PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 110 APRIL 2018 85–101

Improving Topic Coherence Using Entity Extraction Denoising

Ronald Cardenas,[a] Kevin Bello,[a] Alberto Coronado,[a] Elizabeth Villota[b]

[a] Department of Mechanical Engineering, National University of Engineering, Lima, Peru
[b] Section of Mechanical Engineering, Pontifical Catholic University of Peru, Lima, Peru

Abstract

Managing large collections of documents is an important problem for many areas of science, industry, and culture. Probabilistic topic modeling offers a promising solution. Topic modeling is an unsupervised machine learning method and the evaluation of this model is an interesting problem on its own. Topic interpretability measures have been developed in recent years as a more natural option for topic quality evaluation, emulating human perception of coherence with word sets correlation scores. In this paper, we show experimental evidence of the improvement of topic coherence score by restricting the training corpus to that of relevant information in the document obtained by Entity Recognition. We experiment with job advertisement data and find that with this approach topic models improve interpretability in about 40 percentage points on average. Our analysis reveals as well that using the extracted text chunks, some redundant topics are joined while others are split into more *skill-specific* topics. Fine-grained topics observed in models using the whole text are preserved.

1. Introduction

Probabilistic topic models, such as Latent Dirichlet Allocation (Blei et al., 2003) and its many variants (Newman et al., 2006; Blei and Lafferty, 2005, 2006; Teh et al., 2006; Blei et al., 2007), were introduced in an unsupervised setting to discover latent semantic structures in a collection of documents, namely the topics. However, there is no guarantee that the inferred topics – typically modeled as a set of important words – are easily interpretable by humans.

Traditionally, held-out likelihood had been used to perform topic model evaluation. Chang et al. (2009) conducted a study that showed that perplexity actually corre-

Corresponding author: racardenasa@uni.pe

Cite as: Ronald Cardenas, Kevin Bello, Alberto Coronado, Elizabeth Villota. Improving Topic Coherence Using Entity Extraction Denoising. The Prague Bulletin of Mathematical Linguistics No. 110, 2018, pp. 85–101. doi: 10.2478/pralin-2018-0004.

lates negatively with human interpretability of such topics. In other words, choosing the model with the lowest perplexity on unseen data may generate topics that are hardly interpretable. This motivates the search of different evaluation methods for topic modeling, referred in the literature as topic coherence measures (Newman et al., 2010; Musat et al., 2011; Mimno et al., 2011; Stevens et al., 2012; Aletras and Stevenson, 2013; Lau et al., 2014).

In this work, we hypothesize that topic interpretability – as measured by topic coherence – can be improved by training a topic model over text chunks of relevant information instead of the whole text per document, for job advertisement posts published in job-hunting websites. We analyze two scenarios of how categories of skills required for a specific job vacancy span across professional majors. The first scenario is a noisy scenario in which the topics are inferred using all the information available in job ads which includes e.g. company description, payment, working schedule. In the second scenario, the topics are inferred only over specific information about the job itself, such as expected skills, tasks to perform, and professional major of preference, extracted by named entity recognition. We find that this last setup scenario successfully increases coherence scores of inferred topics, obtains much cleaner topics and is able to infer meaningful clusters of majors related by the skills applicants are required to know.

This article is structured as follows. We first present related work on the field. Then, in section 3 we present all the theoretical background necessary to formulate the problem tackled. In section 4, the experimental setup of every module is thoroughly explained, and the dataset used in presented as well. Section 5 presents the results and discussion of our findings. Finally, section 6 presents the conclusions.

2. Related Work

In recent years, several topic coherence measures have been proposed (Newman et al., 2010; Musat et al., 2011; Mimno et al., 2011; Stevens et al., 2012; Aletras and Stevenson, 2013; Lau et al., 2014) in order to automate the method of Chang et al. (2009) and emulate human interpretability. Newman et al. (2010) introduced the notion of coherence and was the first to propose an automatic measure based on pairwise pointwise mutual information (PMI) between the topic words. Subsequent empirical works on topic coherence proposed measures based on word statistics that differ in several details, such as normalization (Lau et al., 2014), aggregation methods (Mimno et al., 2011), and reference corpus (Musat et al., 2011; Aletras and Stevenson, 2013). Röder et al. (2015) proposed a framework for the exploration of all possible coherence measures, modeled as a pipeline where the blocks (e.g. aggregation method, confirmation measure) can be exchanged and create new measures. They combined two complementary lines of research on coherence: scientific coherence and topic modeling.

As the acceptance of topic coherence measures increases as a mean of topic model assessment (Paul and Girju, 2010; Reisinger et al., 2010; Hall et al., 2012), recent research trends focus on proposing fast and efficient models that can be scaled up to big amounts of data (Yang et al., 2015; Nguyen et al., 2015), using the whole text per document for training.

Prior to directly evaluating human interpretability, several approaches were proposed to improve topic quality. Airoldi et al. (2010) analyzed the effect of varying the source text and inference strategies for PNAS biological sciences publications, obtaining a slightly higher number of new categories that better explain nowadays intertwined research fields. The usage of name entities as extra information in a topic model is explored by Newman et al. (2006). They propose a customized probabilistic graphical model that directly learns the entity-topic relationship and making better predictions about entities.

3. Problem Formulation

We define the problem of improving topic coherence as follows. Given a collection of highly noisy documents, we extract only relevant information from each document in the form of custom entities. The extraction task is modeled as a sequence labeling problem, and we tackle it by using the averaged structured perceptron (see Section 3.1).

As test case, we consider the domain of job advertisements. A job ad contains valuable information about what skills applicants are expected to have, but they contain spurious information as well. In order to avoid inferring topics over noise, we extract requirements, functions and preferred major from a job ad using a custom named entity recognition and extraction pipeline.

We now present notation and definitions core to the modules our model is based. We start by formally defining the entity extractor module, followed by the topic modeling. Then, the coherence metric is presented.

3.1. Averaged Structured Perceptron

The structured perceptron and its averaged version was initially introduced by Collins (2002). They differ from the well-known perceptron algorithm in that the output for each training instance pair $(x_t, y_t) \in T$ is a structure $y' \in Y_t$, where Y_t is the space of permissible structured outputs for input x. The inference algorithm to predict y' is problem dependent. In our case, sequence labeling, a first order *Viterbi* decoder is used. In each step, the candidate y' is transformed to a high-dimensional feature representation $f(x, y) \in R^m$ and the prediction is determined by a linear classifier based on the dot product of this representation and a weight vector $w \in R^m$.

In practice, this algorithm can be implemented easily and behaves remarkably well in several problems. These two characteristics make the structured perceptron algorithm a natural first choice for prototyping structured models.

3.2. Latent Dirichlet Allocation

In this section, we briefly describe the graphical model called Latent Dirichlet Allocation (LDA) (Blei et al., 2003), originally proposed for doing topic modeling. LDA is a generative probabilistic model in which the data is in the form of a collection of documents, and each document in the form of a collection of words. The model assumes that each document is a mixture of latent topics, and each topic is modeled as a mixture of words. These random mixture distributions are considered Dirichlet-distributed to be inferred from the data. The generative process of LDA can be described as follow:

1. For all D documents sample $\theta_d \sim \mathit{Dir}(\alpha)$.
2. For all K topics sample $\phi_k \sim \mathit{Dir}(\beta)$.
3. For each of the N_d words v_i in document d:
    - Sample a topic $z_i \sim \mathit{Multinomial}(\theta_d)$
    - Sample a word $v_i \sim \mathit{Multinomial}(\phi_{z_i})$
    - Observe the word

We assume symmetric Dirichlet priors for θ_d and ϕ_k, as suggested by Griffiths and Steyvers (2004).

Regarding inference strategies for the models, we make use of Gibbs Sampling as described in Griffiths and Steyvers (2004) and the Variational Expectation - Maximization (VEM) algorithm as described in Blei et al. (2003).

3.3. Topic Coherence

We use the coherence metric proposed by Mimno et al. (2011), based in conditional log likelihood of co-occurrence of top topic word pairs. We refer to it as *UMass* coherence from now on. It is defined as follows:

$$C_{UMass} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{P(w_i, w_j) + \epsilon}{P(w_j)},$$

where N is the number of top words in a topic to consider.

4. Experimental Setup

4.1. Job Ads Corpus

The job ads corpus (Cardenas Acosta et al., 2016) was built by extracting job ads from several popular job search websites in Peru, and it is divided in two parts, one for entity extraction tasks and the other for topic inference.

The first part consists of 400,000 word tokens spanning 800 posts manually labeled with entity tags following the CoNLL-2000 BIO tagging format (Ramshaw and Marcus, 1995). This amount of data proved to give good results for named entity extraction in Spanish, as reported by Carreras et al. (2002). The custom entities defined for our task are FUN (tasks to be performed at the job), REQ (skills required) and CARR (preferred professional major of the applicant). Table 1 show an example of annotation along with its translation into English, whereas Table 2 shows the proportion of entities in the annotated corpus as well as the average length in words.

Spa	Egresado/*O* en/*O* Ingeniería/*B-CARR* de/*I-CARR* Software/*I-CARR* con/*O* conocimientos/*O* de/*O* base/*B-REQ* de/*I-REQ* datos/*I-REQ* MySQL/*I-REQ* .
Eng	Graduate in Software Engineering with knowledge of MySQL databases

Table 1: Example of tagging of custom entities

Entity	**Number of chunks**	**Avg. number of words per chunk**
FUN	3291	11.09
REQ	4833	1.84
CARR	2097	1.64

Table 2: Defined entities and presence in corpus

The second part consists of only job ads requesting engineering professions published between January and March 2015. We compose each document instance as the concatenation of the title and description fields of each job ad. We consider 23 engineering categories and leave out categories with less than 50 posts. Since the same job ad can be published in more than one website, we consider it as repeated if the same description of the position is found within the last fifteen days in the database. The final topic inference corpus consists of 9,472 job ads, with an average of 91.3 ± 40.8 tokens per document and a total of 476,990 tokens.

The dataset is publicly available in the Lindat repository.[1]

[1] http://hdl.handle.net/11234/1-2673

4.2. Data Preparation

Job ads often contain very sparse information like emails, dates, office hours and salary. We treated this type of tokens as noise and replaced them with appropriate tags (e.g. URL) using regular expressions. Low-frequency words were filtered as well, following Bikel et al. (1999) approach of using generic labels based on orthographic features (e.g. Capitalized, hasDigit, AllCaps).

4.3. Skills and Tasks Extraction

We train one tagger for each entity, each one with the following features. Note that each feature is conditioned to the current label being predicted, unless otherwise specified (e.g. transition features).

- Trigger word features for the current word (Carreras et al., 2002), only for REQ and CARR entities.
- Lowercase form and position of all words in a window of $\pm n$ words (Carreras et al., 2002). For the CARR entity, $n = 2$ and for the others $n = 3$.
- Stemmed form and position of previous, current and next word.
- Part-of-list feature ($\mathtt{list} :: y_i$), if current word is part of a list.
- Orthographic features, including long-word and single-digit (Carreras et al., 2002), for previous, current and next word.
- Suffix and prefix features, last and first 3 characters respectively, for previous, current and next word.
- Word brown-cluster mapping features (Miller et al., 2004) for previous, current and next word.
- Token bigram and trigram emission features (Liang and Collins, 2005) for lowercase and stemmed form of all words, as well as orthographic class, in a window of ± 2 words.
- Relative position of sentence in document, if the current sentence belongs to the document border (first one or last two sentences). Only used for FUN entity.
- Bigram transition features for word cluster mapping (Liang and Collins, 2005), used only for REQ entity.
- Bigram transition features (Liang and Collins, 2005) for lowercase and stemmed form, as well as orthographic class, of each word in the bigram.
- Bigram transition features of last states (labels) predicted.

Preliminary experiments showed that POS information does not contribute significantly to the taggers' performance. Additionally, usage of a Conditional Random Field model (Lafferty et al., 2001) showed no significant improvements with respect to the Averaged Perceptron. We also considered using pre-trained word embeddings as input, but the limited amount of data available would not allow us to obtain reliable estimates. On the other hand, pre-training the embeddings on a large monolingual benchmark and then training over our data would not allow the model to learn terminology not only specific to the domain but to the Spanish dialect spoken in the country in which the ads where published.

The annotated dataset is divided in 70, 15 and 15 percent for training, validation and testing, respectively. The evaluation metrics are the standard precision P (fraction of output chunks that exactly match the reference chunks), recall R (fraction of reference chunks returned by the tagger), and their harmonic mean, the F_1 score, $F_1 = 2 \times P \times R/(P + R)$. The accuracy rate for individual labeling decisions is overoptimistic as an accuracy measure for NER, given that O labels are more frequent. Even so, we report BIO accuracy for reference.

4.4. Topic Modeling

We employ the analysis approach suggested by Airoldi et al. (2010), aimed to explore the effect of varying the data source over model dimensionality and using different hyperparameters inference strategies and algorithms (Variational Inferences vs Gibbs sampling).

We explore models both estimating and fixing the latent categories proportion per document hyperparameter (α), and compare each for the case in which all the text from the ad is used for training versus using only entities extracted by the taggers. Hence, we compare six LDA models in a layout denoted as {VEM with estimated alpha, VEM with fixed alpha, Gibbs with estimated alpha} × {Whole text, Text chunks }.

For the case in which α is estimated during training, we set its initial value to $\alpha = 5/K$ and fix $\beta = 0.1$, as suggested by Griffiths and Steyvers (2004). Then, K is grid-search tuned to minimize perplexity of the model. For the case in which α is fixed, it is grid-search tuned after an optimum K is found. This strategy follows the conclusion that the VEM inference algorithm estimates too low α hyperparameters, as reported by Asuncion et al. (2009). Low α hyperparameters cause the model to assign few topics per document, only one in the worse case.

Dimensionality Selection Each time we fit a mixed-membership model to data, we must specify the number of latent categories, K, in the model. The goal of model selection is to find K^*, the number of latent categories that is optimal in some sense. We use 10-fold cross-validation following the approach described in Airoldi et al. (2010), and widely used in other machine learning applications. First, we split the N job ads into 10 batches. Then, we estimate the model parameters using the ads in nine batches, and we calculate the posterior perplexity of the ads in the tenth held-out batch. This approach leads to summarize how good a model fits for a given $K \in [5, 200]$, on a batch of ads not included in the estimation. We fit each model a total of 60 times (10 times in cross-validation for each of 6 models) for each value of K. Fold splitting during cross-validation was seeded to assure consistency of multiple runs of a model and to assure comparability among different models that use the same data.

For our experiments, we use the LDA R library *topicmodels* by Grün and Hornik (2011), which wraps Blei et al. (2003) C code for VEM inference and Phan et al. (2008) C++ code for Gibbs sampling.

4.5. Topic Coherence

In our coherence experiments, we use the framework proposed by Röder et al. (2015), available online,[2] in which many more scores are available and a reference corpus for probability counts can be specified. Although Mimno et al. (2011) do not use any external reference corpus, Röder et al. (2015) showed that using Wikipedia as an additional reference corpus improved correlation with gold human ratings for this metric. Following this setup, we use as external reference corpus the concatenation of the entire Job Ads dataset (more than 500,000 documents) and the Wikipedia dump in Spanish. Following the literature (Chang et al., 2009; Mimno et al., 2011; Aletras and Stevenson, 2013; Lau et al., 2014), we employ the top 10 words by topic.

5. Results and Discussion

5.1. Skills and Tasks Extraction

Table 3 shows results for the tagger. It can be observed that CARR tagger shows the best performance. This can be explained by the fact that majors are mostly mentioned in determined word patterns in job ads. For the FUN tagger, taking advantage of the fact that functions are not mentioned in the beginning nor the end of the ad improves the precision significantly in comparison to early experiments. In addition, FUN entities mostly appear at the beginning of the sentences.

Entity	# Feat.	P	R	F_1	ACC.
FUN	503701	61.1	62.3	61.7	93.4
REQ	605864	77.6	55.9	65.0	97.1
CARR	215143	87.2	86.9	87.0	99.5

Table 3: Feature set sizes and taggers' performance

5.2. Topic Models Tuning

Following the procedure described in sections 4, we show in Figure 1 the behavior of the held-out perplexity as the number of topics changes. We observe that in general

[2] https://github.com/AKSW/Palmetto

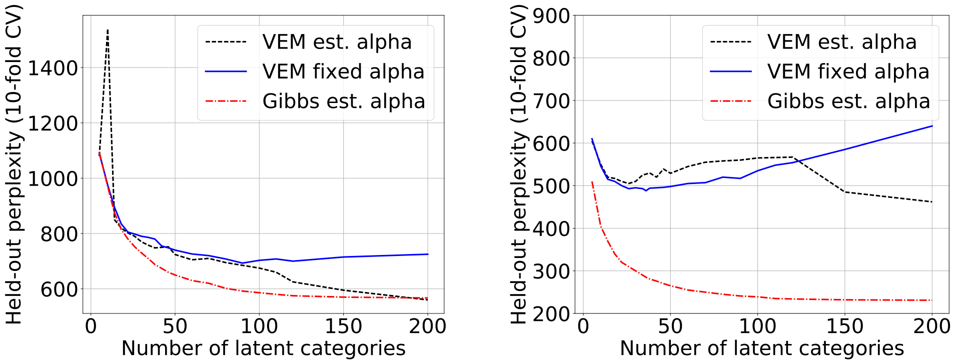


Figure 1: Average held-out perplexity as a function of the number of latent categories K for whole text models 1, 2 and 3 (left), and text chunks models 4, 5 and 6 (right).

there is no agreement among the methods of inference for the optimal number of topics and that in some cases the perplexity does not converge.

Using the *UMass* topic coherence score to measure the quality of the models as the number of topics changes, we observe in Figure 2 that for each method of inference, the optimal number of topics is found between 5 and 18. We choose $K = 10$ as the optimal value for both models, as it gives the best score for models using text chunks (Figure 2, right) regardless of the inference strategy followed. For models using the whole text (left), this value is fairly close to the optimum (15).

5.3. Topic Coherence Improvement

For the optimal number of topics chosen in Section 5.2, 10, the bar plot in Figure 3 shows the improvement of the UMass topic coherence when restricting the text to the chunks extracted by the entity extractors. Also, it can be observed that this happens independently of the method of inference, and that there is at least an improvement of 40% in each case, with *VEM estimated alpha* having the better coherence score when text chunks are used.

5.4. Qualitative and quantitative analysis of inferred categories

Topics are explored by examining the top 10 words (Tables 4, 5 and 6). In addition, the topic proportion for each professional major is investigated. For each major, the mean of posterior membership scores of all documents where this major was required is taken, as proposed by Erosheva et al. (2004). Figure 5 shows this calculation for VEM inference method with fixed alpha. Figure 4 presents matrices for the six

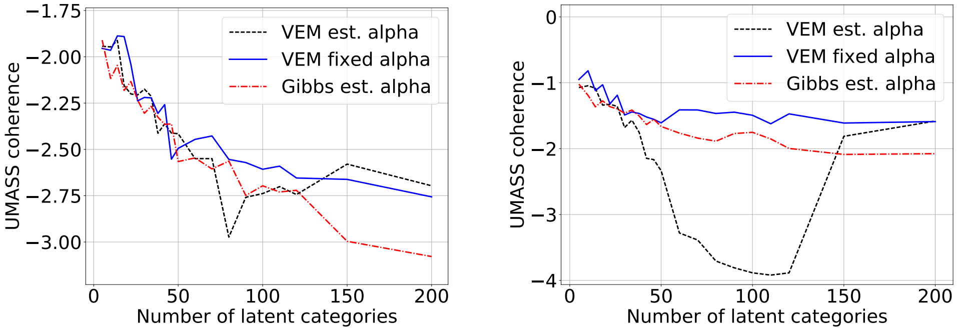


Figure 2: Average *UMass* coherence score (higher is better) as a function of the number of topics K for whole text models 1, 2 and 3 (left), and text chunks models 4, 5 and 6 (right).

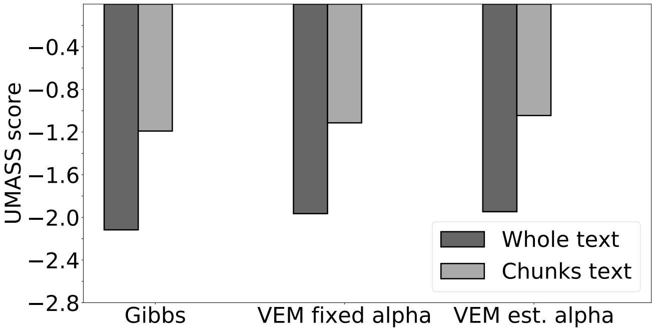


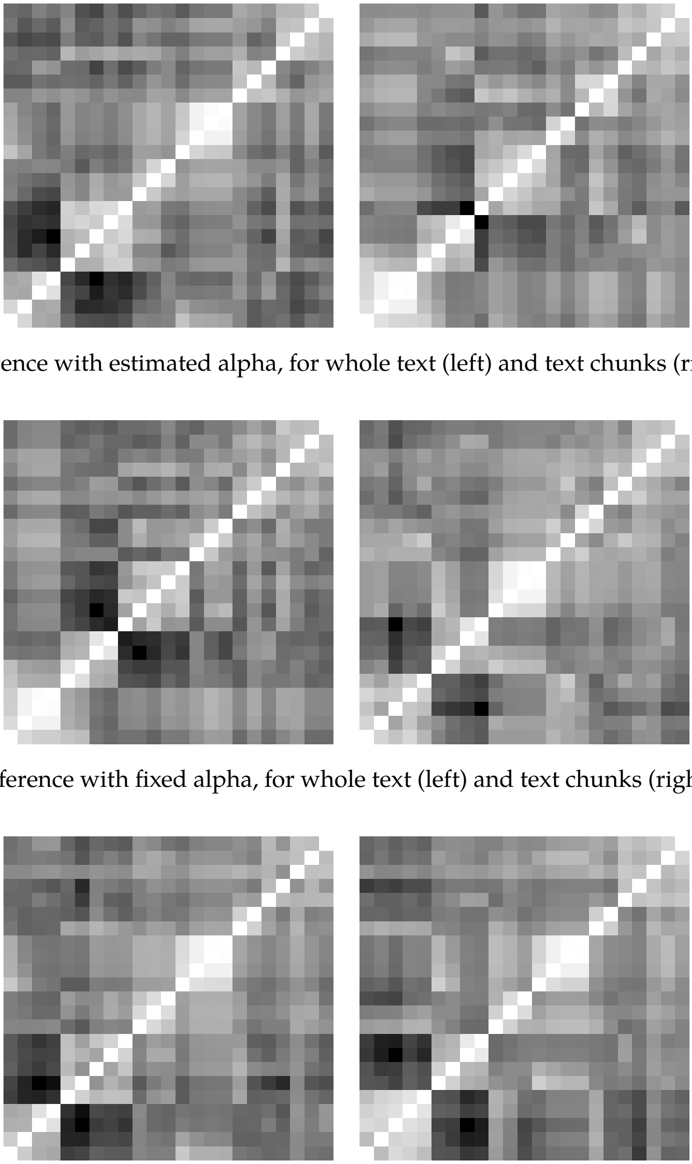
Figure 3: Comparison of the UMass coherence score for each method of inference.

mixed-membership models, which represent the similarity of the probability distributions over categories between all majors. This similarity is calculated using Hellinger distance. Each row and column of each matrix represent a professional major and its similarity with other majors, regarding the text source and inference strategy applied. Major names are not shown because each matrix has different major names order in rows and columns. The purpose of Figure 4 is to unveil the effect of how professional majors are grouped. A similar behavior can be observed in Figure 5 by observing for each topic the majors that have the most vivid colors.

Furthermore, it can be observed in both graphics Figure 5 and 4 that for the case of the text chunks model, getting rid of irrelevant words (ignored by the entity extractors) has the effect of smoothing the probability distribution over topics. For instance, for the whole text model, the job ads for environmental engineering basically just talk about one topic. On the other hand, for the text chunks model, the major now talks about more than one topic with similar proportions.

A closer look at Figure 5 allows to spot three main behaviors under the effect of restricting the source text (whole text versus text chunks).

- Joining of redundant categories
  Consider the major of Electronic Engineering. In Figure 5 for the whole text model, topics 4 and 7 are the predominant ones. See Table 4 for the content of the topics. On the other hand, for the text chunks model, it can be seen that only topic 5 is predominant. Table 4 confirms that topic 5 of the text chunks model contains words (with high probability) from both of the topics of the whole text model.
- Splitting in two or more detailed categories
  Consider the majors of Environmental Engineering and Industrial Hygiene and Safety. In Figure 5 for the whole text model, topic 2 is predominant for both majors. Exploration of this topic reveals that its content is related to industrial, environmental safety and management, as can be appreciated in Table 5. On the other hand, for the text chunks model, it can be observed that categories 2 and 10 are predominant and with almost the same proportion. A closer exploration reveals that topic 2 is related to environmental safety and management but no longer contains the word industrial, which appears in topic 10, i.e. the top two words from topic 2 (whole text model) was split.
- Persistence of latent structure
  There are cases where the number of predominant topics does not change. Consider the majors of Systems and Informatics Engineering. For the whole text model, it can be observed that topic 4 is predominant. Likewise, for the text chunks model, topic 9 present the same behaviour. Table 6 shows that the content of these topics is maintained in both models.

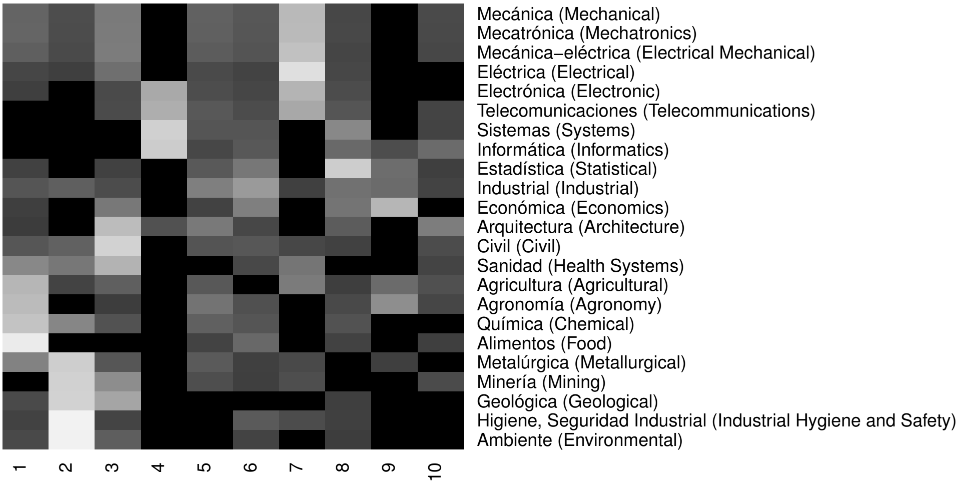


(a) VEM inference with estimated alpha, for whole text (left) and text chunks (right) models.

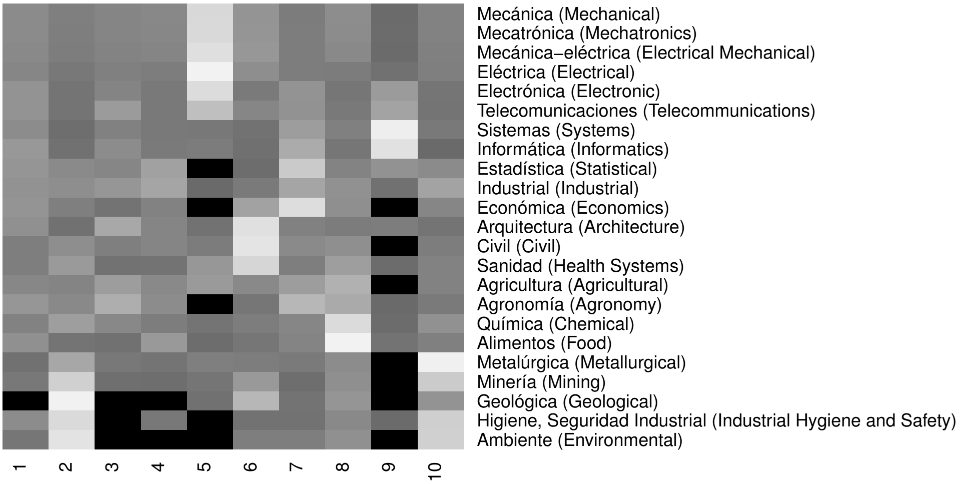
(b) VEM inference with fixed alpha, for whole text (left) and text chunks (right) models.

(c) Gibbs inference, for whole text (left) and text chunks (right) models.

Figure 4: Similarity matrices using Hellinger distance between discrete distributions (topic proportion over majors), for each of the six topic models mentioned in section 4.4. A whiter cell means a shorter distance, i.e. more similar categories.



(a) Whole text



(b) Text chunks

Figure 5: Scaled estimated average membership of engineering majors to 10 categories inferred by VEM with fixed alpha for (a) whole text setup and (b) text chunks setup. The whiter the highest the membership; black denotes zero membership. Original Spanish names for majors are showed with the English gloss in parenthesis.

Whole text		Text chunks
Topic 4	Topic 7	Topic 5
sistemas (*systems*)	técnico (*technician*)	mantenimiento (*maintenance*)
técnico (*technician*)	mantenimiento (*maintenance*)	mecánica (*mechanical*)
informática (*informatics*)	mecánica (*mechanical*)	electrónica (*electronics*)
desarrollo (*development*)	eléctrica (*electrical*)	eléctrica (*electrical*)
computación (*computation*)	electricidad (*electricity*)	electricidad (*electricity*)
sql (*SQL*)	industrial (*industrial*)	técnico (*technician*)
programador (*programmer*)	preventivo (*preventive*)	instalación (*installation*)
analista (*analyst*)	electrónica (*electronics*)	reparar (*repair*)
programación (*programming*)	sistemas (*systems*)	preventivo (*preventive*)
servidor (*server*)	instalación (*installation*)	sistemas (*systems*)

Table 4: Topics behavior for VEM fixed α strategy: joining of redundant categories. Each entry consists of the Spanish token and its English gloss in parenthesis.

Whole text		Text chunks
Topic 2	Topic 2	Topic 10
seguridad (*safety*)	seguridad (*safety*)	industrial (*industrial*)
industrial (*industrial*)	risk	supervisor (*supervisor*)
management	environmental	administración (*management*)
ocupacional (*occupational*)	management	marketing
ambiente (*environment*)	ocupacional (*occupational*)	especialización (*specialization*)
supervisor (*supervisor*)	normas (*norms*)	venta (*selling*)
normas (*norms*)	documentos (*documents*)	economía (*economy*)
capacitación (*capacitation*)	seguimiento (*tracing*)	proactivo (*proactive*)
risk	industrial (*industrial*)	responsable (*responsible*)
iso (*ISO*)	soporte (*support*)	dinámico (*dynamic*)

Table 5: Topics behavior for VEM fixed α strategy: splitting in two or more detailed categories. Each entry consists of the Spanish token and its English gloss in parenthesis when applicable.

Whole text	Text chunks
Topic 4	Topic 9
sistemas (*systems*) técnico (*technician*) informática (*informatics*) desarrollo (*development*) computación (*computation*) sql (*SQL*) programador (*programmer*) analista (*analyst*) programación (*programming*) servidor (*server*)	sistemas (*systems*) informática (*informatics*) analista (*analyst*) programador (*programmer*) sql (*SQL*) desarrollo (*development*) computación (*computation*) programación (*programming*) servidor (*server*) administrador (*administrator*)

Table 6: Topics behavior for VEM fixed α strategy: persistence of latent structure. Each entry consists of the Spanish token and its English gloss in parenthesis.

6. Conclusions

Throughout the analysis of multiple variants of topic models, consistent results confirm our hypothesis that coherence of inferred categories significantly improves when using only relevant text extracted by named entity extraction rather that the whole document. In our case study, the relevant text constitutes expected skills, tasks to perform, and academic background in job ads.

Compared to categories inferred using whole-text models, entities models generate categories that join redundant ones and split to high skill-specific categories. In addition, fine-grained categories are preserved with entity models.

Bibliography

Airoldi, E. M., E. A. Erosheva, S. E. Fienberg, C. Joutard, T. Love, and S. Shringarpure. Reconceptualizing the classification of PNAS articles. In *Proceedings of the National Academy of Sciences of the USA*, volume 107, pages 20899–20904, 2010.

Aletras, Nikolaos and Mark Stevenson. Evaluating Topic Coherence Using Distributional Semantics. In *10th Int. Conf. on Computational Semantics (IWCS'13)*, 2013.

Asuncion, A., M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI'09)*, pages 27–34, 2009.

Bikel, D. M., R. Schwartz, and R. M. Weischedel. An Algorithm that Learns What's in a Name. *Journal of Machine Learning*, 34:211–231, 1999.

Blei, D. M. and J. D. Lafferty. Correlated Topic Models. In Weiss, Y., B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, pages 147–154. MIT Press, Cambridge, 2005.

Blei, D. M. and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference of Machine Learning (ICML '06)*, pages 113–120, August 2006.

Blei, D. M., A. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Blei, D. M., T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57:7.1–7.30, 2007.

Cardenas Acosta, Ronald, Kevin Bello Medina, Alberto Coronado, and Elizabeth Villota. Engineering job ads corpus, 2016. URL `http://hdl.handle.net/11234/1-2673`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Carreras, X., L. Marquez, and L. Padró. Wide-Coverage Spanish Named Entity Extraction. In *VIII Conferencia Iberoamericana de Inteligencia Artificial, IBERAMIA'02*, 2002.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In Bengio, Y., D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc., 2009.

Collins, M. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002.

Erosheva, E. A., S. E. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. In *Proceedings of the National Academy of Sciences of the USA*, volume 101, pages 5220–5227, 2004.

Griffiths, T. L. and M. Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the USA*, volume 101, pages 5228–5235, 2004.

Grün, Bettina and Kurt Hornik. topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13):1–30, 2011. doi: 10.18637/jss.v040.i13.

Hall, Mark M., Paul D. Clough, and Mark Stevenson. Evaluating the Use of Clustering for Automatically Organising Digital Library Collections. In *Second International Conference on Theory and Practice of Digital Libraries (ERCIMDL)*, 2012.

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML*, pages 282–289, San Francisco, CA, USA, 2001. ISBN 1-55860-778-1.

Lau, Jey Han, David Newman, and Timothy Baldwin. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *European Chapter of the Association for Computational Linguistics (EACL'14)*, 2014.

Liang, P. and M. Collins. Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology, 2005.

Miller, S., J. Guinness, and A. Zamanian. Name tagging with word clusters and discriminative training. In *Proceedings of the Proceedings of HLT-NAACL 2004*, pages 337–342, 2004.

Mimno, David M., Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew McCallum. Optimizing Semantic Coherence in Topic Models. In *Empirical Methods in Natural Language Processing (EMNLP 2011)*, 2011.

Musat, Claudiu Cristian, Julien Velcin, Stefan Trausan-Matu, and Marian-Andrei Rizoiu. Improving Topic Evaluation Using Conceptual Knowledge. In *22nd International Joint Conference on Artificial Intelligence (IJCAI-2011)*, 2011.

Newman, D., C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 680–686, August 2006.

Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL http://dl.acm.org/citation.cfm?id=1857999.1858011.

Nguyen, Thang, Jordan L. Boyd-Graber, Jeffrey Lund, Kevin D. Seppi, and Eric K. Ringger. Is Your Anchor Going Up or Down? Fast and Accurate Supervised Topic Models. In *North American Chapter of the Association for Computational Linguistics (NAACL 2015)*, 2015.

Paul, Michael J. and Roxana Girju. A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics. In *24th Annual Conference on Artificial Intelligence (AAAI-10)*, 2010.

Phan, Xuan Hieu, Minh Le Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *17th International World Wide Web Conference (WWW 2008)*, pages 91–100, 2008.

Ramshaw, L. A. and M. P. Marcus. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*. ACL, 1995.

Reisinger, Joseph, Austin Waters, Bryan Silverthorn, and Raymond J. Mooney. Spherical Topic Models. In *27th International Conference on Machine Learning (ICML 2010)*, 2010.

Röder, Michael, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of WSDM*, 2015.

Stevens, Keith, W. Philip Kegelmeyer, David Andrzejewski, and David Buttler. Exploring Topic Coherence over Many Models and Many Topics. In *Proceedings of EMNLP-CoNLL'12*, 2012.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

Yang, Yi, Doug Downey, and Jordan L. Boyd-Graber. Efficient Methods for Incorporating Knowledge into Topic Models. In *Empirical Methods in Natural Language Processing*, 2015.

Address for correspondence:
Ronald Cardenas
racardenasa@uni.pe
National University of Engineering,
Department of Mechanical Engineering
Tupac Amaru Avenue 210, Lima 25, Lima, Peru