Challenge: Processing Web Texts for Classifying Job Offers

Flora Amato*, Roberto Boselli†‡, Mirko Cesarini†‡, Fabio Mercorio†‡
Mario Mezzanzanica†‡ Vincenzo Moscato* Fabio Persia* and Antonio Picariello*
*Department of Computer Science and Systems, University of Naples Federico II, Italy
†Dept. of Statistics and Quantitative Methods, Univerisity of Milan-Bicocca, Italy
‡CRISP Research Centre, Univerisity of Milan-Bicocca, Italy

***Abstract*—Today the Web represents a rich source of labour market data for both public and private operators, as a growing number of job offers are advertised through Web portals and services. In this paper we apply and compare several techniques, namely *explicit-rules*, *machine learning*, and *LDA-based* algorithms to classify a real dataset of Web job offers collected from 12 heterogeneous sources against a standard classification system of occupations.**

I. Introduction and Contribution

The diffusion of Web-centric services for labour market is growing apace, and this enables a significant part of the labour demand to be routed through Web portals, services, and applications. As a consequence, the Web is increasingly used by both employers and job seekers to advertise demand and supply, and to enhance the recruitment activities in new ways (e.g., *social recruitment* and *e-recruitment*, see e.g., [1], [2], [3]). On the other hand, organizations and governments have made a great effort in defining national and international skills/occupations classifiers, that aim to classify and standardize labour *occupations*, *skills* and *competences* over several countries and languages.

Indeed, reconciling Web job offers over an international standard occupation classifier - rather than a proprietary one - will provide labour market analysts and policy makers a sort of *lingua franca* useful for studying the labour market dynamics over several countries, overcoming the linguistic boundaries. In such a context, classifying a Web job offer against a classifier is not a trivial task. Indeed, information in Web vacancies is expressed using raw text (i.e., unstructured or semistructured data), the lexicon used therein can be very different from the one used in the reference classifiers.

The contribution of this work goes towards this direction by classifying Web job offers onto the categories of a well-established classifier. To this end, we propose to apply several (and different) text classification techniques to classify a real dataset of Web job offers onto to the ISTAT classifier, namely CP2011[1]. Furthermore, the effectiveness of each approach is evaluated by comparing classification results against a gold classification manually performed by domain experts at CRISP Research Centre [2], that collect Web job offers to study labour market dynamics[3].

The paper is organised as follows: Sec. II provides a literature review while Sec. III introduces the techniques and tools we used, namely explicit-rules, machine learning and LDA. Then, the experimental results are provided in Sec. IV while the concluding remarks are outlined in Sec. V.

II. Related Work

In the last decade, the huge availability of text information has led to a strong interest in automatic text extraction and processing technology for extracting task-relevant information [4], [5], [6]. In such a scenario, Information Extraction (IE) techniques automatically extract structured information from unstructured and/or semi-structured documents, exploiting different kinds of text analysis, mostly related to Natural Language Processing (NLP) methodologies and to cross-disciplinary perspectives including Statistical and Computational Linguistics, see, e.g., [7], [8], [9]. Moreover, IE techniques may be associated with text mining [10] and semantic technologies activities in order to detect relevant concepts from text data, for information indexing, classification and retrieval aims, as well as long term preservation issues [11], [12].

In the labour market domain the extraction of meaningful information from unstructured texts has been mainly devoted to support the e-recruitment process (see, e.g.,[1]) attempting to automate the *resume* management for matching candidate profiles with job descriptions. To give a few examples, the work [13] proposes a system aiming to screen candidate profiles for jobs, by extracting various pieces of information from the unstructured resumes through the use of probabilistic information extraction techniques as Conditional Random Fields. Similarly, the ”mandatory communications” system realised by the Italian Ministry of Labour and Welfare [14] has been used for studying the Italian Labour Market dynamics performing both data quality [15], [16] and knowledge discovery activities [17]. Differently, in [18] a cascaded Information Extraction model based on SVM for mining resumes is used whilst [19] uses structural relevance models to identify job

[1] The Italian standard classifier for Occupations is based on the logic of the ISCO (International Standard Classification of Occupations), and it is totally cross-linkable with the latter.

[2] CRISP, Inter-university Research Centre on Public Services, Italy www.crisp-org.it

[3] Notice that the ISTAT's classifier is built to be cross-linkable with some other international classifiers, such as ISCO and ESCO

IEEE ICSC 2015, February 7-9, 2015, Anaheim, California, USA
978-1-4799-7935-6/15/$31.00 ©2015 IEEE

descriptions and resumes vocabulary. In [20] the authors develop a job recommender system to dynamically update the job applicant profiles by analysing their historical information and behaviours. Poch et. al [21] aim to match the appropriate candidates to a job offer, and for this task they use supervised classifiers to suggest a ranked list of job offers to job seekers. Among the methods, they use Latent Dirichlet Allocation (LDA here on) to cluster similar job offers. Job clustering is accomplished by authors in terms of candidate classification for one or more classes of jobs according to a model learnt from the matching information and not to a well-established classifier.

Although all these approaches are relevant and effective, they differ from our approach as we aim at processing *job-offers* rather than resumes, and this requires to deal with shorter texts that present a high degree of heterogeneity. Here, the joint use of *different* techniques would be beneficial in evaluating the effectiveness of these approaches in our application domain.

III. Applied Techniques

A. Text Preprocessing and Job Classification Systems

A preprocessing of the texts was performed according to the following pipeline: tokenization, lower case reduction, html special characters substitution, stop words removal, misleading words elimination, and numbers elimination. Then, word stemming was performed using the Italian stemmer provided by the NLTK framework version 3.0a3 [22]. The text preprocessing pipeline was used before applying any of the approaches described in the next subsection, unless otherwise specified.

A dataset of about 40,000 job vacancies were downloaded from 12 Web sources, and then, a subset of 412 job offers was randomly selected to get a sample representative of the 40,000 vacancies. The 412 sampled job vacancies were manually labelled by domain experts at CRISP Research Centre using the qualification codes outlined in the CP2011 classifier. The domain experts considered both offer titles and full descriptions to assign labels. The domain expert classification results were used as a gold reference to evaluate the classification technique output.

B. Explicit Rule based approach

A commercial tool has been used to classify texts using a rule based approach. The user can define rules for classifying texts focusing on taxonomies modelling the terminology of a certain domain. The rules building process is driven by a twofold goal: to identify the relevant terms used in vacancy web ads, and to relate them to the occupation codes used in the CP2011 classifier.

C. Machine Learning approach

The technique described in this subsection focused on the (sampled) 412 job offers set. The Job offer titles and the qualification codes given by the domain experts were used to train and evaluate a (machine learning based) classifier according to the process described below.

The offer titles were preprocessed using the pipeline described in Sec. III-A. Then, each title is turned into a vector of word occurrences according to a bag of words approach. The whole set of job offer titles was turned into a matrix of 412×542 elements whereas each line represents a job offer title, and the columns represents the features (i.e., the word count of the different stemmed words).

Then, two machine learning classifiers were used to perform the text classification purposes: the LinearSVC (an implementation of Support Vector Machine Classification using a linear kernel) and the Perceptron classifier, both built using the Scikit-learn framework [23]. A grid search of the classifier parameters maximizing the classification accuracy was performed on both classifiers.

D. LDA based approach

The developed technique only leverages titles of Web job offers as document set and is based on two different steps: *feature extraction* and *classification*.

The goal of the feature extraction task is to derive for each ISTAT job category a particular data structure, named *Weighted Word Pairs* (WWP) and containing the most relevant pairs of *lexical items* (i.e. single word, a part of a word or a chain of words) together with the probabilistic dependencies characterizing titles of job vacancies [24], [25], [26]..

On the other hand, the classification procedure is based on the matching between terms derived from a job vacancy title and the set of pairs related to different ISTAT categories, according to a distance-based approach.

In a first stage, all possible pairs of relevant terms (without repetitions) from texts related to the titles of job vacancies are extracted exploiting a classical *Natural Language Processing* (NLP) pipeline. Such pairs are then compared with the WWPs of all job categories using as similarity the *Levenshtein* metric.

For each category, the number of *hits* (positive matchings) is determined and each single hit is weighted by the dependency probability of the related pairs in the WWP structure. The category having the highest score is finally selected as *winner* and chosen for the classification.

IV. Experimentation

The 412 sampled job offers were classified against 62 qualification codes, unfortunately 42 of them occurs less than 5 time in the sample. The Tab. IV shows the precision, recall, f-score, and accuracy for all the techniques [4].

The machine learning approach reaches good performances when an ample number of training samples are available: a gap exists between precision and recall computed considering all the occupation codes and excluding the infrequent ones. The *rule-based* approach has a similar gap (although smaller): the rationale is that domain experts paid more attention to some specific occupation codes (by the way, the more frequent ones) during the rules developing process. Furthermore, the cases matching no rule generate a certain number of

[4] Within parenthesis we show the same values computed by excluding the vacancies corresponding to infrequent occupation codes.

	LDA	**Rules**	**Linear SVC**	**Percept.**
Accuracy	.5 (.51)	.444 (.469)	**.556 (.633)**	.483 (.543)
Avg. Precision	**.507 (.587)**	.353 (.432)	.259 (.576)	.284 (.503)
Avg. Recall	**.502** (.538)	.354 (.432)	.272 (**.576**)	.254 (.503)
Avg. FScore	**.471** (.532)	.328 (.462)	.26 (**.607**)	.263 (.564)
Precision Std.Dev.	.41 (.305)	.378 (.328)	**.34 (.239)**	.364 (.243)
Recall Std. Dev.	.391 (.274)	.348 (.224)	.358 (.224)	**.332 (.195)**
FScore Std. Dev.	.366 (.253)	**.326** (.257)	.337 (.216)	.336 (**.201**)

TABLE I
TEXT CLASSIFICATION TECHNIQUES SCORES.

miss-classifications which also negatively affect the results. The LDA approach has satisfying results respect to all the categories (we have an overall precision of 50%), and - as expected - such results are less or not affected by the infrequent occupation codes at all. Here, the problems of incorrect classifications are related to the presence of categories having a similar lexicon and a different number of training samples.

To explore and better understand the determinants of the classification performance values, we asked researchers at CRISP Research Centre to compare the information content of titles with respect to the vacancy full descriptions. They identified that about 30% of the offer titles do not carry enough information to identify the occupation, indeed accessing the vacancy full description is required. Furthermore, in Fig.1 we compared the occupation codes distribution of the Gold Classification Benchmark (solid boxes) against the distribution of the Linear SVC classifier (dashed boxes) that used the *complete* job offers dataset as input, as shown in the lower graph of Fig. 1. Here, the 62 occupation codes have been enumerated on the x-axis in a non-increasing order with respect to how often they appear in the gold classification.

In addition, we exploited a well-suited multidimensional visualisation technique to investigate the classification results, namely the *parallel-coordinates* (abbrv: $\|$-coord see [27]). Informally speaking, $\|$-coord allow one to represent an n-dimensional datum $(x_1, \ldots, x_n)$ as a polyline, by connecting each x_i point in n parallel y-axes. Generally, $\|$-coord tools show their powerfulness when used interactively (i.e., by selecting ranges from the y-axes, by emphasising the lines traversing through specific ranges, etc). For these reasons, the online demo has been made publicly available[5].

Specifically, Fig. 2 shows classification results focusing on occupation code 2.x.x (Intellectual, scientific and highly specialized occupations). The classification techniques scored good results, correctly classifying several advertisements belonging to the 2.x.x class. The misclassified vacancies are mostly placed under the 3.x.x class (Technical professions) by all of the techniques, whilst LinearSVC pushes some of them also under the 6.x.x class (Artisans workers etc.). Considering less skilled workers, Fig. 3 shows classification results focusing on occupation code 4.x.x (clerical workers). The three classification methods scored very good results in this class with very few misclassifications.

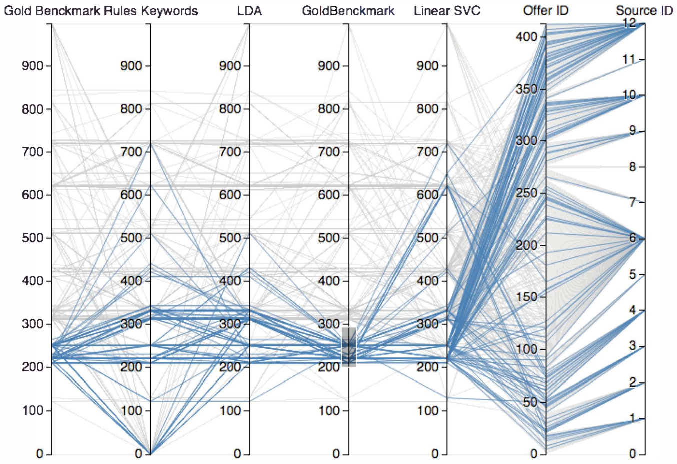

Fig. 2. Parallel Coordinates focusing on CP2011 2.x.x occupation codes.

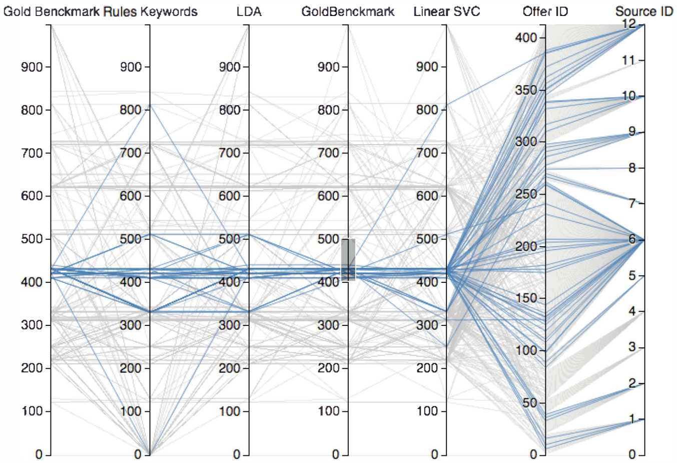

Fig. 3. Parallel Coordinates focusing on CP2011 4.x.x occupation codes.

V. CONCLUSIONS AND FUTURE WORK

In this paper we have compared several techniques to classify vacancies advertised through 12 relevant heterogeneous Web sources: in particular, we have classified Web job vacancies according to a standard and well-established occupations' classification and some preliminary results have been obtained and discussed. Together with the CRISP Research Centre, we are planning the following future research steps.

First, we are trying to include the full descriptions of job vacancies and to tackle the increased complexity of longer texts through computational linguistic approaches.

Second, we are going to classification techniques for extracting other relevant information (in addition to occupation codes) from the vacancy texts, (e.g., the required skills, contract types, business sectors, education levels, etc): such kind of information may be very relevant for job seekers, policy makers, and human resources professionals.

Finally, as vacancies could frequently be related to several occupation codes, we are exploring alternative evaluation

[5] http://goo.gl/6qC5Vj

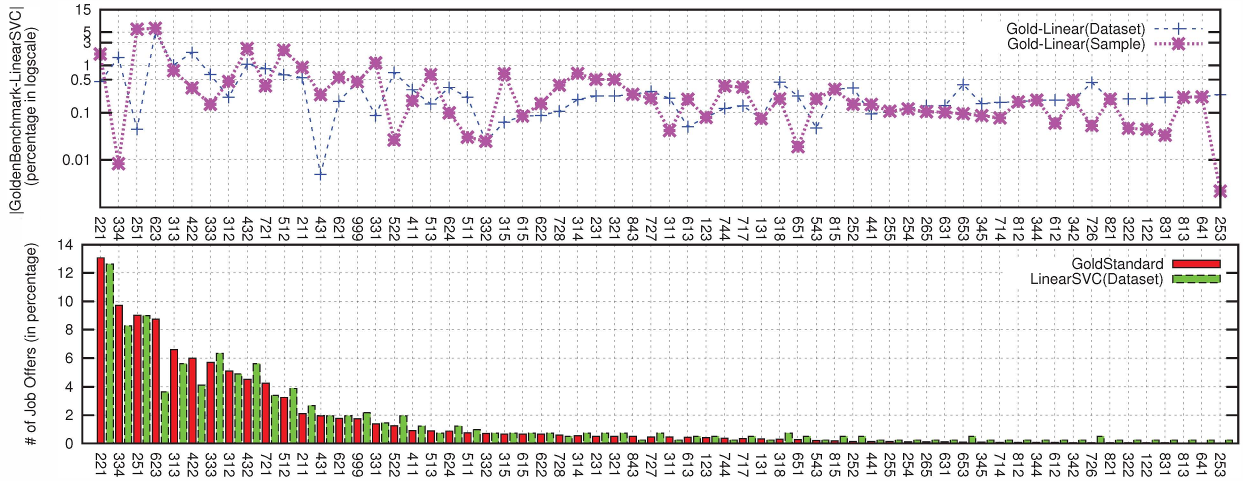

Fig. 1. Distribution of Job Offers over ISTATs' codes

methodologies to deal with multi-classification, e.g. the top-K evaluation: the identification of metrics to measure the distance between different classification outputs would surely be beneficial.

References

[1] I. Lee, "Modeling the benefit of e-recruiting process integration," *Decision Support Systems*, vol. 51, no. 1, pp. 230–239, 2011.

[2] R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, "How the social media contributes to the recruitment process," in *Proceedings of European Conference on Social Media (ECSM 2014)*, 2014, pp. 10–11.

[3] S. Lang, S. Laumer, C. Maier, and A. Eckhardt, "Drivers, challenges and consequences of e-recruiting: a literature review," in *Proceedings of the 49th SIGMIS annual conference on Computer personnel research*. ACM, 2011, pp. 26–35.

[4] R. Grishman, "Information extraction: Capabilities and challenges," 2012.

[5] D. E. Appelt, "Introduction to information extraction," *AI Commun.*, vol. 12, no. 3, pp. 161–172, 1999.

[6] M. F. Moens, *Information Extraction: Algorithms and Prospects in a Retrieval Context*, ser. The Springer international series on information retrieval. Springer, 2006.

[7] C. S. Butler, *Statistics in Linguistics*. Blackwell, Oxford, 1985.

[8] D. Biber, R. Reppen, and S. Conrad, *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, 1998.

[9] G. D. Kennedy, *An introduction to corpus linguistics*. Longman, 1998.

[10] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer, 2012.

[11] F. Amato, A. Mazzeo, A. Penta, and A. Picariello, "Building rdf ontologies from semi-structured legal documents," in *Complex, Intelligent and Software Intensive Systems. CISIS.*, 2008.

[12] F. Amato, A. Mazzeo, V. Moscato, and A. Picariello, "A system for semantic retrieval and long-term preservation of multimedia documents in the e-government domain," *International Journal of Web and Grid Services*, vol. 5, no. 4, pp. 323–338, 2009.

[13] A. Singh, C. Rose, K. Visweswariah, V. Chenthamarakshan, and N. Kambhatla, "Prospect: a system for screening candidates for recruitment," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 659–668.

[14] The Italian Ministry of Labour and Welfare, "Annual report about the CO system, available at http://goo.gl/XdALYd," pp. 1–9, 2012.

[15] R. Boselli, M. Mezzanzanica, M. Cesarini, and F. Mercorio, "Planning meets data cleansing," in *The 24th International Conference on Automated Planning and Scheduling (ICAPS 2014)*. AAAI, 2014, pp. 439–443.

[16] R. Boselli, M. Mezzanzanica, M. Cesarini, and F. Mercorio, "A policy-based cleansing and integration framework for labour and helthcare data," in *Knowledge Discovery and Data Mining, LNCS 8401*. Springer, 2014, pp. 141–168.

[17] M. Mezzanzanica, R. Boselli, M. Cesarini, and F. Mercorio, "A model-based evaluation of data quality activities in KDD," *Information Processing and Management (in press, available at http://dx.doi.org/10.1016/j.ipm.2014.07.007)*, 2014.

[18] K. Yu, G. Guan, and M. Zhou, "Resume information extraction with cascaded hybrid model," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 499–506.

[19] X. Yi, J. Allan, and W. B. Croft, "Matching resumes and jobs based on relevance models," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 809–810.

[20] W. Hong, S. Zheng, and H. Wang, "Dynamic user profile-based job recommender system," in *Computer Science & Education (ICCSE), 2013 8th International Conference on*. IEEE, 2013, pp. 1499–1503.

[21] M. Poch, N. Bel, S. Espeja, and F. Navıo, "Ranking job offers for candidates: learning hidden knowledge from big data," in *Language Resources and Evaluation Conference*, 2014.

[22] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[25] F. Colace, M. De Santo, L. Greco, and P. Napoletano, "Improving text retrieval accuracy by using a minimal relevance feedback," in *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, ser. Communications in Computer and Information Science, A. Fred, J. Dietz, K. Liu, and J. Filipe, Eds. Springer Berlin Heidelberg, 2013, vol. 348, pp. 126–140. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-37186-8_8

[26] F. Colace, M. De Santo, L. Greco, and P. Napoletano, "Text classification using a few labeled examples," *Computers in Human Behavior*, vol. 30, no. 0, pp. 689 – 697, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563213002823

[27] A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1, no. 2, pp. 69–91, 1985.