E-Gen: Automatic Job Offer Processing System for Human Resources

Rémy Kessler[α,∇], Juan Manuel Torres-Moreno[∂,∇,⋆⋆], Marc El-Bèze[∇]

[∇] Laboratoire Informatique d'Avignon, BP 1228 F-84911 Avignon Cedex 9 FRANCE
{remy.kessler,juan-manuel.torres,marc.elbeze}@univ-avignon.fr
http://www.lia.univ-avignon.fr
[∂] École Polytechnique de Montréal - Département de génie informatique
CP 6079 Succ. Centre Ville H3C 3A7, Montréal (Québec), CANADA.
[α] AKTOR Interactive Parc Technologique 12, allée Irène Joliot Curie Bâtiment B3
69800 Saint Priest FRANCE

Abstract. The exponential growth of the Internet has allowed the development of a market of on-line job search sites. This paper aims at presenting the E-Gen system (Automatic Job Offer Processing system for Human Resources). E-Gen will implement two complex tasks: an analysis and categorisation of job postings, which are unstructured text documents (e-mails of job listings possibly with an attached document), an analysis and a relevance ranking of the candidate answers (cover letter and *curriculum vitae*). This paper aims to present a strategy to resolve the first task: after a process of filtering and lemmatisation, we use vectorial representation before generating a classification with Support Vector Machines. This first classification is afterwards transmitted to a "corrective" post-process which improves the quality of the solution.

1 Introduction

The exponential growth of the Internet has allowed the developement of an on-line job-search sites market [1–3]. The mass of information obtained through candidate response represents a lot of information that is difficult for companies to manage [4–6]. It is therefore indispensable to process this information by an automatic or assisted way. The *Laboratoire Informatique d'Avignon* (LIA) and Aktor Interactive have developed the E-Gen system in order to resolve this problem. It will be composed of two main modules:

1. A module to extract information from a corpora of e-mails containing job descriptions.
2. A module to analyse and compute a relevance ranking of the candidate answers (cover letter and *curriculum vitae*).

In order to extract useful information, the system analyses the contents of the e-mails containing job descriptions. In this step, there are many difficulties and

⋆⋆ Corresponding author.

interesting problems to resolve related to Natural Language Processing (NLP), for example, that job postings are written in free-format, strongly unstructured, with some ambiguities, typographic errors, etc. Similar work has been carried out in the recrutment domain [1, 7], but this concerns only the handling of responses and not integration of job offers.

Aktor Interactive[1] is a French communication agency, specialised in e-recruiting. Aktor's key service is the publication of job adverts on different online job boards on behalf of their clients. Therefore a system that is able to automate this is desirable due to the high number and spread of specialised[2], non-specialised[3] or still local sites[4]. To do this, Aktor uses an automatic system to send job offers in XML format (Robopost Gateway) defined with job boards. Therefore, the first step of the workflow is to identify every part of the job posting and extract the relevant information (contract, salary, localization etc.) from the received job offer. Figure 1 shows an overview of the workflow. Until now, the first step of the workflow was a laborious manual task: with users having to copy and paste job offers in the Aktor Information system.

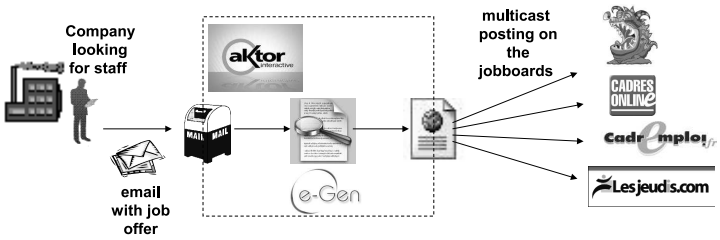

Fig. 1. Aktor's workflow.

We present in this paper only the extraction system i.e. the first module of E-Gen, and its performance on this extraction and categorisation task. Section 2 shows a general system overview. In Section 3, we present the textual corpora and a short description of Aktor Interactive. In Section 4, we describe the algorithms used in the information extraction module. In Section 5, we show some results of our system before concluding and indicating future work.

[1] http://www.aktor.fr
[2] http://www.admincompta.fr (Bookkeeper), http://www.lesjeudis.com (computing jobs), etc.
[3] http://www.monster.com, http://www.cadremploi.fr, http://www.cadresonline.com
[4] http://www.emploiregions.com or http://www.regionsjob.com

2 System overview

The main activity of Aktor Interactive is the processing of job offers on the Internet. As the Internet proposes new means for the recruitment market, Aktor is modifying its procedures to become able to integrate systems which carry out this processing as fast and judiciously as possible. An e-mail-box receives messages (sometimes with an attached file) containing the offer. After identification of the language, E-Gen parses the e-mail and examines attached file. Then the text containing the offer is extracted from the attachment. An external module, wvWare[5] processes MS-Word documents and produces a text document version as an output file, splitting this text into segments[6]. After filtering and lemmatisation, we are able to use a vectorial representation for each segment in order to assign a correct label to each segment using Support Vector Machines (SVM). This label sequence is processed by a corrective process which validates it or proposes a better sequence. At the end of the processing, an XML file is generated and sent to the Aktor Information system. The whole processing chain of E-Gen system is represented in figure 2.

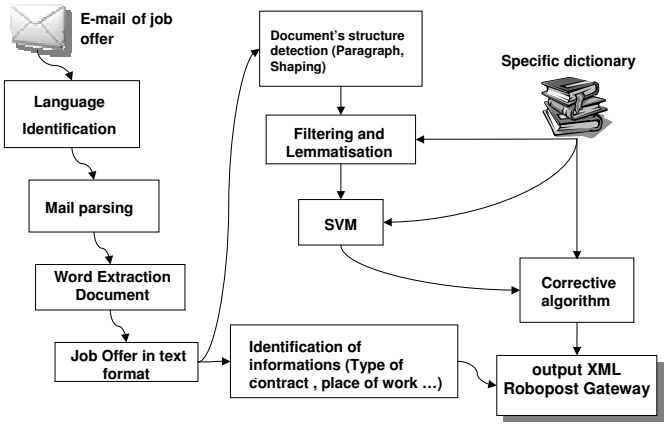

Fig. 2. E-Gen system overview.

Information extraction To post on-line a job vacancy, job boards require certain information. During the publication of job posting , some informations are required by the job board. So we need to find these fields in the job posting

[5] http://wvware.sourceforge.net

[6] In most situations, wvWare division matches the paragraph of the MSWord's document. Segmentation of text is a real issue, so we choose to use an existing tool.

in order to include them in the XML document. We set up different solutions in order to locate each type of information:

- Salary: Regular expressions and rules were created to locate expressions such as "Salary: from X to Y", "Salary: between X and Y" or "X fixed salary with bonus", etc.
- Place of work: A table including area, city and department fields was created to find the location listing in a job posting. Most of the job boards categorise job postings according to the region to help job seekers in their job search.
- Company: In order to be able to integrate logos in job offers, a list of customers was plugged into the system to detect the company's name in the job postings.

Other information is recovered by similar processes (contract, reference, duration of mission, etc.). Finally, a report is sent to the user in order to show the fields correctly detected and the fields not found (either by extraction error or missing information in the job offer).

3 Corpora and modelisation

3.1 Corpora description

We have selected a data subset from Aktor's database. This corpora is a mixture of different job listings in several languages. Initially we concentrated our study on French job posting because the French market represents the Aktor's main activity. This subset has been called the *Reference Corpora*. A job listing example is presented in table 1, translated from French to English (the content of the job offer is free, as we can see in this french example[7], but it stays conventional as we find a rather similar presentation in every job listing and vocabulary according to every part as we shall see later on). The extraction of Aktor database made it possible to have an important corpora, without manual categorisation. A first analysis of this corpora shows that job offers often consist of similar blocks of information that remain, however, strongly unstructured. Each job posting is separated in four classes, as follow:

[7] *Ce groupe français spécialisé dans la prestation d'analyses chimiques, recherche un: RESPONSABLE DE TRANSFERT LABORATOIRE. Sud Est. En charge du regroupement et du transfert d'activités de différents laboratoires d'analyses, vous etudiez, conduisez et mettez en oeuvre le sequencement de toutes les phases necessaires à la realisation de ce projet, dans le respect du budget previsionnel et des delais fixes. Vos solutions integrent l'ensemble des parametres de la demarche (social, logistique, infrastructures et materiels, informatique) et dessinent le fonctionnement du futur ensemble (Production, methodes et accreditations, developpement produit, commercial). De formation superieure Ecole d'ingenieur Chimiste (CPE) option chimie analytique environnementale, vous avez deja conduit un projet de transfert d'activite. La pratique de la langue anglaise est souhaitee. Merci d'adresser votre candidature sous la reference VA 11/06 par e-mail beatrice.lardon@atalan.fr*

This french firm, specialised in chemical analysis, is looking for: PERSON IN CHARGE OF LABORATORY TRANSFER South East You will be in charge of regrouping the transfer activities of different analysis laboratories. You will analyse, conduct and implement the necessary phases of the project, respecting budgets and previously defined, dead lines. Your solution will need to consider different parameters of the project (social, logistic, materials, data processing...) and integrate a roadmap (production, methods, accreditations, development, commercial...). Being a post graduate in chemical engineering with a focus on environmental analytical chemistry, you have already led an activity transfer project. Fluent English required. Please send your CV and cover letter indicating reference number VA 11/06 to beatrice.lardon@atalan.fr

Table 1. Job postings example.

1. "Description_of_the_company": Brief digest of the company that is recruiting.
2. "Title": presumably job title.
3. "Mission": a short job description.
4. "Profile": required skills and knowledge for the position. Contacts are generally included in this part.

Table 2 shows a few statistics about our Reference Corpora.

Number of job postings	D=1000	
Number total of Segments	P=15621	
Number of Segments "Title"	1000	6.34%
Number of Segments "Description_of_the_company"	3966	25.38%
Number of Segments "Mission" description	4401	28.17%
Number of Segment "Profile" description	6263	40.09%

Table 2. Corpora statistics.

A pre-processing task of the corpora was perfomed to obtain a suitable representation in the Vector Space Model (VSM). Mainly deletion of the followings items : verbs and functional words (to be, to have, to be able to, to need,...), common expressions (for example, that is, each of,...), numbers (in numeric and/or textual format) and symbols such as $,#, *, etc. because these terms may introduce noise in the segment classification. Lemmatisation processing has also been performed to obtain an important reduction of the lexicon. It consists of finding the root of verbs and transform plural and/or feminine words to masculine singular form[8]. This process allows to decrease the curse of dimensionality [8] which

[8] So we can transform terms *sing, sang, sung, will sing* and possibly *singer* into sing.

raises severe problems of representing of the huge dimensions [9]. Other reduction mechanisms of the lexicon are also used: compound words are identified by a dictionary, then transformed into a unique term. All these processes allow us to obtain a representation in bag-of-words (a matrix of frequencies/absences of segment texts (rows) and a vocabulary of terms (columns)).

3.2 Markov's Machine

Preliminary experiments show that segment categorisation without segment positioning of a job posting is not enough and may be a source of errors. Figure 4 shows that SVM produces a good classification of segments globally, but the job postings (documents) are rarely classified completely. Therefore due to the huge number of cases, the rules don't seem to be the best way to solve the problem. So we have implemented a machine with 6 states ("Start" (0), "Title" (1), "Description_of_the_company" (2), "Mission" (3), "Profile" (4) and "End" (5)). Thus, we have considered each job posting as a succession of states in a Markov's machine. *Reference Corpora* has been analysed to determine the probabilities to switch from one state to another (transition). Matrix M (eq. 1) shows the values of the probablities.

$$M = \begin{pmatrix} & \text{Start} & \text{Title} & \text{Description} & \text{Mission} & \text{Profil} & \text{End} \\ \text{Start} & 0 & 0,01 & 0,99 & 0 & 0 & 0 \\ \text{Title} & 0 & 0,05 & 0,02 & 0,94 & 0 & 0 \\ \text{Description} & 0 & 0,35 & 0,64 & 0,01 & 0 & 0 \\ \text{Mission} & 0 & 0 & 0 & 0,76 & 0,24 & 0 \\ \text{Profil} & 0 & 0 & 0 & 0 & 0,82 & 0,18 \\ \text{End} & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (1)$$

Observing this matrix[9] allows us to learn a lot of things about the organisation of segments in a job posting. A job posting has a probability $p = 0.99$ to start with a segment "Description" (2) while it is impossible that a job posting starts with a "Mission" or "Profile" segment (null probability). In the same way, each "Mission" segment can only be followed by another "Mission" or "Profil" segment. This matrix allows to build a Markov's machine shows in the figure 3.

4 Segment categorisation algorithms

4.1 Support Vector Machine classification

SVM machines, proposed by Vapnik [10] have been successfully used in several tasks of machine learning. In particular, they offer a good estimation of the minimisation principle of the structural risk. The main ideas behind this method are:

[9] "Description" label corresponding in the matrix to "Description_of_the_company".

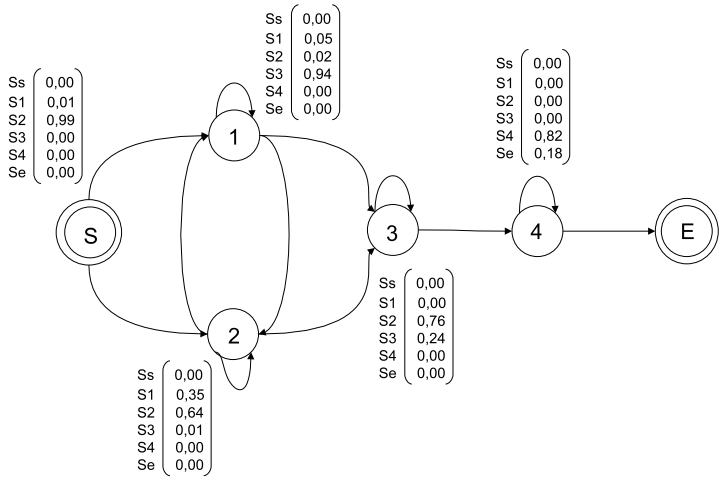

Fig. 3. Markov's machine used to correct wrong labels.

- Data is mapped in a high dimensional space through a transformation based on a linear, polynomial or gaussien kernel.
- Classes are separated (in the new space) by linear classifiers, which maximise the margin (distance between the classes).
- Hyperplans can be determined by a few number of points: each of them is called a support vector.

Thus, the complexity of a classifier SVM depends, not on the dimension of the data space, but on the number of support vectors required to realise the best separation [11]. SVM has been already applied in the domain of the classification of the text in several works [12]. We have chosen to use SVM in this type of particular corpora as we have already seen good results in similar previous work [13]. We have used the implementation Libsvm [14] that allows to treat the multiclass problems in big dimensions.

4.2 Corrective process

Our preliminary results obtained by the SVM method show a performant classification of segments. However, during the classification of a complete job posting, some segments were incorrectly classified, without regular behaviour (a "Description of the company" segment was detected in the middle of a paragraph profile; the last segment of the job posting was identified as a "Title", etc.). In order to avoid this kind of error, we applied a post-processing, based on the Viterbi algorithm [9, 15]. The SVM classification for each segment provides a predicted class, and thus for a job posting, we have a class sequence (example: the sequence 0↦2↦2↦1↦3↦3↦4↦5, i.e "Start" ↦"Description of the company" ↦"Description of the company" ↦"Title" ↦"Mission" ↦"Mission" ↦"Profile"

↦End). A classical Viterbi algorithm will compute the probability of sequence. If the sequence is not probable, Viterbi's algorithm returns 0. When the sequence has a null probability our corrective process returns the sequence with a minimum error and maximal probability (compared to the original sequence generated by SVM).

```
Calcul next symbol()
Processes the current sequence (Viterbi): full sequence, their probability, and
the number of errors
if the error of current sequence > max error found then
    return current sequence
end
if symbol is the last of the sequence then
    if current error < max error then
        maxerror = currenterror;
    end
    return sequence;
end
else
    foreach symbol successor of the sequence do
        current sequence = Calcul next symbol()
        if current sequence is the best sequence then
            bestsequence = currentsequence;
            if current error < max error then
                maxerror = currenterror;
            end
        end
    end
end
```

Algorithm 1: Corrective Process algorithm with Branch and Bound method.

First results were interesting but involved a considerable amount of processing time. We have introduced an improvement using Branch and Bound algorithm [16] for pruning the tree: once an initial solution is found, its error and probability are compared each time that a new sequence is processed. If the solution is not improved, the end of the sequence is not computed. The use of this algorithm enables us to reach an optimal solution, but not the best time (it have an exponential complexity). In test, this strategy computes sequences ≤ 50 symbols in approximately 4 seconds.

5 Results and discussion

We have used a corpora of $D = 1,000$ job postings split into $P = 15,621$ segments. Each test was carried out 20 times with a random distribution between the test corpora and the training corpora.

Figure 4 shows the comparison between the results obtained by the Support Vector Machines and the corrective process. The curves present the number of segments unrecognized according to the size of the training corpora. On the left, we present the results of SVM machines alone (dotted line) applied to the segment classification task. The baseline is computed with the most probably label class, i.e. "Profile" (cf. table 2). The results are good and show that even with a small fraction of learning patterns (20% of total), the SVM classifier obtains a low rate of misclassification (less than 10% error). The corrective process (solid line) always gives better results than SVM whatever fraction of patterns in learning.

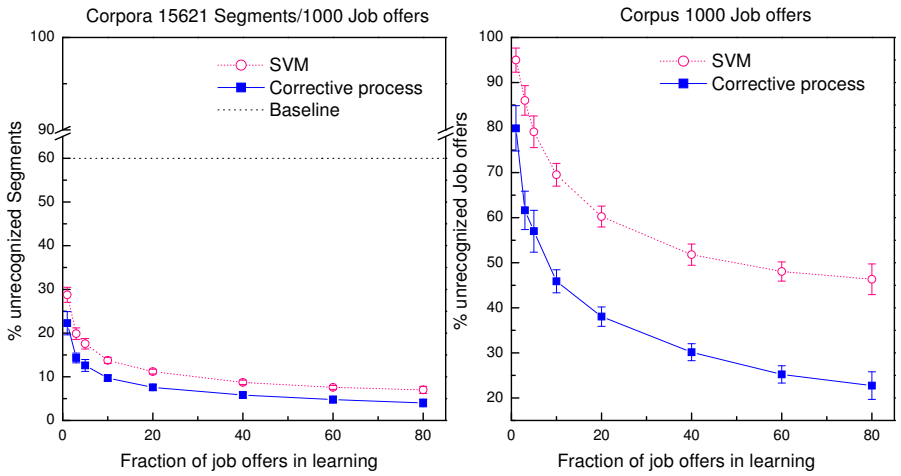

Fig. 4. Results of SVM and corrective process for segments on the left and for jobs offers on the right.

The curve on the right hand of the figure 4, compares the results obtained by each method according to unrecognized job postings. We can also see a considerable improvement of the number of job postings recognized with the corrective process. SVM algorithm reaches a maximum of $\approx 50\%$ of unrecognized job postings while the corrective process gives 20% of unrecognized job postings, so an improvement of more than 50% on the SVM score.

An analysis of wrongly-classified job postings, shows that about 10% of the job postings contains one or two errors. These misclassified segments generally correspond block boundaries between 2 different categories [17], as it is shown in figure 5. So, the obtained sequence, for example 1 (cf. 3.1) is $0\mapsto2\mapsto1\mapsto3\mapsto3\mapsto3\mapsto4\mapsto 4\mapsto5$ but the correct sequence is $0\mapsto2\mapsto1\mapsto3\mapsto3\mapsto4\mapsto4\mapsto4\mapsto5$. The wrong segment is reproduced in table 3.

We observed that some important terms are present in two different categories, leading to a wrong classification. In particular, in this example we are talking

Fig. 5. Block boundary errors.

De formation superieure Ecole d'ingenieur Chimiste (CPE) option chimie analytique environnementale, vous avez deja conduit un ***projet*** *de* ***transfert d'activite****.*
Translation of the wrong classified segment: Being a post graduate in chemical engineering with a focus on analystic environmental chemistry, you have already lead a **project** of **transfer of activity**.

Table 3. Example of a misclassified segment.

about terms like **project**, **activity transfer** that corresponding to "Mission" and "Profile" categories. The segment is classified as "Profile". In fact, this segment occurs at the boundary between the "Mission" and "Profile" segments, and the sequence is probable (Viterby's probability is not null), so this error is not corrected by the corrective process. The improvement of the block boundary's detection is one of the ways that we are currently exploring [18] in order to increase the performance of our system.

6 Conclusion and future work

Processing job posting information is a difficult task because the information flow is still strongly unstructured. In this paper we show the categorisation module, the first component of E-Gen, a modular system to treat jobs listings automatically. The first results obtained through SVM were interesting (approximately 10% error for a training corpora of 80%). The application of the corrective process improves the results by approximately 50% on the SVM score and considerably decreases errors such as "wrongly classified segments that are isolated" with very good computing times. Informations such as salary (minimum, maximum salary and currency), place of work and categorisation of the occupation are correctly detected to send pertinent information about the job posting to the job boards. The first module of E-Gen is currently in test on Aktor's server and offers a considerable time saving in the daily treatment of job offers. E-Gen is a independent and portable database, because it is a modular system with e-mail as input and XML documents as output. The promising results in this paper allow us to continue the E-Gen project with the relevance ranking of candidate responses. Several approaches (information retrieval, machine learning, automatic summarisation) will be considered to resolve these problems with a minimal cost in terms of human intervention.

Acknowledgement

Autors thanks to Jean Arzalier, Eric Blaudez, ANRT *(Agence Nationale de la Recherche Technologique*, France) and Aktor Interactive that partially supported this work (Grant Number CIFRE 172/2005).

References

1. Bizer, C., Heese, R., Mochol, M., Oldakowski, R., Tolksdorf, R., Eckstein, R.: The impact of semantic web technologies on job recruitment processes. In: International Conference Wirtschaftsinformatik (WI 2005), Bamberg, Germany. (2005)
2. Rafter, R., Bradley, K., Smyt, B.: Automated Collaborative Filtering Applications for Online Recruitment Services. (2000) 363–368
3. Rafter, R., Smyth, B.: (Passive Profiling from Server Logs in an Online Recruitment Environment)
4. Bourse, M., Leclère, M., Morin, E., Trichet, F.: Human resource management and semantic web technologies. In: Proceedings, 1st International Conference on Information & Communication Technologies: from Theory to Applications (ICTTA). (2004)
5. Morin, E., Leclère, M., Trichet, F.: The semantic web in e-recruitment (2004). In: The First European Symposium of Semantic Web (ESWS'2004). (2004)
6. Rafter, R., Smyth, B., Bradley, K.: (Inferring Relevance Feedback from Server Logs: A Case Study in Online Recruitment)
7. D. A Zighed, J.C.: Data Mining and CV analysis. (2003) Vol. 17:189–200
8. Bellman, R.: Adaptive Control Processes. Princeton University Press (1961)
9. Manning, D., Schütze, H.: Foundantions of Statistical Natural Language Processing. The MIT Press (2002)
10. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag (1995)
11. Joachims, T.: Making large scale SVM learning practical. Advances in kernel methods: support vector learning. The MIT Press (1999)
12. Grilheres, B., Brunessaux, S., Leray, P.: Combining classifiers for harmful document filtering. In: RIAO. (2004) 173–185
13. Kessler, R., Torres-Moreno, J.M., El-Bèze, M.: Classification automatique de courriers électroniques par des méthodes mixtes d'apprentissage. (2006) 93–112
14. Fan, R.E., Chen, P.H., Lin, C.J.: Towards a Hybrid Abstract Generation System. In: Working set selection using the second order information for training SVM. (2005) 1889–1918
15. Viterbi, A.J.: Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. (1967) 13:260–269
16. A. H. Land, A.G.D.: An Automatic Method of Solving Discrete Programming Problems. (1960) Vol. 28:497–520
17. El-Bèze, M., Torres-Moreno, J., Béchet, F.: Un duel probabiliste pour départager deux Présidents. In: RNTI coming soon. (2007) 1889–1918
18. Reynar, J., Ratnaparkhi, A.: A Maximum Entropy Approach to Identifying Sentence Boundaries. In: In Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington D.C. (1997) 16–19