Shallow Text Clustering Does Not Mean Weak Topics: How Topic Identification Can Leverage Bigram Features

Julien Velcin[1], Mathieu Roche[2], and Pascal Poncelet[3]

[1] Université de Lyon (ERIC, Lyon 2), France.
Julien.Velcin@univ-lyon2.fr,
[2] Cirad (TETIS), Montpellier, France.
mathieu.roche@cirad.fr
[3] Université de Montpellier (LIRMM), France.
Pascal.Poncelet@lirmm.fr

Abstract. Text clustering and topic learning are two closely related tasks. In this paper, we show that the topics can be learnt without the absolute need of an exact categorization. In particular, the experiments performed on two real case studies with a vocabulary based on bigram features lead to extracting readable topics that cover most of the documents. Precision at 10 is up to 74% for a dataset of scientific abstracts with 10,000 features, which is 4% less than when using unigrams only but provides more interpretable topics.

1 Introduction

Text clustering is a huge research area with many applications, such as corpus visualization [12] and document indexing for information retrieval [25]. In addition to the classical task of categorizing similar texts, people are usually interested in characterizing the clusters by the mean of concise descriptions called topics, so that they can easily interpret categories and browse the document collection [24]. Topic extraction (or topic learning) has been widely popularized by the success of Latent Semantic Analysis [4] and Non-negative Matrix Factorization [18]. More recently, probabilistic topic models, such as probabilistic Latent Semantic Analysis [8] and Latent Dirichlet Allocation [1], have emerged as an efficient alternative implemented by many communities, from data mining [19] to natural language processing [7] and social sciences and humanities [21]. They are now used as a routine in many systems dedicated to text analytics [2].

Despite all these numerous works, it turns out that some confusion often subsists between the task of **text clustering** (grouping similar texts, i.e. working on category's *extension*) and the task of **topic identification** (extracting within-category commonalities, their *intension*), as highlighted by [26]. We show here that not-so-good (shallow) clustering does not always mean weak topics. Another observation is related to the vocabulary used by the algorithms: most of the time, groups and topics are estimated from unigram tokens (words) [17], whose number is often arbitrarily fixed, or not fully justified [7]. When considering perplexity-based measures only, that is the goodness-of-fit of the probabilistic model on held-out data, words seems to play the main role [9]. However, it has been shown that n-grams ($n \geq 2$) might be really useful, whether for constructing interpretable topics [23] or for improving topic consistency [16,28].

In: P. Cellier, T. Charnois, A. Hotho, S. Matwin, M.-F. Moens, Y. Toussaint (Eds.): Proceedings of DMNLP, Workshop at ECML/PKDD, Riva del Garda, Italy, 2016.

Based on these two observations, our contribution is twofold.

First, we show that a minimum number of features is necessary but sufficient to achieve a good accuracy, both in term of clustering purity and topic description. It is not as obvious as it seems since too many features might add noise and reduce the generalization ability of the model, which actually happens in supervised settings [13]. If we pay attention to select enough features, it is therefore possible to choose phrases (here, bigrams) instead of single words. To the best of our knowledge, it is the first time that this result is clearly highlighted and quantified.

Second, we show that the bigram-based vocabulary provide really useful topic descriptions at the cost of a reasonable decrease in accuracy. The cost is not that important with a drop of about 10%. Our results highlight that a careful choice for the features allows a much better interpretation of the topics given by topic learning techniques (here, LDA). This work is closely related to the task of topic labeling but, here, the descriptive features are defined *before* the topic learning step. Therefore the extracted topics are characterized by the very terms that constitute their backbone, and not labeled by using one among the many heuristics proposed in the literature [15,29]. Besides, a post-processing can be used afterwards to improve the output, such as selecting one term amongst "data set" and "data sets" (see Section 3).

The paper is organized as follows. Section 2 defines the two complementary tasks of text clustering and topic identification, highlighting their close connection but also their difference. Section 3 shows the impact of making the vocabulary change in term of both size and nature (unigrams versus bigrams). Finally, we conclude and suggest future work in Section 4.

2 Text Clustering and Topic Identification: Two Related Tasks

2.1 Definition of Tasks

The first step consists in showing the clear distinction between the two tasks. As illustrated in Fig. 1 (left), text clustering mainly aims at categorizing objects into clearly separated clusters. Even though the membership can be gradual, like in fuzzy clustering [5], or an object can be associated to several clusters, like in overlapping clustering [3], the common aim is to associate each object to one category so that subsequent decisions can be made. In Fig. 1, we can observe that some texts are central to the categories (e.g., d_c for cluster 1 and d_d for cluster 2) whereas other texts lie between clusters (e.g., d_a, d_b and d_e). It is a natural feature of text clustering to assume that texts can be related to *several* topics at the same time, which is at the basis of most topic models.

By adopting a different viewpoint, topic identification is more dedicated to extracting a set of topics that structure the dataset as shown in Fig. 1 (right). Topics can be viewed as weighted lists of keywords (e.g., with LSA) or distributions over words (e.g., with LDA). In order to give an overview of the whole corpus to the final users, the usual solution is to keep the top words (option 1 in Fig. 1) or the top n-grams (option 2 in Fig. 1, here with n=2).

Obviously, the two tasks are related but not fully aligned. Hence, the documents d_a, d_b and d_e can be misclassified as long as we find the expected topics, more identifiable

on the colored groups of documents in Fig. 1. Let us note that we might easily get the top frequent terms for each cluster as a post-processing stage. However, most of the current state-of-the-art methods such as LSA, NMF and LDA address both tasks at the same time, which explains the confusion that may arise.

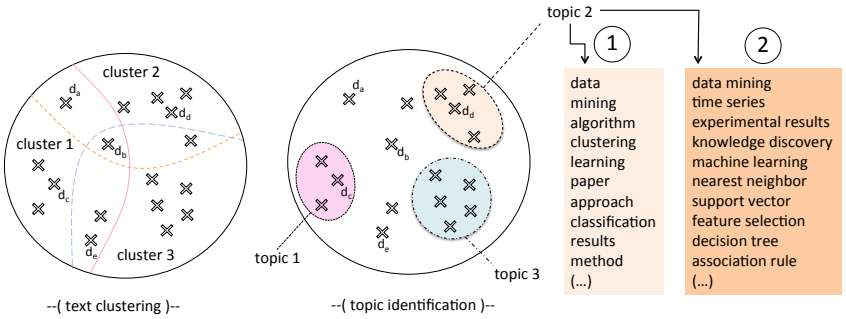

Fig. 1. Distinction between the tasks of text clustering and topic identification (here, topic 2 has been extracted from scientific publications clearly related to the "data mining" field).

Several previous works have used n-grams either during the topic learning process [22,23] or as a post-processing step in order to find automatic labels [10,15,29]. However, they did not study the impact of both the vocabulary size (number of terms) and term nature (unigrams versus bigrams), as we do in this paper.

2.2 Evaluation Measures

In the following sections, we experiment LDA on two datasets in order to address the two tasks simultaneously. For evaluation's sake, we compare the output given by LDA to a gold standard that provides us the real class label of each object. In order to get a partition, we associate each text d to the most likely topic $\hat{z} = \arg\max p(z/d)$. This way, we reduce the expressive power of topic models but we can leverage the usual Adjusted Rand Index (ARI) for assessing the clustering quality. We will see in the next section that the two datasets have been precisely chosen because they fit this crisp clustering assumption. The maximum of 1 with ARI is achieved with a perfect match between the partition and the gold standard.

Establishing a ground truth for the topic identification task is much more challenging. To begin with, we choose to restrict the evaluation to the quality of the top-10 terms associated to each cluster for 10 is the number usually shown to end users. Although we can imagine various ways to extract those lists from the gold standard partition (e.g., selecting the most discriminant terms, etc.), we have chosen to restrict to the most frequent terms for this study. In addition to the simplicity of this solution, we will see that the probability $p(t/z)$ of the term t given the topic z output by LDA clearly favor frequent terms. We then propose to calculate the usual precision for this top-10 terms, noted pre@10. Let us remark that this manner to challenge the list of top K terms is

especially uncommon in the literature, in which the list is always manually evaluated. The mapping between the true category and the topic z is simply derived by taking the category with the higher number of texts related to z. Obviously, pre@10 ranges from 0 (no common term) to 1 (perfect match between the two lists).

2.3 Datasets and Feature Extraction

dataset	#c	#docs	#unigrams	#bigrams
ART	5	18,465	13,778	30,522
20NG	20	18,828	40,142	54,741

Fig. 2. Basic statistics for the two datasets.

The two datasets are the set of scientific abstracts gathered by Tang. et al. [20], noted ART, and 20 Newsgroups, noted 20NG. Both datasets are available online[4]. For both datasets, we perform minimal preprocessing: lowercasing, removing punctuation and English stopwords, removing the terms that occur in only one document. We set the number of expected topics to be the true number of classes #c in the gold standard. Basic statistics can be found in Fig. 2.

In our context, we extract bigrams based on classical patterns in terminology extraction domain (i.e. noun-noun, adjective-noun, and so forth)[5]. Terms extracted from our corpora are then ranked depending on their relative frequency. Other weightings have been experimented (e.g., *TF-IDF*, *Okapi*, *C-value*) but it turns out that the frequency is the more adapted ranking function for both tasks addressed in this study[6].

3 Vocabulary Impact for Both Tasks

We here focus our attention on the importance of vocabulary size and term nature (unigrams versus bigrams). We used the parallel LDA implemented in the MALLET package[7]. The priors α and β are not automatically estimated (default configuration) but we set them both to 0.1 after a preliminary grid search[8]. We set the maximum number of iterations for the Gibb's sampling to 2000, as suggested with this implementation, and run the algorithm ten times. The final mean is only given since the observed standard deviation does not exceed 0.01, so we decided not to overload the figures.

We keep the most frequent K words, K ranging from 500 to 30,000. We then compute the quality of LDA topics both for clustering (Fig. 3) and topic identification

[4] http://arnetminer.org/collaboration and http://qwone.com/ jason/20Newsgroups/

[5] To this end, we used the biotex tool [11], freely available online: http://tubo.lirmm.fr/biotex/.

[6] For instance, the ARI based on the frequency is higher from 0.25 to 0.31 for ART and a vocabulary of 10,000 features.

[7] Homepage of MALLET package: http://mallet.cs.umass.edu

[8] It turns out that, with this amount of data, priors had a limited effect on the final results (± 0.015 on ARI). We are aware that an automatic, dynamic estimation is possible [14] but we do believe that a constant setup of hyperparameters guarantees a fair comparison.

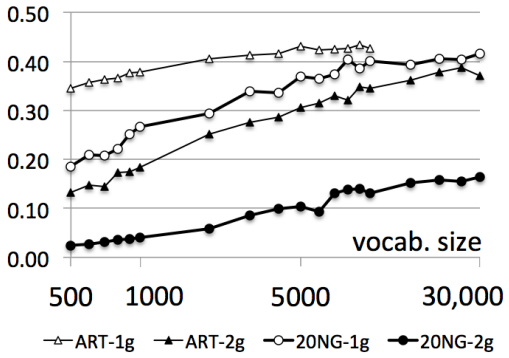

Fig. 3. Evolution of ARI (log scale for x axis).

(Fig. 4). We first observe that ARI increases exponentially below some threshold before converging[9]. This means that a fraction of features is sufficient to get an important gain in ARI (5 000 unigrams for ART achieves 0.434 for a maximum of 0.4346 with 9 000 unigrams ; 10,000 unigrams for 20NG achieves 0.3961 for a maximum of 0.4285 with 30,000 unigrams). For information, recent work [6,27] focusing on text clustering reported 0.397 and 0.425 ARI on 20NG respectively.

In addition, we observe that with three times the number of features, bigram-based vocabulary is able to achieve a really good ARI score for ART, not very far from the maximum with unigrams (0.3865 against 0.4346). This is clearly not the same situation for 20NG with a difference of about 0.26 for the ARI.

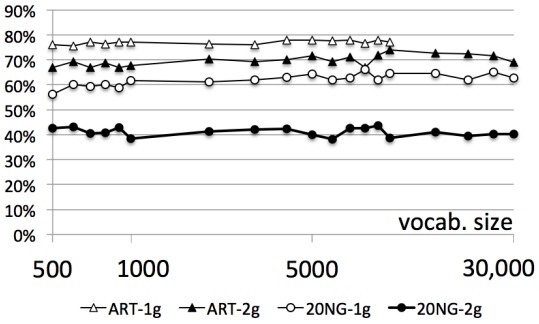

Fig. 4. Evolution of prec@10 (log scale for x axis).

We now take a closer look at the top terms returned by LDA, in comparison to the reference terms extracted from the true classes. The precision achieved by keeping the top-10 terms is shown in Fig. 4. We have been really surprised to notice that the

[9] We stopped the size for ART-1g at the number of words occurring at least twice in the whole corpus (13,778 words).

datamining (ART)		sci.space (20NG)		rec.sport.baseball (20NG)	
1-grams	2-grams	1-grams	2-grams	1-grams	2-grams
data	data mining*	space	solar system*	writes	red sox*
mining	data sets*	earth	henry spencer*	game	san francisco (15)
algorithm	association rules*	launch	physical universe (30)	article	los angeles (18)
clustering	time series*	writes	night sky*	year	st louis*
paper	data streams (13)	shuttle	space shuttle*	team	world series*
approach	experimental results*	nasa	toronto zoology*	games	major league*
learning	knowledge discovery*	mission	oort cloud (12)	good	blue jays*
classification	data set*	orbit	jet propulsion*	players	power play
algorithms	machine learning (12)	system	dick dunn (25)	baseball	mark singer*
results	support vector*	solar	high-speed collision (26)	time	san diego (12)

Fig. 5. top-10 terms of selected topics for ART (columns on the left) and 20NG (columns on the right). * means that the bigram is in the top-10 bigrams extracted from the ground truth, otherwise we note its rank. Henry Spencer posted over 34,000 messages to the sci.space.* newsgroups (source: Wikipedia).

top-10 bigrams are really competitive in comparison to the top-10 words. For ART, the precision achieved is 4% to 7% below only with a score of about 70% (up to 74% for 10,000 terms). We believe that this is a crucial observation: even though we get one term less that with single words in average, the topics are much more readable by using seven bigrams than height unigrams. The bigram "data mining" is more informative than the two words "data" and "mining", even if they are given in the same list.

This advantage is obvious if we take a look at the top terms given in the table of Fig. 5 (top bigrams next to top unigrams). For 20NG, it is even more interesting: despite the ARI collapse, four of the top bigrams are still accurate (six for the unigrams). However, it does not mean that the other terms are unrelated. We have highlighted the terms related to the ground truth with a * in Fig. 5, and noted their rank in the true list otherwise. Let us note that a vocabulary of 500 terms is sufficient to achieve such performances in term of precision. It seems that LDA easily finds the core of topics, without caring much for the result of the text clustering task.

Finally, we have run a last series of experiments in order to see the impact of a mixed unigram-bigram vocabulary. To this end, we fixed the number of features to 10,000 and changed the proportion in steps of 5% (e.g., 80% unigrams with 20% bigrams). The results confirm that bigrams might help increasing the overall clustering accuracy, but the bonus is limited and not significant. The best proportion seems to be highly dependent of the dataset (e.g., we got +0.01 ARI for ART with 5% of bigrams and +0.024 for 20NG with 30%). However, we observed no constant improvement for the precision. When we take a closer look to the top terms, bigrams are overwhelmed by unigrams, which explains the unchanged score.

4 Discussion and Future Work

Despite all the work done so far for integrating phrases into topic learning, we believe that this study is the first to highlight the potentiality of bigrams, not only for improving topic homogeneity (in addition to unigrams) or topic labeling, but for the whole task of

topic identification. Even though we have observed a clear gap between unigram and bigram frequencies, the bigram frequency seems to be sufficient to cover most of topic's aspects, getting rid of the ambiguity carried by unigrams. Hence, it is easy to provide readable topics to end users with a limited energy in the creation of terms (actually, any bigram library is expected to provide interesting features). Our preliminary experiments have shown that this reasoning can be transposed to trigrams as soon as their cumulated frequency is sufficient. We observed a decrease of about 10% for the pre@10 with trigrams (60% for ART and 30% for 20NG).

Interesting work lies ahead. One immediate follow-up is to design a new method that directly focuses on topic identification. By even more weakening our expectations on text clustering, we can find a way to improve the top K terms by favoring the topical core of each category (colored areas in Fig. 1). Improving the input representation, for instance by adding pseudo-counts for complex terms, can be a way to explore this idea. Another exciting, more theoretical question is to question the tradeoff between term frequency, term co-occurrences and performances. During our experiments, we observed a clear logarithmic correlation between the total number of tokens and the performances we can achieve (from $R^2 = 0.94$ for 20NG until $R^2 = 0.98$ for ART described with bigrams). This tells us that we cannot expect much by using too rare terms since they lead to really sparse matrices. However, it seems that the combination of complementary rare terms can compete with more frequent words. Information theory might be used for studying this kind of issues further.

References

1. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.
2. Richard T Carback III, Brad D Gaynor, Nathan R Shnidman, and Sang Hoon Chin. Systems and methods for software analytics, December 17 2015. US Patent 20,150,363,197.
3. Guillaume Cleuziou. An extended version of the k-means method for overlapping clustering. In *Proceedings of the 19th International Conference on Pattern Recognition*, 2008.
4. Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.
5. Joseph C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.
6. Siddharth Gopal and Yiming Yang. Von mises-fisher clustering models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 154–162, 2014.
7. David Hall, Daniel Jurafsky, and Christopher D Manning. Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 363–371. Association for Computational Linguistics, 2008.
8. Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the 15th conference on Uncertainty in Artificial Intelligence*, pages 289–296. Morgan Kaufmann, 1999.
9. Jey Han Lau, Timothy Baldwin, and David Newman. On collocations and topic models. *ACM Trans. Speech Lang. Process.*, 10(3):10:1–10:14, July 2013.
10. Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1536–1545. Association for Computational Linguistics, 2011.
11. Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. BIOTEX: A system for biomedical terminology extraction, ranking, and validation. In *Proceedings of the ISWC 2014 - the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014.*, pages 157–160, 2014.

12. Dongning Luo, Jing Yang, Milos Krstajic, William Ribarsky, and Daniel Keim. Eventriver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics*, 18(1):93–105, 2012.
13. Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.
14. Andrew Mccallum, David M Mimno, and Hanna M Wallach. Rethinking lda: why priors matter. In *Advances in Neural Information Processing Systems*, pages 1973–1981, 2009.
15. Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *13th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 490–499, 2007.
16. Michael Nokel and Natalia Loukachevitch. A Method of Accounting Bigrams in Topic Models. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 1–9, Denver, Colorado, 2015. Association for Computational Linguistics.
17. Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13), 2015.
18. Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2), 1994.
19. Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 306–315, 2004.
20. Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1285–1293, 2012.
21. Christoph Wagner, Vera Liao, Peter Pirolli, Lynn Nelson, and Markus Strohmaier. It's not in their tweets: Modeling topical expertise of twitter users. In *Privacy, Security, Risk and Trust (PASSAT), collocated with the IEEE international conference on Social Computing (SocialCom)*, pages 91–100, 2012.
22. Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 977–984. ACM, 2006.
23. Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM), 2007*, pages 697–702, 2007.
24. Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. TIARA: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 153–162, 2010.
25. Weili Wu, Hui Xiong, and Shashi Shekhar. *Clustering and information retrieval*, volume 11. Springer Science & Business Media, 2013.
26. Pengtao Xie and Eric P Xing. Integrating document clustering and topic modeling. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.
27. Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web (WWW)*, pages 1445–1456. ACM, 2013.
28. Yi Zhang, Guangquan Zhang, Hongshu Chen, Alan L. Porter, Donghua Zhu, and Jie Lu. Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological Forecasting and Social Change*, 2016.
29. Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. Topical keyphrase extraction from twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 379–388, 2011.