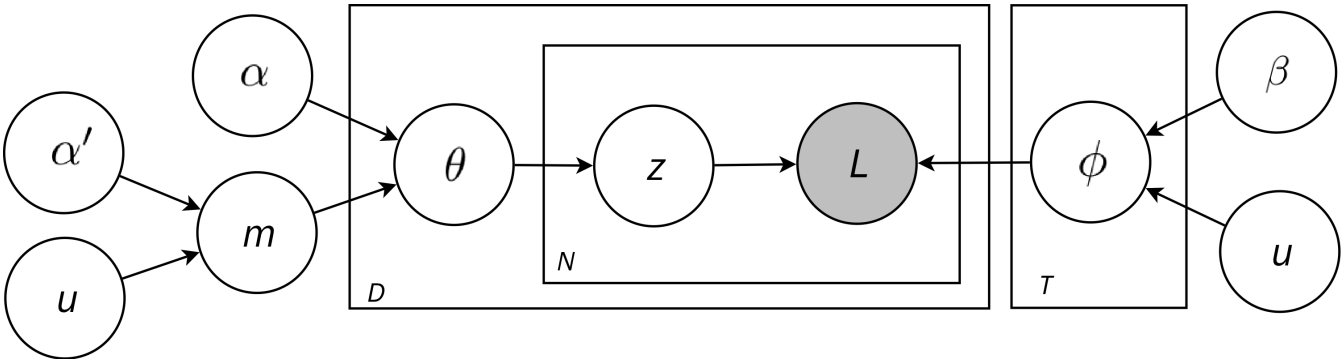

Figure 1: Asymmetric-Symmetric Lexeme-based Approach. L is a lexeme which arise from the morphological analysis of observations w. z is per-word topic assignment, θ is document-specific topic distribution. ϕ means topics. α and β mean Dirichlet concentration parameters of proportions. m and n are base measures.

1 Product Introduction

Topic models such as latent Dirichlet allocation (LDA) [2] have been used to analyze large, unstructured collections of documents. It is a generative model which find the latent structure of the topics to be explained by unobserved groups. Instead of words as vectors in Euclidean space, it can be associated with probability distributions over the contexts. It clusters words into topics and assigns each document a distribution over those topics. Many previous researches widely applied LDA to resolve various issues and tasks including analysis of extending to associate authorship information [9], study of examining temporal dynamics and abstracts to illustrate semantic content [3], topic analysis of news articles [10]. Traditionally, LDA operates over a token-based representation of documents. The problem of common words in natural language which tend to dominate all topics is sometimes dealt with filtering to remove frequent words. In practice, however, we focus that such a token-based document representation and a simple solution like stop words removal are inadequate when analyzing large document sets of morphologically-rich languages. We are typically faced with this immediate problem when applied topic models to a Korean documents, as it fails to capture coherent word patterns, due to the large number of morphological variants.

1.1 Algorithm

The LDA model basically has two sets of unknown parameters. The transition between successive states of the Markov chain results from repeatedly drawing z from its distribution conditioned on all other variables, summing out θ and ϕ using standard Dirichlet integrals. In fact, the word forms in documents significantly affect the topic models because each word is only given as a single set of observed variables which infers the hidden thematic structure using posterior inference.

Morphologically-rich languages like Korean strongly tend to show characteristics of lexical diversity. Both the approaches of algorithmic inference like optimizing parameters [8], [6] and general efforts like stop words removal tried to find better topics of model, high diversity fundamentally makes a

previous topic model harder to achieve appropriate topics yet, so it is helpful to work with lexemes instead of the inflected forms due to the property of bag of words. One of the most important issues in topic model is to decrease the overall perplexity [7] and lexical diversity significantly affects an initial state of perplexities of each input data. The measurements of lexical diversity are well introduced in TTR (Type-Token Ratio), Yule's K [16], and D-estimate [17]. D-estimate is not be affected by the corpus size.

1.2 Characteristics

LDA is for a set of D documents W, and define a topic t as a discrete distribution over words characterized by probability vector ϕ_t. The prior over ϕ_t is a symmetric Dirichlet with concentration parameter β.

$$P(\Phi) = \prod_t Dir(\phi_t; \beta u) = \prod_t \frac{\Gamma(\beta)}{\prod_w \Gamma(\frac{\beta}{W})} \prod_w \phi_{w|t}^{\frac{\beta}{W}-1} \delta(\sum_w \phi_{w|t} - 1) \tag{1}$$

Each document is also associated with a distribution over topics θ_d that is also Dirichlet distributed with concentration parameter α and symmetric base measure.
The word tokens in document d, $w^{(d)} = \{w_n^{(d)}\}_{n=1}^{N_d}$, have corresponding topic assignments $z^{(d)} = \{z_n^{(d)}\}_{n=1}^{N_d}$ drawn i.i.d. from θ_d, and conditioned on $z^{(d)}$ the words are drawn i.i.d. from $\Phi = \{\phi_1, \ldots, \phi_T\}$:

$$P(z^{(d)}|\theta_d) = \prod_n \theta_{z_n^{(d)}|d} \qquad \& \qquad P(w^{(d)}|z^{(d)}, \Phi) = \prod_n \phi_{w_n^{(d)}|z_n^{(d)}} \tag{2}$$

As α' goes to infinity, the asymmetric Dirichlet prior over θ approaches a symmetric Dirichlet prior. That is, symmetric hierarchical Dirichlet prior is a special case of the asymmetric hierarchical Dirichlet prior.

$$\begin{aligned} P(z_{N_{d+1}}^{(d)} = t|Z, \alpha, \alpha' u) &= \int dm P(z_{N_{d+1}}^{(d)} = t|Z, \alpha m) P(m|Z, \alpha' u) \\ &= \frac{N_{t|d} + \alpha \frac{\hat{N}_t + \frac{\alpha'}{T}}{\sum_t \hat{N}_t + \alpha'}}{N_d + \alpha} = \frac{N_{t|d} + \frac{\alpha}{T}}{N_d + \alpha} \end{aligned} \tag{3}$$

A variety of algorithms have been used to estimate the parameters of topic models. The algorithm of EM (Expectation-Maximization) [11] is the basic approach and approximate inference methods like variational EM [2], expectation propagation [12], and collapsed Gibbs sampling [3] applied in topic models. Generic EM algorithm is known for problems with local maxima in LDA [2], suggesting a move to approximate methods in which some of the parameters - such as ϕ and θ - can be integrated out rather than explicitly estimated.
In this paper, we will use collapsed Gibbs sampling, as it provides a simple method for obtaining parameter estimates under Dirichlet priors and allows combination of estimates from several local maxima of the posterior distribution.

$$P(z_n^{(d)}|W, Z_{\backslash d,n}, \alpha u, \beta u) \propto \frac{N^{\backslash d,n}_{w_n^{(d)}|z_n^{(d)}} + \frac{\beta}{W}}{N^{\backslash d,n}_{z_n^{(d)}} + \beta} \frac{N^{\backslash d,n}_{z_n^{(d)}|d} + \frac{\alpha}{T}}{N_d - 1 + \alpha} \tag{4}$$

1.3 Gibbs sampling in the generative model

Markov chain Monte Carlo is a procedure for obtatining samples from complicated probability distributions, allowing a Markov chain to converge to the target distribution and then drawing samples from the Markov chain. [13] Each state of the chain is an assignment of values to the variables being sampled, and transitions between states follow a simple rule.
Gibbs sampling iterates over each word token in the text collection and estimates the probability of assigning the current word token to each topic ($P(z_i = j)$), conditioned on the topic assignments to all other word tokens ($z_{\neg i}$) as [4]. In the original paper 'Finding Scientific Topics', the authors are more interested in text modelling, (find out Z), hence, the Gibbs sampling procedure boils down to estimate
The convergence of Markov chain on the posterior distribution on z and two sets of unknown parameters ϕ and θ are inferred using collapsed Gibbs sampling. [3].

$$P(z_i = j|z_{\neg i}, w) \tag{5}$$

Here, θ, ϕ are intergrated out. Actually, if we know the exact Z for each document, it's trivial to estimate θ, ϕ.

$$p(z_i = j|z_{\neg i}, w_i, d_i, \alpha, \beta) \propto \frac{C^{WT}_{w_i,j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i,j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i,t} + T\alpha} \tag{6}$$

- C^{WT} : W x T matrix of counts.
- C^{WT}_{wj} : the number of times word w is assigned to topic j.
- C^{DT} : D x T matrix of counts.
- C^{DT}_{dj} : the number of times topic j is assigned to some word token in document d.

1.4 Initialization and Iteration

Before iterations of LDA estimation, it is necessary to initialize parameters. Collapsed Gibbs Sampling (CGS) estimation has the following parameters. The most simple initialization is to assign each word to a random topic and increase the corresponding counters a topic of word n of document m, a count of word t with topic z and a word count with topic z. It starts by assigning each word token to a random topic. [4]
The count matrices C^{WT} and C^{DT} are first decremented by one for the entries that correspond to the current topic assignment. Then, a new topic is sampled from the distribution in Equation 3 and the count matrices C^{WT} and C^{DT} are incremented with the new topic assignment. Each Gibbs sample consists the set of topic assignments to all N word tokens in the corpus, achieved by

a single pass through all documents. Perplexity decreases as learnings are progressing until it gets stable.

$$perplexity(\omega) = exp\left\{\frac{1}{N}\sum_{d} log(\sum_{w} \theta_j^d \phi_i^j)\right\} \tag{7}$$

1.5 Evaluation

Each experiment was run with $T \in \{10, \ldots, 100\}$ for 1000 Gibbs sampling iterations. The dataset evaluates the implementation on proprietary data from GS-Shop. It is a large set of korean product reviews which are chosen as the best by their critics. The GS-Shop review corpus[1] was made not public during the project of the data analysis for GS-Shop. This dataset is not officially published because of the user privacy and copyright of company. 1000 iterations of gibbs sampling for each experiment tend to be converged to the optimized perplexity and it is observed the topics with reasonable pairs of words.

Bayesian generative models are typically evaluated by computing the probability of unseen test data w, given training data w^{train} and hyperparameters U. We also calculated the probability of held-out documents using the *left-to-right* evaluation method described by Wallach et al. [15] For models of text, these results are usually reported in terms of the information rate of the test data, measured in bits per word.

We propose our method to evaluate the information rate which is computed from perplexities for topics as follows. In this evaluation metrics, the results of big figures can be interpreted as better modeling.

$$IR = -\sum_{i} \log_2 \left(\frac{exp\left\{\frac{1}{N}\sum_{d} log(\sum_{w} \theta_j^d \phi_i^j)\right\}}{N} \right) \tag{8}$$

We performed experiment on 4 different settings. LDA and collapsed Gibbs sampling with the 2 models (AS, SS) of parameter setting. Most researches use symmetric Dirichlet priors and the parameters are usually set heuristically, but we follow the asymmetric-symmetric setting of parameters. [8]

Table 1 shows that topics with four different types are ranked by the evaluation metrics. The number of k for the first rank means that it is currently evaluated as a top rank by our metrics, but the Δp means a potential to change the further states.

The experimental results of iterations are shown in figure 2. Two vertical pairs are about the comparison of raw and POS-tagged corpora. The upper two results show that the tests of raw corpus start the iterations at high initial perplexity. They have almost 10 times more initial perplexity and 4 times more the number of word types than the lower two results. This means that lexeme-based approach of topic modeling in morphologically-rich languages effectively makes

[1]Language : Korean, May-1-2012 Version of dataset (about 1.4 Gb, 7zipped), Test size : ≒ 31.6 Mb, number of documents : 16972, number of word types : 228767, number of authors : 10792, mean of document length : 116

Table 1: Topics are ranked by the evaluation algorithm. AS means asymmetric-symmetric Dirichlet priors, R means a raw document, M means a documnet which is morphologically analyzed. Perp1000 means a value of $1000th$ iteration. Δp means the exact value for the slope of 900 iterations. (burn in for 100)

Type	Rank	Topics	Perp1000	Eval	Δp
	1	40	1443.860846	5093.44829372	-0.148861618
	2	30	1443.095366	5092.54444623	-0.160838140
$AS - M$	3	20	1471.908241	5074.16771202	-0.149942653
	4	50	1463.029176	5070.27604476	-0.162756086
	5	60	1466.723941	5059.6027747	-0.178713736
	1	50	1371.821908	5213.96837351	-0.010185616
	2	60	1373.561365	5210.98407227	-0.007699750
$SS - M$	3	40	1381.690815	5200.78250046	-0.015603636
	4	80	1394.289635	5189.74987447	-0.008437427
	5	70	1394.153373	5189.62047978	-0.012287782
	1	20	11797.929577	2089.91235513	-0.638262974
	2	30	11796.443899	2086.62796883	-0.710589266
$AS - R$	3	40	11721.960483	2082.50046824	-0.954411517
	4	50	11893.610587	2045.35878549	-1.353196324
	5	10	12367.471012	2035.40250745	-0.238156744
	1	20	12472.586816	2030.83946314	-0.056791648
	2	30	12631.654158	2004.14103997	-0.102808705
$SS - R$	3	40	13003.341743	1972.16765532	0.064325478
	4	10	13020.643990	1970.53839554	-0.056997538
	5	50	13261.577026	1943.28553018	0.034007582

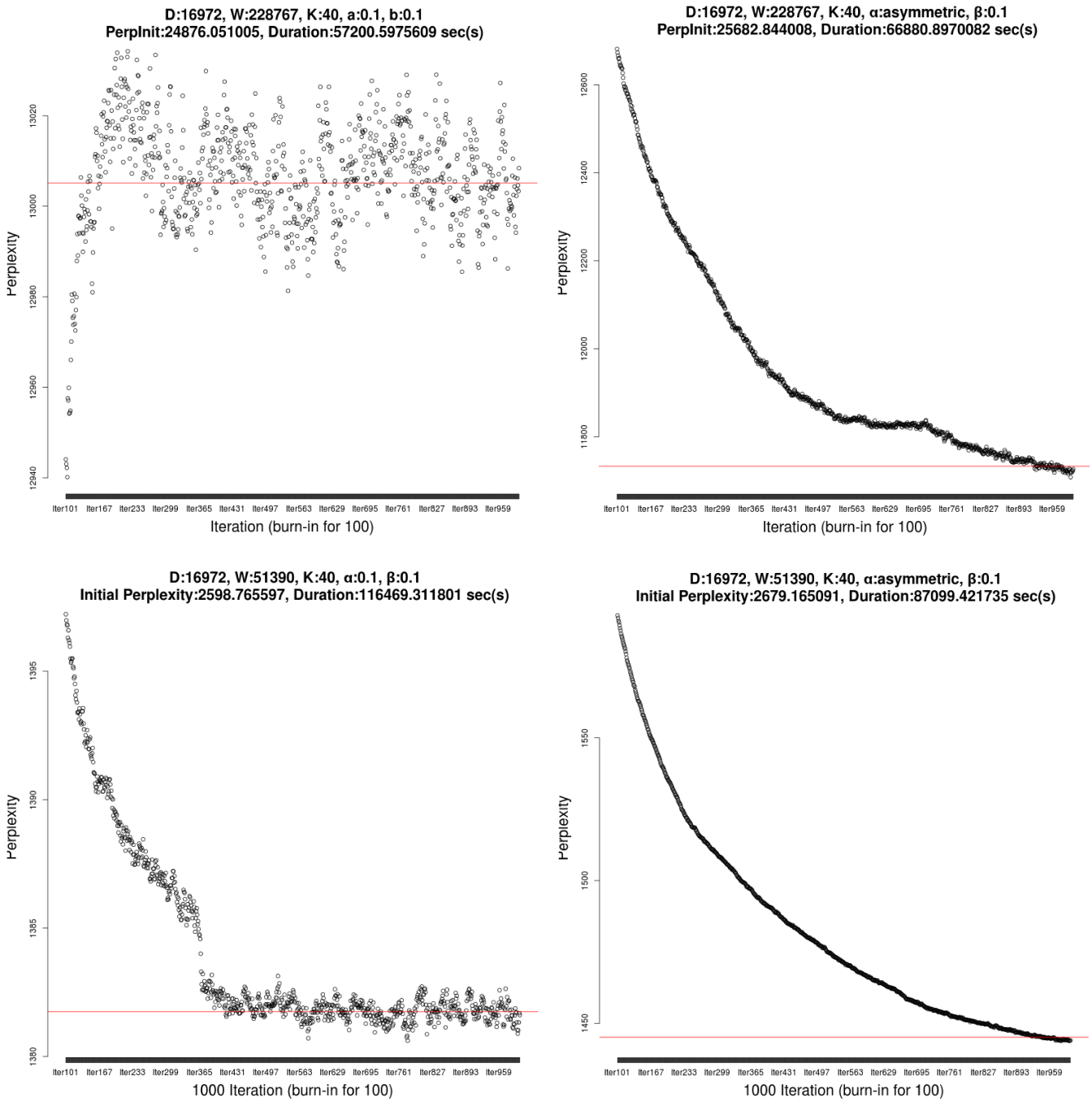

Figure 2: 1000 Iterations of Markov chain for 40 topics. SS-Raw (upper left, 2a): Raw corpus with symmetric parameters, AS-Raw (upper right, 2b): Raw corpus with asymmetric-symmetric parameters, SS-Morph (under left, 2c): POS-tagged corpus with symmetric parameters , AS-Morph (under right, 2d): POS-tagged corpus with asymmetric-symmetric parameters. Red line: Mean point of last 100.

an appropriate state at the beginning and it significantly mitigates the computational burden of iteration.
Two horizontal pairs are about the comparison of symmetric and asymmetric-symmetric parameters. These results are shown in figure 2a-2b and 2c-2d. They exhibit similar patterns to the results with asymmetric parameters over θ. The poor efficiency of iteration with symmetric parameters especially in 2a gives an unpredictable directivity and the hasty convergence in 2c weakens the optimizing potential. The convergence to optimized perplexity of 2b-2d (AS) is more gradual than 2a-2c (SS) at relatively-regular values, but the rate of change Δp at last 100 shows that it tends to be potential to fall more the perplexity. It reveals the clear tendency of optimization in each iteration.
We can predict that both lexeme-based approach and asymmetric-symmetric parameters are important to effectively find the topics in more advantageous condition.

2 Functional Specification

2.1 Dependencies

The following depedencies are required for the installation of the ChiST Algorithm in a Linux (RHEL6/CentOS6/Ubuntu/Debian) 32/64bit environment: gcc version 4.6.3, Cython version 0.15.1. The installation of OWLNEST nlpstat mathematical statistics library is also required. (provided upon installation)

2.2 Methods

- find_most_recent(directory, partial_file_name)
- setStopword(self)
- prettyOutput(content)
- getMorph(sentence)
- wiredSpaceRemover(content)
- cleanText(inline)
- rangeDates(*args)
- getDateRange(startDate, endDate)
- getDocumentsValues(startDate, endDate, INPUT_FILE)
- setDocumentsValues(startDate, endDate, INPUT_FILE)
- getFinalTitle(startDate, endDate, INPUT_FILE)
- getDocumentsDict(startDate, endDate, INPUT_FILE)

- HotTopicTaxonomy(startDate, endDate, INPUT_FILE)
- HotTopicGeneral(startDate, endDate, INPUT_FILE)
- Standardization0to1(NLIST)
- makeInputTopicAnalysisMapKey(startDate, endDate, TU, SGMT)
- getHotTopic(startDate, endDate)
- getMorph(NEWS_INFILE, SGMT, SD, ED, TU)
- cleanText(inline)
- PartitionerByCoreForBigData(InputFile, PartNum)
- RUN_CORR_TOPICLABELER(startDate, endDate, TU, SGMT)
- CORR_TopicLabeler(text_file, TU)
- RUN_TOPICLABELER(startDate, endDate, TU, SGMT)
- TopicLabeler(text_file, TU)
- RUN_TOPICMODEL(startDate, endDate, TU, SGMT)
- TOPIC_INPUT_MAKER(input_file)
- RUN_TOPIC_INPUT_MAKER(startDate, endDate, TU, SGMT)
- CRAWL_DATA_KEYWORD_MAKER(startDate, endDate, TU, SGMT)
- TMResultInputMaker(files, startDate, endDate, TU)
- RUN_TMResultInputMaker(startDate, endDate, TU, SGMT)
- makeInputTopicAnalysisMapKey(startDate, endDate, TU, SGMT)
- ALL_D_RUNNER(startDate, endDate, TU)
- H_RUNNER(startDate, endDate, TU)
- W_RUNNER(startDate, endDate, TU)
- M_RUNNER(startDate, endDate, TU)
- Y_RUNNER(startDate, endDate, TU)
- U_RUNNER(startDate, endDate, TU)
- D_RUNNER(startDate, endDate, TU)

3 Technical Support

3.1 Policy

Topic Analyser must be installed by a technician from OWLNEST on site/via remote. Policies on the source code is only included in the Fully Unlimited license.

3.2 Contact

- Person-in-Charge : Sunghee Kang
- Email : contact@owl-nest.com
- Phone No. : +82-02-742-3021

References

[1] Chang-Seok Oh, Yong-taeck Lee, Minsu Ko, Establishment and Application of ITS Policy Issues Investigation Method in the Road Section using Boundary Analysis and Text Mining, The Journal of The Korea Institute of Intelligent Transport Systems. Vol. 15 (2016)

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent Dirichlet Allocation. The Journal of Machine Learning Research, Volume 3, 993-1022 (2003)

[3] Griffiths, T. L., & Steyvers, M.: Finding scientific topic. Proceedings of the National Academy of Science, 101, 5228-5235 (2004)

[4] Steyvers, M. & Griffiths, T.: Probabilistic topic models. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), Handbook of Latent Semantic Analysis. Hillsdale, NJ: Erlbaum (2007)

[5] Wallach H., Murray I., Salakhutdinov R., & Mimno D.: Evaluation Methods for Topic Models. The Learning Workshop. ICML (2009)

[6] Asuncion A., Welling M., Smyth P., Yee Whye Teh: On Smoothing and Inference for Topic Models. In Proceedings of the Twenty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-09) (2009)

[7] Popel M.: Perplexity of n-Gram and Dependency Language Models. TSD'10 Proceedings of the 13th international conference on Text, speech and dialogue (2010)

[8] Wallach H., Mimno D., McCallum A.: Rethinking LDA: Why Priors matter. NIPS (2009)

[9] Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P.: The author-topic model for authors and documents. Proc. Conf. on Uncertainty in Artificial Intelligence (pp. 487-494) (2004)

[10] Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M.: Analyzing entities and topics in news articles using statistical topic models. In Intelligence and Security Informatics, Lecture Notes in Computer Science (2006)

[11] Hofmann, T.: Probabilistic latent semantic indexing. In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR99), 50-57 (1999)

[12] Minka, T., Lafferty, J.: Expectation-propagation for the generative aspect model. Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (2002)

[13] Gilks, W., Richardson, S., & Spiegelhalter, D.J.: Markov Chain Monte Carlo in Practice. Chapman & Hall, Suffolk (1996)

[14] Kim, D., Lee, S., Choi K., Kim, Gil.: A two-level morphological analysis of Korean. COLING '94 Proceedings of the 15th conference on Computational linguistics - Volume 1 Pages 535-539 (1994)

[15] Wallach, H. M.: Structured Topic Models for Language. Ph.D. thesis, University of Cambridge (2008)

[16] Baayen, R. H.: Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge University Press, NY (2008)

[17] Durán, P., Malvern D., Brian R., Ngoni C.: Development Trends in Lexical Diversity. Applied Linguistics Vol. 25, No. 2 (2004), pp. 220-42.