gourru_2018_united_we_stand_using_multiple_strategies_for_topic_labeling

Year

2018

Author(s)

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

Title

United We Stand: Using Multiple Strategies for Topic Labeling

Venue

Natural Language Processing and Information Systems

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel

Underlying technique

n-gram and sentence labelers based on statistical ranking measures

Topic labeling parameters

The 0-order was computed with both uniform and frequency-based background distribution.

The T-order was computed using the LIDF-value as a termhood measure
We choose not to limit the labeling candidate to be a bigram set, but to keep any length for the labels.

Label generation

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.

The second function of our multi-strategy approach 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.

Topic Labeling Based on n-Grams (based on 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 = (w1, w2 . . . wp).
We consider these candidates as possible labels for a given topic using different new proposed measures.

M-Order Labeler
Aims at improving the 0-order measure (`Automatic labeling of multinomial topic models`) that is computed by

$$\sum_{i=1}^{p} \log \frac{p(w_i|z)}{p(z)}.$$

Instead of normalizing by the marginal we use the odds for the candidate to be generated by the other topics as a background distribution.
With p(wi|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:

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

with Z the set of extracted topics. The denominator penalizes 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” , 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.

Finally, the score is divided by the length len of the candidate. We can now define our new measure:

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

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 Dt 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-based_u(d,z) is based on p(z / d) with a natural bias towards short documents.

The second measure doc-based_n(d,z) is based on p(d / z) ∝ 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.

Topic-Relevant Sentence Extraction

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 “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:

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

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:

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

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 and a simple cosine similarity.
If μ = 0, the θwx 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 |β| we choose to keep) and μ (the amount of context we want to take into account).

Motivation

Improving the performance of existing unsupervised topic labelers by combining multiple unsupervised techniques.

And, in the case of the sentence labeled:

- provide strong surrogate candidates when n-gram topic labelers fall short on providing relevant labels

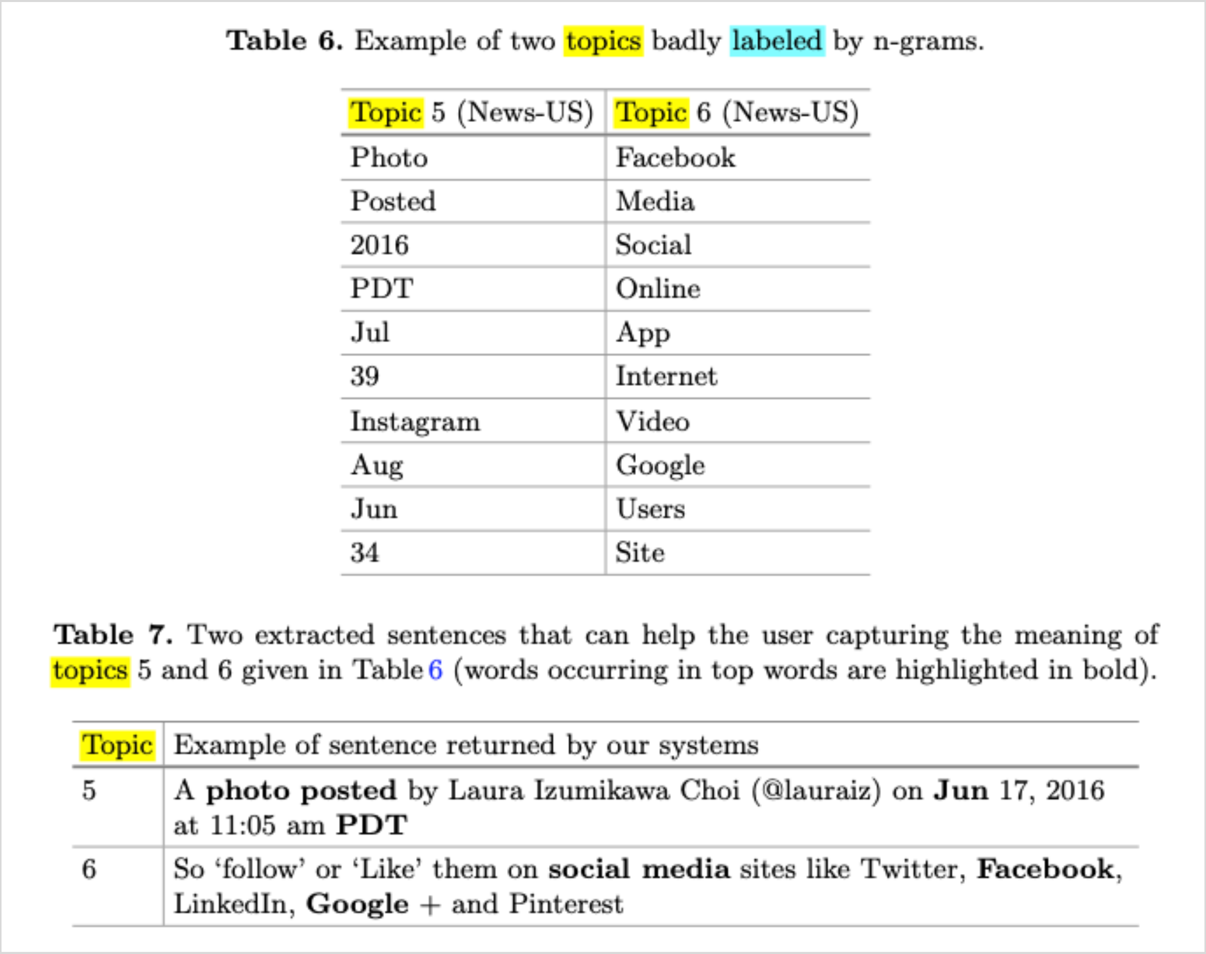

Table 6. Example of two topics badly labeled by n-grams.

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

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

Topic modeling

LDA

Topic modeling parameters

Nr of topics (k): 100

Max nr of iterations: 2000

α and β are automatically tuned

Nr. of topics

145 (100 per dataset, topics with negative NPMI value are removed)

Label

Topic Labeling Based on n-Grams

An n-gram (i.e. a sequence of p consecutive words t = (w1, w2 . . . wp))

Topic-Relevant Sentence Extraction

representative sentences from the corpus as an alternative solution to label a topic

Label selection

The three most highly ranked labels were evaluated, either they have been computed by the basic measures of `Automatic labeling of multinomial topic models` or by our own measures

Label quality evaluation

Every annotator had 48 or 49 tasks to complete (2 annotators × 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 four points Likert scale.

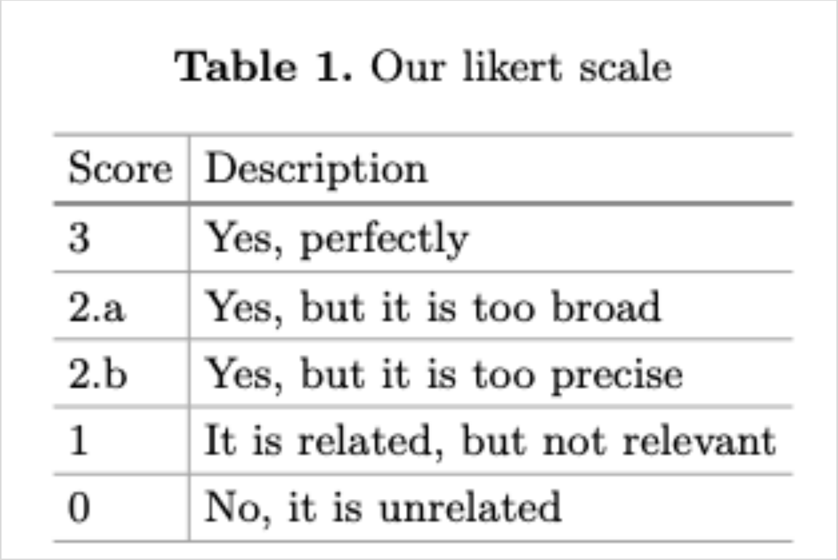

Table 1. Our likert scale

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

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.

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

Results

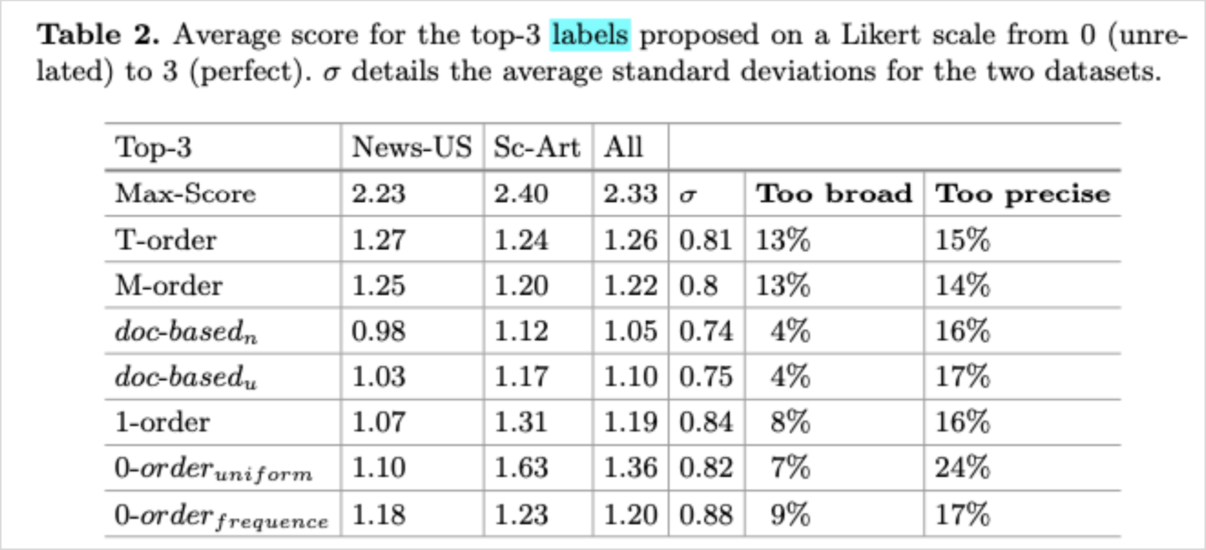

Table 2. Average score for the top-3 labels proposed on a Likert scale from 0 (unrelated) to 3 (perfect). σ details the average standard deviations for the two datasets.

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%
$doc\text{-}based_n$	0.98	1.12	1.05	0.74	4%	16%
$doc\text{-}based_u$	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%

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 (when using two labellers)
An important result is that with 90% of the evaluated topic, a good label (meaning rated 2 or 3) is found.

The presented results mean that even with a very small set of pro- posed 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, three labels produced by two labelers, if there is no overlap).
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.

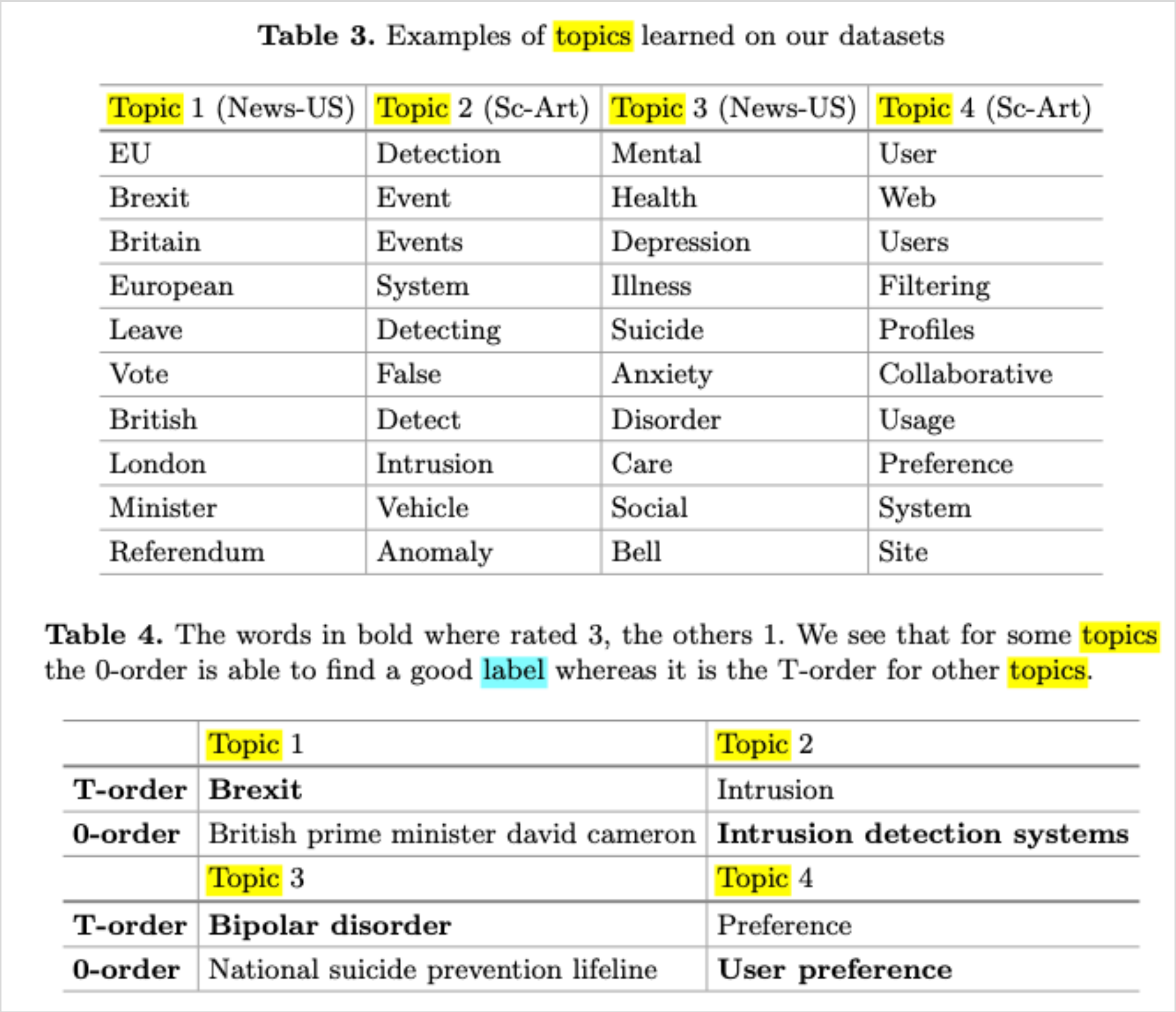

Table 3. Examples of topics learned on our datasets

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
Referendum	Anomaly	Bell	Site

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.

	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**

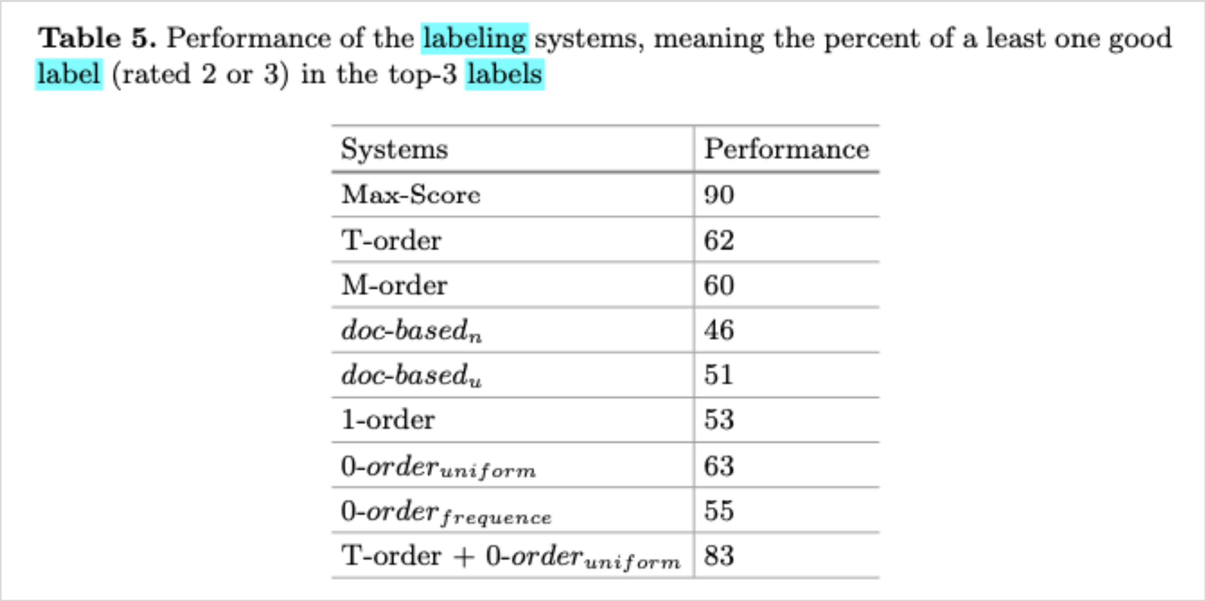

Table 5. Performance of the labeling systems, meaning the percent of a least one good label (rated 2 or 3) in the top-3 labels

Systems	Performance
Max-Score	90
T-order	62
M-order	60
$\textit{doc-based}_n$	46
$\textit{doc-based}_u$	51
1-order	53
0-$order_{uniform}$	63
0-$order_{frequence}$	55
T-order + 0-$order_{uniform}$	83

Evaluation of Topic-Relevant Sentence Extraction

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”. 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.

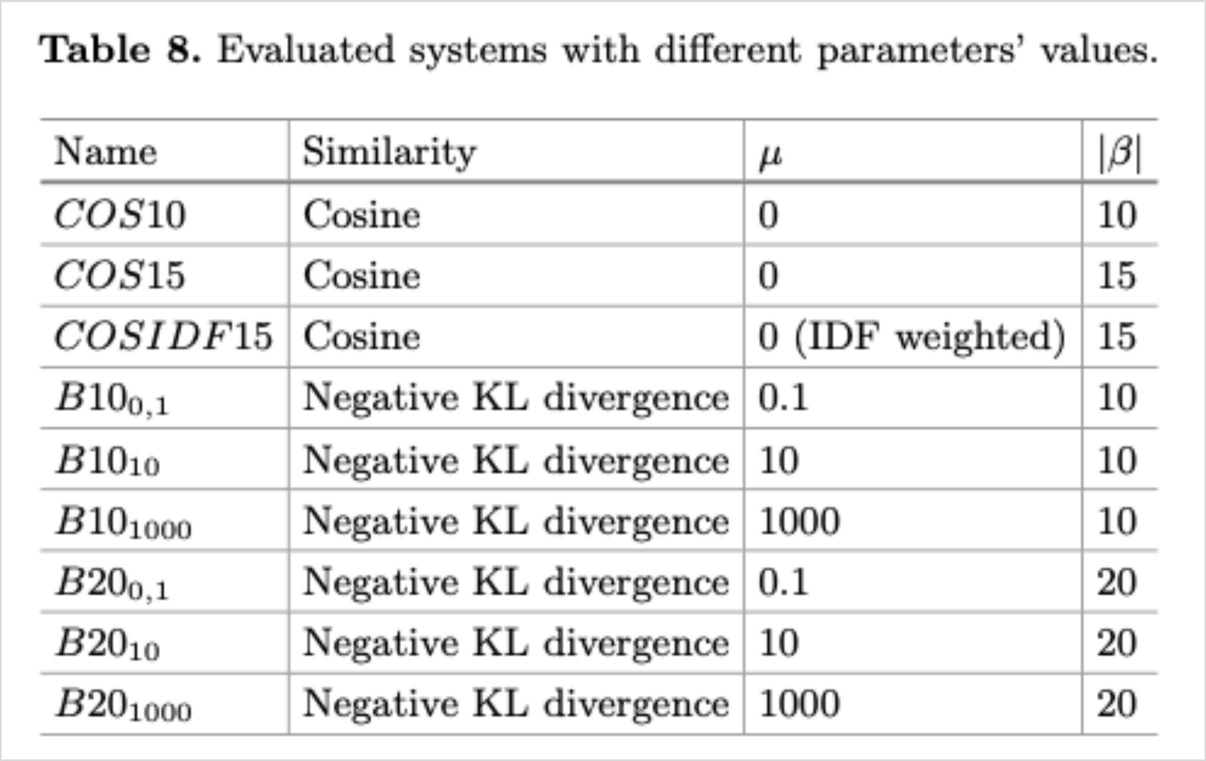

Table 8. Evaluated systems with different parameters' values.

Name	Similarity	μ	$\lvert\beta\rvert$
$COS10$	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

As for the n-grams evaluation in previous section, a weighted Kappa was computed for every annotator pair.

Table9 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?").

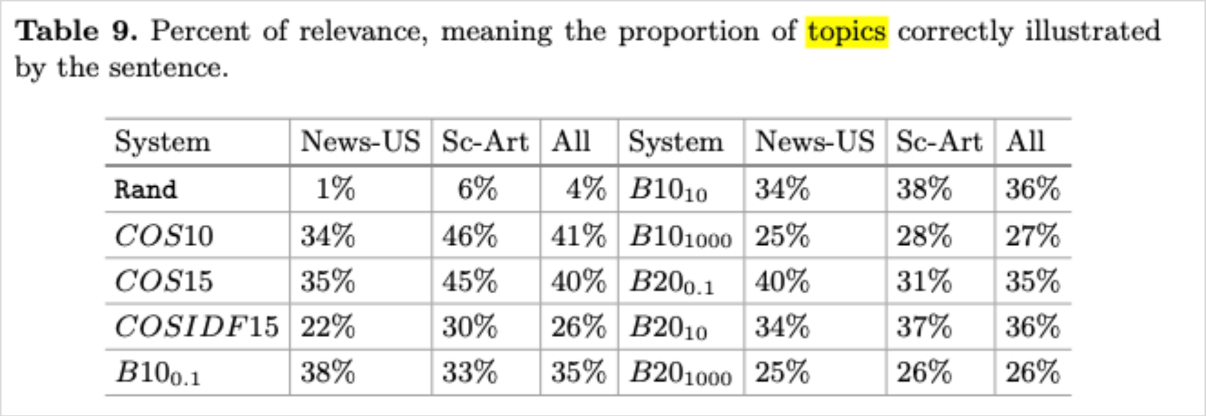

Table 9. Percent of relevance, meaning the proportion of topics correctly illustrated by the sentence.

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%

Assessors

We called six computer scientists as human annotators

Domain

Paper: Topic labeling

Dataset: News and Scientific Literature

Problem statement

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.

Corpus

Sc-art

Origin:

Nr. of documents: 18.465

Details:

- Scientific abstracts gathered over a period of 16 years

News-US

Origin: Huffington Post

Nr. of documents: 12.067

Details:

- News over a period of almost 3 months (from June the 20th until Sept. the 8th, 2016)

Document

Sc-art: Abstract of scientific article

News-US: News article

Pre-processing

```
@inproceedings{gourru_2018_united_we_stand_using_multiple_strategies_for_topic_labeling,
  TITLE = {{United we stand: Using multiple strategies for topic labeling}},
  AUTHOR = {Gourru, Antoine and Velcin, Julien and Roche, Mathieu and Gravier, Christophe and Poncelet, Pascal},
  URL = {https://hal-lirmm.ccsd.cnrs.fr/lirmm-01910614},
  BOOKTITLE = {{NLDB: Natural Language Processing and Information Systems}},
  ADDRESS = {Paris, France},
  VOLUME = {LNCS},
  NUMBER = {10859},
  PAGES = {352-363},
  YEAR = {2018},
  MONTH = Jun,
  DOI = {10.1007/978-3-319-91947-8\_37},
  PDF = {https://hal-lirmm.ccsd.cnrs.fr/lirmm-01910614/file/NLDB_Julien.pdf},
  HAL_ID = {lirmm-01910614},
  HAL_VERSION = {v1},
}
```

#Thesis/Papers/BS