ES-LDA: Entity Summarization using Knowledge-based Topic Modeling

Seyedamin Pouriyeh[1], Mehdi Allahyari[*2], Krys Kochut[1], Gong Cheng[3], and Hamid Reza Arabnia[1]

[1]Computer Science Department, University of Georgia, Athens, GA, USA
{pouriyeh,kkochut,hra@uga.edu}
[2]Department of Computer Science, Georgia Sothern University, Statesboro, USA
{mallahyari@georgiasouthern.edu}
[3]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
{gcheng@nju.edu.cn}

Abstract

With the advent of the Internet, the amount of Semantic Web documents that describe real-world entities and their inter-links as a set of statements have grown considerably. These descriptions are usually lengthy, which makes the utilization of the underlying entities a difficult task. Entity summarization, which aims to create summaries for real world entities, has gained increasing attention in recent years. In this paper, we propose a probabilistic topic model, ES-LDA, that combines prior knowledge with statistical learning techniques within a single framework to create more reliable and representative summaries for entities. We demonstrate the effectiveness of our approach by conducting extensive experiments and show that our model outperforms the state-of-the-art techniques and enhances the quality of the entity summaries.

1 Introduction

With the emergence of Linked Open Data (LOD)[1] as a way of publishing and interacting with the information, many datasets such as DBpedia (Bizer et al., 2009) and YAGO (Hoffart et al., 2013) have been created and are publicly available on the Web. For example, DBpedia as part of LOD is a knowledge base extracted from Wikipedia that consists of Wikipedia resources (entities) described as RDF statements (i.e., RDF triples). The Resource Description Framework (RDF) is the Semantic Web standard data model used for representing information on the Web. An RDF triple is represented in the form of $< subject, predicate, object >$. The latest English version of DBpedia contains over 4.5 million entities collectively described by over 1.6 billion triples. This means that each entity description has an average of 355 RDF triples. Human users and computer applications need to consider these lengthy descriptions while performing various semantic tasks. Thus, *entity summarization*, a task of producing more concise, but still sufficient entity description, has garnered a significant amount of attention.

Recently, with the huge growth of information, summarization techniques are becoming some of the main approaches to making the information more readily available. In fact, summarization techniques aim to facilitate the identification of structure and meaning in data. Researchers in different communities have taken a strong interest in this task and, accordingly, have proposed various methods for a wide variety of summarization techniques in multiple areas. Document summarization (Nenkova and McKeown, 2012), database summarization (Bu et al., 2005), and graph summarization (Navlakha et al., 2008) are just a few examples that have been studied by different communities. RDF data summarization and in particular entity summarization, has attracted considerable attentions in recent years as it can benefit many other tasks in the natural language processing area, including entity recognition (Zhao and Kit, 2008), entity disambiguation (Dai et al., 2011), and many others. Several approaches have been developed to summarize RDF data with respect to entities, including RELIN (Cheng et al., 2011), FACES (Gunaratna et al., 2015), and LinkSUM (Thalhammer et al., 2016). RDF summarization differs from document summarization in the sense that RDF triples are structured and do not have many frequently used words to help the summarization task, which makes RDF

[*] Equal contribution

[1] http://linkeddata.org

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 316–325,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

summarization more challenging.

Topic modeling has become a popular method for uncovering the hidden themes from text corpora. Topic models usually consider each document as a mixture of topics, where a topic is a probability distribution over words. When the topic proportions of documents are estimated, they can be used as the themes (high-level semantics) of the documents. Topic models have been widely used for various text mining tasks, such as machine translation (Su et al., 2015), word embedding (Batmanghelich et al., 2016; Das et al., 2015), automatic topic labeling (Wan and Wang, 2016; Allahyari and Kochut, 2015; Allahyari et al., 2017b), and others(Allahyari et al., 2017a).

In this paper, we propose a novel topic model, called ES-LDA, that integrates prior knowledge with the topic modeling within a single framework for RDF entity summarization. In our approach, each entity, which is considered as a document, is a multinomial distribution over the predicates (properties), where each predicate is a probability distribution over the subjects and objects of the triples in the RDF data. We rank the triples based on their probability distributions and choose the top-k triples that best describe the underlying entity as its summary. We evaluated our approach against state-of-the-art techniques and our experiments indicate that our approach outperforms other methods in terms of the quality of summarization.

The rest of the paper is organized as follows: Section 2 presents an overview of related work. Section 3 introduces the baseline for this paper. In Section 4, we define the main problem and propose our model in detail and afterwards, in Section 5, we explain the configurations of our model and describe the experiments. Finally, in Sections 6 and Section 7, we discuss the results and conclude the paper, respectively.

2 Related Work

Summarization methods can be divided into two main categories, which are called extractive and none-extractive (abstractive) summarization. In extractive approaches, which are usually applicable in text and ontology summarization (Jones, 2007) (Zhang et al., 2007), a set of features is extracted directly from the input data. On the other hand, in non-extractive methods, which generally are employed in graph (Navlakha et al., 2008) and database (Bu et al., 2005) summarization, new sentences from the input data are generated (Hahn and Mani, 2000) to form a summary. In this research, we focus on extractive summarization. The concept of entity summarization in the form of RDF graph data has attracted more attention in recent years. Cheng et al. (Cheng et al., 2011) proposed entity summarization method, called RELIN, based on the PageRank algorithm to extract representative triples, called representative features for RDF graph entities. Because of the centrality based ranking issue, RELIN highlights the most similar and central triples, while in summarization, the diversity of summarized triples is the key point.

SUMMARUM (Thalhammer and Rettinger, 2014) is a system for a better navigation within Linked Data through the ranking of triples. This system also uses the PageRank algorithm to rank triples according to the popularity of resources with the help of Wikipedia pages. Two aforementioned approaches could not meet the diversity requirement in the summarization process. FACES (Gunaratna et al., 2015), on the other hand, tries to keep a balance between the centrality and diversity of the selected triples for each entity. It utilizes a clustering algorithm, called Cobweb (Fisher, 1987), to cluster related triples before ranking them to keep the diversity in the summarization. The recent version of SUMMARUM, which is called LinkSUM (Thalhammer et al., 2016), focused more on the objects instead of the diversity of properties for entities and showed a better result on the same dataset, in comparison with FACES. Beside the aforementioned techniques dedicated to entity summarization, there are various ranking models and tools, including TripleRank (Franz et al., 2009) and TRank (Tonon et al., 2013) that rank triples and concepts, respectively, incorporating ranking algorithms. However, Cheng et al. (2011) indicated that these methods are not appropriate for the entity summarization problem, which needs ranking of feature sets based on their importance to identify the underlying entity.

3 Preliminaries

An RDF data graph is a collection of nodes and edges that connect the nodes together. Nodes are usually recognized by unique IDs which are called *Uniform Resource Identifiers (URIs)* or exact values (i.e. numbers, dates, etc) namely *Lit-*

Table 1: J.C.Penny entity predicates and corresponding objects with the top-5 ES-LDA summary.

Predicate	Object	Top-5
http://dbpedia.org/property/areaServed	http://dbpedia.org/resource/United_States	✗
http://dbpedia.org/ontology/foundedBy	http://dbpedia.org/resource/James_Cash_Penney	✓
http://dbpedia.org/property/founder	http://dbpedia.org/resource/James_Cash_Penney	✗
http://dbpedia.org/ontology/industry	http://dbpedia.org/resource/Retail	✓
http://dbpedia.org/property/keyPerson	http://dbpedia.org/resource/Ron_Johnson	✓
http://dbpedia.org/property/homepage	http://www.jcpenney.com/	✗
http://dbpedia.org/ontology/location	http://dbpedia.org/resource/Plano,_Texas	✓
http://dbpedia.org/ontology/regionServed	http://dbpedia.org/resource/United_States	✗
http://dbpedia.org/property/tradedAs	http://dbpedia.org/resource/S&P_500	✗
http://dbpedia.org/ontology/type	http://dbpedia.org/resource/Public_company	✓

erals. An RDF graph is represented in a form of a collection of triples, each including a *Subject, Predicate, and Object*. In an RDF graph, an entity is defined as a subject with all predicates and corresponding objects to those predicates, collectively forming the entity's description. As Table 1 shows, the *J.C.Penny* entity is represented by its predicates (properties) and the corresponding objects in the triple format. For example, the triple $< J.C.Penny, industry, Retail >$ introduces *J.C.Penny*'s industry as *Retail* (due to space limitations we have dropped the first part of the *URIs*).

Definition 1 (Entity summary): Given an entity *e* and a positive integer *k*, a summary of the entity *e*, denoted *Sum(e, k)*, is the top-*k* subset of all predicates and corresponding objects that are most relevant to that entity. As Table1 shows the top-5 summary for *J.C.Penny* entity, which is represented through *foundedBy, industry, keyPerson, location, and type*.

3.1 Latent Dirichlet Allocation (LDA)

The Latent Dirichlet Allocation (LDA) is a generative probabilistic model for extracting thematic information (topics) from a collection of documents. LDA assumes that each document is made up of various topics, where each topic is a probability distribution over words.

Let $\mathcal{D} = \{d_1, d_2, \ldots, d_{|\mathcal{D}|}\}$ be a corpus of documents and $\mathcal{V} = \{w_1, w_2, \ldots, w_{|\mathcal{V}|}\}$ a vocabulary (words) of the corpus. A topic $z_j, 1 \leq j \leq K$ is represented as a multinomial probability distribution over the $|\mathcal{V}|$ words, $p(w_i|z_j), \sum_i^{|\mathcal{V}|} p(w_i|z_j) = 1$. LDA generates the words in a two-stage process: words are generated from topics and topics are generated by documents. More formally, the distribution of words, given the document, is calculated as follows:

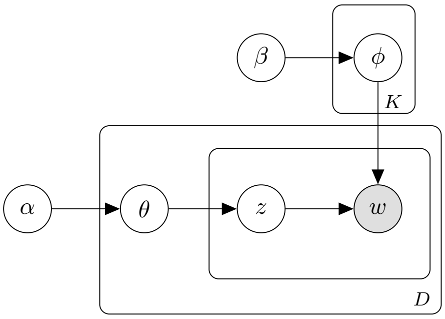

Figure 1: LDA Graphical Representation

$$p(w_i|d) = \sum_{j=1}^{K} p(w_i|z_j)p(z_j|d) \qquad (1)$$

The graphical model of LDA is shown in Figure 1 and the generative process for the corpus $\mathcal{D}$ is:

1. For each topic $k \in \{1, 2, \ldots, K\}$, sample a word distribution $\phi_k \sim \mathrm{Dir}(\beta)$
2. For each document $d \in \{1, 2, \ldots, \mathcal{D}\}$,
   (a) Sample a topic distribution $\theta_d \sim \mathrm{Dir}(\alpha)$
   (b) For each word w_n, where $n \in \{1, 2, \ldots, N\}$, in document d,
      i. Sample a topic $z_i \sim \mathrm{Mult}(\theta_d)$
      ii. Sample a word $w_n \sim \mathrm{Mult}(\phi_{z_i})$

In the LDA model, the word-topic distribution $p(w|z)$ and topic-document distribution $p(z|d)$ are learned entirely in an unsupervised manner, without any prior knowledge about what words are related to the topics and what topics are related to individual documents.

4 Problem Statement

In this section, we first describe the problem and then define how to utilize topic models for RDF graphs. Then, we formally introduce our ES-LDA model and explain how to integrate prior knowledge from RDF data graph within a topic model for entity summarization.

4.1 Problem Definition

Generating summaries for voluminous Semantic Web data, and in particular RDF data, for quick identification of entities has gained considerable attention as a challenging problem in the Semantic Web community. In the literature, *Entity Summarization* is defined as selecting a small but representative subset of the original triples associated with an entity. In this context, given an RDF data set comprising a collection of entities, where each entity is described by a set of its properties (i.e., all triples with the entity as the subject), our goal is to choose *top-k* representative triples for each entity. In other words, since all triples associated with an entity (as its description) share the same subject, our objective is to select *top-k* predicates and their corresponding objects among these triples that best summarize the entity's description.

4.2 Topic Models for RDF Graphs

Topic models were originally introduced for text documents, however, they have been applied to other types of data, such as images (Blei and Jordan, 2003), and recently (Sleeman et al., 2015) used topic modeling for RDF graphs. The first step in applying topic models is to define documents and word-like elements as the basic building blocks of documents. Since an RDF graph is usually represented as a set of triples, where each triple t consists of a subject s, predicate p, and an object o, in the form of $<s, p, o>$, we can consider a collection of such triples as a "document".

Definition 2 (document): A document d is defined as a set of triples, $d = \{t_1, t_2, \cdots, t_n\}$, that describe a single entity e. In other words, all triples of a document d have the same subject.

"Words" of a document can be extracted from different parts of its triples. We define a "**word**" w as the subject or object of a triple t in document d. Therefore, each document is represented by a "bag of words" including all the subjects and objects of its triples. In this paper, all subjects in the triples of a document are the same, because each document corresponds to a single entity, hence, in practice each document is a "bag of objects"[2]

Topic models usually utilize some data preprocessing, such as punctuation removal, downcasting, and abbreviation expansion, etc., to enhance the final performance. We also performed preprocessing on the RDF data and filtered out the schema and dataset dependent predicates, such as *sameAs, wikiPageExternalLink, subject, wikiPageWikiLink*, in addition to *literals*. Since we work with RDF graphs that differ from typical text documents in the sense that RDF data are represented as triples, we need to address several challenges mentioned in (Sleeman et al., 2015) to be able to run topic models on RDF data. These challenges include sparseness, use of unnatural language, and the lack of context. RDF data can be affected by **Sparseness**. We consider documents as sets of triples associated with a single entity. Such a set can be very large, leading to a large bag of words with a semantic theme, or small (sparse), resulting in a poor bag of words with less contextual information. It is also possible that a document with a high number of triples ends up having a small bag of words after pre-processing; for example based on Table 1, *J.C.Penny* entity comes with *United_States, James_Cash_Penney, Retail, Ron_Johnson, Plano,_Texas, United_States, S&P_500 and Public_company* as a bag of words for *J.C.Penny* entity, which shows sparseness in this document. **Unnatural Language** can be problematic for RDF data. A typical text document contains sentences where each sentence has a natural structure. These extra components of a sentence usually provide a further "**context**" for understanding words that are ambiguous or have multiple meanings, such as polysemous or homonymous ones. The aforementioned example for the *J.C.Penny* entity also confirms the unnatural language problem. The "**lack of context**" can further impact RDF data because they are potentially sparse, described by unnatural language, and often using words that have multiple meanings, difficult to differentiate (*J.C.Penny* bag of words example). Additionally, triples are more prone to pre-processing, because it is not uncom-

[2] "bag of words" and "bag of objects" are interchangeably used.

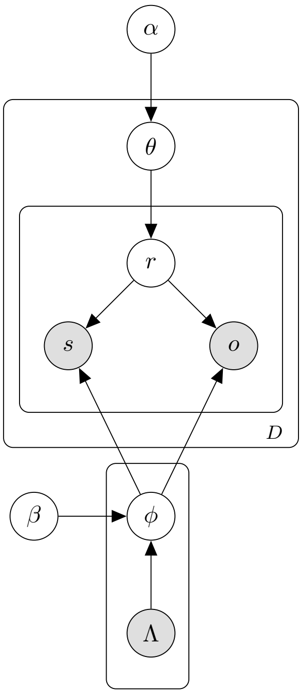

Figure 2: Entity Summarization Model

Algorithm 1: ES-LDA Model

```
foreach predicate $r \in \{1, 2, \ldots, R\}$ do
    Draw an object distribution $\phi_r \sim \mathrm{Dir}(\beta_r \times \Lambda_r)$
end
foreach document $d \in \{1, 2, \ldots, D\}$ do
    Draw a predicate distribution $\theta_d \sim \mathrm{Dir}(\alpha_d)$
    foreach subject $s$ and object $o$ of document $d$ do
        Draw a predicate $r \sim \mathrm{Mult}(\theta_d)$
        Draw a subject $s$ from predicate $r$, $s \sim \mathrm{Mult}(\phi_r)$
        Draw an object $o$ from predicate $r$, $o \sim \mathrm{Mult}(\phi_r)$
    end
end
```

mon for triples to contain unexpected characters. RDF data resemble short texts in terms of the aforementioned challenges. Sparseness in a short text causes the model to be less discriminative to recognize how words are related and the limited context makes it hard for the model to identify the meanings of the words in such short text documents (Yan et al., 2013). In order to alleviate these issues, researchers usually take two approaches. They either augment the short text or design custom versions of the LDA model that address their specific problems. In this paper, we have used both approaches. We describe how to supplement the RDF data in the following section and describe the details of our model in section 4.4.

4.3 Supplementing RDF Data

As topic modeling is based on statistics of the co-occurrence of terms (Sleeman et al., 2015), when we are dealing with short texts with a very limited number of repetitions, which is the case with RDF data, we need to find a way to supplement the data to elevate the performance of the topic modeling approach. We augment the documents using two different methods. In the first method, we increase the frequency of the words in each document. But the question is "*How many times each word of a document should be repeated?*". Entities in DBpedia have been organized into a category network, therefore, every entity has a number of categories associated with it. The relationship between an entity and a category is defined by the "*http://purl.org/dc/terms/subject*" predicate. Since each word of a document is an object of a triple, and accordingly, an entity in DBpedia, it is related to several categories. We assume that objects (words) of a document that have more categories are likely more important. Thus, We expand each document by increasing the frequency of each object by the number of its categories. In the second method, instead of repeating each object a certain number of times, we enlarge each document by adding categories of the objects as extra words, directly to the document. There are multiple advantages of supplementing each document by adding object categories: (i) the sparseness in the document, related to each entity, is lowered as we are adding a number of related words to it; (ii) we reduce the ambiguity in the document, because adding extra categories alleviates the lack of context and helps distinguish the appropriate meanings of the words with multiple connotations; and lastly (iii), adding object categories makes the documents semantically more relevant to their topical themes. We evaluated our model using both methods and the results demonstrate that the first method gives significantly better summaries than the second method.

4.4 Proposed Model

ES-LDA is a probabilistic generative model for modeling entities in RDF graphs. The key idea behind our model is twofold: (1) we exploit statistical topic models as the underlying quantitative framework for entity summarization; and (2) ES-LDA incorporates the prior knowledge from the RDF knowledge base directly into the topic model. The plate notation is shown in Figure 2.

In our model, each document is a multinomial distribution over the predicates. If we consider predicates as topics, at the document level, our model is the same as standard LDA. However, we set the number of topics in ES-LDA to be the number of unique predicates in the corpus. Unlike the standard LDA, where each topic is a multinomial distribution over the vocabulary from the Dirichlet prior β, in our model each predicate is a multinomial distribution over all the subjects and objects of the RDF graph. In our approach, a document consists of a set of triples describing a single entity, i.e. all these triples share the same subject. Thus, we constrain the documents to only have the objects of related triples and also restrict the predicates to be defined only over the objects. In addition, for each predicate r, we further smooth its distribution by Λ_r. Λ is a matrix that has encoded the background knowledge about predicate-object values from DBpedia. Section 4.5 explains how Λ is constructed. The generative process of ES-LDA is shown in Algorithm 1.

Following this process, the joint probability of generating a corpus $D = \{d_1, d_2, \ldots, d_{|D|}\}$, the predicate assignments $\mathbf{r}$ given the hyperparameters α, β and the prior matrix Λ is:

$$P(\mathbf{o}, \mathbf{s}, \mathbf{r}|\alpha, \beta, \Lambda)$$
$$= \int_\phi P(\phi|\beta; \Lambda) \prod_d \sum_{r_d} P(\mathbf{o_d}|r_d, \phi) P(\mathbf{s_d}|r_d, \phi)$$
$$\times \int_\theta P(\theta|\alpha) P(\mathbf{r_d}|\theta, \phi) d\theta d\phi \qquad (2)$$

4.5 Constructing Predicate-Object Prior Matrix Λ

In the ES-LDA model, each predicate has a probability distribution over the objects of the RDF graph. Entity summarization is the task of choosing the top-k predicate-object pairs that best describe an entity. Presumably, if an object is associated with more categories in DBpedia, it is likely more important. We create the the Λ matrix to encode the prior weight of the predicate-object pairs and utilize it to smooth the predicate-object distributions ϕ by incorporating this domain knowledge into the topic model. We build the Λ matrix of size $R \times O$, where R is the number of predicates and O is the number of objects in the RDF graph. Let f be an indicator function where $f(i, j) = 1$ if there is a triple in RDF graph with predicate i and object j, and 1 otherwise, for $1 \leq i \leq R$ and $1 \leq j \leq O$. Additionally, let c be the number of categories assigned to object j. Then, we define Λ_{ij} as follows:

$$\Lambda_{ij} = \begin{cases} c & \text{if } f(i,j) = 1 \\ 1 & \text{otherwise.} \end{cases} \qquad (3)$$

For example, the "*Barack_Obama*" entity has multiple predicate-object pairs in DBpedia, including "*profession-author*", "*profession-lawyer*" and "*profession-professor*" pairs. According to DBpedia, $c_{author} = 2$, $c_{lawyer} = 4$ and $c_{professor} = 2$. It is reasonable to expect a higher probability for the "*profession-lawyer*" pair as it seems to be slightly more important than the other two pairs for "*Barack_Obama*". As a result, $\Lambda_{profession-lawyer} = 4$, which promotes "*profession-lawyer*" in Eq. 5.

4.6 Inference using Gibbs Sampling

Since the posterior inference of the LDA is intractable, we need to find an algorithm for estimating the posterior inference. A variety of algorithms have been used to estimate the parameters of topic models, such as variational EM (Blei et al., 2003) and Gibbs sampling (Griffiths and Steyvers, 2004). In this paper we use the collapsed Gibbs sampling procedure for our ES-LDA topic model. Collapsed Gibbs sampling (Griffiths and Steyvers, 2004) is a Markov Chain Monte Carlo (MCMC) (Robert and Casella, 2004) algorithm, which constructs a Markov chain over the latent variables in the model and converges to the posterior distribution, after a number of iterations. In our case, we aim to construct a Markov chain that converges to the posterior distribution over $\mathbf{r}$ conditioned on observed subjects $\mathbf{s}$, objects $\mathbf{o}$, hyperparameters α, β, and the prior matrix Λ.

In our modified version of the learning algorithm to infer $p(o_i|r_j)$ and $p(r_j|d)$, we (1) constrain the objects that are not paired with a predicate to have 0 probability, i.e. $p(o_i|r_j) = 0$, if $(r_i, o_j) \notin$ RDF graph, and (2) $P(s|r_j) = 1$, since all the triples of a document have the same subject s. We derive the posterior inference from Eq. 2 as follows:

$$P(\mathbf{r}|\mathbf{o},\mathbf{s},\alpha,\beta,\Lambda) = \frac{P(\mathbf{r},\mathbf{o},\mathbf{s}|\alpha,\beta,\Lambda)}{P(\mathbf{o}|\alpha,\beta,\Lambda)} \propto P(\mathbf{r},\mathbf{o}|\alpha,\beta,\Lambda) \propto P(\mathbf{r})P(\mathbf{o}|\mathbf{r})P(\mathbf{s}|\mathbf{r}) \tag{4}$$

$$P(r_i = r|o_i = o, \mathbf{r}_{-i}, \mathbf{o}_{-i}, \alpha, \beta, \Lambda) \propto \frac{n^{(d)}_{r,-i} + \alpha_r}{\sum_{r'}(n^{(d)}_{r',-i} + \alpha_{r'})} \times \frac{n^{(r)}_{o,-i} + \Lambda_{ro}\beta_o}{\sum_{o'}(n^{(r)}_{o',-i} + \Lambda_{ro}\beta_o)} \tag{5}$$

where $n^{(r)}_o$ is the number of times object o is assigned to predicate r. $n^{(d)}_r$ denotes the number of times predicate r is associated with document d. The subscript $-i$ indicates that the contribution of the current object o_i being sampled is removed from the counts. After Gibbs sampling, we can use the sampled predicate to estimate the probability of a predicate, given a document, θ_{dr} and the probability of an object, given a predicate, ϕ_{ro}:

$$\theta_{dr} = \frac{n^{(d)}_r + \alpha_r}{\sum_{r'}(n^{(d)}_{r'} + \alpha_{r'})} \tag{6}$$

$$\phi_{ro} = \frac{n^{(r)}_o + \Lambda_{ro}\beta_o}{\sum_{o'}(n^{(r)}_{o'} + \Lambda_{ro}\beta_o)} \tag{7}$$

5 Experiments

We evaluated our ES-LDA model against the state-of-the-art LinkSUM (Thalhammer et al., 2016) and FACES (Gunaratna et al., 2015) systems. Our goal was to show that the ES-LDA model produces results that are closer to human judgment, in comparison with the other approaches. We used the same dataset[3] that was used in the experiments conducted with FACES, as well as LinkSUM models. The dataset contained 50 entities randomly selected from DBpedia (English version 3.9) in domains including *politician, actors, scientist, song, film, country, city, river, company, game, etc.*. 15 people in the field of Semantic Web were selected as reviewers and each entity was evaluated by at least 7 reviewers to produce the top-5 and top-10 summaries. The average number of properties for each entity was 44.

Based on the two types of RDF supplement methods we discussed in 4.3, we applied two different configurations for the proposed model. In the first experiment, ES-LDA @config-1, we configured the system to supplement each entity (document) by repeating each object based on the *number of categories* that the object has in the DBpedia knowledge base. For example, for the triple $< J.C.Penney, industry, Retail >$ we repeated *Retail* object, 5 times in that document, as *Retail* has five different categories in DBpedia (i.e. *"Retailers, Retailing, French words and phrases, Merchandising, Marketing"*)

In the second experiment, ES-LDA @config-2, we configured the system to supplement each entity (document) by adding the corresponding category(ies) of each object into the document. In this case, each entity is defined as a bag of words including objects and categories of each object. For example, for the aforementioned triple, in addition to the *Retail* we included *"Retailers, Retailing, French words and phrases, Merchandising, Marketing"* as the corresponding categories to the *Retail* object.

For the other parameters, we assumed a symmetric Dirichlet prior and set $\beta = 0.01$ and $\alpha = 50/R$, where R is the total number of unique predicates. We ran the Gibbs sampling algorithm for 1000 iterations and computed the posterior inference after the last sampling iteration. We selected the top-5 and top-10 most probable properties for each entity and calculate the quality of the summary for each entity through equation 8.

$$Quality(Sum(e)) = \frac{1}{n}\sum_{i=1}^{n}|Sum(e) \cap Sum^{I}_{i}(e)| \tag{8}$$

In our experiments, we used the quality of the summary proposed in (Cheng et al., 2011), in which n ideal summaries $Sum^I_i(e)$ generated by expert users for $i = 1, ..., n$ and the summaries generated by the system $Sum(e)$ were compared. The average of the overlap between an ideal summary and a summary generated by the system is denoted as the quality of the summary, which is $0 \leq Quality(Sum(e)) \leq k$ in the top-k settings.

5.1 Experiment Results

The summary in our model is defined as sets of representative triples that can summarize each entity (sets of triples with the same subject) in a way close to a human-created summary. We decided to use the last part of a *URI* to compare the generated summaries with the expert summaries and produce

[3] http://wiki.knoesis.org/index.php/FACES

Table 2: Overall quality results of different models. Best result are bold.

Model	**Top-5**	**Top-10**
ES-LDA @ config-1	**1.20**	**3.50**
ES-LDA @ config-2	1.10	3.26
LinkSUM@ config-1	1.20	3.15
LinkSUM@ config-2	1.20	3.20
FACES	0.93	2.92

the Summary Quality for each entity and average them. As (Thalhammer et al., 2016) reproduced the FACES overall Summary Quality based on this criteria and also applied it to their model, we decided to use their result as it was completely aligned with our summary definition.

In Table 2, we compare the quality of the results from LinkSUM, FACES, and ES-LDA with two distinct configurations (supplementing by object reputation and object categories). As Table 2 shows, the quality of our model outperforms the FACES approach, in both cases. The ES-LDA @ config-2 demonstrates a comparable result with the two configurations of LinkSUM, while ES-LDA @ config-1 outperforms LinkSUM. For some of the entities, the predicates that ES-LDA selected as top-5 most probable did not exist in the FACES dataset. It forced us to calculate the quality of summary for some of the entities with just 4 predicates instead of 5. We believe to be the only reason why top-5 Quality of Summary was lower than or equal to LinkSUM. Although, we had the same issue for the top-10 results, overall, ES-LDA shows a better performance in two configurations.

6 Discussion

We evaluated our approach against the state-of-the-art summarization techniques, including LinkSUM and FACES. LinkSUM primarily focuses on the most relevant facts for each entity, while FACES tries to keep a balance between diversity and relevancy in entity summarization. There is usually a trade-off between diversity and relevancy of the selected predicates. Our ES-LDA model maintains both diversity and relevancy, while representing each entity through *top-k* predicates. As shown in Table 2, our model outperforms the state-of-the-art approaches. Table 3 illustrates a sample of entities from the dataset along with their top-10 predicates, for all approaches. As Table 3 shows, the LinkSUM model is focusing more on the *objects*, while predicate repetition is permitted. For example, *<Marie_Curie, birthPlace, Warsaw>*, *<Marie_Curie, birthPlace, Russian_Empire>*, and *<Marie_Curie, birthPlace, Congress_Polandare>* are representing *Marie_Curie*'s birth place. Although, they differ in terms of objects, it is arguable that referring to the same predicate with multiple objects that are more likely relevant reduces the chance of other important triples that could potentially appear in the summary. It should be noted that in the current ES-LDA configuration, we have not considered predicate repetition, thus, all the predicates of the triples appearing in the resultant summary are unique. FACES on the other hand, considers predicate diversity and tries to keep a balance between the diversity and relevancy but the overall quality of the FACES model is lower than LinkSUM and ES-LDA. In the FACES model, there are selected predicates which seems to be less informative in the sense to be top-10 representative for a particular entity. For example, *<Marie_Curie, thumbnail, 200px-Marie_Curie_c1920.png>*, which is referring to a *png* file, could be replaced with more descriptive one. Additionally, our proposed technique features several unique characteristics: (1) the ES-LDA is a *knowledge-based probabilistic* model that combines prior knowledge with statistical learning technique into a unified framework for entity summarization; (2) for each entity, it ranks all predicates based on their importance by computing marginal probabilities for the predicates. Table 4 illustrates the top-5 predicates for a sample of two entities; and finally (3), each predicate can be represented as a probability distribution over objects in the ES-LDA model, which allows us to describe the relations (predicates) of the RDF graph based on its nodes as shown in Table 5.

7 Conclusions

We have proposed a knowledge-based probabilistic topic model, called ES-LDA, based on the RDF entity representation for entity summarization. In our experiments, we have applied two different configurations: one based on object repetitions and the other based on adding object's categories, to alleviate common RDF data problems including

Table 3: Top-10 predicates for three randomly selected entities after applying three different models.

MARIE CURIE			REIGN OF FIRE			SEYCHELLES		
ES-LDA	**LinkSUM**	**FACES**	**ES-LDA**	**LinkSUM**	**FACES**	**ES-LDA**	**LinkSUM**	**FACES**
doctoralStudents	birthPlace	spouse	starring	country	starring	leaderName	largestCity	leaderName
doctoralAdvisor	birthPlace	field	producer	starring	country	governmentType	governmentType	governmentType
deathPlace	field	workInstitutions	music	starring	distributor	leaderTitle	governmentType	largestCity
children	field	birthPlace	director	starring	musicComposer	officialLanguage	governmentType	sovereigntyType
knownFor	knownFor	deathPlace	cinematography	studio	director	capital	governmentType	source
spouse	almaMater	doctoralAdvisor	country	producer	editing	currency	sovereigntyType	capital
almaMater	birthPlace	knownFor	distributor	producer	studio	timeZone	source	leaderTitle
birthPlace	knownFor	almaMater	studio	director	music	legislature	capital	language
field	doctoralAdvisor	doctoralStudents	editing	artist	producer	anthem	language	languages
establishedEvent	knownFor	thumbnail	screenplay	producer	thumbnail	callingCode	timeZone	legislature

Table 4: Probabilities of top-5 predicates for two randomly selected entities.

LEXUS		MORTAL KOMBAT TRILOGY	
Predicate	**Probability**	**Predicate**	**Probability**
foundedBy	0.21	platforms	0.30
owner	0.17	publisher	0.18
location	0.15	developer	0.17
keyPerson	0.06	computingMedia	0.07
service	0.04	designer	0.05

Table 5: Distributions of two randomly selected predicates over top-5 objects.

PARTY		STARRING	
Object	**Probability**	**Object**	**Probability**
Democratic Party (United States)	0.36	Arnold Schwarzenegger	0.05
Republican Party (United States)	0.17	Angelina Jolie	0.04
Democratic-Republican Party	0.12	Raven Symone	0.03
Communist Party of the Soviet Union	0.08	Matthew McConaughey	0.02
Independent(politician)	0.08	Alan Arkin	0.02

sparseness, unnatural language, and lack of context. We conducted extensive experiments, which show the quality of the top-10 triples in both configurations outperforms the state-of-the-art techniques, LinkSUM and FACES, while for the top-5 quality we surpassed FACES and equaled the LinkSUM results.

There are many interesting future research directions of this work. It would be interesting to investigate how this model and a much richer set of topic models that combine prior knowledge with statistical learning techniques could be used for various tasks in the Semantic Web domain, such as ontology summarization, ontology tagging, and finding similar ontologies.

Acknowledgments

Gong Cheng was partially funded by the NSFC under Grant 61572247.

References

Mehdi Allahyari and Krys Kochut. 2015. Automatic topic labeling using ontology-based topic models. In *14th International Conference on Machine Learning and Applications (ICMLA), 2015*. IEEE.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017a. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919* .

Mehdi Allahyari, Seyedamin Pouriyeh, Krys Kochut, and Hamid R Arabnia. 2017b. A knowledge-based topic modeling approach for automatic topic labeling. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS* 8(9):335–349.

Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. 2016. Nonparametric spherical topic modeling with word embeddings. *arXiv preprint arXiv:1604.00126* .

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web* 7(3):154–165.

David M Blei and Michael I Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, pages 127–134.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Shaofeng Bu, Laks VS Lakshmanan, and Raymond T Ng. 2005. Mdl summarization with holes. In *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, pages 433–444.

Gong Cheng, Thanh Tran, and Yuzhong Qu. 2011. Relin: relatedness and informativeness-based centrality for entity summarization. *The Semantic Web–ISWC 2011* pages 114–129.

Hong-Jie Dai, Richard Tzong-Han Tsai, Wen-Lian Hsu, et al. 2011. Entity disambiguation using a markov-logic network. In *IJCNLP*. pages 846–855.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian lda for topic models with word embeddings. In *ACL (1)*. pages 795–804.

Douglas H Fisher. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine learning* 2(2):139–172.

Thomas Franz, Antje Schultz, Sergej Sizov, and Steffen Staab. 2009. Triplerank: Ranking semantic web data by tensor decomposition. *The Semantic Web-ISWC 2009* pages 213–228.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America* 101(Suppl 1):5228–5235.

Kalpa Gunaratna, Krishnaprasad Thirunarayan, and Amit P Sheth. 2015. Faces: diversity-aware entity summarization using incremental hierarchical conceptual clustering .

Udo Hahn and Inderjeet Mani. 2000. The challenges of automatic summarization. *Computer* 33(11):29–36.

Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence* 194:28–61.

Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Information Processing & Management* 43(6):1449–1481.

Saket Navlakha, Rajeev Rastogi, and Nisheeth Shrivastava. 2008. Graph summarization with bounded error. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, pages 419–432.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining text data*, Springer, pages 43–76.

Christian P Robert and George Casella. 2004. *Monte Carlo statistical methods*, volume 319. Citeseer.

Jennifer Sleeman, Tim Finin, and Anupam Joshi. 2015. Topic modeling for rdf graphs. In *LD4IE@ ISWC*. pages 48–62.

Jinsong Su, Deyi Xiong, Yang Liu, Xianpei Han, Hongyu Lin, Junfeng Yao, and Min Zhang. 2015. A context-aware topic model for statistical machine translation. In *ACL (1)*. pages 229–238.

Andreas Thalhammer, Nelia Lasierra, and Achim Rettinger. 2016. Linksum: using link analysis to summarize entity data. In *International Conference on Web Engineering*. Springer, pages 244–261.

Andreas Thalhammer and Achim Rettinger. 2014. Browsing dbpedia entities with summaries. In *European Semantic Web Conference*. Springer, pages 511–515.

Alberto Tonon, Michele Catasta, Gianluca Demartini, Philippe Cudré-Mauroux, and Karl Aberer. 2013. Trank: Ranking entity types using the web of data. In *International Semantic Web Conference*. Springer, pages 640–656.

Xiaojun Wan and Tianming Wang. 2016. Automatic labeling of topic models using text summaries. In *ACL (1)*.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, pages 1445–1456.

Xiang Zhang, Gong Cheng, and Yuzhong Qu. 2007. Ontology summarization based on rdf sentence graph. In *Proceedings of the 16th international conference on World Wide Web*. ACM, pages 707–716.

Hai Zhao and Chunyu Kit. 2008. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *IJCNLP*. pages 106–111.