Topic Coherence Evaluation- Selecting Key Topic Features using Random Forest and Segmenting Topics using K-means

Muhammad Omar*, Byung-Won On**, Ingyu Lee***, and Gyu Sang Choi**
* Department of Information and Communication, Yeungnam University
** Department of Software Convergence Engineering, Kunsan National University
*** Sorrell College of Business, Troy University

Abstract

Inputs to topic coherence formulae are words of a topic and output is a real value indicating quality of the topic. By treating topics as objects and word similarity formulae as topic features we propose to categorize topics. Using the idea from machine learning, we selected key topic features using an iterative method based on supervised classification (Random Forest) and then apply k-means clustering algorithm to segment topics. We evaluated the clustering results against human ratings. We find that, incoherent topics can be filtered out using PMI based formula and we can also segment coherent topics from incoherent using two different types of formulae. Article is directly related with evaluation of topics and hence evaluation of topic models but it has applications in other IR related tasks. To check robustness in the future, we need to test on corpora from different domains.

1. Introduction

Topic coherence formulae [1-3] take top-n words of a topic and output a real valued score- based on some threshold one can identify that the topic is coherent/interpretable. Point wise mutual information (PMI) based formulae are well-known because of their high correlation with human judgments [1-5]. In [1], PMI and its variations are used as term weights whereas in [3-5] used as semantic similarity measure. Mimno et al. [2] used their coherence-formula for two purposes- to filter out low quality topics and then incorporates into topic model. By [3], WordNet based coherence formulae are less correlated with human ratings but [4] reported their importance when used with others. In [4], topics are represented by coherence formulae and this representation is justified by supervised classification of represented topics.

Getting human annotations, to train supervised models, is cumbersome, conflicting, expensive and don't work in real time. Our question is that can we take human out of the loop specifically for the task of accepting and/or rejecting low/high quality topics without the use of human annotations [4] and manual threshold value used in all the state of the art formulae. We are curious to perform unsupervised clustering of topics into two groups- coherent and incoherent.

We performed clustering of topics using k-means algorithm and evaluated the clusters by mapping clustering results on human annotated topic labels.

2. Materials and Methods

Following sections explain methodology (Figure 1).

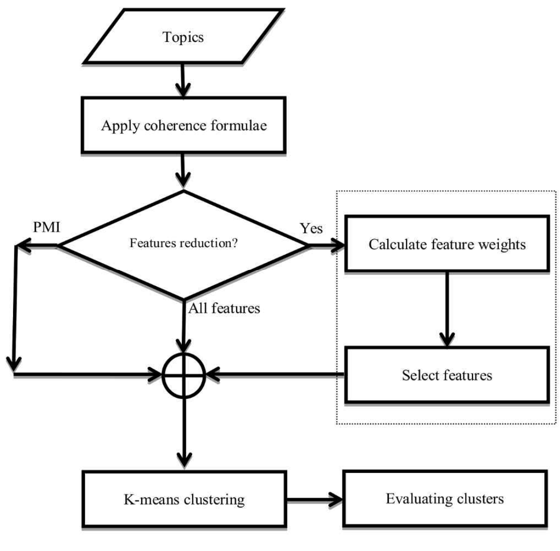

Figure 1. Proposed Methodology.

Here Information gain ratio (IGR), an Entropy based measure, is used as feature weighting and in building Random Forest (RF). In RF, nodes are split based on IGR into smaller nodes until the nodes are more homogeneous [6]. IGR avoids biasness of Information gain (IG) toward multi-valued attributes because it considers intrinsic information of split. PMI [7-9], also Entropy based, is used to find words collocations.

2.1. Data

We investigated the topic representation scheme of [4] with the objective to automate topic categorization task of [4]. We used 10,000 UPI news articles [4]. For the preprocessing- words of English alphabets of lengths 3~25 were retained, stop words removed, lemmatization and filtering low/high frequency words. To extract the topics, Latent Dirichlet allocation (LDA) [10] applied to extract 120 topics (64 coherent and 56 incoherent) using the toolbox provided by [11].

2.2. Coherence Formulae and Represented Topics

We used topic coherence formulae as topic descriptive features (Figure 2). Each row in the dataset numerically represents top-10 words of a topic and columns associate with coherence formulae. Two types of formulae are used- distributional (word count based) and WordNet [14, 15, 16]. See [1, 3, 4] for details. WordNet based formulae were derived using NLTK WordNet interface [16].

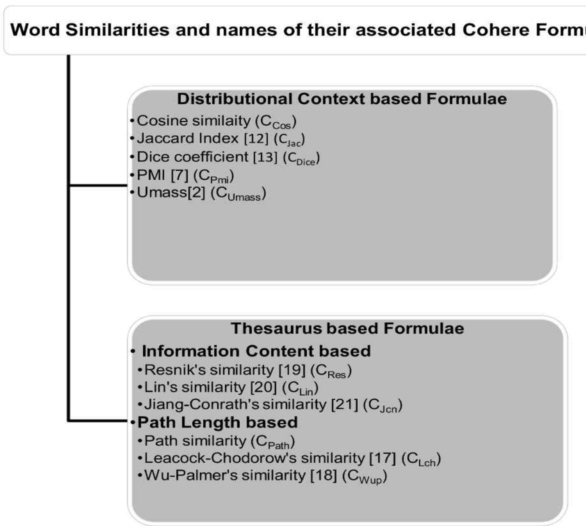

Figure 2. Two types of words similarities and names of their associated topic coherence formulae.

2.3. Features Reduction

It's a 2-step iterative procedure- features weighting and subset selection. Based on IGR we used normalized-weights to avoid biasness due to correlation of input features with output feature (labels). Subset selection starts with the highest weighted feature as the only member of the feature set, then to determine the fitness of the current feature set we evaluated performance vector (precession, recall, accuracy) using RF [22] based on 10-fold cross validation. The procedure stops if last feature added to feature set does not improve classification or all features are added. RF is good at avoiding over fitting.

2.4. K-means Clustering

Clustering is independent of object classes i.e. topic labels. Clustering needs to define- objects (topics), purpose (semantic categorization of topics in to coherent/incoherent), descriptive features of objects/topics (coherence formulae), object similarity measure (inner product), suitable clustering algorithm (k-means for hard clustering), evaluation of clustering.

2.5. Evaluation of Clusters

To evaluate clustering, binary topic labels were assigned to topics based on annotators ratings- 1 for coherent topics where one can guess its title else 0. We approximate cluster mapping between the clustering labels and predictions by adjusting the predicted clusters with the given labels to estimate the best fitting pairs. We employed various performance metrics to cover all values of 2x2 contingency table. Sensitivity (Recall rate or TPR) and specificity (TNR) are useful to judge the performance of a binary classifier. Sensitivity, measures the proportion of actual positives (coherent topics) which are correctly identified as such. Specificity, measures the proportion of negatives (incoherent topics) which are correctly identified as such. There is usually a trade-off between these measures that may be represented as receiver operating characteristic curve ROC.

3. Results and Discussions

In Table, notice the reduced subset of topic representative features {C_{Pmi}, C_{Jcn}, C_{Path}, C_{Wup}}. Using C_{Pmi} formula alone we got good clustering (Table 2).

Table 1. Features weighted by IGR selected by RF.

Topic Features	Weighted by IGR	Selected By RF
C_{PMI}	1	Yes
C_{Jcn}	0.42	Yes
C_{Path}	0.39	Yes
C_{Wup}	0.3	Yes
C_{Cos}	0.23	No
C_{Dice}	0.18	No
C_{Jac}	0.16	No
C_{Res}	0.08	No
C_{Lch}	0.07	No
C_{Lin}	0.02	No

C_{Umass}	0.00	No

Table 2. k-means clustering of topics in two clusters.

Feature sets	Clustering Results			
	Precision %	Recall %	Accuracy %	TNR %
$\{C_{PMI}\}$	81	61	72	84
All features (Fig 2)	73	63	68	73
[a]$\{C_{PMI}, C_{Jcn}, C_{Path}, C_{Wup}\}$	84	80	81	82

[a]Subset obtained by weighted by IGR and optimized by RF.

Next we tried to improve the clustering by using all formulae (Figure 2) and by reduced subset (Table 2). For the reduced set, we beat $\{C_{Pmi}\}$ in Precision, Recall, Average, but $\{C_{Pmi}\}$ always has the highest true negative rate (TNR). PMI shows a unique characteristic to recognize incoherent topics and it was counter checked by ROC curves not shown here.

4. Conclusions

Superiority of PMI based topic coherence formula is found as [1-5] and specifically its rule in identification of incoherent topics. Clustering, for both topic categories, was improved by using C_{Pmi} and WordNet formulae C_{Jcn}, C_{Path} and C_{Wup}. Applications related with topic visualizations (say TopicExplorer can hide incoherent topics to avoid confusion created by less informative topics. Other IR tasks can get benefit by filtering incoherent topics- improving the precision of keyword search and selecting advertising links related to a web page. Selecting the most coherent combinations of words of a phrase (for example in automated machine translation) can also get benefit.

5. Acknowledgements

This research is supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under the Industrial Technology Innovation Program, No. 10063130, by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1A2B4007498), and the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-2016-0-00313) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

[1] Aletras N and Stevenson M. Evaluating topic coherence using distributional semantics. In: Proceedings of the 10th International Conference on Computational Semantics (IW*CS) – Long Papers,* 2013, pp. 13-22.

[2] Mimno D, Wallach HM, Talley E, Leenders M and McCallum A. Optimizing semantic coherence in topic models. In: *Proceedings of the conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 2011, pp. 262-272.

[3] Newman D, Lau JH, Grieser K and Baldwin T. Automatic evaluation of topic coherence. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 100-108.

[4] Muhammad Omar, Byung-Won On, Ingyu Lee and Gyu Sang Choi. LDA Topics: Representation and Evaluation. Journal of Information Science 1–4, accepted by the editor.

[5] Lau JH, Newman D and Baldwin T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: *European Chapter of the Association for Computational Linguistics (EACL),* 2014, 530.

[6] http://www.ke.tu-darmstadt.de/lehre/archiv/ws0809/mldm/dt.pdf

[7] Church KW and Hanks P. Word association norms, mutual information, and lexicography. *Computational linguistics*, 1990; 16(1), 22-29.

[8] *Bouma, Gerlof (2009). "Normalized (Pointwise) Mutual Information in Collocation Extraction". Proceedings of the Biennial GSCL Conference.*

[9] *Fano, R M (1961). "chapter 2". Transmission of Information: A Statistical Theory of Communications. MIT Press, Cambridge, MA. ISBN 978-0262561693.*

[10] Blei DM, Ng AY and Jordan MI. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 2003; 3: 993–1022.

[11] Griffiths, T., & Steyvers, M. (2004). Finding Scientific Topics. Proceedings of the National Academy of Sciences, 101 (suppl. 1), 5228-5235.

[12] Jaccard, P. (1901) Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 241-272.

[13] Dice, Lee R. (1945). "Measures of the Amount of Ecologic Association Between Species". Ecology 26 (3): 297–302. doi:10.2307/1932409. JSTOR 1932409

[14] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[15] Princeton University "About WordNet." WordNet. Princeton University. 2010. http://wordnet.princeton.edu

[16] http://www.nltk.org/howto/wordnet.html

[17] Leacock C, Miller GA and Chodorow M. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistic*, 1998, 24(1); 147-165.

[18] Wu Z and Palmer M. Verb Semantics and Lexical Selection. In: *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistic*s (ACL'94), Las Cruces, New Mexico, 1994, pp. 133-138.

[19] Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint*, 1995, cmp-lg/9511007.
[20] Lin D. An information-theoretic definition of similarity. *ICML*, July 1998; 98: 296-304.
[21] Jiang JJ and Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint*, 1997, cmp-lg/9709008.
[22] Breiman L. Random forests. *Machine learning*, 2001; 45(1): 5-32.