Unsupervised Terminological Ontology Learning based on Hierarchical Topic Modeling

Abstract

In this paper, we present hierarchical relationbased latent Dirichlet allocation (hrLDA), a data-driven hierarchical topic model for extracting terminological ontologies from a large number of heterogeneous documents. In contrast to traditional topic models, hrLDA relies on noun phrases instead of unigrams, considers syntax and document structures, and enriches topic hierarchies with topic relations. Through a series of experiments, we demonstrate the superiority of hrLDA over existing topic models, especially for building hierarchies. Furthermore, we illustrate the robustness of hrLDA in the settings of noisy data sets, which are likely to occur in many practical scenarios. Our ontology evaluation results show that ontologies extracted from hrLDA are very competitive with the ontologies created by domain experts.

Introduction

Although researchers have made significant progress on knowledge acquisition and have proposed many ontologies, for instance, WordNet BIBREF0 , DBpedia BIBREF1 , YAGO BIBREF2 , Freebase, BIBREF3 Nell BIBREF4 , DeepDive BIBREF5 , Domain Cartridge BIBREF6 , Knowledge Vault BIBREF7 , INS-ES BIBREF8 , iDLER BIBREF9 , and TransE-NMM BIBREF10 , current ontology construction methods still rely heavily on manual parsing and existing knowledge bases. This raises challenges for learning ontologies in new domains. While a strong ontology parser is effective in small-scale corpora, an unsupervised model is beneficial for learning new entities and their relations from new data sources, and is likely to perform better on larger corpora.

In this paper, we focus on unsupervised terminological ontology learning and formalize a terminological ontology as a hierarchical structure of subject-verb-object triplets. We divide a terminological ontology into

two components: topic hierarchies and topic relations. Topics are presented in a tree structure where each node is a topic label (noun phrase), the root node represents the most general topic, the leaf nodes represent the most specific topics, and every topic is composed of its topic label and its descendant topic labels. Topic hierarchies are preserved in topic paths, and a topic path connects a list of topics labels from the root to a leaf. Topic relations are semantic relationships between any two topics or properties used to describe one topic. Figure FIGREF1 depicts an example of a terminological ontology learned from a corpus about European cities. We extract terminological ontologies by applying unsupervised hierarchical topic modeling and relation extraction to plain text.

Topic modeling was originally used for topic extraction and document clustering. The classical topic model, latent Dirichlet allocation (LDA) BIBREF11 , simplifies a document as a bag of its words and describes a topic as a distribution of words. Prior research BIBREF12 , BIBREF13 , BIBREF14 , BIBREF15 , BIBREF16 , BIBREF17 , BIBREF18 has shown that LDA-based approaches are adequate for (terminological) ontology learning. However, these models are deficient in that they still need human supervision to decide the number of topics, and to pick meaningful topic labels usually from a list of unigrams. Among models not using unigrams, LDA-based Global Similarity Hierarchy Learning (LDA+GSHL) BIBREF13 only extracts a subset of relations: “broader" and “related" relations. In addition, the topic hierarchies of KB-LDA BIBREF17 rely on hypernym-hyponym pairs capturing only a subset of hierarchies.

Considering the shortcomings of the existing methods, the main objectives of applying topic modeling to ontology learning are threefold.

To achieve the first objective, we extract noun phrases and then propose a sampling method to estimate the number of topics. For the second objective, we use language parsing and relation extraction to learn relations for the noun phrases. Regarding the third objective, we adapt and improve the hierarchical latent

Dirichlet allocation (hLDA) model BIBREF19 , BIBREF20 . hLDA is not ideal for ontology learning because it builds topics from unigrams (which are not descriptive enough to serve as entities in ontologies) and the topics may contain words from multiple domains when input data have documents from many domains (see Section SECREF2 and Figure FIGREF55). Our model, hrLDA, overcomes these deficiencies. In particular, hrLDA represents topics with noun phrases, uses syntax and document structures such as paragraph indentations and item lists, assigns multiple topic paths for every document, and allows topic trees to grow vertically and horizontally.

The primary contributions of this work can be specified as follows.

The rest of this paper is organized into five parts. In Section 2, we provide a brief background of hLDA. In Section 3, we present our hrLDA model and the ontology generation method. In Section 4, we demonstrate empirical results regarding topic hierarchies and generated terminological ontologies. Finally, in Section 5, we present some concluding remarks and discuss avenues for future work and improvements.

Background

In this section, we introduce our main baseline model, hierarchical latent Dirichlet allocation (hLDA), and some of its extensions. We start from the components of hLDA - latent Dirichlet allocation (LDA) and the Chinese Restaurant Process (CRP)- and then explain why hLDA needs improvements in both building hierarchies and drawing topic paths.

LDA is a three-level Bayesian model in which each document is a composite of multiple topics, and every topic is a distribution over words. Due to the lack of determinative information, LDA is unable to distinguish different instances containing the same content words, (e.g. “I trimmed my polished nails" and

“I have just hammered many rusty nails"). In addition, in LDA all words are probabilistically independent and equally important. This is problematic because different words and sentence elements should have different contributions to topic generation. For instance, articles contribute little compared to nouns, and sentence subjects normally contain the main topics of a document.

Introduced in hLDA, CRP partitions words into several topics by mimicking a process in which customers sit down in a Chinese restaurant with an infinite number of tables and an infinite number of seats per table. Customers enter one by one, with a new customer choosing to sit at an occupied table or a new table. The probability of a new customer sitting at the table with the largest number of customers is the highest. In reality, customers do not always join the largest table but prefer to dine with their acquaintances. The theory of distance-dependent CRP was formerly proposed by David Blei BIBREF21 . We provide later in Section SECREF15 an explicit formula for topic partition given that adjacent words and sentences tend to deal with the same topics.

hLDA combines LDA with CRP by setting one topic path with fixed depth INLINEFORM0 for each document. The hierarchical relationships among nodes in the same path depend on an INLINEFORM1 dimensional Dirichlet distribution that actually arranges the probabilities of topics being on different topic levels. Despite the fact that the single path was changed to multiple paths in some extensions of hLDA - the nested Chinese restaurant franchise processes BIBREF22 and the nested hierarchical Dirichlet Processes BIBREF23 , - this topic path drawing strategy puts words from different domains into one topic when input data are mixed with topics from multiple domains. This means that if a corpus contains documents in four different domains, hLDA is likely to include words from the four domains in every topic (see Figure FIGREF55). In light of the various inadequacies discussed above, we propose a relation-based model, hrLDA. hrLDA incorporates semantic topic modeling with relation extraction to integrate syntax and has the capacity to provide comprehensive hierarchies even in corpora containing mixed topics.

Hierarchical Relation-based Latent Dirichlet Allocation

The main problem we address in this section is generating terminological ontologies in an unsupervised fashion. The fundamental concept of hrLDA is as follows. When people construct a document, they start with selecting several topics. Then, they choose some noun phrases as subjects for each topic. Next, for each subject they come up with relation triplets to describe this subject or its relationships with other subjects. Finally, they connect the subject phrases and relation triplets to sentences via reasonable grammar. The main topic is normally described with the most important relation triplets. Sentences in one paragraph, especially adjacent sentences, are likely to express the same topic.

We begin by describing the process of reconstructing LDA. Subsequently, we explain relation extraction from heterogeneous documents. Next, we propose an improved topic partition method over CRP. Finally, we demonstrate how to build topic hierarchies that bind with extracted relation triplets.

Relation-based Latent Dirichlet Allocation

Documents are typically composed of chunks of texts, which may be referred to as sections in Word documents, paragraphs in PDF documents, slides in presentation documents, etc. Each chunk is composed of multiple sentences that are either atomic or complex in structure, which means a document is also a collection of atomic and/or complex sentences. An atomic sentence (see module INLINEFORM0 in Figure FIGREF10) is a sentence that contains only one subject (INLINEFORM1), one object (INLINEFORM2) and one verb (INLINEFORM3) between the subject and the object. For every atomic sentence whose object is also a noun phrase, there are at least two relation triplets (e.g., “The tiger that gave the excellent speech is handsome" has relation triplets: (tiger, give, speech), (speech, be given by, tiger), and (tiger, be, handsome)). By contrast, a complex sentence can be subdivided into multiple atomic sentences. Given that the syntactic verb in a relation triplet is determined by the subject and the

object, a document INLINEFORM4 in a corpus INLINEFORM5 can be ultimately reduced to INLINEFORM6 subject phrases (we convert objects to subjects using passive voice) associated with INLINEFORM7 relation triplets INLINEFORM8 . Number INLINEFORM9 is usually larger than the actual number of noun phrases in document INLINEFORM10 . By replacing the unigrams in LDA with relation triplets, we retain definitive information and assign salient noun phrases high weights.

We define INLINEFORM0 as a Dirichlet distribution parameterized by hyperparameters INLINEFORM1 , INLINEFORM2 as a multinomial distribution parameterized by hyperparameters INLINEFORM3 , INLINEFORM4 as a Dirichlet distribution parameterized by INLINEFORM5 , and INLINEFORM6 as a multinomial distribution parameterized by INLINEFORM7 . We assume the corpus has INLINEFORM8 topics. Assigning INLINEFORM9 topics to the INLINEFORM10 relation triplets of document INLINEFORM11 follows a multinomial distribution INLINEFORM12 with prior INLINEFORM13 . Selecting the INLINEFORM14 relation triplets for document INLINEFORM15 given the INLINEFORM16 topics follows a multinomial distribution INLINEFORM17 with prior INLINEFORM18 . We denote INLINEFORM19 as the list of relation triplet lists extracted from all documents in the corpus, and INLINEFORM20 as the list of topic assignments of INLINEFORM21 . We denote the relation triplet counts of documents in the corpus by INLINEFORM22 . The graphical representation of the relation-based latent Dirichlet allocation (rLDA) model is illustrated in Figure FIGREF10 .

The plate notation can be decomposed into two types of Dirichlet-multinomial conjugated structures: document-topic distribution INLINEFORM0 and topic-relation distribution INLINEFORM1 . Hence, the joint distribution of INLINEFORM2 and INLINEFORM3 can be represented as DISPLAYFORM0

where INLINEFORM0 is the number of unique relations in all documents, INLINEFORM1 is the number of occurrences of the relation triplet INLINEFORM2 generated by topic INLINEFORM3 in all documents, and INLINEFORM4 is the number of relation triplets generated by topic INLINEFORM5 in document

INLINEFORM6 . INLINEFORM7 is a conjugate prior for INLINEFORM8 and thus the posterior distribution is a new Dirichlet distribution parameterized by INLINEFORM9 . The same rule applies to INLINEFORM10 .

Relation Triplet Extraction

Extracting relation triplets is the essential step of hrLDA, and it is also the key process for converting a hierarchical topic tree to an ontology structure. The idea is to find all syntactically related noun phrases and their connections using a language parser such as the Stanford NLP parser BIBREF24 and Ollie BIBREF25 . Generally, there are two types of relation triplets:

Subject-predicate-object-based relations,

e.g., New York is the largest city in the United States INLINEFORM0 (New York, be the largest city in, the United States);

Noun-based/hidden relations,

e.g., Queen Elizabeth INLINEFORM0 (Elizabeth, be, queen).

A special type of relation triplets can be extracted from presentation documents such as those written in PowerPoint using document structures. Normally lines in a slide are not complete sentences, which means language parsing does not work. However, indentations and bullet types usually express inclusion relationships between adjacent lines. Starting with the first line in an itemized section, our algorithm scans the content in a slide line by line, and creates relations based on the current item and the item that is one level higher.

Acquaintance Chinese Restaurant Process

As mentioned in Section 2, CRP always assigns the highest probability to the largest table, which assumes customers are more likely to sit at the table that has the largest number of customers. This ignores the social reality that a person is more willing to choose the table where his/her closest friend is sitting even though the table also seats unknown people who are actually friends of friends. Similarly with human-written documents, adjacent sentences usually describe the same topics. We consider a restaurant table as a topic, and a person sitting at any of the tables as a noun phrase. In order to penalize the largest topic and assign high probabilities to adjacent noun phrases being in the same topics, we introduce an improved partition method, Acquaintance Chinese Restaurant Process (ACRP).

The ultimate purposes of ACRP are to estimate INLINEFORM0 , the number of topics for rLDA, and to set the initial topic distribution states for rLDA. Suppose a document is read from top to bottom and left to right. As each noun phrase belongs to one sentence and one text chunk (e.g., section, paragraph and slide), the locations of all noun phrases in a document can be mapped to a two-dimensional space where sentence location is the x axis and text chunk location is the y axis (the first noun phrase of a document holds value (0, 0)). More specifically, every noun phrase has four attributes: content, location, one-to-many relation triplets, and document ID. Noun phrases in the same text chunk are more likely to be “acquaintances;" they are even closer to each other if they are in the same sentence. In contrast to CRP, ACRP assigns probabilities based on closeness, which is specified in the following procedure.

Let INLINEFORM0 be the integer-valued random variable corresponding to the index of a topic assigned to the INLINEFORM1 phrase. Draw a probability INLINEFORM2 from Equations EQREF18 to EQREF25 below for the INLINEFORM3 noun phrase INLINEFORM4 , joining each of the existing INLINEFORM5 topics and the new INLINEFORM6 topic given the topic assignments of previous INLINEFORM7 noun phrases, INLINEFORM8 . If a noun phrase joins any of the existing k topics, we denote the corresponding

topic index by INLINEFORM9 .

The probability of choosing the INLINEFORM0 topic: DISPLAYFORM0

The probability of selecting any of the INLINEFORM0 topics:

if the content of INLINEFORM0 is synonymous with or an acronym of a previously analyzed noun phrase INLINEFORM1 INLINEFORM2 in the INLINEFORM3 topic, DISPLAYFORM0

else if the document ID of INLINEFORM0 is different from all document IDs belonging to the INLINEFORM1 topic, DISPLAYFORM0

otherwise, DISPLAYFORM0

where INLINEFORM0 refers to the current number of noun phrases in the INLINEFORM1 topic, INLINEFORM2 represents the vector of chunk location differences of the INLINEFORM3 noun phrase and all members in the INLINEFORM4 topic, INLINEFORM5 stands for the vector of sentence location differences, and INLINEFORM6 is a penalty factor.

Normalize the (INLINEFORM0) probabilities to guarantee they are each in the range of [0, 1] and their sum is equal to 1.

Based on the probabilities EQREF18 to EQREF25 , we sample a topic index INLINEFORM0 from INLINEFORM1 for every noun phrase, and we count the number of unique topics INLINEFORM2 in the end. We shuffle the order of documents and iterate ACRP until INLINEFORM3 is unchanged.

Nested Acquaintance Chinese Restaurant Process

The procedure for extending ACRP to hierarchies is essential to why hrLDA outperforms hLDA. Instead of a predefined tree depth INLINEFORM0 , the tree depth for hrLDA is optional and data-driven. More importantly, clustering decisions are made given a global distribution of all current non-partitioned phrases (leaves) in our algorithm. This means there can be multiple paths traversed down a topic tree for each document. With reference to the topic tree, every node has a noun phrase as its label and represents a topic that may have multiple sub-topics. The root node is visited by all phrases. In practice, we do not link any phrases to the root node, as it contains the entire vocabulary. An inner node of a topic tree contains a selected topic label. A leaf node contains an unprocessed noun phrase. We define a hashmap INLINEFORM1 with a document ID as the key and the current leaf nodes of the document as the value. We denote the current tree level by INLINEFORM2 . We next outline the overall algorithm.

We start with the root node (INLINEFORM0) and apply rLDA to all the documents in a corpus.

Collect the current leaf nodes of every document. INLINEFORM0 initially contains all noun phrases in the corpus. Assign a cluster partition to the leaf nodes in each document based on ACRP and sample the cluster partition until the number of topics of all noun phrases in INLINEFORM1 is stable or the iteration reaches the predefined number of iteration times (whichever occurs first).

Mark the number of topics (child nodes) of parent node INLINEFORM0 at level INLINEFORM1 as INLINEFORM2 . Build a INLINEFORM3 - dimensional topic proportion vector INLINEFORM4 based on INLINEFORM5 .

For every noun phrase INLINEFORM0 in document INLINEFORM1 , form the topic assignments INLINEFORM2 based on INLINEFORM3 .

Generate relation triplets from INLINEFORM0 given INLINEFORM1 and the associated topic vector INLINEFORM2 .

Eliminate partitioned leaf nodes from INLINEFORM0 . Update the current level INLINEFORM1 by 1.

If phrases in INLINEFORM0 are not yet completely partitioned to the next level and INLINEFORM1 is less than INLINEFORM2 , continue the following steps. For each leaf node, we set the top phrase (i.e., the phrase having the highest probability) as the topic label of this leaf node and the leaf node becomes an inner node. We next update INLINEFORM3 and repeat procedures INLINEFORM4 .

To summarize this process more succinctly: we build the topic hierarchies with rLDA in a divisive way (see Figure FIGREF35). We start with the collection of extracted noun phrases and split them using rLDA and ACRP. Then, we apply the procedure recursively until each noun phrase is selected as a topic label. After every rLDA assignment, each inner node only contains the topic label (top phrase), and the rest of the phrases are divided into nodes at the next level using ACRP and rLDA. Hence, we build a topic tree with each node as a topic label (noun phrase), and each topic is composed of its topic labels and the topic labels of the topic's descendants. In the end, we finalize our terminological ontology by linking the extracted relation triplets with the topic labels as subjects.

We use collapsed Gibbs sampling BIBREF26 for inference from posterior distribution INLINEFORM0 based on Equation EQREF11 . Assume the INLINEFORM1 noun phrase INLINEFORM2 in parent node INLINEFORM3 comes from document INLINEFORM4 . We denote unassigned noun phrases from document INLINEFORM5 in parent node INLINEFORM6 by INLINEFORM7 , and unique noun phrases in parent node INLINEFORM8 by INLINEFORM9 . We simplify the probability of assigning the INLINEFORM10 noun phrase in parent node INLINEFORM11 to topic INLINEFORM12 among INLINEFORM13 topics as DISPLAYFORM0

where INLINEFORM0 refers to all topic assignments other than INLINEFORM1 , INLINEFORM2 is multinational document-topic distribution for unassigned noun phrases INLINEFORM3 , INLINEFORM4 is the multinational topic-relation distribution for topic INLINEFORM5 , INLINEFORM6 is the number of occurrences of noun phrase INLINEFORM7 in topic INLINEFORM8 except the INLINEFORM9 noun phrase in INLINEFORM10 , INLINEFORM11 stands for the number of times that topic INLINEFORM12 occurs in INLINEFORM13 excluding the INLINEFORM14 noun phrase in INLINEFORM15 . The time complexity of hrLDA is INLINEFORM16 , where INLINEFORM17 is the number of topics at level INLINEFORM18 . The space complexity is INLINEFORM19 .

In order to build a hierarchical topic tree of a specific domain, we must generate a subset of the relation triplets using external constraints or semantic seeds via a pruning process BIBREF27 . As mentioned above, in a relation triplet, each relation connects one subject and one object. By assembling all subject and object pairs, we can build an undirected graph with the objects and the subjects constituting the nodes of the graph BIBREF28 . Given one or multiple semantic seeds as input, we first collect a set of nodes that are connected to the seed(s), and then take the relations from the set of nodes as input to retrieve associated subject and object pairs. This process constitutes one recursive step. The subject and object pairs become the input of the subsequent recursive step.

Implementation

We utilized the Apache poi library to parse texts from pdfs, word documents and presentation files; the MALLET toolbox BIBREF29 for the implementations of LDA, optimized_LDA BIBREF30 and hLDA; the Apache Jena library to add relations, properties and members to hierarchical topic trees; and Stanford Protege for illustrating extracted ontologies. We make our code and data available . We used the same empirical hyper-parameter setting (i.e., INLINEFORM0 , INLINEFORM1 , and INLINEFORM2) across all our experiments. We then demonstrate the evaluation results from two aspects: topic hierarchy and

ontology rule.

Hierarchy Evaluation

In this section, we present the evaluation results of hrLDA tested against optimized_LDA, hLDA, and phrase_hLDA (i.e., hLDA based on noun phrases) as well as ontology examples that hrLDA extracted from real-world text data. The entire corpus we generated contains 349,362 tokens (after removing stop words and cleaning) and is built from articles on INLINEFORM0 INLINEFORM1 . It includes 84 presentation files, articles from 1,782 Wikipedia pages and 3,000 research papers that were published in IEEE manufacturing conference proceedings within the last decade. In order to see the performance in data sets of different scales, we also used a smaller corpus Wiki that holds the articles collected from the Wikipedia pages only.

We extract a single level topic tree using each of the four models; hrLDA becomes rLDA, and phrase_hLDA becomes phrase-based LDA. We have tested the average perplexity and running time performance of ten independent runs on each of the four models BIBREF31 , BIBREF32 . Equation EQREF41 defines the perplexity, which we employed as an empirical measure. DISPLAYFORM0

where INLINEFORM0 is a vector containing the INLINEFORM1 relation triplets in document INLINEFORM2 , and INLINEFORM3 is the topic assignment for INLINEFORM4 .

The comparison results on our Wiki corpus are shown in Figure FIGREF42 . hrLDA yields the lowest perplexity and reasonable running time. As the running time spent on parameter optimization is extremely long (the optimized_LDA requires 19.90 hours to complete one run), for efficiency, we adhere to the fixed parameter settings for hrLDA.

Superiority

Figures FIGREF43 to FIGREF49 illustrates the perplexity trends of the three hierarchical topic models (i.e., hrLDA, phrase_hLDA and hLDA) applied to both the Wiki corpus and the entire corpus with INLINEFORM0 “chip" given different level settings. From left to right, hrLDA retains the lowest perplexities compared with other models as the corpus size grows. Furthermore, from top to bottom, hrLDA remains stable as the topic level increases, whereas the perplexity of phrase_hLDA and especially the perplexity of hLDA become rapidly high. Figure FIGREF52 highlights the perplexity values of the three models with confidence intervals in the final state. As shown in the two types of experiments, hrLDA has the lowest average perplexities and smallest confidence intervals, followed by phrase_hLDA, and then hLDA.

Our interpretation is that hLDA and phrase_hLDA tend to assign terms to the largest topic and thus do not guarantee that each topic path contains terms with similar meaning.

Robustness

Figure FIGREF55 shows exhaustive hierarchical topic trees extracted from a small text sample with topics from four domains: INLINEFORM0 , INLINEFORM1 INLINEFORM2 , INLINEFORM3 , and INLINEFORM4 . hLDA tends to mix words from different domains into one topic. For instance, words on the first level of the topic tree come from all four domains. This is because the topic path drawing method in existing hLDA-based models takes words in the most important topic of every document and labels them as the main topic of the corpus. In contrast, hrLDA is able to create four big branches for the four domains from the root. Hence, it generates clean topic hierarchies from the corpus.

Gold Standard-based Ontology Evaluation

The visualization of one concrete ontology on the INLINEFORM0 INLINEFORM1 domain is presented in Figure FIGREF60 . For instance, Topic packaging contains topic integrated circuit packaging, and topic label jedec is associated with relation triplet (jedec, be short for, joint electron device engineering council).

We use KB-LDA, phrase_hLDA, and LDA+GSHL as our baseline methods, and compare ontologies extracted from hrLDA, KB-LDA, phrase_hLDA, and LDA+GSHL with DBpedia ontologies. We use precision, recall and F-measure for this ontology evaluation. A true positive case is an ontology rule that can be found in an extracted ontology and the associated ontology of DBpedia. A false positive case is an incorrectly identified ontology rule. A false negative case is a missed ontology rule. Table TABREF61 shows the evaluation results of ontologies extracted from Wikipedia articles pertaining to European Capital Cities (Corpus E), Office Buildings in Chicago (Corpus O) and Birds of the United States (Corpus B) using hrLDA, KB-LDA, phrase_hLDA (tree depth INLINEFORM0 = 3), and LDA+GSHL in contrast to these gold ontologies belonging to DBpedia. The three corpora used in this evaluation were collected from Wikipedia abstracts, the same text source of DBpedia. The seeds of hrLDA and the root concepts of LDA+GSHL are capital, building, and bird. For both KB-LDA and phrase_hLDA we kept the top five tokens in each topic as each node of their topic trees is a distribution/list of phrases. hrLDA achieves the highest precision and F-measure scores in the three experiments compared to the other models. KB-LDA performs better than phrase_hLDA and LDA+GSHL, and phrase_hLDA performs similarly to LDA+GSHL. In general, hrLDA works well especially when the pre-knowledge already exists inside the corpora. Consider the following two statements taken from the corpus on Birds of the United States as an example. In order to use two short documents “The Acadian flycatcher is a small insect-eating bird." and “The Pacific loon is a medium-sized member of the loon." to infer that the Acadian flycatcher and the Pacific loon are both related to topic bird, the pre-knowledge that “the loon is a species of bird" is required for hrLDA. This example explains why the accuracy of extracting ontologies from this kind of corpus is low.

Concluding Remarks

In this paper, we have proposed a completely unsupervised model, hrLDA, for terminological ontology learning. hrLDA is a domain-independent and self-learning model, which means it is very promising for learning ontologies in new domains and thus can save significant time and effort in ontology acquisition.

We have compared hrLDA with popular topic models to interpret how our algorithm learns meaningful hierarchies. By taking syntax and document structures into consideration, hrLDA is able to extract more descriptive topics. In addition, hrLDA eliminates the restrictions on the fixed topic tree depth and the limited number of topic paths. Furthermore, ACRP allows hrLDA to create more reasonable topics and to converge faster in Gibbs sampling.

We have also compared hrLDA to several unsupervised ontology learning models and shown that hrLDA can learn applicable terminological ontologies from real world data. Although hrLDA cannot be applied directly in formal reasoning, it is efficient for building knowledge bases for information retrieval and simple question answering. Also, hrLDA is sensitive to the quality of extracted relation triplets. In order to give optimal answers, hrLDA should be embedded in more complex probabilistic modules to identify true facts from extracted ontology rules. Finally, one issue we have not addressed in our current study is capturing pre-knowledge. Although a direct solution would be adding the missing information to the data set, a more advanced approach would be to train topic embeddings to extract hidden semantics.

Acknowledgments

This work was supported in part by Intel Corporation, Semiconductor Research Corporation (SRC). We are obliged to Professor Goce Trajcevski from Northwestern University for his insightful suggestions and discussions. This work was partly conducted using the Protege resource, which is supported by grant

GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.