More Than Words: Collocation Retokenization for Latent Dirichlet Allocation Models

Jin Cheevaprawatdomrong
Chulalongkorn University
jin236248@gmail.com

Alexandra Schofield
Harvey Mudd College
xanda@cs.hmc.edu

Attapol T. Rutherford *
Chulalongkorn University
attapol.t@chula.ac.th

Abstract

Traditionally, Latent Dirichlet Allocation (LDA) ingests words in a collection of documents to discover their latent topics using word-document co-occurrences. Previous studies show that representing bigrams collocations in the input can improve topic coherence in English. However, it is unclear how to achieve the best results for languages without marked word boundaries such as Chinese and Thai. Here, we explore the use of retokenization based on chi-squared measures, *t*-statistics, and raw frequency to merge frequent token ngrams into collocations when preparing input to the LDA model. Based on the goodness of fit and the coherence metric, we show that topics trained with merged tokens result in topic keys that are clearer, more coherent, and more effective at distinguishing topics than those of unmerged models.

1 Introduction

Latent Dirichlet allocation (LDA) models provide useful insights into themes and trends in a large text collection through the unsupervised inference of *topics*, or probability distributions over unigram word types in the corpus (Blei et al., 2003). Topics from these models are often interpreted based on their highest-probability words, with documents expressed as vectors of proportions of each topic. Unfortunately, the context in which these tokens arise can be obscured in the bag-of-words rendering of text as unigram counts in documents. For instance, a topic with high probabilities of both "coffee" and "table" is tempting to interpret as focusing on the furniture item "coffee table", but both words could be frequent in a discussion of cafes containing no coffee tables. This problem is amplified in languages without marked word boundaries, such as Chinese and Thai: while existing tokenizers in these languages can segment characters into words, there is always a question about to what extent the tokenizers should group words together. Words that have been segmented by tokenizers may not express the concept of the original text if they were found as parts of collocations. Meaningful interpretation of topics can be lost without careful recombination of these words.

We hypothesize that the morphology of the language should play an important role in determining the suitable pre-processing steps that would improve the results of topic models. The main morphological types we consider are *synthetic language* and *analytic language*. Synthetic languages use many morphemes to compose a word and can be further divided into fusional and agglutinative languages. Fusional languages such as German differ from agglutinative languages such as Korean and Japanese: a single morpheme in fusional languages can code for many morphosyntactic features. On the other hand, analytic languages such as Thai and Chinese convey meanings by relating many words together, and morphological devices are more rarely used. Under our hypothesis, analytic languages should benefit from token merging, but synthetic languages might not because the meaning is conveyed by inflection (through bound morphemes) and agglutination (through free morphemes).

In this project, we investigate the effects of token merging as a pre-processing step, and study how those effects vary based on the writing systems and the morphological features of the languages. We evaluate three measures to determine when to merge multiple adjacent words into conceptually-unified phrasal tokens prior to LDA model training: chi-squared statistics, *t*-statistics, and raw frequency counts of phrases. We test these merging strategies on English, German, Chinese, Japanese, Korean, Thai, and Arabic. This set of languages is drawn from various writing systems and different morphological typology to see which type of

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2022, pages 2696 - 2704

language favors which type of merging strategy.

The main contributions of this paper are as follows:

- We determine through empirical studies that a *t*-statistic and raw-frequency approach to token merging improves the topic modeling results across all language types and writing systems for the corpora that do not differ much from the collocation training data.
- We also show the positive consequences of token merging: the percentage of merged tokens in the LDA training data is correlated with the quality of the topic modeling results.
- Finally, we provide evidence that the popular approach of applying a χ^2measure to token merging tends to overfit to the collocation training data and result in a low percentage of merged tokens in a number of languages, making it a less suitable general-purpose approach than *t*-statistics.

2 Related Work

Pre-processing steps can substantially alter the results of the LDA models even in languages with good tokenization heuristics such as English (Schofield and Mimno, 2016; May et al., 2016). We believe that languages that do not have clear tokenization standards deserve investigation into what kind of processing is appropriate. Many works recognize that LDA results can be improved when input are including phrases (Lindsey et al., 2012; Lau et al., 2013; Yu et al., 2013; El-Kishky et al., 2014; Wang et al., 2016; Bin et al., 2018; Li et al., 2018). We consider it valuable to specifically assess approaches to determining these phrases.

Despite their popularity in analyzing large amounts of text data, LDA models are notoriously complex to evaluate. One must evaluate both the statistical fit of a model and the human-registered thematic coherence of the words found to arise in the high-probability words, or *keys*, of a topic, which may not correlate (Chang et al., 2009). Analyses often combine evaluations of fit (Wallach et al., 2009) and automated approximations of human judgments of coherence (Bouma, 2009; Mimno et al., 2011) based on mutual information, even with the expectation these may only somewhat correlate with true human judgments (Lau et al., 2014). A limitation of these existing approaches, however, is that they expect the vocabulary and tokenization to remain constant between the two models. For our evaluation, we use a normalized log-likelihood approach to capture fit while accounting for changes in vocabulary (Schofield and Mimno, 2016).

3 Collocations as LDA Token

Collocations consist of two or more words that express conventional meaning, which can convey information about multi-word entities, context, and word usage. We hypothesize that the introduction of multi-word tokens, which capture collocations as bigrams or trigrams by way of concatenation of adjacent tokens, can help achieve more useful and coherent topic models. For languages without clear word boundaries, there is a possible additional benefit to multi-word tokens: it can be hard to intuit whether inferred word boundaries will have a large impact on the final results. Merging adjacent words into 'multi-word' tokens may help remedy the potential problem of a segmentation that is not optimal for topic modeling purposes.

Many methods are possible to select collocations to merge from tokenized text (Manning and Schutze, 1999). In this paper, we evaluate the chi-squared statistics (χ^2), the *t*-statistic and raw frequency as approaches to develop a threshold for merging collocations into multi-word tokens prior to topic model training. The chi-squared measure $\chi^2(w_1, w_2)$ and $t(w_1, w_2)$ *t*-statistic for two adjacent tokens w_1 and w_2 are defined as:

$$\chi^2(w_1, w_2) = \frac{(P(w_1, w_2) - P(w_1)P(w_2))^2}{P(w_1)P(w_2)} \quad (1)$$

$$\begin{aligned} t(w_1, w_2) &= \frac{\bar{x} - \mu}{\frac{s^2}{N}} \\ &\approx \frac{P(w_1, w_2) - P(w_1)P(w_2)}{\sqrt{\frac{P(w_1, w_2)}{N}}} \end{aligned} \quad (2)$$

We first compute the collocation measures for all bigrams on a large collocation training corpus. Then we select the top bigrams that score the highest on the collocation measures and add those to our lexicon. After we tokenize and pre-process the collection of documents on which we would like to train LDA, we retokenize the data based on the collocation training corpus. We find all of the bigrams in the LDA training data that are also found in the top bigram lexicons that we obtain from the

collocation training corpus. Then, the LDA training process proceeds as usual but with some of the original tokens merged into multi-word tokens as defined from the collocation training data.

4 Evaluation Metrics

We consider two primary evaluation metrics for exploring the effect of merging tokens: one based on log-likelihood, and one based on silhouette coefficients.

Held-Out Likelihood. When multi-word phrases are converted to individual tokens, the number of tokens in the document decreases while the size of the corpus vocabulary increases. It is therefore illogical to compare the likelihoods of the word-token model and collocation-token model directly. In order to normalize the scores between the two models that do not have the exact same vocabulary and tokens, we use the log-likelihood ratio between the LDA model likelihood and the null (unigram) likelihood for each model. In other words, we normalize the LDA model likelihood ($\mathcal{L}_{\text{model}}$) by dividing it with the unigram likelihood ($\mathcal{L}_{\text{unigram}}$) as introduced by Schofield and Mimno (2016). Therefore, the normalized loglikelihood per token ($\text{PTLL}_{\text{norm}}$) is

$$\text{PTLL}_{\text{norm}} = \frac{\log \mathcal{L}_{\text{model}} - \log \mathcal{L}_{\text{unigram}}}{N} \quad (3)$$

where N is the number of tokens. Since likelihood per token has been normalized by the unigram likelihood per token, the higher the PTLL, the better the model.

Concatenation-based Embedding Silhouette (CBES) Previous measures of topic coherence rely on statistics from the training data and assume that the vocabularies are identical for both models, which is not the case for our settings. To address this, we propose a new application of the silhouette coefficients (Rousseeuw, 1987), a common clustering evaluation metric to measure topic coherence.

A good topic should have all of its topic keys close to each other and away from other words that do not belong in the same topic. Therefore, the word embeddings of these topic keys should have shorter cosine distances within the same topic, and longer distances to the topic keys in other topics. When words are represented as a vector, this is exactly what the silhouette coefficients measure. To compute them, we first compute the $a(i)$, which is the mean cosine distance between topic-key i and other topic-keys in the same topic.

$$a(i) = \frac{1}{\mid C_i \mid -1} \sum_{j \in C_i, i \neq j} d(i,j) \quad (4)$$

where $d(i,j)$ is the distance between ith and jth topic-key and $\mid C_i \mid$ is the number of topic-keys in topic i. Then for each other topic, we compute the mean of the distance of topic-key i to topic-keys in that other topic. And $b(i)$ is the smallest of such mean among other topics.

$$b(i) = \min_{k \neq i} \frac{1}{\mid C_k \mid} \sum_{j \in C_k} d(i,j) \quad (5)$$

After obtaining $a(i)$ and $b(i)$, the silhouette coefficient for topic-key i is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \text{if } \mid C_i \mid > 1 \quad (6)$$

and

$$s(i) = 0, \text{if } \mid C_i \mid = 1 \quad (7)$$

The silhouette coefficient for the entire model is the average $s(i)$ over all i. The larger silhouette coefficient means that topic-keys are relatively similar within their topic and different from other topics.

In order to compare the distances among words merged by different criteria, all compared word embeddings must be in the same space. Since merged tokens will modify the vocabulary of the corpus, we create four versions of the word embedding training corpus: the original version and the three other versions where tokens are merged based on χ^2, t and frequency collocation measures. We train the word embeddings on these four versions of the corpus so we can then compare word embeddings on a consistent vocabulary in each retokenization scheme.

5 Experiments

We hypothesize that morphology should play an important role in determining the suitable preprocessing steps. We test our methods on one fusional language (German), two agglutinative languages (Japanese and Korean), three analytic languages (Chinese, Thai, and Arabic), and English, which can be thought of as either analytic or fusional. These languages also represent languages drawn from all writing systems: logograms (Chinese), syllabic system (Japanese), featural system (Korean), abugida (Thai), abjad (Arabic), and true alphabets (English and German).

	Domains	Docs (K)	Tokens (M)	%Merged		
				CHI	T	FREQ
EN-NYTimes	News	53	0.7	1.64	12.71	12.72
EN-SOTU	Speeches	42	0.8	0.86	9.76	10.33
EN-Yelp	Restaurants	67	2.1	0.16	7.85	8.97
DE-10kGNAD	News	222	1.9	0.09	7.46	7.68
CN-Chinanews	News	49	0.8	0.00	11.61	11.64
CN-Dianping	Restaurants	40	0.8	0.01	2.82	2.80
CN-Douban	Movies	98	0.6	0.03	4.17	4.23
JA-JapanNews	News	528	3.6	21.74	21.95	21.85
KO-KAIST	Misc	20	0.2	19.82	20.71	21.27
TH-Prachathai	News	32	4.4	0.07	15.97	14.06
TH-Wongnai	Restaurants	40	1.2	0.00	8.52	6.09
TH-BEST	Misc	7	2.1	0.03	14.94	13.09
TH-TNC	Misc	4	1.0	0.03	13.65	12.00
AR-ANT	News	60	1.1	0.16	26.13	27.45

Table 1: A survey of corpora providing the number of documents and tokens, as well as the percentage of unigram tokens merged using each approach.

The English corpora are drawn from The New York Times (Sandhaus, 2008), the Yelp Dataset[1], and United States State of the Union addresses (1790 to 2018) divided into paragraphs[2]. The German data come from Ten Thousand German News Articles Dataset[3]. The Chinese data come from three corpora: the news articles from Chinanews[4], restaurant reviews from Dianping[5], and the movie reviews from Douban[6]. The Japanese data is from the Webhose's Free Datasets[7]. The Korean data come from the KAIST Corpus[8]. The Thai data come from the news articles in Prachathai[9], the restaurant reviews from Wongnai[10], the BEST corpus[11], and the Thai National Corpus (Aroonmanakun, 2007). The Arabic data come from the Antcorpus (Chouigui et al., 2017). Each corpus is separated into 75% training documents and 25% test documents (Table 1).

We train the χ^2, t, and frequency-based tokenizers for each language on Wikipedia articles for that language. For all languages, we use the reduced version of Wikipedia database, except for English we use the filtered Wiki103 dataset (Merity et al., 2016). English, German, Chinese, Japanese, Korean, Thai and Arabic documents are tokenized with NLTK (Bird, 2006), SoMaJo (Proisl and Uhrig, 2016), Stanford Word Segmenter (Tseng et al., 2005), Fugashi (McCann, 2020), KoNLPy (Park and Cho, 2014), Attacut (Chormai et al., 2020) and Camel-tools (Obeid et al., 2020) respectively. For each criterion, we create a list of 50,000 top bigrams that have the highest scores. These lists of top bigrams will be used to merge words in the input of the LDA, effectively training a new tokenizer.

[1]www.yelp.com/dataset
[2]www.kaggle.com/rtatman/state-of-the-union-corpus-1989-2017
[3]github.com/tblock/10kGNAD
[4]www.chinanews.com
[5]github.com/zhangxiangxiao/glyph
[6]www.kaggle.com/utmhikari/doubanmovieshortcomments
[7]webhose.io/free-datasets/japanese-news-articles/
[8]semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus
[9]github.com/PyThaiNLP/prachathai-67k
[10]www.kaggle.com/c/wongnai-challenge-review-rating-prediction
[11]thailang.nectec.or.th/downloadcenter

	χ^2-t	χ^2-freq	t-freq
English	8.90	7.78	74.87
German	0.00	0.00	83.06
Chinese	0.00	0.00	86.48
Japanese	29.06	22.60	73.34
Korean	10.56	7.34	71.95
Thai	0.22	0.06	67.25
Arabic	1.22	1.20	66.89

Table 2: The percentage of overlapping merged tokens between two methods of retokenization computed on the retokenization training data. t and χ^2yield similar results for all languages.

To train word embeddings, we use the `gensim` (Řehůřek and Sojka, 2010) implementation with the Continuous Bag-of-Word (CBOW) algorithm (Mikolov et al., 2013) to obtain word embeddings. The training corpora and their collocation versions are prepared based on the tokenizers that we discuss above. We preprocess the word embedding training data and the LDA training data the same way. For English, we lemmatize and lowercase the data. For Korean, Japanese, and Arabic, we lemmatize the data. For German, Chinese, and Thai, we do not do any normalization.

We use `MALLET` (McCallum, 2002) implementation of LDA with the default hyperparameters to train and evaluate topic models in both word and multi-word (collocation) documents with 10, 50, 100 topics. We run the experiment 3 times for each combination of corpus, type of retokenization (no retokenization, χ^2, t or frequency) and number of topics to compute the means of the normalized held-out likelihood and CBES, discussed in section 4.

6 Results and Discussion

The normalized log-likelihood per token of the t and frequency-based retokenization is significantly higher than the baseline for English, German, Chinese, Japanese, Korean, and Arabic for all text collections and the number of topics except EN-Yelp, TH-BEST, and TH-TNC (Table 3). Frequency-

	10 topics				50 topics				100 topics			
	Word	χ^2	*t*	freq	Word	χ^2	*t*	freq	Word	χ^2	*t*	freq
EN-NYTimes	.3646	.3675	.4119	**.4386**	.5214	.5225	.5766	**.6128**	.5588	.5533	.6050	**1.0492**
EN-SOTU	.2699	.2660	.2967	**.3145**	.3809	.3809	.4122	**.4430**	.4135	.4101	.4367	**.4705**
EN-Yelp	.1597	.1607	.1833	**.2021**	.2589	.2599	.2893	**.3169**	.3357	.2822	.3130	**.3412**
DE-10kGNAD	.4982	.5001	.5233	**.5251**	.7272	.7272	.7622	**.7651**	.7784	.7809	.8122	**.8188**
CN-Chinanews	.5033	.5046	.5510	**.5592**	.7647	.766	.8170	**.8344**	.8427	.8394	.8847	**.9044**
CN-Dianping	.2557	.2574	.2644	**.2659**	.3899	.3906	.3965	**.4013**	.4188	.4212	.4255	**.4263**
CN-Douban	.2966	.2955	.3076	**.3092**	.4048	.4073	.4144	**.4173**	.4294	.4301	.4332	**.4374**
JA-JapanNews	.4540	**.7803**	.5942	.6342	.7173	.9268	.9339	**.9926**	.8088	1.0325	1.0316	**1.1003**
KO-KAIST	.2901	**1.0315**	.4589	.5442	.6446	.6833	.7152	**.8390**	.4755	.7437	**1.3443**	.9221
TH-Prachathai	.4367	.4331	**.4756**	.4743	.7052	**.8458**	.7699	.7719	.7854	.7854	.8537	**.8548**
TH-Wongnai	.2048	.2013	**.2225**	.2192	.3237	.3222	**.3472**	.3399	.3467	.3463	**.3720**	.3636
TH-BEST	**.6995**	**.6995**	.6704	.6838	.9148	.9190	.9279	**.9389**	.9812	.9819	.9967	**1.0100**
TH-TNC	.7420	**.7422**	.7079	.7239	.9969	.9952	1.0079	**1.0219**	1.0508	1.0473	1.0608	**1.0758**
AR-ArabicNews	.3183	.3152	.4676	**.5663**	.4923	.4913	.7175	**.8742**	.5417	.5409	.7681	**.9355**

	10 topics				50 topics				100 topics			
	Word	χ^2	*t*	freq	Word	χ^2	*t*	freq	Word	χ^2	*t*	freq
EN-NYTimes	.0143	.0153	.0246	**.0453**	-.0582	-.0625	-.0544	**-.0487**	-.0876	-.0875	-.0783	**-.0780**
EN-SOTU	.0034	-.0013	.0070	**.0100**	-.0602	-.0597	-.0595	**-.0527**	-.0812	-.0823	-.0793	**-.0743**
EN-Yelp	-.0634	-.0548	-.0465	**-.0337**	-.1117	-.1085	-.1023	**-.0952**	-.1299	-.1290	-.1179	**-.1153**
DE-10kGNAD	-.0209	-.0244	-.0190	**-.0134**	-.0804	-.0860	-.0785	**-.0680**	-.0753	-.0730	-.0655	**-.0599**
CN-Chinanews	.0002	.0018	.0152	**.0162**	-.0523	-.0559	-.0456	**-.0388**	-.0699	-.0712	-.0665	**-.0620**
CN-Dianping	**-.0708**	-.0854	-.0714	-.0744	**-.1278**	-.1316	-.1317	-.1339	**-.1373**	-.1439	-.1446	-.1439
CN-Douban	-.0226	-.0140	**-.0078**	-.0095	**-.0847**	-.0854	-.0864	-.0850	**-.1037**	-.1041	-.1073	-.1053
JA-JapanNews	-.0925	-.0655	-.0562	**-.0133**	-.1503	-.1010	-.0977	**-.0716**	-.1644	-.1120	-.1106	**-.0915**
KO-KAIST	-.0608	-.0315	-.0317	**-.0191**	-.0895	-.0691	-.0664	**-.0503**	-.0868	-.0698	-.0726	**-.0592**
TH-Prachathai	**-.0039**	-.0092	-.0040	.0160	-.0806	-.0797	-.0684	**-.0623**	-.1137	-.1121	-.0939	**-.0896**
TH-Wongnai	**-.0667**	-.0672	-.0733	-.0726	-.1468	-.1530	**-.1462**	-.1505	-.1761	**-.1709**	-.1738	-.1767
TH-BEST	-.0278	-.0187	-.0248	**-.0095**	-.0987	-.0977	-.0987	**-.0927**	-.1145	-.1153	-.1086	**-.1007**
TH-TNC	-.0284	-.0324	**-.0133**	-.0271	-.1079	-.1053	-.1332	**-.0964**	-.1281	-.1274	-.1297	**-.1175**
AR-ArabicNews	-.0695	-.0673	-.0496	**.0124**	-.1255	-.1129	-.0834	**-.0434**	-.1355	-.1309	-.1010	**-.0735**

Table 3: Normalized unigram log-likelihood per token (top) and Concatenation-based Embedding Silhouette (CBES) scores (bottom) for between the baseline and retokenization models: χ^2 , textitt, and raw frequency. Shaded cells mean that the results are inferior to the baseline, while bolded cells show the best results for each corpus and number of topics.

based retokenization gives the best results for most settings but not significantly higher than *t* retokenization. However, we observe mixed results from χ^2retokenization for some languages. This is quite surprising because raw frequency was previously found to be an inferior measure of collocation. This suggests that *t* and frequency-based retokenization might be a more reliable method for improving the goodness of fit of the LDA model. This also suggests that Japanese and Korean might have some specific quality that interacts well with all three types of retokenization.

Similarly, we observe a general improvement in coherence for the *t* and frequency retokenization (Table 3). The higher CBES score indicates that topic-keys are more semantically coherent and topics are more distinct. The coherence improves after *t* and frequency-based retokenization for English, Japanese, Korean, and Arabic corpora regardless of the number of topics. The improvement for Thai is spotty, and Chinanews is the only Chinese corpus in which we see improvement. This suggests that the choice of retokenization strategy might depend on the language types or the content of corpora itself. Consistent with the normalized log-likelihood results, Japanese and Korean corpora interact well with all three types of retokenization, suggesting that the morphology or typology of these two languages consistently benefit from collocation before training LDA models.

What could account for this discrepancy across languages and corpora? First, we observe a large variation of percentages of merged tokens across corpora. Because we fix the number of bigrams types to merge during the tokenizer training process to 50,000 for all three criteria (Table 1), we can use this analysis to find trends in the relative frequency of merged tokens. We see that χ^2 retokenizer only merges barely 1% of all the tokens before training the LDA models for English, Chinese,

χ^2: **dvenadsat apostolov, jormp jomp, malwae tweep, aboul gheit, achduth vesholom, adavari matalaku, adeste fideles**, afforementionede oughtt, **agoraf drws, aht urhgan, akanu ibiam, aksak maboul, alberthiene endah, alfava metraxis, alfonsas eidintas, allasani peddana, alteram partem, amantes clandestinos, amarin winitchai, amel oluna**
***t*: united states, new york, world war**, km h, take place, miles km, **los angeles, united kingdom**, first time, high school, tropical storm, **new zealand, war ii**, video game, mph km, h mph, **north america, air force**, two years, peak number
frequency: united states, new york, world war, km h, take place, miles km, first time, **los angeles, united kingdom**, high school, tropical storm, **new zealand**, video game, **war ii**, mph km, two years, h mph, **north america, air force**, peak number

χ^2: うそ寒い 肌寒, **ぎっこん ばったん**, ざらり ぐらり, へへへへ へへへ, **アウレオルス ボンバストゥス, アジタ ケサカンバリン, アッシャルク アルアウサト, アトミズム アドリアシン, アドリアシン アドリアマイシン, アルパイ オザラン, アワサカ ツマオ, イブリツモマブ チウキセタン, ウダヤン プラサッド, ウラマツ サミタロウ, エウグランディナ ロセア, エストラムスチン エストラサイト, オクタクロルテトラヒドロ メタノフタラン, オドネ センデロル, オランバヤル ビャンバジャブ, クツミ ソクチュウ**
***t*:** 年 月, る 居る, 月 日, る 事, 其の 後, 成る 居る, 昭和 年, 事 出る, 年 昭和, 於く り, 年 年, 成る, 事 有る, 事 成る, 使用 る, 物 有る, 存在 る, 平成 年, 第 回, る 年
frequency: る 居る, 年 月, 月 日, る 事, る 年, 年 年, 成る 居る, 居る 年, 其の 後, 事 有る, 昭和 年, る , る 其の, 事 成る, 事 出る, 年 昭和, 有る 年, 成る, 使用 る, 於く り

χ^2: **가닛 알훤소, 가욋일 봇일, 가츠테루 우루샤, 가톨리콘 앰뷸, 갈뀨 가실뀨, 갈라람 알부담, 감민 월민, 감성채 널@21, 갑복 갑규, 강첸 키숑, 강취완 강취일, 강홍업 강효업, 강홍선 강홍익, 개영 궤영, 개초항 거륜항, 개튀의얄 똥퍼먹는, 객렬액 겁렬액, 갤러 리@KCUA**, 갤런에서 갤런으로, **거대유방증 대유방**
***t*:** 적 인, 하다 수, 한 다, 위 한, 말 하다, 시작 하다, 사용 하다, 못 하다, 수 없다, 위치 한, 하다 않다, 사용 되다, 하다 위해, 가지 고, 기도 하다, 일반 적, 되다 않다, 존재 하다, 기록 하다, 은 대한민국
frequency: 적 인, 하다 수, 하다 하다, 한 다, 사용 하다, 말 하다, 시작 하다, 하다 않다, 위 한, 못 하다, 수 없다, 위치 한, 하다 위해, 하다 는, 사용 되다, 기록 하다, 되다 않다, 하다 되다, 기도 하다, 활동 하다

Figure 1: The top 20 collocations from each retokenization methods. χ^2 favor proper names (bold-faced) more heavily than the other two methods.

German, Arabic, and Thai corpora, possibly introducing noise in the data that yield the results similar to or worse than the baseline. In contrast, the *t* and frequency-based retokenizers merge around 8%- 15% of all the tokens for English, German, and Chinese. Arabic has seen the highest merging percentage of 26%-27%. Notably, around 20 % of tokens are retokenized by all three retokenizers in Japanese and Korean. The truncation of the top χ^2bigrams list might cause this different behavior. The number of χ^2collocations that pass the hypothesis testing is significantly larger than that of *t* collocations. For example, there are 3.73 million χ^2collocations versus 231 thousand *t* collocations in Thai for the same significance level $\alpha = 0.005$. This full list of χ^2collocations includes all the top collocations from the *t* score and frequency treatments, implying that were we to use this significance threshold, the percentage of merged word would be at least as high as the two methods. However, the large vocabulary that the χ^2approach induces is impractical in many applications, suggesting it is an inefficient approach if the goal is primarily to merge frequent ngrams.

Another possible effect these results may show is that the writing system or the morphology could account for this notable discrepancy in retokenization percentage across languages. For English, the top 20 χ^2collocations are primarily specific named entities, but the *t* and frequency-based retokenizers yield more general compound nouns and common phrases (Figure 1). As the top 50,000 χ^2collocations contain primarily rare words, these are expected to co-occur rarely enough that even a few co-occurrences can trigger significance. Therefore, when we use this truncated list of rarely-occurring χ^2collocations, we generally see a very low merged token percentage.

The quality of retokenization impacts both the goodness of fit the model, as indicated by the normalized log-likelihood score, and the coherence of the model, as indicated by the CBES score. Within the same language, news corpora have higher percentages of merged words when merged with *t* and frequency collocations, while corpora containing restaurant and movie reviews tend to see lower percentages (Table 1). This could be because the news corpora are in a similar domain to that of the Wikipedia which we use to build the list of co-occurring words. A good retokenizer (in our cases, trained on Wikipedia data) should generalize well and recognize many collocations in a new corpus, which differs somewhat from the retokenizer training data. We found a significant positive correlation between merge percentage and the margin of improvement over the baseline (the difference between the PTLL of the model without retokenization and the PTLL or CBES of the model

	Word	$\chi 2$	t	Freq
EN-SOTU	security social program system benefit welfare legislation need must reform propose congress health retirement administration meet enact national work insurance	health care security social insurance welfare work americans reform system cost benefit program must make need help plan pay retirement	americans **social_security** **health_care** cost families benefit pay plan save system american help reform care retirement tax medicare coverage work make	**social_security** welfare **health_care** system benefit families insurance reform cost care health save americans retirement medicare work must pay coverage workers
DE-10kGNAD	de Spanien Madrid spanischen El Barcelona Mexiko Messi Brasilien Chile Rousseff spanische Valencia Venezuela Kuba La USA Präsidentin Real Luis	FC Der Madrid Barcelona Bayern Real Gruppe City Manchester League München Hinspiel Tore United In Die Spanien Minute Trainer Champions	Der Hinspiel Madrid Bayern Spanien Barcelona spanischen Valencia **Real_Madrid** Atletico **Champions_League** Messi Real **FC_Barcelona** Trainer Tore Saison Liverpool Gesamtscore Arsenal	Der Trainer Hinspiel Die Janko **Champions_League** Alaba Bayern Valencia Minute Saison Tore **Real_Madrid** Madrid Atletico Messi Real David Barcelona **FC_Barcelona**
CN-Chinanews	世界杯 巴西 国际 足联 时间 南非 比赛 足球 预选赛 球队 届 中新网 凌晨 抽签 进行 北京 强 小组 欧洲 支	世界杯 巴西 南非 时间 足球 欧洲 比赛 届 场 球队 德国 杯 支 球场 球迷 葡萄牙 非洲 法国 阿根廷 凌晨	世界杯 **国际_足联** 巴西 足球 南非 球迷 主席 **巴西_世界杯** 体育 法国 球场 天 俄罗斯 **南非_世界杯** 民众 次 卡塔尔 布拉特 标志 德国	世界杯 巴西 **国际_足联** **巴西_世界杯** 女足 足球 南非 国足 昨天 **北京_时间** 球迷 国家队 男足 中国 中国队 今天 预选赛 小组赛 强 球队
JA-JapanNews	月 日 御 年 為る 期間 限定 居る 発売 中 商品 販売 時 キャンペーン 購入 円 成る 下さる 頂く 掲載	**億_円** 大手 事業 商品 社 利益 日 社長 企業 市場 同社 米 投資 サービス 同 ブランド 中国 因る 海外 販売	月 成る 円 店 今年 同 **億_円** 年 比 前年 増 調査 日本 増える 販売 利益 **年_月** 日 料金 別	御 キャンペーン 商品 ロッテ 限定 円 纏め 掲載 **期間_限定** 此の 下記 下さる 無料 実況 **期間_中** クーポン セール **開催_為る_居る** 買う 頂く
KO-KAIST	- 교수 하다 대 연구 는 팀 한 되다 연 황 서울대 현 광고 미국 연구원 신 밝히다 검증 이라고 수첩	교수 서울대 황 대 **교수_는** 연구 팀 **말_하다** 인 도 줄기세포 취재 검증 본보다 대한 수첩 연구원 이르다 논문 논란	교수 황 서울대 팀 연구 광고 대 취재 수첩 인 줄기세포 **교수_는** 김 본보다 연구원 한편 프로그램 한 논문 **이라고_밝히다**	교수 서울대 황 팀 연구 대 **교수_는** 줄기세포 연구원 수첩 의학 취재 신 한 인 본보다 검증 논문 **말_하다** 대한
TH-Prachathai	แดง เสื้อ คน แกนนำ ชุมนุม นปช. สี ไทย นิรโทษกรรม เมือง เหลือง เวที รัฐบาล ทหาร สลาย จำนวน เคลื่อนไหว ปราศรัย เสื้อแดง ประชาชน	ชุมนุม คน เสื้อ แดง แกนนำ เวลา นปช. ตำรวจ หน้า สลาย น. เวที สถานการณ์ พื้นที่ ประกาศ ปราศรัย เหตุการณ์ บริเวณ ประชาชน ฉุกเฉิน	ชุมนุม **คน_เสื้อ_แดง** แกนนำ คน สลาย รัฐบาล **เสื้อ_แดง** ประชาชน เหตุการณ์ รุนแรง นปช. พื้นที่ เรียกร้อง เวที เจรจา เคลื่อนไหว ประกาศ สถานการณ์ กปปส. ตำรวจ	ชุมนุม แกนนำ คน **คน_เสื้อ_แดง** นปช. เวที **เวลา_น.** ปราศรัย **เสื้อ_แดง** รัฐบาล สลาย กปปส. หน้า เคลื่อนไหว ประกาศ เดินทาง เรียกร้อง ประชาชน ตำรวจ พันธมิตรฯ
AR-ANT	شَرِكَة فِي غاز طاقَة أَنَّ تُونِسِيّ كَهْرَباء إِنْتاج مِن نَشاط مُدِير كَهْرَبائِيّ عَن سَبَب مَصْنَع مَنْجَم عامِل عامّ بِتْرُولِيّ مُتَجَدِّد	شَرِكَة فِي غاز طاقَة مِن إِنْتاج تُونِسِيّ كَهْرَباء عَن نَشاط مُدِير مَصْنَع أَنَّ عامِل مَنْجَم صِناعَة بِتْرُولِيّ قِطاع مُتَجَدِّد عامّ	شَرِكَة إِنْتاج تُونِسِيّ نَشاط غاز مُؤَسَّسَة مَصْنَع وَطَنِيّ عامِل مَنْجَم قَفَص بِتْرُولِيّ مَجْمَع إِلَى الفسفاط **طاقَة_مُتَجَدِّد** مُدِير_عامّ اِسْتِغْلال مادَّة **ماء_مَعْدِنِيّ**	تُونِسِيّ غاز شَرِكَة كَهْرَباء نَشاط قَفَص مَنْجَم إِنْتاج مَصْنَع إِفْلِيم سَبَب **صَباح_إِلَى_ساعَة_ثانِي_بُعْد_زَوال** بِتْرُولِيّ **طاقَة_مُتَجَدِّد** طاقَة أُفْرِيل شُغْل الفسفاط **يَوْم_أَحَد سِيرَة_ذاتِيّ**

Figure 2: Topic keys comparison in languages.

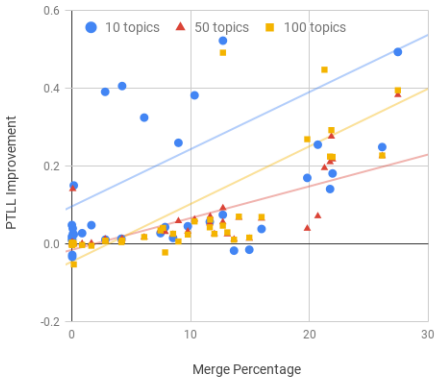

Figure 3: PTLL improvement vs. merged percentage.

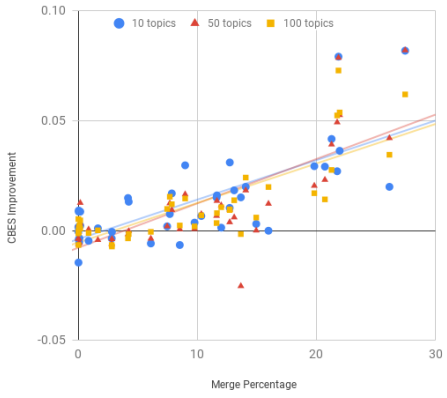

Figure 4: CBES improvement vs. merged percentage.

with retokenization). Pooling across all languages and corpora, we found the correlation coefficients of 0.41, 0.77, and 0.68 for the models with 10, 50, and 100 topics respectively for PTLL. As for the coherence metric, we found the correlation coefficients of 0.73, 0.76, and 0.79 for the models with 10, 50, and 100 topics respectively for CBES. This means the models with higher merge percentages are better than their corresponding word models in reproducing the statistics of the held-out data. This suggests that the quality of the LDA models depends on the generalizability of the retokenizers.

The LDA model results become more understandable when certain tokens are retokenized. We see merged tokens in the topic key sets of almost all topics in all corpora when retokenized based on *t* or raw frequency. Many of these represent non-compositional meanings that might have been lost without retokenization: for example, the collocation "social security" is not fully represented by the individual tokens "social" or "security" separately. More strikingly, the collocation '*kōn sŭa dāng*' refers to a political movement group in Thailand. When it is separated into *kōn* (people) *sŭa* (shirt) *dāng* (red), the key meaning is totally lost. When we compare by looking at the topic-keys of the word and multi-word models, we can come up with similar topics because we as a human who understands English and has general knowledge of the world can make the connection based on surrounding topic-keys even though they are not explicitly merged. However, if we want to use these topic keys as input to other downstream tasks such as information retrieval or text classification, the merged tokens help retain the specificity of the "red shirt people" as a meaningful entity distinct from the phrase's constituting parts.

7 Conclusion

In this work, we improve the quality of LDA models by better processing the input text before training the model. We found that the retokenizers trained based on *t* statistics and raw frequency yield an improvement across all languages considered in this study, while the χ^2approach was a less efficient approach that focuses more on rare named entities than common noun phrases. Using retokenizers ensures that LDA models can fit better to the data, the topic keys are more coherent, and the topics are more distinct. Outputs from retokenization with *t* statistics and frequency approaches yield common noun phrases in the most frequent terms of topics that represent a significant aid to both direct topic interpretation and expected utility of these topics in downstream tasks.

Acknowledgments

This project is partially supported by Grants for Development of New Faculty Staff, Ratchadaphiseksomphot Endowment Fund. The authors would like to thank Vincent Ng, who provided us with very insightful comments as a Student Research Workshop mentor. We are also grateful for the suggestions from the anonymous reviewers from the previous submission.

References

Wirote Aroonmanakun. 2007. Creating the thai national corpus. *MANUSYA: Journal of Humanities*, 10(3):4–17.

GE Bin, Chun-hui HE, Sheng-ze HU, and GUO Cheng. 2018. Chinese news hot subtopic discovery and recommendation method based on key phrase and the lda model. *DEStech Transactions on Engineering and Technology Research*, (ecar).

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Pattarawat Chormai, Ponrawee Prasertsom, Jin Cheevaprawatdomrong, and Attapol Rutherford. 2020. Syllable-based neural thai word segmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4619–4637.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2017. Ant corpus: an arabic news text collection for textual classification. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 135–142. IEEE.

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han. 2014. Scalable topical phrase mining from text corpora. *Proc. VLDB Endow.*, 8(3):305–316.

Jey Han Lau, Timothy Baldwin, and David Newman. 2013. On collocations and topic models. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(3):1–14.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Bing Li, Xiaochun Yang, Rui Zhou, Bin Wang, Chengfei Liu, and Yanchun Zhang. 2018. An efficient method for high quality and cohesive topical phrase mining. *IEEE Transactions on Knowledge and Data Engineering*, 31(1):120–137.

Robert Lindsey, William Headden, and Michael Stipicevic. 2012. A phrase-discovering topic model using hierarchical pitman-yor processes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 214–222.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Chandler May, Ryan Cotterell, and Benjamin Van Durme. 2016. An analysis of lemmatization on topic models of morphologically rich language. *arXiv preprint arXiv:1608.03995*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Paul McCann. 2020. fugashi, a tool for tokenizing Japanese in python. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 44–51, Online. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Eunjeong L. Park and Sungzoon Cho. 2014. Konlpy: Korean natural language processing in python. In *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, Chuncheon, Korea.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics (ACL).

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Huihsin Tseng, Pi-Chuan Chang, Galen Andrew, Dan Jurafsky, and Christopher D Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1105–1112, New York, NY, USA. Association for Computing Machinery.

Minmei Wang, Bo Zhao, and Yihua Huang. 2016. Ptr: phrase-based topical ranking for automatic keyphrase extraction in scientific publications. In *International Conference on Neural Information Processing*, pages 120–128. Springer.

Zhiguo Yu, Todd R Johnson, and Ramakanth Kavuluru. 2013. Phrase based topic modeling for semantic information processing in biomedicine. In *2013 12th International Conference on Machine Learning and Applications*, volume 1, pages 440–445. IEEE.