United we stand: Using multiple strategies for topic labeling

Antoine Gourru, Julien Velcin, Mathieu Roche, Christophe Gravier, Pascal Poncelet

United we stand: Using multiple strategies for topic labeling

Antoine Gourru[1], Julien Velcin[1], Mathieu Roche[2,5], Christophe Gravier[3], and Pascal Poncelet[4]

[1] Université de Lyon, Lyon 2, ERIC EA 3083, France,
antoine.gourru@univ-lyon2.fr julien.velcin@univ-lyon2.fr
[2] TETIS, Univ. Montpellier, APT, Cirad, CNRS, Irstea, Montpellier, France,
mathieu.roche@teledetection.fr
[3] Université Jean Monnet, Laboratoire Hubert Curien UMR CNRS 5516, France,
christophe.gravier@univ-st-etienne.fr
[4] Univ. Montpellier, LIRMM, France,
pascal.poncelet@lirmm.fr
[5] Cirad, TETIS, Montpellier, France,
mathieu.roche@cirad.fr

Abstract. Topic labeling aims at providing a sound, possibly multi-words, label that depicts a topic drawn from a topic model. This is of the utmost practical interest in order to quickly grasp a topic informational content – the usual ranked list of words that maximizes a topic presents limitations for this task. In this paper, we introduce three new unsupervised n-gram topic labelers that achieve comparable results than the existing unsupervised topic labelers but following different assumptions. We demonstrate that combining topic labelers - even only two - makes it possible to target a 64% improvement with respect to single topic labeler approaches and therefore opens research in that direction. Finally, we introduce a fourth topic labeler that extracts representative sentences, using Dirichlet smoothing to add contextual information. This sentence-based labeler provides strong surrogate candidates when n-gram topic labelers fall short on providing relevant labels, leading up to 94% topic covering.

1 Introduction

With the ever-growing information flow, we are facing unprecedented difficulties to quickly grasp the informational content on large streams of textual documents. This happens in our daily life when we are browsing social media or syndicated contents, but also in professional processes for which the stakes are to categorize and understand large chunks of documents in an unsupervised setting. Topic models are a solution to tackle this issue. They have been successfully applied to many situations, including historical record analysis, machine translation, sentiment analysis, to name a few [1]. While topic models make it possible to build sound representations of the different main *topics* discussed in the corpus they are draw from, they hardly provide human-friendly outputs. As a consequence,

gaining a quick understanding of topics content is a cumbersome task. Interactive topic visualization [2] does help in this matter, though they require an in-depth and time-consuming cognitive effort to grasp each topic informational content. This is the compelling rationale for the advent of topic models based on n-grams [3] and topic labeling techniques [4, 5]. Such techniques aim at providing a topic label, which means a single or multi-word title that is relevant to represent the informational content being discussed in the documents falling into that topic.

With the existing topic labeling techniques that we will discuss in Section 2, the popular approach for topic labeling is to rely on a single topic labeling technique. This implies a single measure to rank label candidates – therefore assuming the hypothesis that a single measure exists that befits any kind of topic regardless their number, their informational content type, their size, etc. While a single strategy may be overall satisfactory, reaching about 2.0 on average on a 0-3 Likert scale in a unsupervised setting [6], it also leaves many topics with very poor labels (0 or 1 on such scale).

In this paper, we investigate how to improve topic labeling. In this context, a new multi-strategy method is proposed with two main functions.

First, we propose in Section 3 three new n-gram topic labeling techniques, called M-Order, T-Order, and document-based labelers. M-Order and T-Order both leverage the odds for label candidates to be generated by the other topics as a background distribution as a penalty, while T-order also demotes label candidates with high score when they are nested in another significant candidate. Document-based labeler investigates the possibility that the best label may be found in a very few number of documents that are central to the category. We discuss how each labeler performs but also how they complement each other by showing that it is possible to reach a 64% improvement when using at least 2 labelers.

The second function of our multi-strategy approach (see Section 4) consists in surrogating labels using sentence information retrieval and showing that they provide a complementary approach for some topics that cannot find a proper fit with n-gram labels. Section 5 concludes and provides further lines of research.

2 Related work

Topic labeling aims at finding a good label or title to provide a better understanding of what constitutes the homogeneity of a given topic [4]. Several measures have been proposed to associate either a term or a phrase that roughly summarizes the top-K words associated to a given topic. More recently, [6] explored new solutions relying on letter trigram vectors and word embeddings.

Both information-based topic labeling strategies [4, 7] and those based on external semantic resources [5, 8] only provide a satisfactory labeling efficiency in a subset of topics in a corpus. On a 0-3 Likert scale, [5] achieves 2.03 at best while [6] reaches 2.13 at best – and less than 2 in most of the cases. It does not mean that every topic finds a good match (score 2), but that a set of topics achieves 3 whereas another set achieves 1 or even 0. This has been confirmed

in our own experiments that show standard deviations from 0.74 to 0.88 (see Table 2). We demonstrate in this paper that in an unsupervised setting, without the need of external semantic resources or user supervision, it is nonetheless possible to achieve results that let most of the topics be associated to a good label.

Another option is to work with representative sentences, which has been successfully integrated into topic-modeling oriented applications [9]. However, their usefulness in how the topic is understood has not been evaluated yet. In particular, in this paper we evaluate to which extent the sentences can complement the n-gram labels.

Other work investigated the use of other modalities for topic labeling, such as images. It turns out that previous work proved that labels based on n-grams showed the best performances when evaluated by human beings [10]. This explains why we focus on n-gram labeling here, with an additional interest in the possibility to enhance topic understanding with sentences as a surrogate.

3 Topic labeling based on n-grams

First, our multi-strategy approach proposes several labels (i.e., phrases) by using different statistical ranking measures. In what follows, we score a candidate term t that is a sequence of p consecutive words, also called n-grams: $t = (w_1, w_2 \dots\ w_p)$. We consider these candidates as possible labels for a given topic using different new proposed measures. The probability $p(w/z)$ of word w given topic z and the probability $p(z/d)$ of topic z given document d are given by the topic modeling algorithm (here, LDA).

3.1 T-order, M-order and document-based labelers

M-order labeler. Our first contribution aims at improving the 0-order measure [4] that is computed by $\sum_{i=1}^{p} \log \frac{p(w_i|z)}{p(z)}$. Instead of normalizing by the marginal probability, we use the odds for the candidate to be generated by the other topics as a background distribution. With $p\,(w_i|z)$ the probability for the topic z to generate the ith word of t, we define a first score of relevance $M-order$ as follows:

$$M-order(t,z) = \sum_{i=1}^{p} \log \frac{p\,(w_i|z)}{\frac{1}{|Z|-1}\sum_{z' \neq z} p\,(w_i|z')}$$

with Z the set of extracted topics. The denominator penalize the candidates that are also likely to be generated by the other topics than topic z.

T-order labeler. To introduce the notion of "termhood" [11], we define that a term t is a short term if it is nested in a longer term t' that has a bigger value for some base measure of termhood (e.g., c-value). For example, in a computer science corpus, "Gibbs" would be a short term, because it is usually nested in "Gibbs Sampling" that has a higher termhood. In this case, the term t can be

ignored. This post-processing method is akin to the "completeness" measure usage in [7]. Finally, the score is divided by the length *len* of the candidate. We can now define our new measure:

$$T - order(t, z) = \begin{cases} 0 & \text{if } t \text{ is a short term} \\ \frac{1}{len(t)} \cdot M - order(t, z) & \text{else} \end{cases}$$

The score adds a notion of distance maximization between topics to 0-order, and dividing by the term length prevents from favoring long terms too much. The use of a termhood base measure prevents the labels to be terms that are not semantically relevant.

Document-based labeler. In some cases, the best label can be found in a very few number of documents that are central to the category. We define our second new measure by averaging the importance of the set of documents featuring a given term:

$$Doc - Based(t, z) = \Big(\prod_{d \in D_t} imp_z(d) \Big)^{\frac{1}{|D_t|}}$$

where D_t is the set of documents in which the term t can be found and $imp_z(d)$ stands for the importance of document d in z. We decided to estimate $imp_z(d)$ in two ways. First, $doc\text{-}based_u(d, z)$ is based on $p(z/d)$ with a natural bias towards short documents. The second measure $doc\text{-}based_n(d, z)$ is based on $p(d/z) \propto p(z/d) * p(d)$ for a given z. We decided to approximate $p(d)$ by the ratio between the length of d and the total length of the corpus. The rationale of the new measure is therefore to find terms very specific to the topic although they exhibit moderate topic covering.

3.2 Evaluation

Methodology. We experiment on two case studies with the LDA model [12] and a fixed number of topics $k = 100$. The maximum number of iteration is set to 2000 and the hyperparameters α and β are automatically tuned as explained in [13]. We may do other choices, such as using other topic learning algorithms or setting k to other values. However, we think that the following experimental design is sufficient for supporting the claim of this paper.

Datasets. Topic models are drawn from the following datasets:
• **Sc-art:** a set of 18 465 scientific abstracts gathered by [14] over a period of 16 years. Many contributions in topic visualization and labeling techniques have been applied on similar datasets, on which n-grams seem to provide very relevant labels.
• **News-US:** a set of 12 067 news we gathered automatically from the Huffington Post RSS feeds (US version). This set spans a period of almost 3 months (from June the 20th until Sept. the 8th, 2016). Monitoring news feed is another case study of the utmost practical interest.

After learning topic models, we perform standard post-processing that is removing the topics that present a poor homogeneity. Actually, topics of poor quality would bias our experiments since no labeler will be capable of finding a satisfactory solution. To this end, we compute the Normalized Pointwise Mutual Information (NPMI) of the top 10 words, and we eventually remove the topics with a negative score, as it is shown in [15] that this score is correlated with human measure of the topic coherence. Therefore, we keep 145 topics (on 200 initial topics) and every annotator had 48 or 49 tasks to complete (2 annotators $\times$ 145 topics / 6). Each task corresponds to the evaluation of two types of elements: (i) evaluation of candidate labels (i.e. words and/or phrases), (ii) evaluation of representative sentences.

As in previous works, the evaluation consists in measuring how well an n-gram candidate labels the topic on a Likert scale. A scale of five points is used in [7] and four in [16, 6]. Our four points scale is presented on Table 1; it was constructed so that we can easily compare our results to the literature. We called six computer scientists as human annotators. The annotation task aims at evaluating the three main n-gram labels provided by the different labelers for a given tuple (dataset, topic). For any given tuple to annotate candidates were ranked randomly and the annotators were blind to the kind of labeler which generated each label. Each annotation was given to two annotators in order to calculate an agreement score.

Score	**Description**
3	Yes, Perfectly
2.a	Yes, but it is too broad
2.b	Yes, but it is too precise
1	It is related, but not relevant
0	No, it is unrelated

Table 1: Our Likert Scale

For a given annotation task, we provided five documents that maximize $p(d|z)$, three documents that maximize $p(z|d)$, plus the thirty top words with their associated probabilities, as in [16].

The three most highly ranked labels were evaluated, either they have been computed by the basic measures of [4] or by our own measures presented in Section 3.1. The 0-order was computed with both uniform and frequency-based background distribution. The T-order was computed using the LIDF-value from [17] as a termhood measure, which seems to provide better results than the other termhood scoring measures after a preliminary screening. We choose not to limit the labeling candidate to be a bigram set, but to keep any length for the labels.

Top-3	**News-US**	**Sc-Art**	**All**			
Max-Score	2.23	2.40	2.33	σ	**Too Broad**	**Too Precise**
T-order	1.27	1.24	1.26	0.81	13%	15%
M-order	1.25	1.20	1.22	0.8	13%	14%
$doc-based_n$	0.98	1.12	1.05	0.74	4%	16%
$doc-based_u$	1.03	1.17	1.10	0.75	4%	17%
1-order	1.07	1.31	1.19	0.84	8%	16%
$0-order_{uniform}$	1.10	1.63	1.36	0.82	7%	24%
$0-order_{frequence}$	1.18	1.23	1.20	0.88	9%	17%

Table 2: Average score for the top-3 labels proposed on a Likert scale from 0 (unrelated) to 3 (perfect). σ details the average standard deviations for the two datasets.

The candidate generation was performed using the BioTex API[6] [17].

Results. Table 2 shows the average score (from 0 to 3) of the top-3 results of the labeling systems. Max-Score is the upper bound, which corresponds to selecting the score of the system that achieves the highest score for every topic. As it was presented earlier, the annotator was able to give some qualitative information when rating with 2: "it is too broad" or "it is too precise". The distribution by labeler is provided in Table 2. The kappa shows a fair agreement (0.34,0.49,0.51) that allows us to look for an automated labeling recommender system.

One first observation is that the simple 0-order$_{uniform}$ is clearly better than the others for the Sc-Art dataset (1.63 against 1.24), but the T-order measure outperforms it for the News-US dataset (1.27 against 1.1). As it is well illustrated in Table 2, labeling systems are not always good (maximum 1.36 on average), but there is (almost) always a labeling system that is able to provide a good label (2.33/3 on average). This means that we can expect an improvement of about 64% in the labeling task. Tables 3 and 4 give a more concrete illustration of that idea on a selected set of topics.

An important result is that with 90% of the evaluated topic, a good label (meaning rated 2 or 3) is found. We name this amount as the *covering* of the labeling strategy. When we reduce the labels to those produced by the T-order and the $0-order_{uniform}$ only, there is a good label 83% of the time. A labeler alone can only achieve a good score at best for 63% ($0-order_{uniform}$) and 62% ($T-order$) of all topics.

Discussion. The presented results mean that even with a very small set of proposed labels, one can access the inner semantic content of a given topic. In the case of the two datasets we experiment on, we only need six labels (meaning,

[6] http://tubo.lirmm.fr/biotex/

Topic 1(News-US)	Topic 2(Sc-Art)	Topic 3(News-US)	Topic 4(Sc-Art)
eu	detection	mental	user
brexit	event	health	web
britain	events	depression	users
european	system	illness	filtering
leave	detecting	suicide	profiles
vote	false	anxiety	collaborative
british	detect	disorder	usage
london	intrusion	care	preference
minister	vehicle	social	system
referundum	anomaly	bell	site

Table 3: Examples of topics learned on our datasets

	Topic 1	Topic 2
T-order	**brexit**	intrusion
0-order	british prime minister david cameron	**intrusion detection systems**
	Topic 3	**Topic 4**
T-order	**bipolar disorder**	preference
0-order	national suicide prevention lifeline	**user preference**

Table 4: The words in bold where rated 3, the others 1. We see that for some topics the 0-order is able to find a good label whereas it is the T-order for other topics.

three labels produced by two labelers, if there is no overlap). On average, among these six labels, 15% are rated as unrelated, 46% as related, 29% as good label and 9%as perfect label, 1% have been rated "I don't know".

If we consider 2 and 3 as good labeling scores, the precision is about 38% (29%+9%), even though the other answers are not totally wrong. For instance, the topic 3 presented in Table 3 gets the following labels (0-order followed by T-order labels): "bipolar disorder, depression, anxiety, national suicide prevention lifeline, mental health disorder, mental health care".

The presented results can be thought as over-optimistic: they need further experiment on other various datasets (e.g., book series or blog posts) and we know that within the labels given to the users there is still unrelated/non relevant items (in our case, about 15% of the proposed labels). This is the reason why we need to find advanced strategies to a) improve the quality of the recommended

Systems	Performance
Max-Score	90
T-order	62
M-order	60
$doc - based_n$	46
$doc - based_u$	51
1-order	53
$0 - order_{uniform}$	63
$0 - order_{frequence}$	55
T-order+$0 - order_{uniform}$	83

Table 5: Performance of the labeling systems, meaning the percent of a least one good label (rated 2 or 3) in the top-3 labels

labels, b) complete the labels when the n-gram based approach is not sufficient to fully capture the inner semantics.

Regarding (a), we can naturally think at integrating other features than the relation between topic-word probabilities and the candidate label. For instance, we can add features related to the topic (importance of the topic, skewness of the word distribution, etc.) or to the dataset (for instance, longer n-grams can be favored in scientific datasets). We can also follow the work of Lau et al. [5] in leveraging more supervision in the labeling process.

Regarding (b), it seems that for some cases, the n-gram labelers can never achieve a satisfactory output. This is the reason why we propose, in the next section, to leverage information retrieval techniques to find relevant *sentences*.

4 Topic-relevant sentence extraction

In the second function of our multi-strategy approach, different representative sentences are proposed in order to identify the semantic content of the topics.

4.1 Rationale

Would no n-gram labeler yields an acceptable label candidate, looking for representative sentences from the corpus is another solution to label a topic. For example, the labels returned by all the n-gram based systems for the two topics in Table 6 were rated low (the maximum score being 0 for the first and 1 for the second). But we can find sentences that were well rated, as shown in Table 7.

4.2 Sentence extraction solution

With this fourth new labeling technique, we assume that an information retrieval procedure can be used to post-process the top documents (considered as the

Topic 5(News-US)	Topic 6(News-US)
photo	facebook
posted	media
2016	social
pdt	online
jul	app
39	internet
instagram	video
aug	google
jun	users
34	site

Table 6: Example of two topics badly labeled by n-grams.

Topic	Example of sentence returned by our systems
5	A **photo posted** by Laura Izumikawa Choi (@lauraiz) on **Jun** 17, 2016 at 11:05am **PDT**
6	So 'follow' or 'Like' them on **social media** sites like Twitter, **Facebook**, LinkedIn, **Google** + and Pinterest

Table 7: Two extracted sentences that can help the user capturing the meaning of topics 5 and 6 given in Tab. 6 (words occurring in top words are highlighted in bold).

"context") and look for representative sentences. The top documents, i.e. the documents that maximize $p(d/z)$, are split into sentences. We propose to use a Dirichlet smoothing to add contextual information.

We define β, the context distribution of a document collection, by:

$$\beta_w = \frac{c(w, C)}{\sum_{w \in V} c(w, C)} \tag{1}$$

where $c(w, C)$ counts the frequency of word w in the context C. With μ as a positive real number, we obtain the following language model:

$$\theta_w^x = \frac{c(w, x) + \mu\beta_w}{len(x) + \mu} \tag{2}$$

where $c(w, x)$ stands for the frequency of word w in the candidate sentence x. We can then compute different distance measures between the sentence vector representation and the topic. We choose to compute a negative Kullback-Leibler distance, like [4], and a simple cosine similarity. If $\mu = 0$, the θ_w^x calculated is a simple TF representation of the sentence. The greater μ is, the more importance

we give to the context (the top documents). Our model is parameterized by: β (more precisely, the number of top documents $|\beta|$ we choose to keep) and μ (the amount of context we want to take into account).

4.3 Evaluation

We experiment with the same models and datasets than the n-gram evaluation of the previous section. We choose to ask the following question: "Does the sentence give a clear understanding of the topic content?". Then, the rater could choose between "yes", "no", or "don't know". The systems are presented in Table 9. We choose to compare our systems with random sentences, extracted from documents that do not maximize $p(d/z)$. We call `Rand` this system based on random sentences.

name	**similarity**	μ	$\|\beta\|$
$COS10$	cosine	0	10
$COS15$	cosine	0	15
$COSIDF15$	cosine	0 (IDF weighted)	15
$B10_{0,1}$	negative KL divergence	0.1	10
$B10_{10}$	negative KL divergence	10	10
$B10_{1000}$	negative KL divergence	1000	10
$B20_{0,1}$	negative KL divergence	0.1	20
$B20_{10}$	negative KL divergence	10	20
$B20_{1000}$	negative KL divergence	1000	20

Table 8: Evaluated systems with different parameters' values.

As for the n-grams evaluation in previous section, a weighted Kappa was computed for every annotator pair. The results are similar: it is not really high (0,23, 0.36, 0.04), but sufficient for a significant agreement. Table 9 presents the average proportion of extracted sentences tagged as 1 (answer "yes" to the question: "Does the sentence give a clear understanding of the topic content?"). It shows that a simple cosine based on a TF vector using the top 10 documents is better, without the need of smoothing. However, a closer look shows that $B20_{0.1}$ (meaning a really small smoothing) is slightly better for News-US.

We can now wonder whether sentences can be combined with n-gram labels to improve the overall topic understanding. For instance, we can estimate the proportion of topics for which we can find at least one good n-gram label (with 0-order or T-order) and, if we cannot, one sentence otherwise. The performance goes from 83% covering up to 93% with the COS_{15} labeler. This improvement

System	News-US	Sc-Art	All	System	News-US	Sc-Art	All
`Rand`	1%	6%	4%	$B10_{10}$	34%	38%	36%
$COS10$	34%	46%	41%	$B10_{1000}$	25%	28%	27%
$COS15$	35%	45%	40%	$B20_{0.1}$	40%	31%	35%
$COSIDF15$	22%	30%	26%	$B20_{10}$	34%	37%	36%
$B10_{0.1}$	38%	33%	35%	$B20_{1000}$	25%	26%	26%

Table 9: Percent of relevance, meaning the proportion of topics correctly illustrated by the sentence.

can be seen even with no agreement among the last annotator pair. If we ignore the annotations attributed by the last annotator pair (weighted Kappa of only 0.04), the covering goes even until 94%.

5 Conclusion

Finding a suitable textual description of the output of a statistical model is still a difficult task that can be related to interpretable machine learning. In this paper, we introduced three new n-gram topic labelers that are at least on par with the existing labeling technique. A key observation is that those new labelers are complementary labelers to the best known so far (0-*order* labeler) so that when one of them is combined with the 0-*order* labeler, the resulting combined labeler provides labels scoring 2 or more on a 0-3 Likert scale. A direct application of this consists in a simple recommender system that suggests a limited set of labels (e.g., the labels produced by T-order and $0 - order_{uniform}$) that are mostly of good quality. When expanding such a combined n-gram labeler with extracted sentences from the corpus as surrogate labels for difficult labeling cases, it is possible to reach the same performance for 94% of the topics to label.

There is still room for future research. First of all, we can consider the covering we get as a good recall, but the precision (meaning, the proportion of really good labels among the labels returned by the system) needs to be improved. If we are able to automatically choose the perfect labeler for each case, this will constitute an important improvement[7]. If we assume that the user can still have difficulties for selecting the best labels for *some* topics, we might adopt a semi-supervised system following [5]. Another track would be to define a label as a combination of n-grams and sentences that are complementary for they propose different views over the targeted topic. The must would be to generate a small paragraph that summarizes the underlying topic semantics.

Acknowledgments: This work is partially funded by the SONGES project (Occitanie and FEDER).

[7] We plan to publicly release the annotations made by our human judges.

References

1. Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296, 2017.
2. Daeil Kim, Benjamin F Swanson, Michael C Hughes, and Erik B Sudderth. Refinery: An open source topic modeling web platform. *Journal of Machine Learning Research*, 18(12):1–5, 2017.
3. Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Seventh IEEE International Conference on Data Mining (ICDM)*, pages 697–702. IEEE, 2007.
4. Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 490–499. ACM, 2007.
5. Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of Annual Meeting of ACL-HLT - Vol. 1*, pages 1536–1545. Association for Computational Linguistics, 2011.
6. Wanqiu Kou, Fang Li, and Timothy Baldwin. Automatic labelling of topic models using word vectors and letter trigram vectors. In *Asia Information Retrieval Symposium*, pages 253–264. Springer, 2015.
7. Marina Danilevsky, Chi Wang, Nihit Desai, Xiang Ren, Jingyi Guo, and Jiawei Han. Automatic construction and ranking of topical keyphrases on collections of short documents. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 398–406. SIAM, 2014.
8. Davide Magatti, Silvia Calegari, Davide Ciucci, and Fabio Stella. Automatic labeling of topics. In *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*, pages 1227–1232. IEEE, 2009.
9. Mennatallah El-Assady, Rita Sevastjanova, Fabian Sperrle, Daniel Keim, and Christopher Collins. Progressive learning of topic modeling parameters: A visual analytics framework. *IEEE Trans. on Visualization and Computer Graphics*, 2017.
10. Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology*, 68(1):154–167, 2017.
11. Katerina T Frantzi, Sophia Ananiadou, and Junichi Tsujii. The c-value/nc-value method of automatic recognition for multi-word terms. In *International Conference on Theory and Practice of Digital Libraries*, pages 585–604. Springer, 1998.
12. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.
13. Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In *Advances in Neural Information Processing Systems*, pages 1973–1981, 2009.
14. 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.
15. Nikolaos Aletras and Mark Stevenson. Evaluating topic coherence using distributional semantics. In *IWCS*, volume 13, pages 13–22, 2013.
16. Zhixing Li, Juanzi Li, Yi Liao, Siqiang Wen, and Jie Tang. Labeling clusters from both linguistic and statistical perspectives: A hybrid approach. *Knowledge-Based Systems*, 76:219–227, 2015.
17. Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, 19(1):59–99, 2016.