UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Automatic Extraction of Nursing Tasks from Online Job Vacancies

Kobayashi, V.; Mol, S.T.; Kismihók, G.; Hesterberg, M.

Publication date
2016
Document Version
Final published version
Published in
Professional Education and Training through Knowledge, Technology and Innovation

Link to publication

Citation for published version (APA):
Kobayashi, V., Mol, S. T., Kismihók, G., & Hesterberg, M. (2016). Automatic Extraction of Nursing Tasks from Online Job Vacancies. In M. Fathi, M. Khobreh, & F. Ansari (Eds.), *Professional Education and Training through Knowledge, Technology and Innovation : Proceedings of the Symposium of Professional Nursing Education and Training (Pro-Nursing Project) : 24 June 2016, Bonn, Germany* (pp. 51-56). Universi. http://www.pro-nursing.eu/web/download/show

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:09 Nov 2026

Automatic Extraction of Nursing Tasks from Online Job Vacancies

Vladimer Kobayashi[*], Stefan T. Mol[*], Gábor Kismihók[*] and Maria Hesterberg[**]

[*] Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands

[**] UKB Universy Hospital Bonn, Bonn, Germany

Popular ways of collecting job Information for purposes of job analysis are the interview, questionnaires, observation, and participant diary or logs (Dessler, 2004, pp. 114–117). One potential source of job information that has been gaining a lot of research traction lately is the job vacancy(Gallivan, Truex, & Kvasny, 2002; Litecky, Aken, Ahmad, & Nelson, 2010; Smith & Ali, 2014; Sodhi & Son, 2010). A typical job vacancy may contain information related to human requirements, job context, and work activities among others. However, extracting these information types and sorting them to categories from job vacancies are challenging because the vacancies are written in free text and to arrive at a generalizable model one may need to examine thousands of vacancies. Current advances in automated text analysis may be able to offer a helping hand in developing an automatic and accurate procedure for the extraction and classification process.

In this study we explored the use of text mining procedures (Aggarwal & Zhai, 2012) to automatically extract a specific job information type, namely *tasks/activities* from online nursing job vacancies. We developed the method following the approach outlined in (Solka, 2008) consisting of text preprocessing, feature extraction, application of classification algorithm and classification evaluation. The online job vacancies were provided by an online recruitment agency called Monster[1].For each job vacancy, the first step is to remove unnecessary elements such as HTML tags (since the original vacancies are in HTML format). Only the free text of the job description were considered in the succeeding analyses. The output from this step is the job vacancy containing only the relevant text. The text is then fed to a sentence segmentor where individual sentences are extracted. The sentence segmentor is a rule based segmentation algorithm. The rules were logical if-then statements that were constructed to detect sentence boundaries. For example, if there is a line break (or a newline character) between two successive words then separate the two words; one word goes to one sentence and the other to another sentence. Identifying individual sentences in the German language is quite similar in the English language, thus, we borrowed some rules from the English language (e.g. sequence of period, space, and upper case letter signifies the end of one sentence and the beginning of the next). Moreover, we introduced specific rules that are idiosyncratic to job vacancies such as taking into account the enumeration of required tasks.

Each sentence is then represented based on the vector-space model (Salton, Wong, & Yang, 1975). In this representation, words are treated as the features and sentences are represented as vectors in which the elements are the weights of the individual words.

[1] http://www.monsterboard.nl/

The weights can be a simple 1 or 0, where the value is 1 if the word occurs in that sentence and 0 if not. In this study we chose the raw frequency as the weight, i.e. the raw count of words in each sentence. In this approach we treated the sentences as being independent from each other for simplicity. Since a labelled data set is needed for the training, we prepared a training corpus by manually labeling a number of sentences (approximately 2000 sentences). Each sentence is labelled with 1 if the sentence expresses job activity and 0 if not. Before we run classification algorithms we applied dimensionality reduction techniques. The dimensionality reduction served two main purposes, to merge similar words together (e.g. synonyms) and to reduce the number of features. Three techniques were tested, namely, Latent Semantic Analysis, thresholding based on term frequency–inverse document frequency (Tf-Idf), and Random Projection. The technique based on Tf-Idf does not really merge features but can nevertheless eliminate non-relevant features. The best dimensionality reduction was chosen by running Support Vector Machines (SVM) on the reduced feature set from each technique. The reduced feature set that resulted to the highest precision and recall on SVM was the final set of features used in the classification step. We selected precison and recall as performance metrics instead of the standard accuracy since the proportions of sentences in the categories are not balanced. In our training data, only 7% of the sentences are labelled as tasks.

We then run three text classification algorithms, namely, Random Forest, Support Vector Machines, and Naïve Bayes (Duda, Hart, & Stork, 2001) on the training corpus (with the reduced feature set). Instead of considering the prediction of each classifier, we found out that we got a better performance if we combine the predictions of the three classifiers. We combined their prediction through a majority vote. The combination of classification plus the dimensionality reduction constitute the classification model. We evaluated the model using 10 fold cross-validation.

Results from applying the three dimensionality reduction techniques is shown in Figure 2. The bar chart shows the comparison of Precision and Recall among the three techniques trained using Support Vector Machines. The three techniques have comparable Precision but LSA has the lowest recall. This means that LSA failed to detect many task sentences (i.e. many false negatives). This is to be expected since LSA generally does not work well with short sentences. Consequently, the choice was focused between Random Projection and Tf-Idf thresholding. Both have comparable performance in terms of the Precision and Recall but we preferred Random Projection due to the resulting reduction of the number of features. Using Tf-Idf thresholding we retained around 10,000 of the original features whereas Random Projection gave only 500 features. Note that we can vary the number of features for the Random Projection as this is usually set by the researcher. In our case, 500 features were enough to generate an acceptable performance. Thus, the final number of features was 500 which was an almost 13,000 reduction from the original. We then run the three classification models on the reduced data set and obtained the results as presented Figure 3. Naïve Bayes has the least performance. As can be seen from the figure, the voting approach that combined the predictions of the other three classifiers has yield the highest recall while maintaining precision. From the

results, the best model is the one that uses Random Projection for dimensionality reduction and instead of using just the individual prediction of the standard classifiers we combine them in a voting approach.

We then applied the classification on unlabelled sentences (numbering to 18,000). There were 2000 new tasks added after applying the model on unlabelled sentences. The prediction of the model was validated by having an expert check the correctness of the labels. Based on the input of the expert we reran the model and obtained new parameters. We ran several iterations of training and expert validation until no further improvement was been obtained. The final precision was 80%. This implies that 80% of the sentences identified as tasks by the classifier are true tasks as validated by an expert.

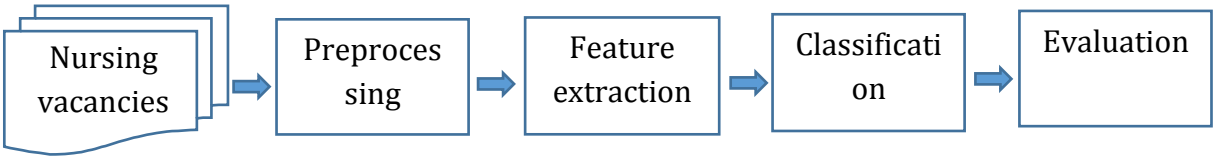

Figure 1. Text classification process for the extraction of nursing tasks from nursing job vacancies.

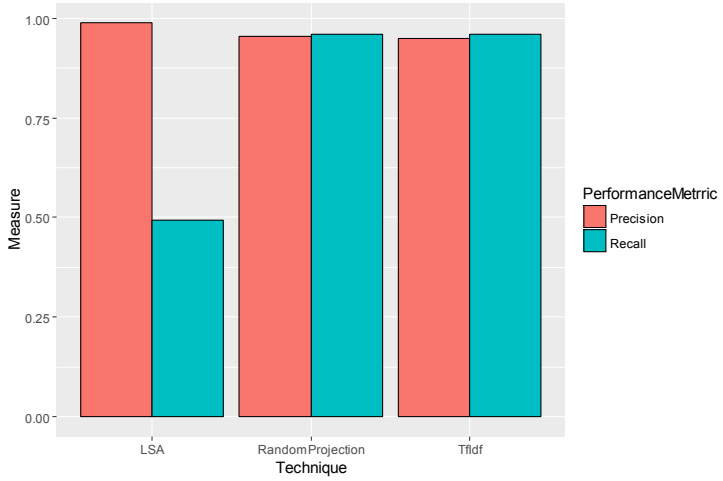

Figure 2. Comparison of different dimensionality reduction techniques on two performance metrics, namely, precision and recall. The performance metrics were derived from running SVM on the derived feature set from each technique.

The final step was to cluster (Jain, Murty, & Flynn, 1999) the extracted task sentences in order to get a more compact list of nursing job tasks and also to merge duplicated tasks. The clusters were obtained using topic modeling. We run Latent Dirichlet Allocation model (Blei, Ng, & Jordan, 2003) and constructed around 100 topics and each sentence was assigned to each topic. Here the topics were considered as the clusters. The constructed clusters were subsequently examined by an expert to further investigate whether it is still possible to merge clusters or break-up some clusters. We want to ensure that the clusters are as homogenous as possible. Moreover, the expert assigned cluster names which represent the general type of task in each cluster. Examples of labels include Basic Care, Medical Care, Internal Management, and Teamwork. The task clusters will be further validated by including them in a survey which aims to compare

nursing tasks obtained from traditional job analysis and tasks from text mining. As of writing the survey is still on-going.

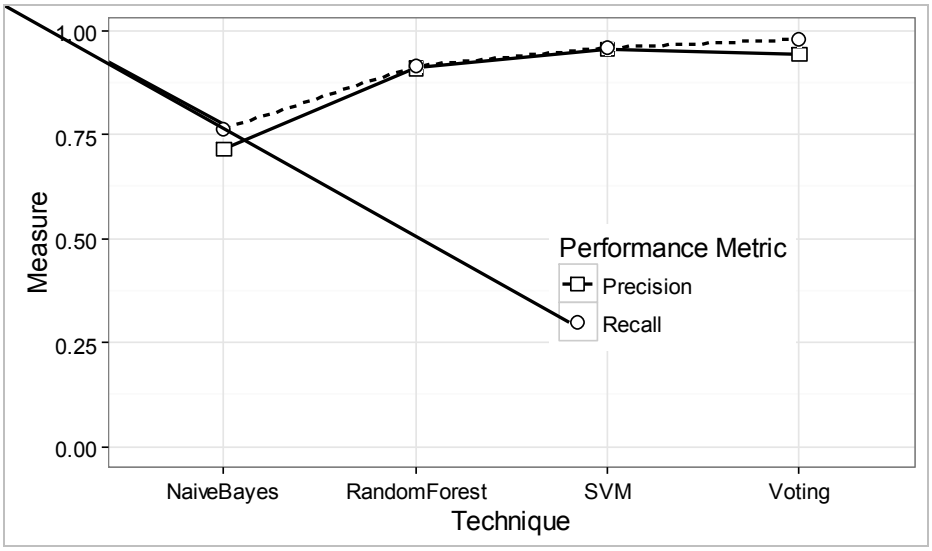

Figure 3. Comparison of four classifiers on the nursing task extraction. The voting approach combined the predictions of the other three classifiers using majority voting.

Based on the study, we find that applying text mining with expert validation on nursing job vacancies can offer an alternative, fast and efficient way of extracting job information. Also, the classification model is straightforward to apply in practice. The results are useful for a number of purposes such as job analysis for the nursing profession and for designing recruitment strategies. One lesson from this research is the importance of employing experts (e.g. job experts and job holders) in the model tuning loop since as we have found out, the precision from the training could be high but when applied to unseen sentences there was a sizable reduction in the precision. During training we got a precision of 90% but when applied to unseen sentences we got only 70%. Expert contributed about 10% improvement in the performance of the classification model.

We identified a number of limitations for our current approach. First we assumed that sentences were independent from each other. This is definitely not the case since at least sentences that come from the same vacancy must be somewhat related. Moreover, we found out that in many vacancies, task sentences are usually written close to each other. Another limitation is in the choice of features, we only considered word-based features though in our investigation we observed that structural and grammatical based features have the potential to substantially enhance the classification performance. Example is, task sentences are usually longer compared to non-task sentences and task sentences usually have more verbs than non-task sentences. Aside from the technical concerns, there is also the challenge of validating the extracted tasks. As of the moment we are dealing only with content validity by having subject matter experts assess the output of the classification, however, it is also desirable to address predictive validity. In the future we will deal with questions such as: How can these information about nursing tasks could better prepare aspiring nurses for this profession? These and other limitations gave us insight on how we can improve the models in the next iteration.

We plan to continue improving the classifier by relaxing the independence assumption for the sentences and exploring algorithms that incorporate order of sentences as they are written in the vacancies (e.g. HMM, CRF). We are convinced that the order by which sentences are written in the vacancies definitely play a role in what type of job information is expressed in each sentence. We also plan to explore other strategies for combining the prediction of individual classifiers and try approaches in semi-supervised learning for dealing with the limited number of labelled data.

References

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In Mining text data (pp. 163–222). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4614-3223-4_6

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3, 993–1022.

Dessler, G. (2004). Human Resource Management (10 edition). Upper Saddle River, N.J: Prentice Hall.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. 2nd. Edition. New York.

Gallivan, M., Truex, D. P., III, & Kvasny, L. (2002). An Analysis of the Changing Demand Patterns for Information Technology Professionals. In Proceedings of the 2002 ACM SIGCPR Conference on Computer Personnel Research (pp. 1–13). New York, NY, USA: ACM. http://doi.org/10.1145/512360.512363.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM Computing Surveys (CSUR), 31(3), 264–323.

Litecky, C., Aken, A., Ahmad, A., & Nelson, H. J. (2010). Mining for Computing Jobs. IEEE Software, 27(1), 78–85. http://doi.org/10.1109/MS.2009.150.

Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. Commun. ACM, 18(11), 613–620. http://doi.org/10.1145/361219.361220.

Smith, D., & Ali, A. (2014). Analyzing Computer Programming Job Trend Using Web Data Mining. Issues in Informing Science and Information Technology, 11. Retrieved from http://iisit.org/Vol11/IISITv11p203-214Smith0494.pdf

Sodhi, M. S., & Son, B.-G. (2010). Content analysis of OR job advertisements to infer required skills. Journal of the Operational Research Society, 61(9), 1315–1327. http://doi.org/10.1057/jors.2009.80.

Solka, J. L. (2008). Text Data Mining: Theory and Methods. Statistics Surveys, 2(0), 94–112. http://doi.org/10.1214/07-SS016.

More from these Authors

Khobreh, M., Ansari, F., Fathi, M., Vas, R., Mol, S., Berkers, H., & Varga, K. (2016). An Ontology-based Approach for the Semantic Representation of Job Knowledge. IEEE Transactions on Emerging Topics in Computing, 4 (3), pp. 462-473

Kismihók, G., Vas, R., & Mol, S. T. (2012). An innovative ontology-driven system supporting personnel selection: the OntoHR case. International Journal of Knowledge and Learning, 8 (1–2), 41–61. http://doi.org/10.1504/IJKL.2012.047549

Kobayashi, V., Maret, P., Muhlenbach, F., & Lhérisson, P.-R. (2013). Integration and Evolution of Data Mining Models in Ubiquitous Health Telemonitoring Systems. In I. Stojmenovic, Z. Cheng, & S. Guo (Eds.), Mobile and Ubiquitous Systems: Computing, Networking, and Services, Springer International Publishing, pp. 705–709. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-11569-6_