

Labor market demand analysis for engineering majors in Peru using Shallow Parsing and Topic Modeling

Ronald A. Cardenas
Department of Mechanical
Engineering
Universidad Nacional de
Ingeniería
Lima, Peru
roncardenas-
acosta@gmail.com

Kevin S. Bello
Department of Mechanical
Engineering
Universidad Nacional de
Ingeniería
Lima, Peru
ksbm190493@gmail.com

Alberto M. Coronado
Department of Mechanical
Engineering
Universidad Nacional de
Ingeniería
Lima, Peru
am.coronado@gmail.com

Elizabeth R. Villota
Section of Mechanical
Engineering
Pontificia Universidad Católica
del Peru
Lima, Peru
elvillota@gmail.com

ABSTRACT

Talent shortage is a well-known problem that industry has been dealing with for several years now, with a large number of engineering vacant positions being difficult to fill. We present an analysis of the Peruvian labor market demand based on data collected from job ads available on the web.

We use topic models, a type of mixed-memberships models, in order to infer semantic categories in co-occurrence of words by focusing in the functions, requirements and professions requested in the job position. This information is extracted by applying shallow parsing over the ads before feeding them to the models. Models using the whole text from the ads are compared to models using only the text chunks extracted with shallow parsing.

Our analysis reveal that using the extracted text chunks, some redundant categories are joined while others are split into more skill-specific categories. Fine-grained categories observed in models using the whole text are preserved.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing—*Language parsing and understanding, Text analysis*;
K.3.2 [Computers and Education]: Computer and Information Science Education—*curriculum*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

TM'15 October 19 2015, Melbourne, VIC, Australia
Copyright 2015 ACM ISBN 978-1-4503-3784-7/15/10...\$15.00.
<http://dx.doi.org/10.1145/2809936.2809945>.

General Terms

Design, Experimentation

Keywords

topic modeling, shallow parsing, education engineering

1. INTRODUCTION

The economic growth many South American countries have experienced during the last years has boosted the demand of high-specialized job positions, which often require multidisciplinary expertise. Many local employers are having a hard time finding qualified workers, especially in technical fields [16].

For educational institutions around the world, and South America is not the exception, it is not always clear what majors they should offer or what competencies they should emphasize in their curricula [17, 21]. Xue [11] analyzes if there is in fact a “STEM crisis or a STEM surplus” in the US, reporting that depending on the major and sector, it can be one or another.

This paper contributes an analysis of the Peruvian labor market demand, in an effort to unveil the influence of the academic background in the search for the ideal job applicant, regarding to what skills engineering professionals should have according to the industry.

To model the hidden relationship between academic background and skills required for a specific job vacancy, we extract requirements, functions for the job, and undergraduate background information by applying shallow parsing. The effects of using the above information versus using the whole text in the job ad as input for a carefully tuned topic model, are analyzed.

The machine learning techniques used in this analysis are explained in detail in sections 4 and 5. Section 6 explains the experiments performed and section 7 presents the discussion of results.

2. RELATED WORK

Mixed membership models, especially topic models, have received a lot of attention in the past few years regarding efficient inference algorithms and proper syntonization of hyperparameters [2], as well as recommended experiments [8].

Latent Dirichlet Allocation, the initial topic model proposed by Blei et al. [7] has been extended and modified in several ways. Variations in the graphical model include the correlated model [5], the hierarchical Dirichlet process model [25], the hierarchical topic model [4]; and the dynamic topic model [6]. Gruber et al. [14] modifies the original graphical model to include a Markov chain in the topic assignment of words, in the so called Hidden Topic Markov Model, leaving the bag-of-words approach and modeling the document as a Markov chain. As for distributed and parallelized solutions, Smola and Narayanamurthy [24] proposed an architecture for large scale parallel topic modeling with good results.

Airoldi et al. [1] analysed the effect of varying the source text and inference strategies for PNAS biological sciences publications. Newman et al. [20] proposed a customized graphic model which includes named entity words as extra source of text but with its own hyperparameter estimation, directly learning the entity-topic relationship and making better predictions about entities.

We investigate the effect of varying the source text extracting meaningful information with shallow parsing over the resulting latent structure of categories, as well as their quality. Shallow parsing is frequently reduced to sequence labeling problems, and most of the previous work uses machine learning approaches. Most of the work tackles the task of extracting noun phrases with structured predictors [23, 18]. Although other chunk types have not received the same attention as NP chunks, Buchholz et al. [9] present excellent results for NP, VP, PP, ADJP and ADVP chunks.

3. JOB ADS CORPUS

The job ads corpus was built by extracting job ads from several popular job search websites in Peru. The data was extracted from January to March 2015. We consider job ads published only in January, consisting of 60,000 ads, for prototyping and preliminary experiments. 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. Furthermore, as our data of interest corresponds to engineering positions, our implemented profession recognizer, as described in Section 4, was used to select job ads requesting engineering professions, obtaining roughly 9,000 job ads. Our initial number of engineering majors was 30, however, we discarded majors with less than 50 job ads, ending up with a total of 23 engineering majors.

4. SHALLOW PARSING

The task of extracting requirements, functions and professions from a job ad is reduced to a sequence labeling problem.

Since natural language processing models are very corpus sensitive, a sample of 800 job ads, more than 130 000 words. This amount of data gave good results for shallow parsing (specifically for named entity extraction) in Spanish, as reported by Carreras et al. [10]. This subset was manually tagged following the CoNLL-2000 BIO tagging format, ini-

Table 1: Defined chunks and presence in corpus

Chunk	Base phrase	# of chunks	Avg. word length
FUN	VP	3291	11.09
REQ	NP & VP	4833	1.84
CARR	NP	2097	1.64

tially proposed by Ramshaw and Marcus (1995) [22], but without annotating part-of-speech (POS) tags. Preliminary experiments showed that POS information does not contribute significantly to chunkers performance.

The base phrases defined are FUN (functions of the job, VP), REQ (requirements, NP and VP) and CARR (majors, NP). Table 1 shows the proportion in the annotated corpus for each chunk defined, as well as the average length in words.

4.1 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. [3] approach.

4.2 Averaged Structured Perceptron

The structured perceptron and its averaged version was initially introduced by Collins, 2002 [11]. They differ from the well-known perceptron algorithm in that the output for each training instance pair $(x_i, y_i) \in T$ is a structure $y' \in Y_i$, where Y_i 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.

The set of features used in the chunkers are the following.

- Trigger word features for the current word [10], only for REQ and CARR chunkers.
- Lowercase form and position of all words in a window of $\pm n$ words [10]. For the CARR chunker, $n=2$ and for the other ones $n=3$.
- Stemmed form and position of previous, current and next word.
- Part-of-list feature ($list :: y_i$), if current word is part of a list.
- Orthographic features, including long-word and single-digit [10], 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 [19] for previous, current and next word.

- Bigram and trigram emission features [15] for lowercase and stemmed form of all words, as well as orthographic class, in a windows of ± 2 words.
- Bigram transition features for word cluster mapping [15], used only in REQ chunker.
- Bigram transition features [15] for lowercase and stemmed form, as well as orthographic class, of each word in the bigram.
- Relative position of sentence in document, if current sentence belongs to the document border (first one or last two sentences). Only used in FUN chunker.

5. TOPIC MODELING

5.1 Latent Dirichlet Allocation

In this section, we briefly describe the graphical model called Latent Dirichlet Allocation (LDA) [7], 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 $\theta_d \sim \text{Dir}(\alpha)$.
2. For all T topics sample $\phi_t \sim \text{Dir}(\beta)$.
3. For each of the N_d words v_i in document d :
 - Sample a topic $z_i \sim \text{Multinomial}(\theta_d)$
 - Sample a word $v_i \sim \text{Multinomial}(\phi_{z_i})$
 - Observe the word

We assume symmetric Dirichlet priors for θ_d and ϕ_t , as suggested by Griffiths and Steyvers [13].

Nowadays, in the literature there are several learning algorithms for LDA. Regarding inference strategies for the models, we make use of Gibbs Sampling as described in [13] and the Variational Expectation - Maximization (VEM) algorithm as described in [7].

5.2 Experiment configuration

We compare six LDA models in a 2x3 layout, as summarized in Table 2. This analysis approach, suggested by Airola et al. [1], aims to explore the effect over model dimensionality, of varying the data source (all the text from the ad versus text extracted by the chunkers) and using different hyperparameters inference strategies.

We explore models both estimating and fixing the latent categories proportions per document hyperparameter (α). When estimating α , the initial value is set to $5/K$, as suggested by Griffiths and Steyvers [13]. Once a value of K in which the global or a local minimum is spotted for the estimated α case, a grid search over α is performed for that value of K in order to find the best α for the fixed-case. This strategy follows the conclusion that the VEM inference algorithm estimates too low α hyperparameters, as reported by Asuncion et al. [2]. Low α hyperparameters cause the model to assign few topics per document, only one in the worse case.

Table 2: Mixed-membership models in the analysis

	VEM with estimated alpha	VEM with fixed alpha	Gibbs with estimated alpha
Full text	model 1	model 2	model 3
Text chunks	model 4	model 5	model 6

Table 3: Feature set sizes and chunkers' performance

Chunker	# Feat.	P	R	F_1	BIO 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

5.3 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 [1], 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 four batches, and we calculate the posterior perplexity of the ads in the fifth held-out batch. This approach leads to summarize how good a model fits for a given K , on a batch of ads not included in the estimation. We consider a wide range of number of latent categories as follows: increments of 5 for $5 \leq K \leq 10$, increments of 10 for $60 \leq K \leq 120$, and increments of 50 for $150 \leq K \leq 200$. Thus, we fit the models a total of 60 times (10 times in cross-validation for each of 6 models) for each of 24 values of K .

6. EXPERIMENTS

6.1 Shallow Parsing

The annotated dataset is divided in 70, 15 and 15 percent for training, validation and testing, respectively. The standard evaluation metrics for a chunker are precision P (fraction of output chunks that exactly match the reference chunks), recall R (fraction of reference chunks returned by the chunker), 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 over-optimistic as an accuracy measure for shallow parsing, given that O labels are more frequent. Even so, we report BIO accuracy for reference.

Table 3 shows results for the chunkers. It can be observed that CARR chunker 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 chunker, 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 chunks mostly appear at the beginning of the sentences.

6.2 Topic modeling

6.2.1 Dimensionality

Following the procedure described in Section 5.2, first, we compute the average held-out perplexity for each of the

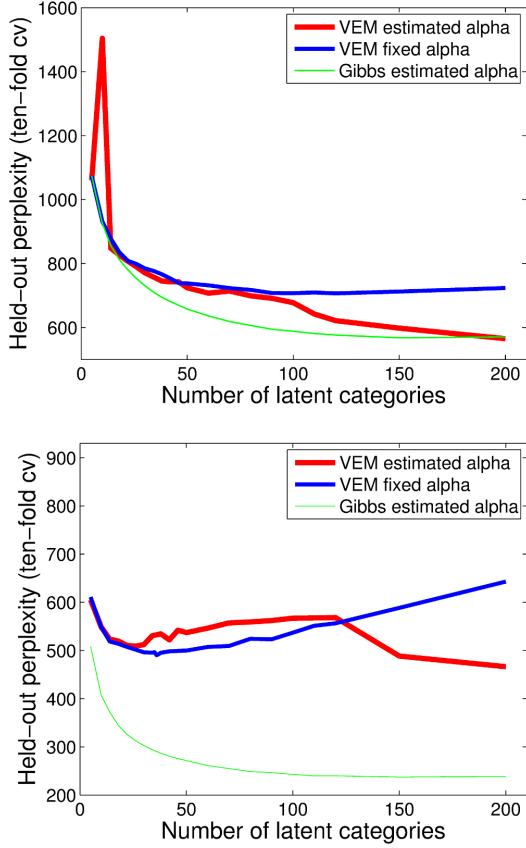


Figure 1: Average held-out perplexity as a function of the number of latent categories K for whole text models 1, 2 and 3 (top), and text chunks models 4, 5 and 6 (bottom).

different values of the number of latent categories, as mentioned in Section 5.3, for the models with the hyperparameter α estimated (lines red and green in Figure 1). Then, as can be observed in Figure 1, among the green and red lines, only the red line (bottom) has a local minimum with the number of latent categories set to 26. We use this local minimum to find an optimal value for the hyperparameter α , as shown in Figure 2. Finally, we use this optimal value of the hyperparameter to reveal the behavior of the models with fixed hyperparameters (blue lines in Figure 1).

Although the blue line (bottom) in Figure 1 has a global minimum with the number of latent categories set to 36, we use the local minimum found in the red line (bottom) with K set to 26. The latter because the number of engineering majors in Peru is less than 36, and according to our criterion, specializations converge to a mixture of common pools of knowledge and required skills. Therefore, we think that a value of 26 is a more reasonable choice.

6.2.2 Qualitative and quantitative analysis of inferred categories

Regarding the dimensionality analysis, we set K to 26 as explained in Section 6.2.1 for the analysis of inferred categories, for every LDA inference strategy.

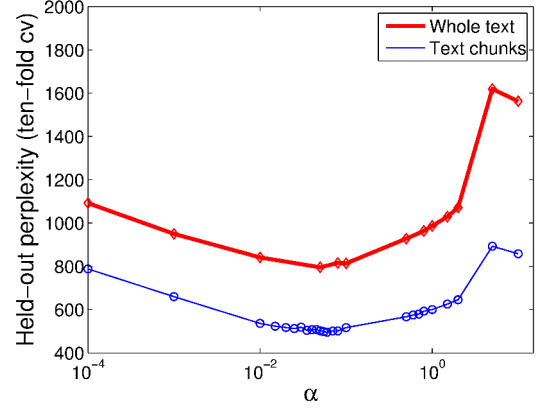


Figure 2: Average held-out perplexity as a function of α , for mixed-membership models using all the words from the ad (red) and using words from the extracted chunks (blue). Both models were fitted by using a number of latent categories set to 26.

Latent categories are explored examining high probability words in each category (Tables 4, 5 and 6). In addition, the latent category proportion in each major is investigated. For each major, the mean of posterior membership scores of all documents where this major was found is taken (Figure 3), as proposed by Erosheva et al. [12]. An inspection to the latter plot provides an idea of how related the engineering majors are to one another by observing for each topic the majors that have the most vivid colors.

Furthermore, we show in Figure 4 matrices for the six 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.

In both graphics (Figures 3 and 4), it can be observed that for the case of the text chunks models (models 4, 5 and 6), getting rid of irrelevant words (ignored by the chunkers) has the effect of smoothing the probability distribution over topics. For instance, in Figure 3, for whole-text models 1, 2 and 3, the job ads for agricultural engineering basically just talks about one topic.

On the other hand, for text-chunk models 4, 5 and 6, the major now talks about more than one topic with similar proportions. This phenomenon also appears in the Hellinger matrices (Figure 4) in which by using words from all the text causes the majors have high proportions in few topics. Thus the plot ends up with lots of blue squares which means there is no relationship between the two majors. Whereas for the case of using words from the extracted chunks the topic proportions are more uniformly distributed. Thus the plot contains less blue squares and reveal more significant relationships between the majors.

A closer look at Figure 3 allows to spot three main cate-

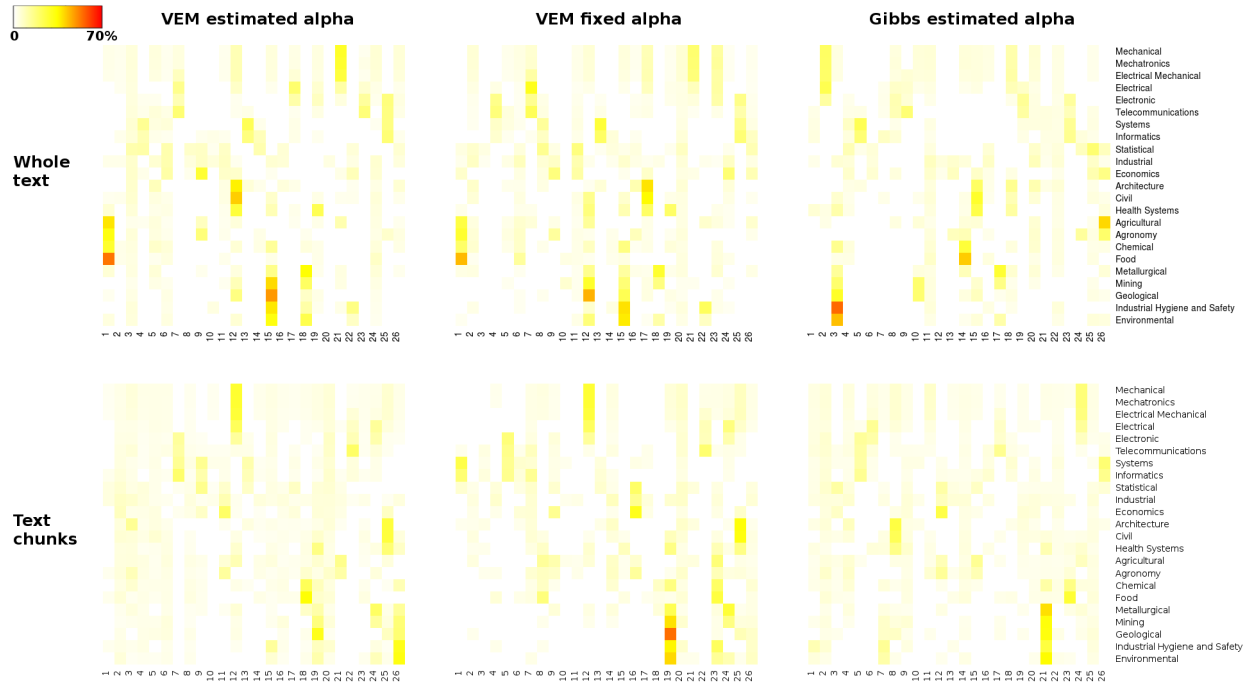


Figure 3: Estimated average membership of engineering majors in the 26 latent categories for mixed-membership models 1-6 in Table 2.

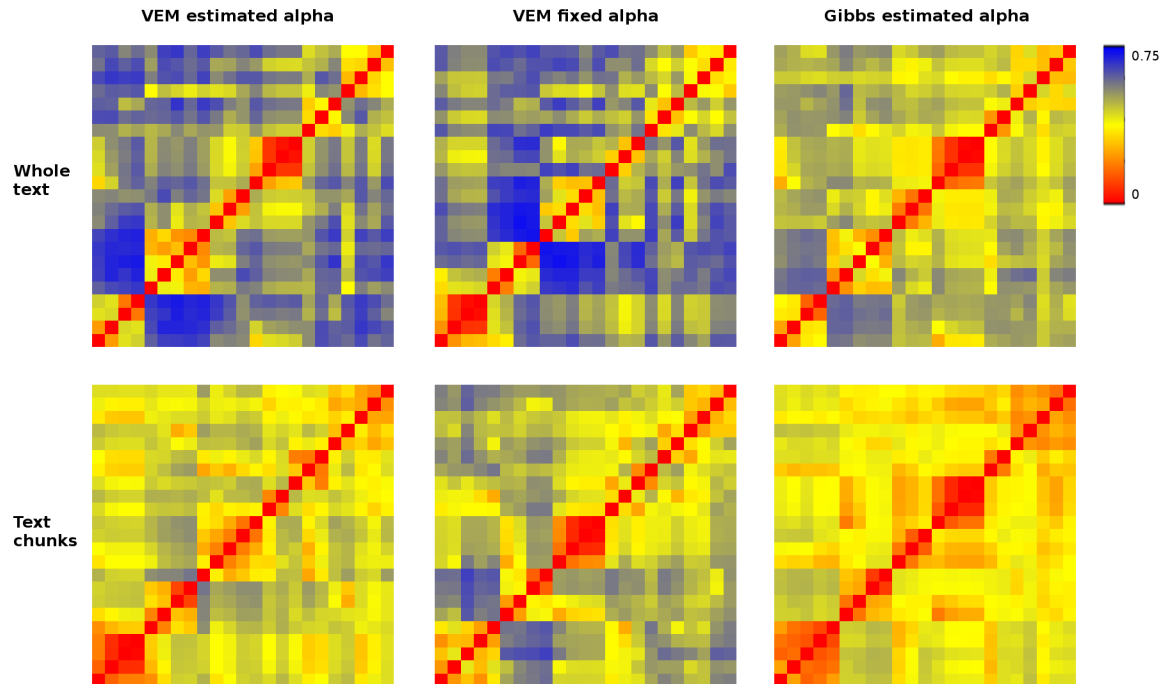


Figure 4: Similarity matrices using Hellinger distance between discrete distributions (categories proportion over majors), for each of the six mixed-membership models in Table 2.

gories behaviors under the effect of variation of source text (whole text versus text chunks), as follows.

- **Joining of redundant categories**

Consider the majors of Mechanics and Mechatronics Engineering for models with VEM inference algorithm with fixed α (models 2 and 5). In Figure 3, for whole-text model 2, categories 21 and 23 are predominant and have almost the same high category proportion. It can be seen in Table 4 that these categories are technically equivalent.

On the other hand, for text-chunk model 5, it can be seen that only category 12 is predominant in these majors. Table 4 confirms that its content is the combination of the categories separated in the whole text model. The same behaviour is observed when exploring Mining and Geological Engineering, where categories 3 and 10 for model 2 are combined into category 21 in model 5.

- **Splitting in two or more detailed categories**

Consider the major of Geological Engineering for models with VEM inference algorithm with estimated α (models 1 and 4). In Figure 3 for whole-text model 1, category 15 is predominant in the major. Exploration of this category revealed that its contents are related to personnel and risk management, as can be appreciated in Table 5. On the other hand, for text-chunk model 4, it can be observed that categories 19 and 26 are predominant and with almost the same proportion. A closer exploration revealed that category 19 is related to personnel management and 26 to risk management, i.e. category 15 in the previous model was splitted.

- **Persistence of latent structure**

There are cases where the number of predominant categories does not change. Consider the majors of Systems and Informatics Engineering for the models with Gibbs sampling (3 and 6). For whole-text model 3, it can be observed that categories 5 and 23 are predominant and with almost the same proportion. Likewise, for text-chunk model 6, categories 5 and 26 present the same behaviour. Table 6 shows that the content of these categories is maintained in both models.

The behaviour of categories over Industrial Engineering deserves closer attention. It can be observed that in all the analysed models, the latent structure of categories over this major persist, i.e. their proportions over this major present low variance. This can be interpreted as the fact that knowledge and skills, once exclusive of this major, are nowadays required for most majors and areas.

7. CONCLUSIONS

Throughout the analysis of multiple variants of mixed-membership models, consistent results support the fact that quality of inferred categories significantly improves when using only text extracted by shallow parsing rather than the whole document. In our case study, the relevant text constitutes academic background and skills required in job ads.

Compared to categories inferred using whole-text models, text-chunk models generate categories that join redundant

Table 4: Categories behaviour: joining of redundant categories

VEM with fixed α & whole text (model 2)		VEM with fixed α & text chunks (model 5)
Cat.21	Cat.23	Cat.12
maintenance	maintenance	maintenance
technician	technician	mechanic
mechanic	industrial	electricity
garage	plant	electronic
machinery	mechanic	industrial
repair	production	electric
replacement	control	plant
service	electricity	supervisor
customer	machinery	machinery
preventive	electronic	replacement

Table 5: Categories behaviour: splitting in two or more detailed categories

VEM with estimated α & whole text (model 1)	VEM with estimated α & text chunks (model 4)	
Cat.15	Cat.19	Cat.26
assurance	management	assurance
management	specialization	environmental
industrial	system	industrial
occupational	entitled	occupational
standard	coordinator	management
risk	iso	standard
iso	industrial	risk
supervisor	mastery	prevention
procedure	operation	sanitation
prevention	manager	incident

Table 6: Categories behaviour: persistence of latent structure

Gibbs with estimated α & whole text (model 3)		Gibbs with estimated α & text chunks (model 6)	
Cat.5	Cat.23	Cat.5	Cat.26
system	technician	system	system
development	system	informatic	analyst
analyst	informatic	computer	programmer
sql	computer	electronic	sql
programmer	configuration	windows	informatic
programming	windows	communication	programming
informatic	user	operative	server
java	service	office	java
server	installation	technician	oracle
application	electronic	configuration	visual

ones and split to high skill-specific categories. In addition, fine-grained categories are preserved with text-chunk models.

A more in-depth study of semantic patterns, inferred from bigger job ads data (e.g. a whole year time span) could lead to reveal connections that get closer to the way industry treats engineering majors in reality. In addition, these results could be used by educational institutions to train professionals in what the labor market really demands. And help professionals to build a more successful career path in industry.

8. REFERENCES

- [1] E. M. Airolidi, 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.
- [2] A. Asuncion, 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.
- [3] D. M. Bikel, R. Schwartz, and R. M. Weischedel. An algorithm that learns what's in a name. *Journal of Machine Learning*, 34:211–231, 1999.
- [4] D. M. Blei, 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.
- [5] D. M. Blei and J. D. Lafferty. Correlated topic models. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, pages 147–154. MIT Press, Cambridge, 2005.
- [6] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference of Machine Learning (ICML '06)*, pages 113–120, August 2006.
- [7] D. M. Blei, A. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [8] J. Boyd-Graber, D. Mimno, and D. Newman. Care and feeding of topic models: Problems, diagnostics, and improvements. In E. M. Airolidi, D. Blei, E. A. Erosheva, and S. E. Fienberg, editors, *Handbook of Mixed Membership Models and Its Applications*, pages 225–254. CRC Press, 2014.
- [9] S. Buchholz, J. Veenstra, and W. Daelemans. Cascaded grammatical relation assignment. In *Proceedings of EMNLP/VLC-99*, pages 239–246. ACL, 1999.
- [10] X. Carreras, L. Marquez, and L. Padró. Wide-coverage spanish named entity extraction. In *VIII Conferencia Iberoamericana de Inteligencia Artificial, IBERAMIA'02*, 2002.
- [11] M. Collins. 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.
- [12] E. A. Erosheva, 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.
- [13] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the USA*, volume 101, pages 5228–5235, 2004.
- [14] A. Gruber, Y. Weiss, and M. Rosen-Zvi. Hidden topic markov models. In *Proceedings of the Conference on Artificial Intelligence and Statistics*, 2007.
- [15] P. Liang and M. Collins. Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology, 2005.
- [16] ManpowerGroup. The talent shortage continues. 2014.
- [17] I. Markes. A review of literature on employability skill needs in engineering. *European Journal of Engineering Education*, 31:637–650, 2006.
- [18] R. McDonald, K. Crammer, and F. Pereira. Flexible text segmentation with structured multilabel classification. In *Proceedings of HLT/EMNLP-05*, pages 987–994, 2005.
- [19] S. Miller, 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.
- [20] D. Newman, 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.
- [21] H. J. Passow. *What Competencies Should*

Undergraduate Engineering Programs Emphasize? A Dilemma of Curricular Design that Practitioners' Opinion Can Inform. PhD thesis, The University of Michigan, 2008.

- [22] L. A. Ramshaw and M. P. Marcus. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*. ACL, 1995.
- [23] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proceedings of HLT/NAACL-03*, volume 1, pages 134–141, 2003.
- [24] A. Smola and S. Narayanamurthy. An architecture for parallel topic models. In *Proceedings of the VLDB Endowment*, volume 3, pages 1–2, 2010.
- [25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.