adhitama_2017_topic_labeling_towards_news_document_collection_based_on_latent_dirichlet_allocation_and_ontology

Year

2017

Author(s)

Adhitama, Rifki and Kusumaningrum, Retno and Gernowo, Rahmat

Title

Topic Labeling Towards News Document Collection Based on Latent Dirichlet Allocation and Ontology

Venue

ICICoS

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel approach

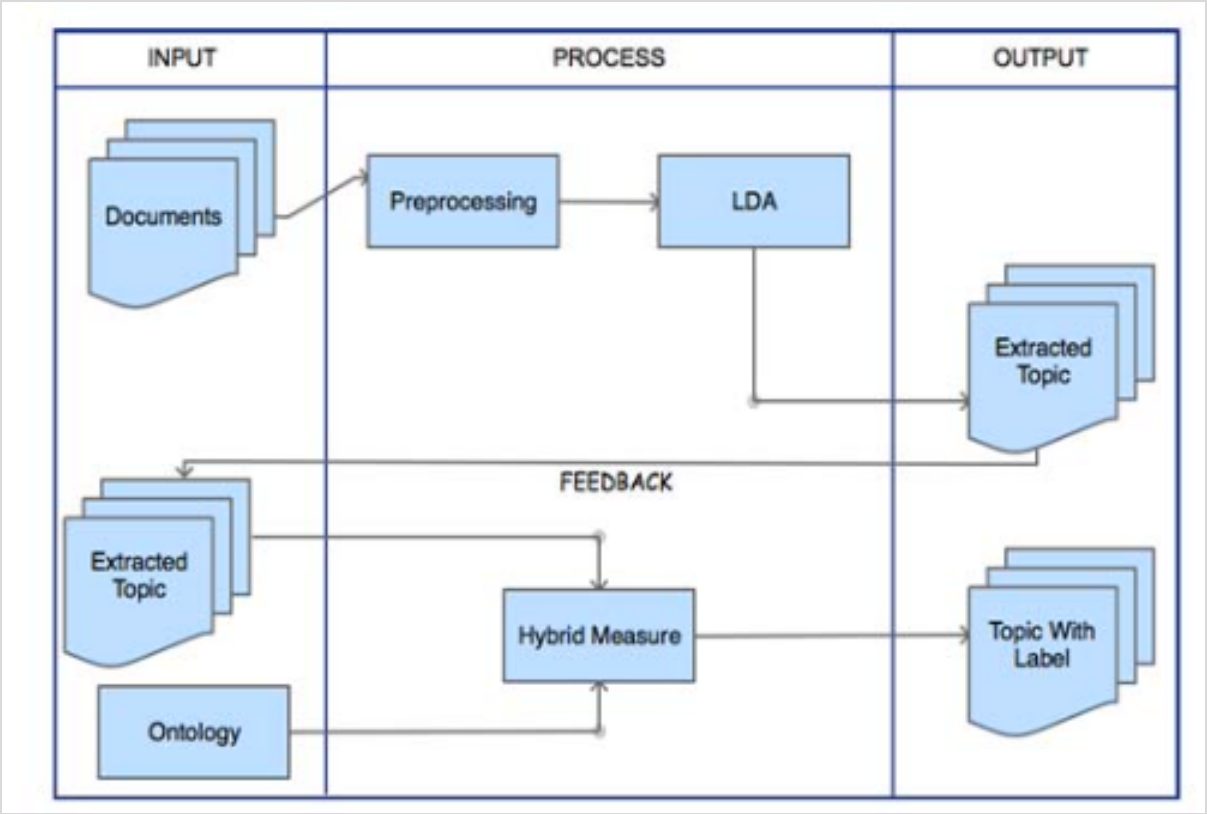
Underlying technique

Ontology-based topic labeling

Topic labeling parameters

Zhou similarity weight factor: 0.5

Label generation



Topic labeling

This study uses ontology scheme to build word-to-word relationship and Zhou semantic similarity to measure the similarities between words in each topic to generate the topic label.

The formula of Zhou similarity is described below:

$$sim_{zhou}(c_i, c_j) = 1 - k\left(\frac{\log(len(c_i, c_j) + 1)}{\log(2 * (deep_{max} - 1))}\right) - (1 - k) * ((IC(c_i) + IC(c_j) - 2 * IC(lso(c_i, c_j)))/2) \quad (4)$$

where:

$len(c_i, c_j)$: the shortest path from c_i to c_j

$lso(c_i, c_j)$: the lowest common *subsumer* of c_i and c_j

$deep_{\max}$: the maximum depth of the taxonomy

$IC(c)$: information content of c

k : weight factor

Motivation

The weakness of the LDA method is the inability to label the topics that have been formed. This research combines LDA with ontology scheme to overcome the weakness of labeling topic on LDA.

This study aims to automatically create a generic label in clustered news documents for easier interpretations.

Topic modeling

LDA

Topic modeling parameters

Nr of topics: 15

Nr. of topics

15

Label

Ontology class

Label selection

\

Label quality evaluation

Cohen's kappa coefficient is used to measure the reliability of the label based on the agreement of two linguistic experts, while the mean relevance rate is used to measure the average of the relevant value of linguistic experts on a label with particular words representation that has more than 41% of the kappa value.

Evaluation is made for labels made on topics with {5, 10, 15, 20, 25, 30, 35} words
High kappa values indicate the labels of the topics have a high consistency of relevant agreement between experts.

TABLE I. RELEVANCE VALUE AND RELEVANCE RATE FOR EACH WORDS REPRESENTATION

Words representation	Kappa Value	Relevance Value		Relevance Rate		Mean relevance Rate
		Expert 1	*Expert 2*	*Expert 1*	*Expert 2*	
5 Words	1.00	12	12	0.80	0.80	0.80
10 Words	0.50	10	6	0.67	0.40	0.53
25 Words	0.84	11	10	0.73	0.67	0.70
30 Words	0.59	13	11	0.87	0.73	0.80
35 Words	0.55	12	9	0.80	0.60	0.70
Average Value	0.61	11.6	9.6	0.77	0.64	0.71

12 relevant labels out of 15, etc…

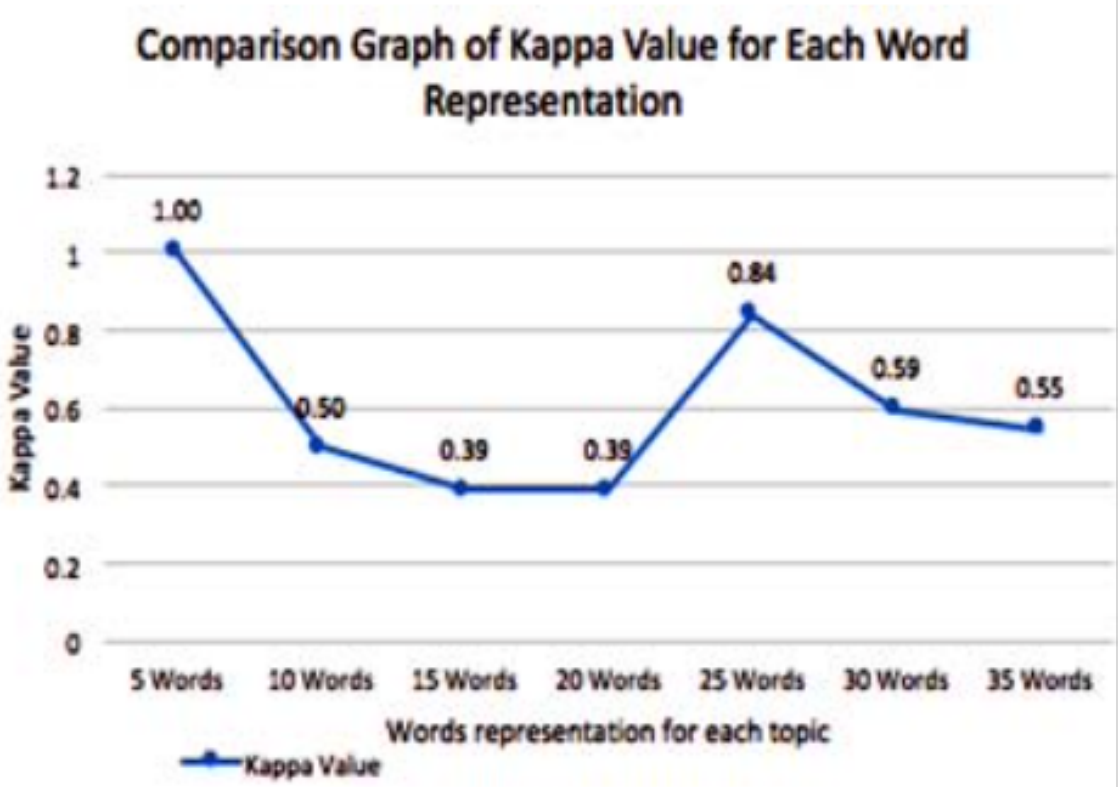


Fig. 3. Comparison Graph of Kappa Value for each words representation

Assessors

Two linguistic experts

Domain

Paper: Topic labeling
Dataset: News

Problem statement

This study uses datasets of 50 news documents taken from the online news portal.
The ontology scheme used in this study is based on the dictionary of the field contained in "Kamus Besar Bahasa Indonesia (KBBI)".
The experiment aims to find the best word count representation for each topic in order to produce the relevant label name for the topic.
Cohen's kappa coefficient is used to measure the reliability of the label based on the

agreement of two linguistic experts, while the mean relevance rate is used to measure the average of the relevant value of linguistic experts on a label with particular words representation that has more than 41% of the kappa value.

Corpus

Origin: News
Nr. of documents: 50
Details:

Document

Text of a news article

Pre-processing

- Tokenization
- Stopwords removal
- N-Gram Splitting

```
@INPROCEEDINGS{adhitama_2017_topic_labeling_towards_news_document_collection_based_on_latent_dirichlet_allocation_and_ontology,
  author={Adhitama, Rifki and Kusumaningrum, Retno and Gernowo, Rahmat},
  booktitle={2017 1st International Conference on Informatics and Computational Sciences (ICICoS)},
  title={Topic labeling towards news document collection based on Latent Dirichlet Allocation and ontology},
  year={2017},
  volume={},
  number={},
  pages={247-252},
  doi={10.1109/ICICOS.2017.8276370}}
```

#Thesis/Papers/BS