***allahyari_2017_a_knowledge_based_topic_modeling_approach_for_automatic_topic_labeling

Year

2017

Author(s)

Mehdi Allahyari and Seyedamin Pouriyeh and Krys Kochut and Hamid Reza Arabnia

Title

A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling

Venue

International Journal of Advanced Computer Science and Applications

---

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel approach

Underlying technique

Knowledge (Ontology)-based topic model (KB-LDA)

Topic labeling parameters

Label generation

In the proposed model, we define another latent (i.e. hidden) variable called, concept, between topics and words.

Thus, each document is a mixture of topics, while each topic is made up of concepts, and finally, each concept is a probability distribution over the vocabulary.

TABLE IV. EXAMPLE OF TOPIC-WORD REPRESENTATION LEARNED BY LDA AND TOPIC-CONCEPT REPRESENTATION LEARNED BY KB-LDA.

LDA		KB-LDA	
Human Label: Sports		Human Label: American Sports	
Topic-word	**Probability**	**Topic-concept**	**Probability**
team	(0.123)	oakland raiders	(0.174)
est	(0.101)	san francisco giants	(0.118)
home	(0.022)	red	(0.087)
league	(0.015)	new jersey devils	(0.074)
games	(0.010)	boston red sox	(0.068)
second	(0.010)	kansas city chiefs	(0.054)

We define a labeling approach for topics considering the semantics of the concepts that are included in the learned topics in addition to existing ontological relationships between the concepts of the ontology.

In other words, our aim is to use the semantic knowledge graph of concepts in an ontology (e.g., DBpedia) and their diverse relationships with unsupervised probabilistic topic models (i.e. LDA), in a principled manner and exploit this information to automatically generate meaningful topic labels.

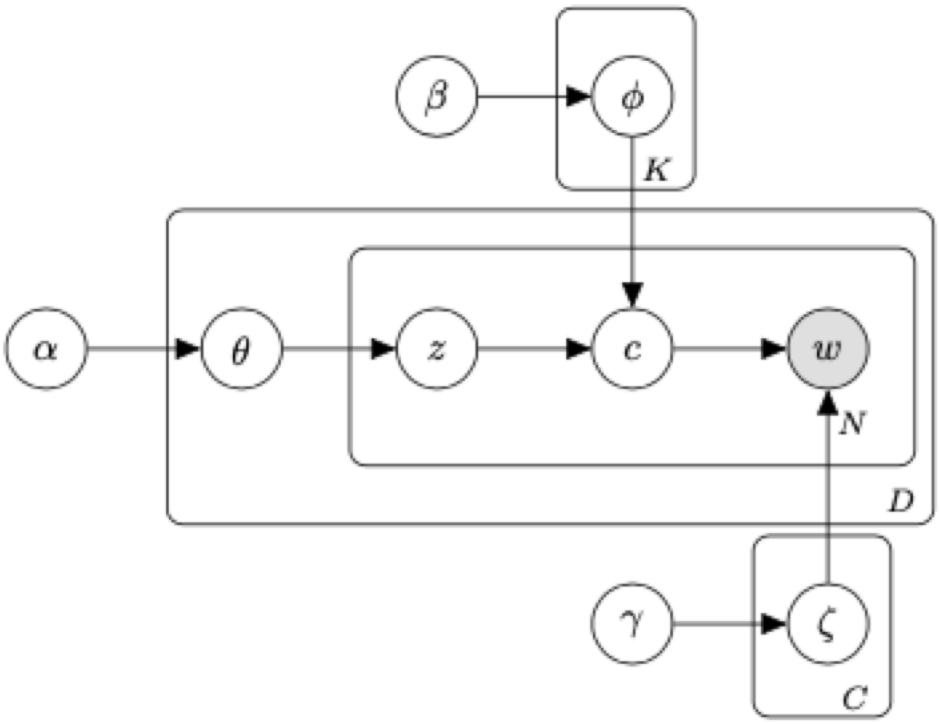
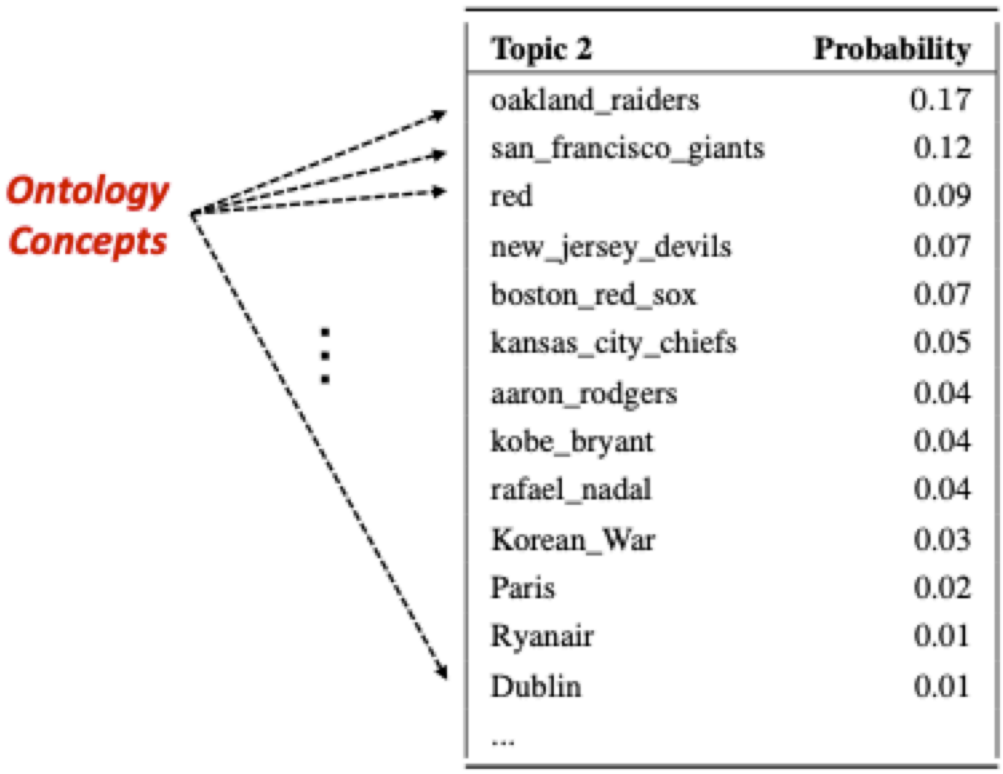


Fig. 2. Graphical representation of KB-LDA model.

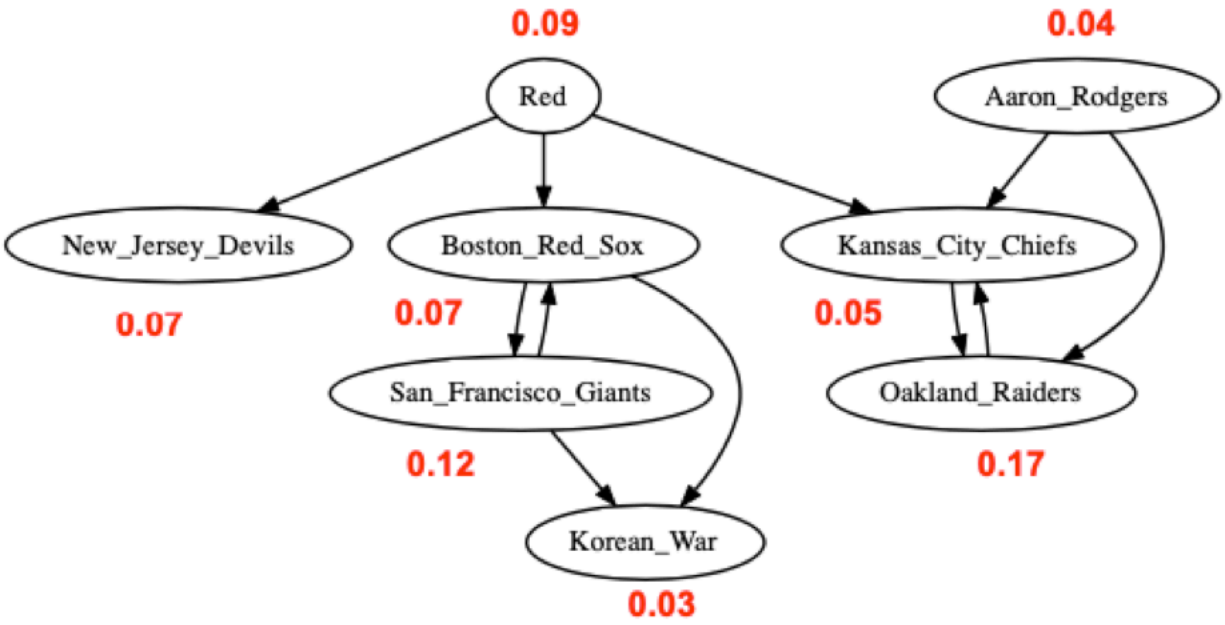
Algorithm 1: KB-LDA Topic Model

foreach concept $c \in \{1, 2, \ldots, C\}$ **do**
  Sample a word distribution $\zeta_c \sim \text{Dir}(\gamma)$
end
foreach topic $k \in \{1, 2, \ldots, K\}$ **do**
  Sample a concept distribution $\phi_k \sim \text{Dir}(\beta)$
end
foreach document $d \in \{1, 2, \ldots, D\}$ **do**
  Sample a topic distribution $\theta_d \sim \text{Dir}(\alpha)$
  foreach word w of document d **do**
    Sample a topic $z \sim \text{Mult}(\theta_d)$
    Sample a concept $c \sim \text{Mult}(\phi_z)$
    Sample a word w from concept $c, w \sim \text{Mult}(\zeta_c)$
  end
end

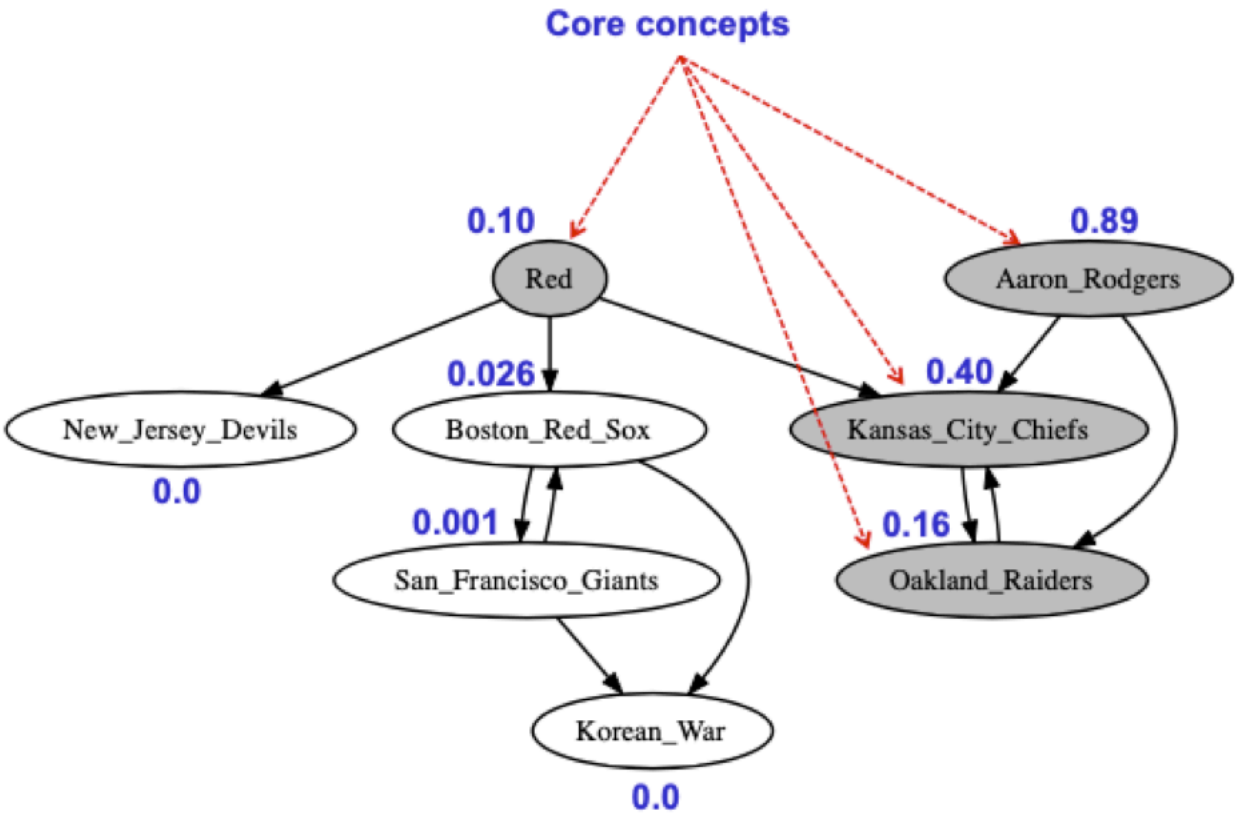
1. constructs the semantic graph from top concepts from topic-concept distribution for the given topic;



2. selects and analyzes the dominant thematic graph, a semantic graph's subgraph;



3. Extract the set of the the most authoritative and central (core) concepts in the dominant thematic graph



4. Extracts the topic label graph from the core thematic graph concepts
    1. Extract the topic label graph by traversing the ontology from each core concept and retrieving all the nodes laying at most three hops away from the core ones.

5. Computes the semantic similarity between topic and the candidate labels of the topic label graph.

EXAMPLE OF A TOPIC WITH TOP-10 CONCEPTS (FIRST COLUMN) AND TOP-10 LABELS (SECOND COLUMN) GENERATED BY OUR PROPOSED METHOD

Topic 2	Top Labels
oakland_raiders	National_Football_League_teams
san_francisco_giants	American_Football_League_teams
red	American_football_teams_in_the_San_Francisco_Bay_Area
new_jersey_devils	Sports_clubs_established_in_1960
boston_red_sox	National_Football_League_teams_in_Los_Angeles
kansas_city_chiefs	American_Football_League
nigeria	American_football_teams_in_the_United_States_by_league
aaron_rodgers	National_Football_League
kobe_bryant	Green_Bay_Packers
rafael_nadal	California_Golden_Bears_football

Motivation

Addressing the fact that:
“interpreting the label of the topics based on the distributions of words derived from the text collection is a challenging task for the users and it becomes worse when they do not have a good knowledge of the domain of the documents. Usually, it is not easy to answer questions such as “What is a topic describing?” and “What is a representative label for a topic?””

Additionally, using ontological concepts:
“as an extra latent variable (i.e. represent- ing topics over concepts instead of words) are advantageous in several ways including: (1) it describes topics in a more extensive way; (2) it also allows to define more specific topics according to ontological concepts, which can be eventually used to generate labels for topics; (3) it automatically incorporates topics learned from the corpus with knowledge bases.”

TABLE VII. SAMPLE TOPICS OF THE BAWE CORPUS WITH TOP-6 GENERATED LABELS FOR THE MEI METHOD AND KB-LDA + CONCEPT LABELING, ALONG WITH TOP-10 WORDS

Mei07				
Topic 1	**Topic 3**	**Topic 12**	**Topic 9**	**Topic 6**
rice production	cell lineage	nuclear dna	disabled people	mg od
southeast asia	cell interactions	eukaryotic organelles	health inequalities	red cells
rice fields	somatic blastomeres	hydrogen hypothesis	social classes	heading mr
crop residues	cell stage	qo site	lower social	colorectal carcinoma
weed species	maternal effect	iron sulphur	black report	cyanosis oedema
weed control	germline blastomeres	sulphur protein	health exclusion	jaundice anaemia
KB-LDA + Concept Labeling				
Topic 1	**Topic 3**	**Topic 12**	**Topic 9**	**Topic 6**
agriculture	structural proteins	bacteriology	gender	aging-associated diseases
tropical agriculture	autoantigens	bacteria	biology	smoking
horticulture and gardening	cytoskeleton	prokaryotes	sex	chronic lower respiratory
model organisms	epigenetics	gut flora	sociology and society	inflammations
rice	genetic mapping	digestive system	identity	human behavior
agricultur in the united kingdom	teratogens	firmicutes	sexuality	arthritis
Topic top-10 words				
Topic 1	**Topic 3**	**Topic 12**	**Topic 9**	**Topic 6**
soil	cell	bacteria	health	history
water	cells	cell	care	blood
crop	protein	cells	social	disease
organic	dna	bacterial	professionals	examination
land	gene	immune	life	pain
plant	acid	organisms	mental	medical
control	proteins	growth	medical	care
environmental	amino	host	family	heart
production	binding	virus	children	physical
management	membrane	number	individual	information

TABLE VIII. SAMPLE TOPICS OF THE REUTERS CORPUS WITH TOP-6 GENERATED LABELS FOR THE MEI METHOD AND KB-LDA + CONCEPT LABELING, ALONG WITH TOP-10 WORDS

Mei07				
Topic 20	**Topic 1**	**Topic 18**	**Topic 19**	**Topic 3**
hockey league	mobile devices	upgraded falcon	investment bank	russel said
western conference	ralph lauren	commercial communications	royal bank	territorial claims
national hockey	gerry shih	falcon rocket	america corp	south china
stokes editing	huffington post	communications satellites	big banks	milk powder
field goal	analysts average	cargo runs	biggest bank	china sea
seconds left	olivia oran	earth spacex	hedge funds	east china
KB-LDA + Concept Labeling				
Topic 20	**Topic 1**	**Topic 18**	**Topic 19**	**Topic 3**
national football league teams	investment banks	space agencies	investment banking	island countries
washington redskins	house of morgan	space organizations	great recession	liberal democracies
sports clubs established in 1932	mortgage lenders	european space agency	criminal investigation	countries bordering the philippine sea
american football teams in maryland	jpmorgan chase	science and technology in europe	madoff investment scandal	east asian countries
american football teams in virginia	banks established in 2000	organizations based in paris	corporate scandals	countries bordering the pacific ocean
american football teams in washington d.c.	banks based in new york city	nasa	taxation	countries bordering the south china sea
Topic top-10 words				
Topic 20	**Topic 1**	**Topic 18**	**Topic 19**	**Topic 3**
league	company	space	bank	china
team	stock	station	financial	chinese
game	buzz	nasa	reuters	beijing
season	research	earth	stock	japan
football	profile	launch	fund	states
national	chief	florida	capital	south
york	executive	mission	research	asia
games	quote	flight	exchange	united
los	million	solar	banks	korea
angeles	corp	cape	group	japanese

Topic modeling

KB-LDA

Topic modeling parameters

Nr of topics: 20

Beta: 0.01

Gamma: 0.01

Gibbs sampling iterations: 500

Nr. of topics

20

Label

Ontological concept

Label selection

\

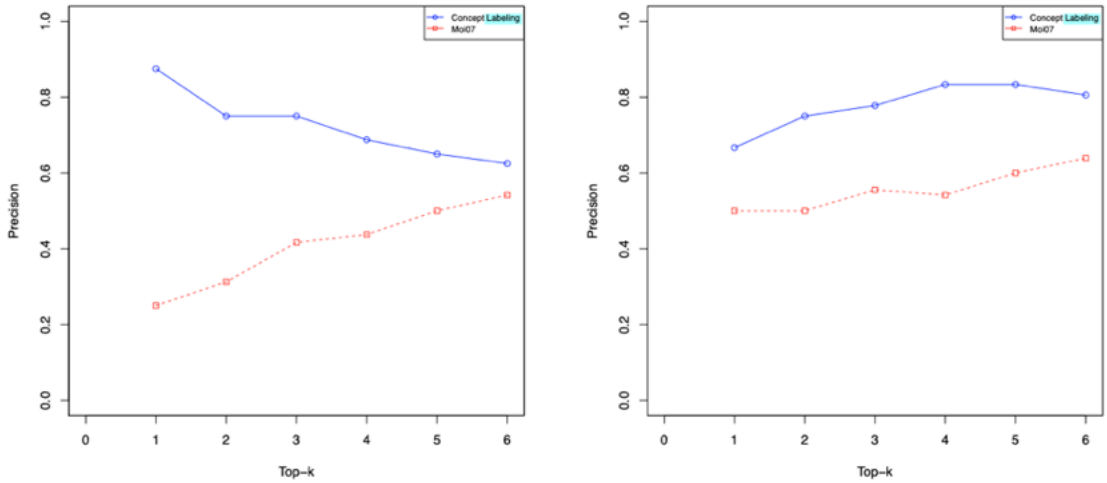
Label quality evaluation

In regards to quantitative evaluation for two aforementioned methods three human experts are asked to compare the generated labels and choose between “Good” and “Unrelated” for each one.

We compared the two different methods using the *Precision*@k, by considering the top-1 to top-6 generated labels. The Precision factor for a topic at top-k is represented as follows:

$$Precision@k = \frac{\text{\# of “Good” labels with rank} \leq k}{k}$$

Figure 8 illustrates the averaged the precision over all the topics for each individual corpus.



(a) Precision for Reuters Corpus

(b) Precision for BAWE Corpus

Fig. 8. Comparison of the systems using human evaluation

Assessors

three human experts

Domain

Paper: Topic labeling

Dataset: News, Literature

Problem statement

In this paper, we are taking concepts of ontology into consideration instead of words alone to improve the quality of generated labels for each topic.

We have highlighted some aspects of our approach including:

1. we have incorporated ontology concepts with statistical topic modeling in a unified framework, where each topic is a multinomial probability distribution over the concepts and each concept is represented as a distribution over words
2. a topic labeling model according to the meaning of the concepts of the ontology included in the learned topics. The best topic labels are selected with respect to the semantic similarity of the concepts and their ontological categorizations.

We demonstrate the effectiveness of considering ontological concepts as richer aspects between

topics and words by comprehensive experiments on two different data sets.

Corpus

Origin: the British Academic Written English Corpus (BAWE)

Nr. of documents: 683

Details:

Origin: Reuters

Nr. of documents: 1414

Details:

- subset of the Reuters news article

Document

Pre-processing

- removing punctuation stopwords, numbers, and words occurring fewer than 10 times in each corpus

---

```
@article{2017_allahyari_a_knowledge_based_topic_modeling_approach_for_automatic_topic_labeling,
    author = {Mehdi Allahyari and Seyedamin Pouriyeh and Krys Kochut and Hamid Reza Arabnia},
    date-added = {2023-03-30 17:29:16 +0200},
    date-modified = {2023-03-30 17:29:16 +0200},
    doi = {10.14569/IJACSA.2017.080947},
    journal = {International Journal of Advanced Computer Science and Applications},
    number = {9},
    publisher = {The Science and Information Organization},
    title = {A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling},
    url = {http://dx.doi.org/10.14569/IJACSA.2017.080947},
    volume = {8},
    year = {2017}}
```

#Thesis/Papers/To Complete/BS#