDA. Then, a topic is sampled from the topic distribution of the document, and a word or entity is sampled from the topic-word or topic-entity distribution. To better model the correlation between words and entities, CorrLDA2 (Newman, Chemudugunta, and Smyth 2006) models word topics and entity topics separately, where word (entity) topics are used to generate words (entities). In the generative process, words are generated first, and then entity topics are sampled uniformly from all sampled word topics. Some authors propose an ETM for entity linking (Han and Sun 2012); though it also considers entities in topic modeling, it is designed for entity linking, thus not directly comparable with our model.

In our work, we propose two variants of a topic model: one models words and entities with a single observed variable, while the other uses two observed variables to distinguish entities from words. The advantage of our model lies in the fact that we do not only consider entities as part of observed variables, but also incorporate entity salience information in the document generation process. In this way, our model can make the best use of available entity (salience) information.

2.3 Entities as entity topics

Entities can also be treated as special topics and contribute to the generation of documents together with general topics. For example, the ETM (Kim *et al.* 2012) learns the topical nature of entities. Similar to topics, entities are represented by a multinomial distribution over words.

For each document, a topic distribution is drawn from a Dirichlet prior and a joint multinomial distribution over words Φ is obtained by linearly combining entities and topics of a document. To generate a word, a topic is sampled from Φ and a word is sampled from the topic word distribution. Though ETM seems to be a valid baseline for our work, it is not applicable because of scalability issues. It is applicable to short texts with few entities, such as abstracts of academic papers or small collections of news articles but not for long web documents. In contrast, the models that we propose do scale to large documents.

Another disadvantage of ETM is that it treats all entities equally, while in reality, salient entities are more important than non-salient entities. Compared to ETM, our model only introduces salient entities into the document generation process, which is more realistic.

2.4 Topic modeling versus topic labeling

Existing work on topic modeling can be roughly classified into two categories. The first category proposes novel topic models for resolving particular applications, such as document classification (Rubin *et al.* 2012), entity linking (Kataria *et al.* 2011; Han and Sun 2012), and question answering (Ji *et al.* 2012).

The second category focuses on improving topic modeling by incorporating new information. Kim *et al.* (2012) propose an ETM for mining documents associated with entities. Xu *et al.* (2017) incorporate Wikipedia concepts and categories into topic models. Andrzejewski, Zhu, and Craven (2009) incorporate domain knowledge into topic modeling and conducts qualitative analysis on both synthetic and real world dataset. These works explore a new paradigm of improving over existing topic models, rather than solving a particular important downstream application. Our work aligns to this category.

Topic labeling, on the other hand, is to make topic representations learned by topic models more interpretable. Topics are conventionally represented by their top N words or terms (Blei *et al.* 2003; Griffiths and Steyvers 2004). Recent work on topic labeling proposes to label topics using phrases (Lau *et al.* 2011), structured knowledge base data (Hulpus *et al.* 2013), entities (Lauscher *et al.* 2016), and even images (Aletras and Mittal 2017). Compared to topic labeling, which label topics mined by topic models, our work is focused on improving topic model itself by incorporating entity saliency.

3. Salient entity topic model

3.1 Overview

We present two variants of the Salient Entity Topic Model (SETM), such as SETM-Word-Only (SETM-WO) and SETM-Word-and-Entity (SETM-WE). SETM-WO is a simplified version of SETM-WE, where documents are represented by a bag of words and a bag of entities. The reason to have two variants is two-fold. First, we want to understand the effect of differentiating between words and entities as observed variables, if any. Second, there may be situations that such a separation provides flexibility; for instance, in academic articles, references can be viewed as entities, and hence considered separately from words, while in news articles, words and entities can be mixed together because they appear in the same context.

In what follows we focus on SETM-WE; SETM-WO can be considered to be a simplified version of SETM-WE and it is described only when this simplification affects the proposed algorithms.

The input used to train SETM-WE is a collection of documents, which consist of a bag of words and a bag of entities. An entity is a real world thing that has a corresponding entry in a knowledge base and is represented by a unique identifier. An entity can have multiple surface forms, which could be a unigram or n-gram. Entities are recognized by entity linking tools and this preprocessing step is not considered in our work. In other words, we take recognized entities as inputs. Each entity in a document has a binary label indicating whether the entity is salient or not.

Table 1. Notations

Symbol	description
D	Document collection
S_d	Bag of salient entities in document d
E_d	Bag of entities in document d
N_d	Bag of words in document d
K	Number of topics
θ_d	Topic distribution of document d
ϕ_k	Multinomial distribution over salient entities of topic k
φ_e	Multinomial distribution over entities of topic k
ψ_k	Multinomial distribution over words of topic k
ρ_s	Multinomial distribution over topics of entity $s \in S_d$
Φ	Multinomial distribution over topics

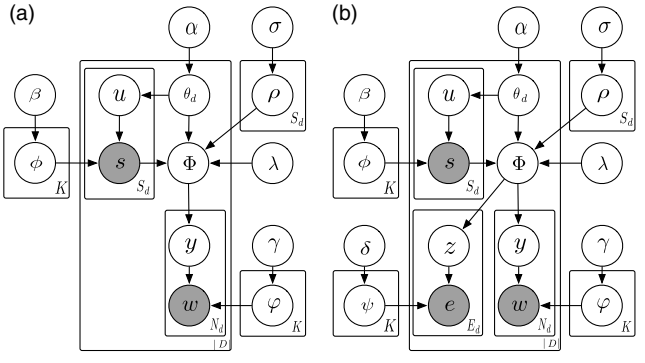

Figure 1. A graphical representation of the Salient Entity Topic Model (SETM). (a) SETM-WO. (b) SETM-WE.

Formally, a document d is represented by a word vector N_d, where each $w_{d,i} \in N_d$ is chosen from the vocabulary of words W, and an entity vector E_d, where each $e_{d,j}$ is chosen from the vocabulary of entities E. Since we have salience labels for each entity, we have the set of salient entities in d, denoted as S_d. The goal is to discover word patterns of topics, and learn topic distributions of documents and entities. The notation used in the paper is summarized in Table 1.

After model training, we need to infer the topic distribution of a newly incoming document using SETM. However, we might not have salience labels for new-coming documents. This is similar to the scenario considered by Labeled LDA (Ramage *et al.* 2009). We adopt their strategy and perform inference by assuming that no entity is salient in the document.

3.2 The SETM model

Graphical representations of SETM-WO and SETM-WE are shown in Figure 1(a) and (b), respectively. A detailed explanation follows.

3.2.1 Hypotheses

The main hypotheses of our model are: (1) salient entities are derived from the topics of a document; and (2) salient entities themselves affect the generation of words and other entities in a document. The intuition behind the first hypothesis is that the topics of a document are

decided before choosing salient entities. When composing a story in an article, one first has some abstract story line indicating the main theme of the story. For example, to write a news report on a football game, one first decides the topics, for example, sports, and then adds teams, players, and their interactions. The second hypothesis comes from the fact that non-salient entities may have some connection to salient entities, but they are loosely related to the theme of the document. For example, in the news item *Liberia: Former football striker George Weah wins presidential election*,[b] football club *Manchester City* is mentioned because the person of interest used to play for the club, though the club is not very important for this particular news article.

Algorithm 1 Generative Process of the SETM-WE Model.

for each topic k **do**
  Draw $\phi_k \sim Dir(\beta)$
  Draw $\varphi_k \sim Dir(\gamma)$
  Draw $\psi_k \sim Dir(\delta)$
end for
for each entity e **do**
  Draw $\rho_e \sim Dir(\sigma)$
end for
for each document d **do**
  Draw $\theta_d \sim Dir(\alpha)$
  for each salient entity s **do**
    Draw topic $u \sim \theta_d$
    Draw salient entity $s \sim \phi_u$
  end for
  Obtain $\Phi_d = \lambda\theta_d + (1-\lambda)\frac{1}{|S_d|}\sum_{s\in S_d}\rho_s$
  for each entity e **do**
    Draw topic $z \sim \Phi_d$
    Draw entity $e \sim \varphi_z$
  end for
  for each word w **do**
    Draw topic $y \sim \Phi_d$
    Draw word $w \sim \varphi_y$
  end for
end for

3.2.2 Generative process

The generative process is shown in Algorithm 1. For each topic k, a topic salient entity distribution ϕ_k, a topic entity distribution ψ_k and topic word distribution φ_k are drawn from a Dirichlet prior with parameters β, δ and γ, respectively. For each document d, a multinomial distribution θ_d over topics is drawn from a Dirichlet prior with parameter α. Then, each salient entity $s \in S_d$ in the document is generated by first sampling a topic u from θ and then drawn from the topic salient entity distribution ϕ_u. To generate words and observed entities in document d, a joint topic distribution Φ is obtained by combining θ_d and the topic distribution of all salient entities of the document ρ_s ($s \in S_d$). Finally, words (or entities) are generated by first sampling a topic y (z), and then sampling a word (or an entity) from the topic word (or entity) distribution φ (ψ).

[b] https://en.wikinews.org/wiki/Liberia:_Former_football_striker_George_Weah_wins_presidential_election.

Note that ϕ and ρ are obtained from the same matrix, while from different perspective. ϕ is a matrix with size $K \times V_E$. From a row viewpoint, it is a list of topics (ϕ_k), with each topic represented by a multinomial distribution over entities. When viewed from a column perspective, it is a list of entities, with each entity represented by a topic distribution (ρ_e).

One could also assume a switch distribution after Φ is derived, which is used to generate either words or entities. A similar switch distribution can be found in Switch LDA (Newman *et al.* 2006), as illustrated by the *Binomial*(φ_{z_i}). However, we do not consider switch distribution in our model for the following reasons. First, we want to keep the flexibility of our model, so that it is still valid in cases where there is no direct connection between words and entities. For example, when analyzing scientific publications, documents (papers) are represented by bag of words in abstracts and list of references of papers. In this case, it is inappropriate to have the switch probability since words in abstracts are very different from references. Second, given our model and the extension from CI-LDA to Switch LDA, we consider it straightforward to extend our model by taking the switch distribution into account when necessary.

4. Model inference

Gibbs sampling is used for parameter estimation. Specifically, we repeatedly sample the topic assigned to each salient entity, word and entity in the document collection, given the topic assignment of the remaining salient entities, words, and entities, as well as the priors. The inference process for SETM is detailed first, followed by clarification of the difference between the inference process of SETM-WE and SETM-WO.

4.1 Inference of SETM-WE

4.1.1 Sampling salient entity topics s

The conditional posterior of assignment u_i to the i-th salient entity in document d is:

$$P(u_i = j|\mathbf{u}_{-i}, \mathbf{s}) \propto P(s_i|u_i = j, \mathbf{u}_{-i}, \mathbf{s}_{-i})P(u_i = j|\mathbf{u}_{-i}), \tag{1}$$

where $\mathbf{u}_{-i}$ is the topic assignments of all salient entities except the i-th one. The first item on the right-hand side is a likelihood and the second is a prior.

For the first term in Equation 1, we have

$$P(s_i|u_i = j, \mathbf{u}_{-i}, \mathbf{s}_{-i}) \propto \int P(s_i|u_i = j, \phi^{(j)})P(\phi^{(j)}|\mathbf{u}_{-i}, \mathbf{s}_{-i})d\phi^{(j)}, \tag{2}$$

where $\phi^{(j)}$ is the multinomial distribution over salient entities associated with topic j, and the integral is over all such distributions. We can obtain the rightmost item from Bayes' rule

$$P(\phi^{(j)}|\mathbf{u}_{-i}, \mathbf{s}_{-i}) \propto P(\mathbf{s}_{-i}|\phi^{(j)}, \mathbf{u}_{-i})P(\phi_{(j)}). \tag{3}$$

Since $P(\phi^{(j)})$ is Dirichlet(β) and conjugate to $P(\mathbf{s}_{-i}|\phi^{(j)}, \mathbf{u}_{-i})$, the posterior distribution $P(\phi^{(j)}|\mathbf{u}_{-i}, \mathbf{s}_{-i})$ will be Dirichlet($\beta + n_{-i,j}^{(s_i)}$), where $n_{-i,j}^{(s_i)}$ is the number of instances of salient entity s assigned to topic j, not including the current salient entity.

Since the first term on the right-hand side of Equation 2 is just $\phi_{s_i}^{(j)}$, we can complete the integral to obtain

$$P(s_i|u_i = j, \mathbf{u}_{-i}, \mathbf{s}_{-i}) = \frac{n_{-i,j}^{(s_i)} + \beta}{n_{-i,j}^{(\cdot)} + V_S\beta}, \tag{4}$$

where $n_{-i,j}^{(\cdot)}$ is the total number of salient entities assigned to topic j, not including the current one.

For the second item in Equation 1, we have

$$P(u_i = j|\mathbf{u}_{-i}) = \int P(u_i = j|\theta_d)P(\theta_d|\mathbf{u}_{-i})d\Phi_d = \frac{n_{-i,j}^{(d_i,s_i)} + \alpha}{n_{-i,\cdot}^{(d_i,s_i)} + K\alpha}, \tag{5}$$

where θ_d is the topic distribution of document d, $n_{-i,j}^{(d_i,s_i)}$ is the number of times salient entities from document d_i assigned to topic j except the current salient entity, and $n_{-i,\cdot}^{(d_i,s_i)}$ is the total number of salient entities in document d_i except the current one.

Putting together the results in Equations 4 and 5, we obtain the conditional probability

$$P(u_i = j|\mathbf{u}_{-i}, \mathbf{s}) \propto \frac{n_{-i,j}^{(s_i)} + \beta}{n_{-i,j}^{(\cdot)} + V_S\beta} \frac{n_{-i,j}^{(d_i,s_i)} + \alpha}{n_{-i,\cdot}^{(d_i,s_i)} + K\alpha}. \tag{6}$$

4.1.2 Sampling word topics y

The conditional posterior of assignment y_i to the i-th word in document d is:

$$P(y_i = j|\mathbf{y}_{-i}, \mathbf{w}) \propto P(w_i|y_i = j, \mathbf{y}_{-i}, \mathbf{w}_{-i})P(y_i = j|\mathbf{y}_{-i}), \tag{7}$$

where $\mathbf{y}_{-i}$ is the topic assignments of all words except the i-th one. The first item on the right-hand side is a likelihood and the second is a prior. By following a similar line of reasoning as from Equations (2) to (4), we have

$$P(w_i|y_i = j, \mathbf{y}_{-i}, \mathbf{w}_{-i}) = \frac{n_{-i,j}^{(w_i)} + \gamma}{n_{-i,j}^{(\cdot)} + V_W\gamma}. \tag{8}$$

For the second item in Equation (7), by integrating over the multinomial distribution over topics for the document from which w_i is drawn, specified by Φ_d, we obtain

$$P(y_i = j|\mathbf{y}_{-i}) = \int P(y_i = j|\Phi_d)P(\Phi_d|\mathbf{y}_{-i})d\Phi_d, \tag{9}$$

where Φ_d is a combination of the document and salient entities in the document. In particular, the influence of the topic distribution of the document is weighted by λ compared with the influence from salient entities, and the topic distribution of salient entities is equally weighted. Finally, we have Φ_d represented as

$$\Phi_d = \lambda\theta_d + (1 - \lambda)\frac{1}{|S_d|}\sum_{s\in S_d}\rho_s.$$

Since $P(\theta_d)$ and $P(\rho_s)$ are Dirichlet priors Dir(α) and Dir(σ), the prior distribution $P(\Phi_t)$ is $\lambda\alpha + (1 - \lambda)\sigma$. Since Φ_d is conjugate to the likelihood function (the first item in Equation 9), the posterior distribution in Equation (9) is as follows:

$$Dir(\lambda\alpha + (1 - \lambda)\sigma + \lambda n_{-i,j}^{(d_i,w_i)} + (1 - \lambda)\frac{1}{|S_d|}\sum_{s\in S_d} n_{-i,j}^{s}),$$

where $n_{-i,j}^{(d_i,w)}$ is the number of words assigned to topic j in document d_i except the current instance, and $n_{-i,j}^{s}$ is the number of instances of salient entity s assigned to topic j, except the current

instance. Then by Dirichlet-multinomial conjugate, we have

$$P(y_i = j|\mathbf{y}_{-i}) = \frac{\lambda\alpha + (1-\lambda)\sigma + \lambda n_{-i,j}^{(d_i,w_i)} + (1-\lambda)(\frac{1}{|S_{d_i}|}\sum_{s\in S_{d_i}} n_{-i,j}^{s})}{K(\lambda\alpha + (1-\lambda)\sigma) + \lambda n_{-i,\cdot}^{(d_i,w_i)} + (1-\lambda)\frac{1}{|S_{d_i}|}\sum_{s\in S_{d_i}} n_{-i,\cdot}^{s}}. \tag{10}$$

4.1.3 Sampling entity topics z

The conditional posterior of assignment z_i to the i-th entity in document d is:

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{e}) \propto P(e_i|z_i = j, \mathbf{z}_{-i}, \mathbf{e}_{-i})P(z_i = j|\mathbf{z}_{-i}), \tag{11}$$

where $\mathbf{z}_{-i}$ is the topic assignments of all entities except the i-th one. The first item on the right-hand side is a likelihood and the second is a prior.

By following a similar line of reasoning as from Equations (2) to (4), we have

$$P(e_i|z_i = j, \mathbf{z}_{-i}, \mathbf{e}_{-i}) = \frac{n_{-i,j}^{(e_i)} + \delta}{n_{-i,j}^{(\cdot)} + V_E\delta}, \tag{12}$$

where $n_{-i,j}^{(\cdot)}$ is the total number of entities assigned to topic j, not including the current one.

By following the steps we followed to derive Equation (10) from Equation (9), we have

$$P(z_i = j|\mathbf{z}_{-i}) = \frac{\lambda\alpha + (1-\lambda)\sigma + \lambda n_{-i,j}^{(d_i,e_i)} + (1-\lambda)(\frac{1}{|S_{d_i}|}\sum_{s\in S_{d_i}} n_{-i,j}^{s})}{K(\lambda\alpha + (1-\lambda)\sigma) + \lambda n_{-i,\cdot}^{(d_i,e_i)} + (1-\lambda)\frac{1}{|S_{d_i}|}\sum_{s\in S_{d_i}} n_{-i,\cdot}^{s}}. \tag{13}$$

4.2 Inference of SETM-WO

The Gibbs sampling process for SETM-WO is similar to SETM-WE, except that there is no sampling process for entity topic assignments in SETM-WO. In other words, the process of sampling entity topics does not exist in the inference process for SETM-WO because entities are not distinguished from non-entity words in SETM-WO.

5. Experimental setup

In the remainder of the paper we address the following research questions: (RQ1) How does SETM compare to state-of-the-art ETMs in terms of perplexity? (RQ2) How does SETM perform in the task of entity salience detection? (RQ3) Why can SETM achieve better performance in distinguishing salient entities from non-salient entities?

5.1 Datasets

The dataset used in the experiments aimed at answering our research questions is the New York Times corpus, with salience annotations provided by Dunietz and Gillick (2014). We refer to this dataset as the NYT-Sal dataset. Annotations were automatically generated by aligning the entities in the abstract and the document and assuming that every entity occurring in the abstract is salient. The New York Times dataset consists of two partitions. Documents from 2003 to 2006 are used as the training set, while documents in 2007 are used as the test set. The number of documents in the training set and test set are 80,667 and 9,706, respectively. We further

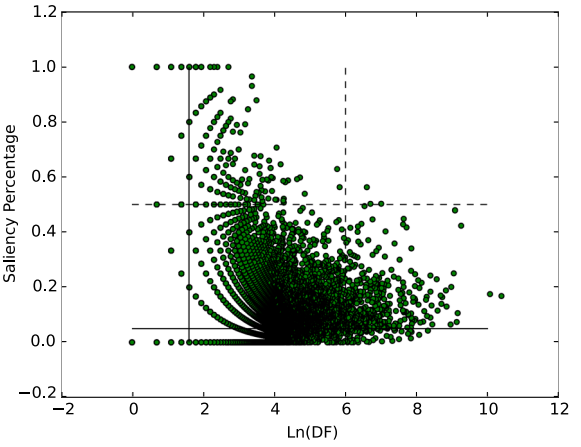

Figure 2. Scatter plot of log DF and SP.

split the training set into a smaller training set (80%) for model training and a validation set (20%) for parameter selection. The size of the word vocabulary is 621, 724, including 189, 480 entities.

To analyze the performance on different type of entities, we categorize entities based on their document frequency (DF) and salience. In particular, we define the salience percentage (SP) of an entity e, $sp_e = \frac{SDF_e}{DF_e}$, as the percentage of the documents in which entities appear and are labeled as salient, where SDF_e is the number of documents in which entity e is salient. The SP and the log of DF for each entity in the collection are shown as a scatter plot in Figure 2. We choose two threshold values to define high and low salience entities and high and low frequency entities. The lower and upper salience thresholds are set to 0.05 and 0.5 respectively, indicated by the red solid line $y = 0.05$ and the red dashed line $y = 0.05$. We define entities whose DF higher than 400 (approximately 5% of all training documents) or lower than 5 as head and tail entities respectively. The thresholds are indicated by the solid blue line $x = 1.6$ ($\ln(5) = 1.6$) and the dashed blue line $x = 6$ ($\ln(400) = 6$).

We consider entities that satisfy the following conditions as "torso" entities: (1) entities for which SP is above 0.05 and below 0.5; (2) entities for which DF is above 400 and below 5. In other words, torso entities fall into the square formed by the lines in Figure 2.

5.2 Intrinsic evaluation

The first type of evaluation we conduct is an intrinsic evaluation. We quantify the ability of the SETM to represent entities and documents better than baseline ETMs by computing the similarity between topically similar entities, and the similarity between topically similar documents. We further quantify the ability of the SETM to generate new documents by computing the perplexity of our model. Instead of analyzing all entities, we focus on entities that are neither highly frequent High Document Frequency (HDF) nor rare Low Document Frequency (LDF). This way, we avoid any possible bias introduced by head or tail entities. We want to perform analysis on entities with neither abundant nor limited information.

5.2.1 Entity-to-entity topical similarity

First, we test the ability of our topic model to produce an effective representation of entities compared to the baseline models. We make the assumption that two ("torso") entities are topically similar if both entities are salient in more than 50% of the documents they co-occur. Out of all entity pairs, 141 fulfill this condition. We test our model against baseline models by computing

the cosine similarity of these entity pairs; the higher the computed similarity is, the better the topic model.

5.2.2 Document-to-document topical similarity

Second, we test the ability of our topic model to produce an effective representation of documents compared to baseline models. Given an entity e, we denote with D_e^s the set of documents where entity e is salient, and with D_e^{ns} the set of documents where entity e is not salient. To measure the topical coherence of a set of documents, we follow the definition of coherence score in Kulkarni *et al.* (2009), and define the topical coherence of a set of documents D related to e as

$$\textit{topical-coherence}(e, D) = \sum_{m=2}^{D} \frac{1}{m-1} \sum_{l=1}^{m-1} \cos(d_m, d_l).$$

Our hypothesis is that the topical coherence calculated by using the document representations learned by the SETM will be higher than baseline models, which means that our learned document representations are better in capturing topical similarity. We use the set of 567 "torso" entities.

5.2.3 Model perplexity

Perplexity is a standard measure for estimating the performance of a probabilistic model. We evaluate SETM by estimating the perplexity of unseen held-out documents given a set of training documents. A better model will have a lower perplexity of held-out documents on average. We follow the perplexity definition in Blei *et al.* (2003). For a test set of M documents, perplexity is defined as follows:

$$\textit{perplexity}\,(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}. \tag{14}$$

5.3 Extrinsic evaluation

The second type of evaluation we conduct is an extrinsic evaluation. We first quantify the usefulness of the document representations learned by the SETM by the task of entity salience detection. For each document, we measure the similarity between the document and its set of salient entities, and that of its set of non-salient entities. We further measure the divergence between these two similarities to identify the capability of our model in capturing the topical differences.

5.3.1 Entity salience detection

To evaluate the learned topic distributions of entities, we test our model on the task of entity salience detection. The goal of entity salience detection is to iterate over each entity in a document and identify whether the entity is salient or not.

Our classification setup is as follows. First of all, we train an SETM model using the training set and the information about the salience of entities in that set. Then, for each training instance (an entity document pair), the topic distribution representations of the entity and the document are used as features to train a classifier. For each test entity document pair, we infer the topic distribution of the document and make predictions about whether the entity is salient or not in the document. Since entity saliency information is document specific, we have no prior information about the saliency of an entity in the test documents during classification.

We assume that if a model learns better entity and document representations, it should achieve higher classification performance. It is important to note that in this work we do not compare our proposed method with the current state-of-the-art entity saliency systems, such as SWAT (Ponza,

Ferragina, and Piccinno 2018). This is due to the fact that the focus of this work is to model text in a more faithful way, around topics and salient entities, and use the task of salient entity detection as a way to compare the learned topic distributions of our model with that of baseline topic models, rather than improving the state-of-the-art performance over entity salience detection (which is approximately around 0.56 F1 score for part of our dataset).

Following Dunietz and Gillick (2014), we use a set of standard binary classification metrics, that is, recall, precision and F1, to quantify the classification performance. Note that since the majority of entities are non-salient, our metrics are calculated only over the positive class, that is, salient entities. Statistical significance of the observed differences between the performance of two methods is tested using a two-tailed paired t-test and is denoted by $^{\blacktriangle}$ for strong significance for $\alpha = 0.01$, and $^{\triangle}$ for weak significance for $\alpha = 0.05$. In our experiments, all models are tested for significance against the best performing baseline, CorrLDA2. In addition to evaluate the performance on all entities, we also analyze over head and tail entities as defined in Figure 2.

5.3.2 Document-entity similarity divergence analysis

To intuitively understand the performance, we analyze the topical similarity between salient entities and non-salient entities within individual documents.

The reason to perform an analysis on the basis of individual documents is that entity saliency is document specific. In other words, an entity could be salient in one document while not salient in another, which makes analyzing salient entities across document impossible. Ideally, we expect that the similarity between salient entities and documents is higher than that of non-salient entities and documents. By visualizing the divergence between these two similarities for each document, we can see how close we are to the ideal situation compared to baseline models.

Given a document d, we denote with E_s the set of salient entities, and with E_{ns} the set of non-salient entities. We calculate the similarity between each salient entity $s \in E_s$ and document d, and obtain the average similarity $avg\text{-}sim(E_s, d)$ across all salient entities and the document. We do the same for E_{ns} and obtain $avg\text{-}sim(E_{ns}, d)$. The assumption is that the better a model is, the larger the difference between $avg\text{-}sim(E_s, d)$ and $avg\text{-}sim(E_{ns}, d)$. Then, we calculate the *se-ne-divergence* as $avg\text{-}sim(E_s, d) - avg\text{-}sim(E_{ns}, d)$, and rank documents based on the divergence (which ranges from 1 to -1) in descending order. The higher the divergence value is, the better the model.

5.3.3 Entity topic analysis

Given an entity e, we have a collection of documents D_s where e is salient in $d \in D_s$ and another collection of documents D_n where e appears in $d \in D_n$ and is not salient. We first compute the average topic distribution of documents in D_s and D_n respectively to find topics that are most relevant with e. Then we present the top words under those relevant topics to see its relevance with the given entity. We choose entity *New York Jets*, a professional American football team located in New York, as an example. The size of D_s and D_n is 407 and 403, respectively.

5.4 Baselines and parameter settings

Table 2 lists the entity salience detection methods considered in our experiments. Since our goal is to evaluate the effectiveness of our topic model, we compare with existing topic models, such as LDA (Blei *et al.* 2003), LLDA (Erosheva *et al.* 2004), and CorrLDA2 (Newman *et al.* 2006). LDA is used as a simple baseline to showcase how a standard model without considering entities works in our setting.

Beyond the baselines mentioned, there is a growing body of work on topic models that involve entities (Jeong and Choi 2012). However, their focus is on sequential topic flows of entities and entity groups in a single document (Jeong and Choi 2012) or on dynamic topic hierarchies and

Table 2. Methods and baselines used for comparison

Acronym	Description	Ref.
LDA	Latent Dirichlet Allocation, which use latent topic distributions to represent documents	Blei *et al*. 2003
LLDA	Link-LDA, similar with LDA, except that it considers words and entities in documents separately	Erosheva *et al.* 2004
CorrLDA2	Correlated topic model, which models the correlation between word topics and entity topics	Newman *et al*. 2006
SETM-WO	Our proposed model with only one observed variable, that is, words	This paper
SETM-WE	Our proposed model with two observed variables, that is, words and entities	This paper

Table 3. Entity representation analysis

Model	Similarity
LDA	0.6960
LLDA	0.7240
CorrLDA2	0.1392
SETM-WO	0.7271
SETM-WE	0.7336

timeliness of news data (Hu *et al.* 2015). Our task and our focus are not on the dynamics of topics. Therefore, such methods are not included as baselines.

Finally, there is a work in the literature that explicitly focuses on entity salience detection, such as Dunietz and Gillick (2014). This work is not included in our comparison since they target developing discriminative models with a specific focus on entity salience detection. Our goal is different, that is, to evaluate topic distributions learned by topic models. A comparison with such algorithms is beyond the scope of this work.

Following the standard practice (Kim *et al.* 2012), we set the hyperparameters of the baseline methods and our models to predefined values. In LDA, LLDA, CorrLDA2, and our models, we set both α and β as 0.1. The number of iterations of Gibbs Sampling is set to 1,000 for all topic models. For perplexity analysis, we set the number of topics to 5, 10, 15, 20, 30, 40, 50, 80, and 100. For model analysis and extrinsic evaluation, we use the corresponding model trained with the number of topics set to 100.

6. Results

6.1 Intrinsic evaluation

6.1.1 Entity-to-entity similarity

The results on the entity representation analysis are presented in Table 3. The average similarity of LLDA is higher than LDA, indicating that by distinguishing words from entities as observed variables we obtain better entity representations. This is also demonstrated by the comparison between SETM-WO and SETM-WE. Further, we can observe that SETM-WO outperforms LDA and that SETM-WE outperforms LLDA. This demonstrates that incorporating entity salience information into the topic models can be helpful in learning good entity representations, regardless of the setting of observed variables in topic models. Here the entity representation learned by

Table 4. Document representation analysis

Model	Similarity SD	Similarity NSD	Ratio of difference
LDA	0.6585	0.4943	1.3322
LLDA	0.6906	0.5405	1.2778
CorrLDA2	0.9293	0.9301	0.9991
SETM-WO	0.6722	0.5141	1.3075
SETM-WE	0.6641	0.4916	1.3509

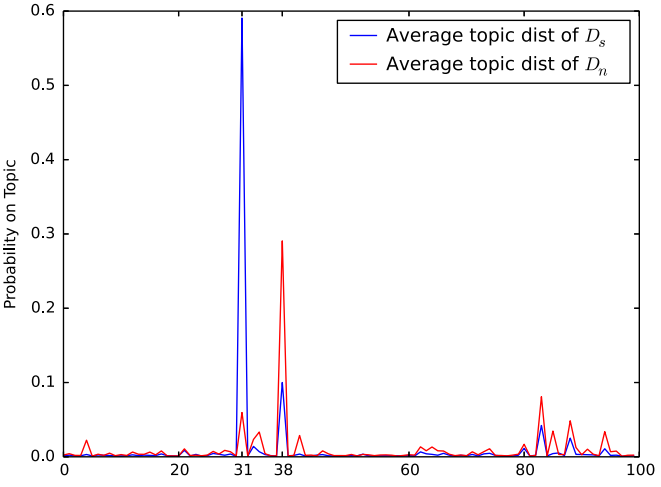

Figure 3. Average topic distribution between documents where entity *New York Jets* is salient and documents where it is not.

CorrLDA2 is not performing well. The reason might be that the entity topics are forced to align with word topics in documents, which makes entity representations meaningless.

6.1.2 Document-to-document similarity

The results on the document representation analysis are presented in Table 4. We expect the documents in D_e^s to be topically coherent, while documents in D_e^{ns} not. Therefore, the higher the value of *Similarity SD* the better, while the lower the value of *Similarity NSD* the better. To combine these two metrics, we calculate the ratio between them, and the higher the ratio the better. As we can see in the results, the ratio achieved by SETM-WE is the highest, which means that by considering entity salience information, our learned document representations can actually capture the similarity between similar documents better, and make dissimilar documents more distinguishable.

6.1.3 Entity topic analysis

We first present the average topic distribution between D_s and D_n in Figure 3. As we can see, topic 31 stands out in the blue line, indicating the relevance between entity *New York Jets* and the collection of documents where it is salient. Similarly, topic 38 is the most relevant topic in the red line. Note that the probability of topic 31 is close to 0.6 and much higher than that of topic 38, which indicates higher coherence within the salient documents of *New York Jets*.

Table 5. Top 10 words under topic 31 and topic 38 in a SETM model trained on the NYT-Sal dataset

Topic 31	Topic 38
Jets	Giants
West	Football
Team	Game
Edwards	Season
Stadium	Bowl
Club	Coach
Diamond	Team
South	nfl
East	Super
Game	Players

We present the top 10 words under topic 31 and 38 in Table 5. It is obvious that both topics are closely related to sports and American football. The difference is that topic 38 is a more general topic about National Football League (NFL), where words such as "super," "bowl," and "season" appear frequently. On the other hand, topic 31 is more relevant to entity *New York Jets*. "Jets" is one word in the name of the team, while "edwards" is the surname of a professional player of the team.[c] By analyzing on the basis of individual entity, we find that it is possible to explain the learned topics. Therefore, we consider it helpful to take entity salience into account in topic modeling whenever possible.

6.1.4 Perplexity

Figure 4 shows the perplexity values of our models and the baselines under different number of topics. Since the baseline models do not have entity salience information in their models, they cannot take advantage of salience labels. As we can see in Figure 4(a), our models and Link-LDA outperform LDA and CorrLDA2. For Link-LDA, the reason is that it distinguishes entities from words when learning topic distributions in documents. For the case of our models, it is better because the entity salience information is incorporated into the generative process of documents. Link-LDA performs slightly better than our models. This might be because during inference we assume no entity salience information, which has a negative impact on the inferred topic distributions of documents.

To study the perplexity of different observed variables, we present the perplexity of words, salient entities, and non-salient entities in Figure 4(b), (c), and (d), respectively. For LDA, the perplexity is lower for words, while much higher for salient or non-salient entities. This is not surprising since the number of words is larger than the number of entities in documents, and LDA is biased to be better at generating words than entities. For LLDA, CorrLDA2, and SETM-WE, the perplexity of entities is obviously lower than that of words, demonstrating the effective of distinguishing entities from words. Both of our model variants are better than the baseline models, showing that our model incorporates entity salience information into a topic model in an effective manner.

[c] https://en.wikipedia.org/wiki/Lac_Edwards.

Table 6. Performance of entity salience detection methods on the NYT-Sal dataset

	P	R	F1	P	R	F1
	All entities			Seen entities		
LDA	0.1362▲	0.8875▲	0.2361▲	0.1372▲	0.8334△	0.2348▲
LLDA	0.1606	0.5896	0.2493	0.1718▲	0.4673▲	0.2509
CorrLDA2	0.1544	0.6664	0.2507	0.1551	0.6659	0.2516
SETM-WO	0.1700▲	0.5184▲	0.2560▲	0.1717▲	0.5256▲	0.2589▲
SETM-WE	0.1718▲	0.5038▲	0.2562△	0.1736▲	0.5046▲	0.2583▲
	Head entities			Tail entities		
LDA	0.1598▲	0.8860▲	0.2708▲	0.1221▲	0.9067▲	0.2152▲
LLDA	0.1998	0.6273	0.3005	0.1294	0.4680	0.1990
CorrLDA2	0.1854	0.7484	0.2972	0.1269	0.5787	0.2081
SETM-WO	0.2348▲	0.5123▲	0.3220▲	0.1340▲	0.5261▲	0.2136
SETM-WE	0.2372▲	0.4878▲	0.3192▲	0.1347▲	0.4967▲	0.2120

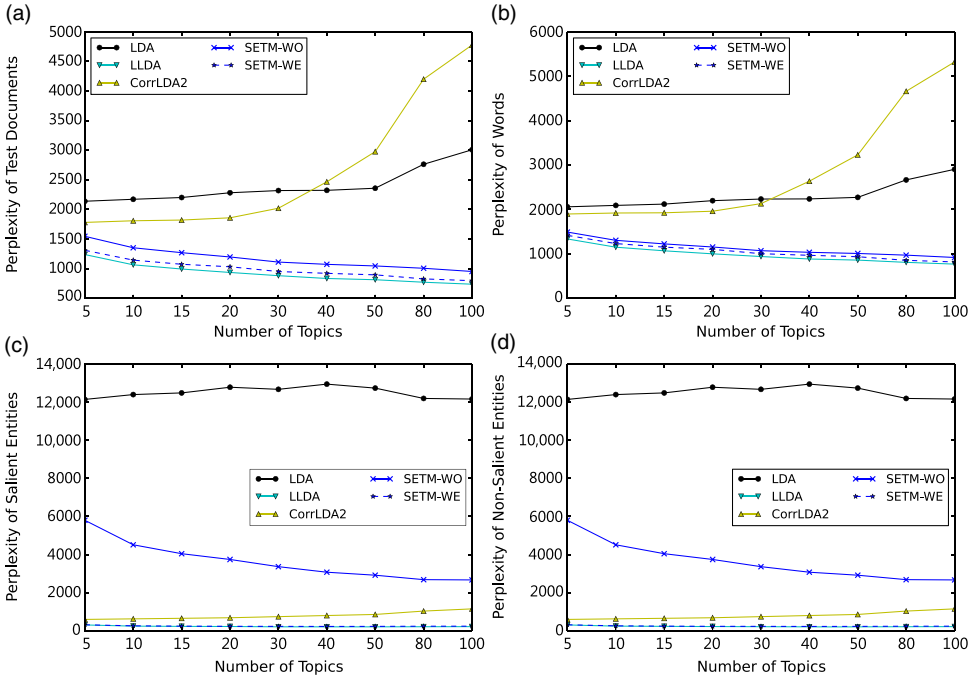

Figure 4. Perplexity of per document, salient entities, non-salient entities, and words.

6.2 Extrinsic evaluation: Entity salience detection

The overall results on the entity salience detection tasks are shown in Table 6. As we can see, the performance of our models on all entities is better than other methods in terms of F1. It demonstrates the effectiveness of our model by learning better topic distributions for entities and documents. Our model has the highest precision, but lower recall, which means that our model makes fewer positive predictions. This makes sense, since the dataset is biased to negative

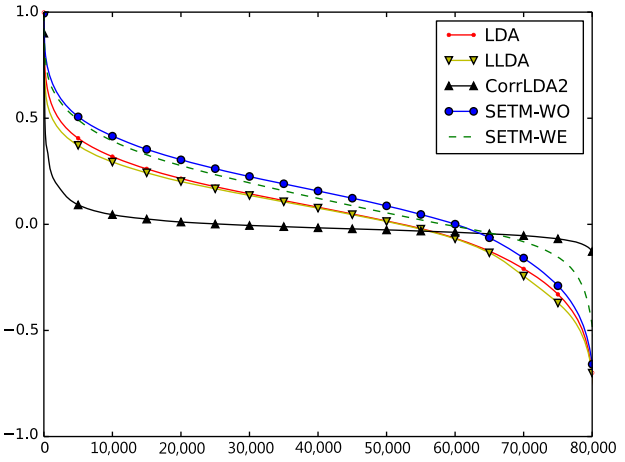

Figure 5. Topical similarity analysis on documents in training set. Each point is the *se-ne–divergence* of a document, as described in Section 5.3. Documents are ranked by their divergence values in descent order.

instances. Note that we are not comparing our work with the work in Dunietz and Gillick (2014) because their goal is to optimize for the task of entity salience detection, while our goal is to compare the entity and document representations.

Results on seen, head and tail entities are also shown in Table 6. As we expect, the overall performance on seen entities is better for all models. Compared to baseline models, the recall of our models is higher while sharing similar precision. The precision of head entities is significantly better than baseline models. The reason is that we have more training examples on positive and negative examples on entity salience for head entities. This demonstrates that with more training examples, our model learns the salience of entities better by showing better capability at predicting entity salience. For tail entities, the performance of all models are similar. This is because little information is available for tail entities, and the strength of our models cannot be leveraged by tail entities.

The result of topical similarity analysis within individual documents is shown in Figure 5. Ideally, we expect that all lines are above zero and as close to $y = 1$ as possible, indicating that for each document, the average similarity between salient entities in the document and the document is higher than that of non-salient entities. We can observe that the lines of our models: (1) are higher than baseline models, especially in the beginning; (2) cross the $y = 0$ line later than baseline models. This demonstrates that our models are more capable in distinguishing salient entities from non-salient entities. As we can see, CorrLDA2 shows relatively consistent behavior across documents. Together with the results of LLDA, they demonstrate that modeling entities in topic models might not help learning the salience of entities.

7. Conclusions

In this paper, we have proposed to incorporate entity salience information into topic models. A novel Salient Entity Topic Model (SETM) is proposed, which can explicitly model the generation of documents with salient entities under consideration. A Gibbs sampling-based algorithm is proposed for the parameter estimation of the model. We compare our model with several state-of-the-art baselines in terms of the generative capability. The evaluation shows that our model is better than the baselines, which demonstrates the effectiveness of incorporating entity salience information into document generative process. We also evaluate the learned document representations and entity representations by the task of entity salience detection. The results show that the representations of document and entities using our model can better distinguish salient entities out of non-salient entities compared to baseline representations.

Our model can be used for topic analysis with the increasingly available entity salience information, extracted from either web log (Gamon *et al.* 2013) or news corpus (Dunietz and Gillick

2014). As a potential application, by performing clustering on documents where a particular entity is salient, we might find different aspects of the entity by detecting the difference in learned topic distributions of documents.

One of the limitations of our model lies in the fact that training our model requires large scale and high quality labels of entity salience. However, this can be approximated by automatically mining salience information from existing data, such as the soft labeling approach in (Gamon *et al*. 2013), which we leave as future work.

Acknowledgements. We would like to thank our anonymous reviewers for helpful suggestions.

Financial support. This research was partially supported by Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), the China Scholarship Council, and the Innovation Center for Artificial Intelligence (ICAI). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Aletras N. and Mittal A. (2017). Labeling topics with images using a neural network. In *Advances in Information Retrieval 39th European Conference on IR Research*. Springer, pp. 500–505.

Andrzejewski D., Zhu X. and Craven M. (2009). Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *The 26th International Conference on Machine Learning*. Association for Computing Machinery (ACM), pp. 25–32.

Balog K. (2018). *Entity-Oriented Search*. Cham, Switzerland: Springer.

Bicalho P., Pita M., Pedrosa G., Lacerda A. and Pappa G.L. (2017). A general framework to expand short text for topic modeling. *Information Sciences* **393**, 66–81.

Blei D.M., Ng A.Y. and Jordan M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022.

Dunietz J. and Gillick D. (2014). A new entity salience task with millions of training examples. In *The European Chapter of the ACL*, vol. 14. Association for Computational Linguistics (ACL), pp. 205–209.

Erosheva E., Fienberg S. and Lafferty J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences* **101**(suppl 1), 5220–5227.

Escoter L., Pivovarova L., Du M., Katinskaia A. and Yangarber R. (2017). Grouping business news stories based on salience of named entities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), pp. 1096–1106.

Gamon M., Yano T., Song X., Apacible J. and Pantel P. (2013). Identifying salient entities in web pages. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM)*. Association for Computing Machinery (ACM), pp. 2375–2380.

Griffiths T.L. and Steyvers M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* **101**(suppl 1), 5228–5235.

Han X. and Sun L. (2012). An entity-topic model for entity linking. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics (ACL), pp. 105–115.

Hu L., Li J., Zhang J. and Shao C. (2015). o-hetm: An online hierarchical entity topic model for news streams. In *Advances in Knowledge Discovery and Data Mining 19th Pacific-Asia Conference*. Springer, pp. 696–707.

Hulpus I., Hayes C., Karnstedt M. and Greene D. (2013). Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery (ACM), pp. 465–474.

Jeong Y.-S. and Choi H.-J. (2012). Sequential entity group topic model for getting topic flows of entity groups within one document. In *Advances in Knowledge Discovery and Data Mining 16th Pacific-Asia Conference*. Springer, pp. 366–378.

Ji Z., Xu F., Wang B. and He B. (2012). Question-answer topic model for question retrieval in community question answering. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM)*. Association for Computing Machinery (ACM), pp. 2471–2474.

Kataria S.S, Kumar K.S., Rastogi R.R., Sen P. and Sengamedu S.H. (2011). Entity disambiguation with hierarchical topic models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery (ACM), pp. 1037–1045.

Kim H., Sun Y., Hockenmaier J. and Han J. (2012). Etm: Entity topic models for mining documents associated with entities. In *12th IEEE International Conference on Data Mining (ICDM)*. Institute of Electrical and Electronics Engineers, pp. 349–58.

Kulkarni S., Singh A., Ramakrishnan G. and Chakrabarti S. (2009). Collective annotation of Wikipedia entities in web text. In *Proceedings of the 18th ACM International Conference on Information and Knowledge Management (CIKM)*. Association for Computing Machinery (ACM), pp. 457–66.

Lau J.H., Grieser K., Newman D. and Baldwin T. (2011). Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), pp. 1536–1545.

Lauscher A., Nanni F., Fabo P.R. and Ponzetto S.P. (2016). Entities as topic labels: combining entity linking and labeled lda to improve topic interpretability and evaluability. *IJCol-Italian Journal of Computational Linguistics* **2**(2), 67–88.

Levit M., Parthasarathy S., Chang S., Stolcke A. and Dumoulin B. (2014). Word-phrase-entity language models: Getting more mileage out of n-grams. In *15th Annual Conference of the International Speech Communication Association*. International Speech Communication Association, pp. 666–670.

Li X., Ouyang J. and Zhou X. (2015a). Centroid prior topic model for multi-label classification. *Pattern Recognition Letters* **62**, 8–13.

Li X., Ouyang J. and Zhou X. (2015b). Supervised topic models for multi-label classification. *Neurocomputing* **149**, 811–819.

Li X., Wang Y., Zhang A., Li C., Chi J. and Ouyang J. (2018). Filtering out the noise in short text topic modeling. *Information Sciences* **456**, 83–96.

McCallum A. (1999). Multi-label text classification with a mixture model trained by em. In *AAAI workshop on Text Learning*. Association for the Advancement of Artificial Intelligence, pp. 1–7.

McCallum A., Corrada-Emmanuel A. and Wang X. (2005). The author-recipient-topic model for topic and role discovery in social networks, with application to enron and academic email. In *Proceedings of Workshop on Link Analysis, Counterterrorism and Security*, p. 33.

Newman D., Chemudugunta C. and Smyth P. (2006). Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery (ACM), pp. 680–686.

Ponza M., Ferragina P. and Piccinno F. (2018). Swat: A System for Detecting Salient Wikipedia Entities in Texts. arXiv preprint arXiv:1804.03580.

Qiu Z. and Shen H. (2017). User clustering in a dynamic social network topic model for short text streams. *Information Sciences* **414**, 102–116.

Ramage D., Hall D., Nallapati R. and Manning C.D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics (ACL), pp. 248–256.

Rosen-Zvi M., Griffiths T., Steyvers M. and Smyth P. (2004). The author-topic model for authors and documents. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence (2004)*. AUAI Press, pp. 487–494.

Rubin T.N., Chambers A., Smyth P. and Steyvers M. (2012). Statistical topic models for multi-label document classification. *Machine Learning* **88**(1–2), 157–208.

Shen W., Wang J., Luo P. and Wang M. (2013). Linking named entities in tweets with knowledge base via user interest modeling. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery (ACM), pp. 68–76.

Tran T.A., Niederée C., Kanhabua N., Gadiraju U. and Anand A. (2015). Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *Proceedings of the 24th ACM international Conference on Information and Knowledge Management (CIKM)*. Association for Computing Machinery (ACM), pp. 1201–1210.

Wang S., Chen Z. and Liu B. (2016). Mining aspect-specific opinion using a holistic lifelong topic model. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conference Committee, pp. 167–176.

Xie R., Liu Z., Jia J., Luan H. and Sun M. (2016). Representation learning of knowledge graphs with entity descriptions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence and the Twenty-Eighth Innovative Applications of Artificial Intelligence Conference*. Association for the Advancement of Artificial Intelligence, pp. 2659–2665.

Xiong C., Liu Z., Callan J. and Liu T.-Y. (2018). Towards better text understanding and retrieval through kernel entity salience modeling. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery (ACM), pp. 575–584.

Xu K., Qi G., Huang J. and Wu T. (2017). Incorporating Wikipedia concepts and categories as prior knowledge into topic models. *Intelligent Data Analysis* **21**(2), 443–461.

Zhang Y., Mao W. and Zeng D. (2016). A non-parametric topic model for short texts incorporating word coherence knowledge. In *Proceedings of the 25th ACM international Conference on Information and Knowledge Management (CIKM)*. Association for Computing Machinery (ACM), pp. 2017–2020.

Cite this article: Wu C, Kanoulas E and de Rijke M (2020). It all starts with entities: A Salient entity topic model. *Natural Language Engineering* **26**, 531–549. https://doi.org.10.1017/S1351324919000585