E-Gen: automatic profiling system for ranking candidates answers in Human Resources

Rémy Kessler[∇],Nicolas Béchet[α],
Juan Manuel Torres-Moreno[∂,∇],Mathieu Roche[α], Marc El-Bèze[∇]

[∇] LIA, BP 1228 F-84911 Avignon Cedex 9 - France
[∂] École Polytechnique de Montréal - Département de génie informatique
CP 6079 Succ. Centre Ville H3C 3A7, Montréal (Québec), Canada.
[α] LIRMM - UMR 5506, CNRS - Univ. Montpellier 2 - France
{remy.kessler,juan-manuel.torres,marc.elebeze}@univ-avignon.fr
{nicolas.bechet,mathieu.roche}@lirmm.fr

Abstract. The exponential growth of Internet allowed the development of a market of online job search sites. This paper aims at presenting the E-Gen system (Automatic Job Offer Processing system for Human Resources). E-Gen will implement several complex tasks: an analysis and categorization of jobs offers which are unstructured text documents (e-mails of job offers possibly with an attached document), an analysis and a relevance ranking of the candidate answers. We present a strategy to resolve the last task: After a process of filtering and lemmatisation, we use vectorial representation and different similarity measures. The quality of ranking obtained is evaluated using ROC curves.

1 Introduction

The exponential growth of Internet allowed the developement of an online job-search sites market [1, 2]. The answers of candidates represent a lot of information that can not be managed efficiently by companies [3, 4]. It is therefore indispensable to process this information by an automatic or assisted way. Thus, we develop the E-GEN system to resolve this problem. It will be composed of three main modules:

1. A module of information extraction from a corpora of e-mails of job offers.
2. A module to analyse the candidate answers (what part of the e-mails is cover letter, curriculum vitae).
3. A module to analyse and compute a relevance ranking of the candidate answers (cover letter and curriculum vitae).

Our previous works present the first module [5], the identification of different parts of a job offer and the second module [6] which analyses the contents of a candidate e-mail with Support Vector Machine [7] and n-grams tools. We present in this paper a strategy to resolve the last module. The large number of candidates answers for a job generates a lenghty process of reading for the

recruiting consultant. To facilitate this task, we want to set up a system capable of providing an initial evaluation of candidates answers according to various criteria. We show which document (curriculum vitae or cover letter) contains the most relevant information and the location of this relevant information in each documents. We use different similarity measures between a job offer and candidates answers to rank them. Section 2 shows a general system overview. In section 3, we describe the pre-processing task and the different's measures used to rank the candidates answers. In section 4, we present statistics about textual corpora, examples (job offer, cover letter and curriculum vitae), experimental protocol, results, and discussions before concluding and indicating future work.

2 System overview

The main activity of Aktor Interactive is the processing of job offers on the Internet. As the Internet proposes new ways to the employment market, Aktor modifies its procedures to integrate a system which answers as fast and judiciously as possible to this processing. An e-mail-box receives messages containing the offer. After language identification, E-GEN parses the e-mail, splits the offer in segment, and retrieves relevant information (contract, salary, location, etc.) to put on line job offer. During the publication of jobs offer, Aktor generates an e-mail address for apply to the job. Each e-mail is so redirected

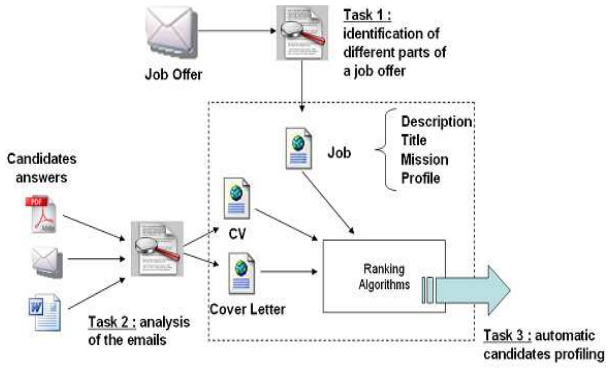


Fig. 1. System overview.

to a Human Resources software, Gestmax[1] to to be read by a recruiting consultant. In this step, E-GEN analyses candidates answers to identify each part of the candidature and extract text from email and enclosed files (wvWare[2] for document MS-Word and pdftotext[3] for pdf document). In this last step, E-GEN

[1] http://www.gestmax.fr
[2] http://wvware.sourceforge.net
[3] http://www.bluem.net/downloads/pdftotext_en

analyses each part of each candidate answer before ranking candidates for the mission. The whole processing chain of E-GEN System is represented in figure 1.

3 Ranking algorithms

3.1 Corpora pre-processing

A pre-processing task of the corpora was perfomed to obtain a suitable representation in the Vector Space Model (VSM). Mainly deletion of the following items: Verbs and functional words (to be, to have, to be able to, to need,...), common expressions with a stoplist[4](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 our model. Lemmatisation processing has been also performed to obtain an important reduction of the lexicon. It consists in finding the root of verbs and transform plural and/or feminine words to masculine singular form[5]. This process allows to decrease the curse of dimensionality [8] which raises severe problems of representation of the huge dimensions [9]. Other mechanisms of reduction of the lexicon are also used: compounds 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 Mesures of similarity

We decided to use a number of similarity measures to determine which is most effective. In first, we use Enertex. Textual energy (Enertex) measure was successfully used in different NLP (Natural Language Processing) tasks as automatic summarization and topic segmentation [10]. Based on energy of the Ising magnetic model, it considers a document of N terms as a chain of N binary units called spins (Ising units). Upward spins are the present words and downwards are the absentees. In our model, we are particularly interested in textual energy between the job offer and each candidate as:

$$J^{i,j} = \sum_{\mu=1}^{P} s_\mu^i s_\mu^j \tag{1}$$

$$E_{\mu,\nu} = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} s_\mu^i \; J^{i,j} \; s_\nu^j \tag{2}$$

s_μ^i =term i of job offer,s_ν^j =term j of candidate,Jij = interaction of terms i and j in the P documents

[4] http://sites.univ-provence.fr/ veronis/donnees/index.html

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

The next measure is the Needleman-Wunsch algorithm, commonly used in bioinformatics to align sequences of proteins or nucleotides. We consider vector of job offer j and vector of candidate answer d as sequence of terms and we compute the best score $H(j,d)$ between the two sequence with the Needleman-Wunsch algorithm[11]. We also tested different typically similarity measures for vector as defined in [12]: *cosine*, formula 3, which allows us to calculate angle between job offer and each candidate answer, the Minkowski distances, formula 4 ($p = 1$ for Manhattan, $p = 2$ for euclid) and overlap 5.

$$sim_{cosine}(j,d) = \frac{j_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} |j_i|^2 \cdot \sum_{i=1}^{n} |d_i|^2}} \tag{3}$$

$$sim_{Minkowski}(j,d) = \frac{1}{1 + (\sum_{i=1}^{n} |j_i - d_i|^p)^{\frac{1}{p}}} \tag{4}$$

$$sim_{Overlap}(j,d) = \frac{j_i \cdot d_i}{Min\left(\sum_{i=1}^{n} |j_i|^2, \sum_{i=1}^{n} |d_i|^2\right)} \tag{5}$$

with j job offer, d a candidate answer, i a term.
The last measure used is Okabis[13]. Based on okapi formula, measure often used in Information Retrieval, we defined as :

$$Okabis(j,d) = \sum_{w \in d \cap j} \frac{TF_{w,d}}{TF_{w,d} + \frac{\sqrt{|d|}}{M_S}} \tag{6}$$

with j job offer, d a candidate answer, w a term, $TF_{w,d}$ occurrence of w in S, N total numbers of candidates and M_S their average size. To combine these measures, we use an algorithm decision[14], which will weigh the values obtained by each measure of similarity. Two averages are calculated: the positive tendency, (that is $\lambda > 0.5$), and the negative tendency, for ($\lambda < 0.5$).

4 Experiments

4.1 Corpora description

We have selected a data subset from Aktor's database. This corpora, called *Reference Corpora*, is a set of job offers with various thematics (jobs in accountancy, business enterprise, computer science, etc...) associated with his candidates. Each candidates is tagging **positive** or **negative**. A **positive** value is a potential candidate for a given job by the recruiting consultant. A **negative** value is for a irrelevant candidate for the job (decision of the company). Table 1 shows a few statistics about our *Reference Corpora*.

Total number of jobs offers	25
Number of jobs offers with less 10 candidates	2
Number of jobs offers with more 10 candidates	8
Number of jobs offers with more 50 candidates	6
Number of jobs offers with more 100 candidates	9
Total Number of candidates	2916
Number of candidates with tagging **positive**	220
Number of candidates with tagging **negative**	2696

Table 1. Corpora statistics.

4.2 Example of job offer

Each job offers is separated in four classes, as defined in [5] :

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

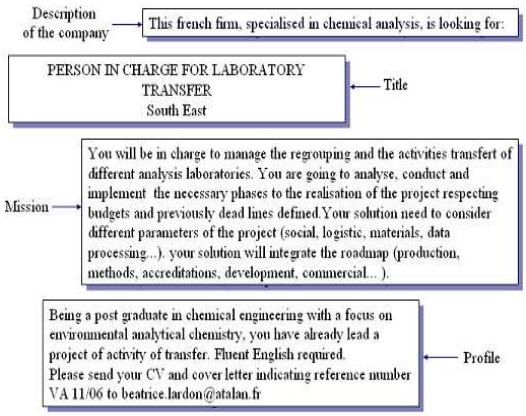
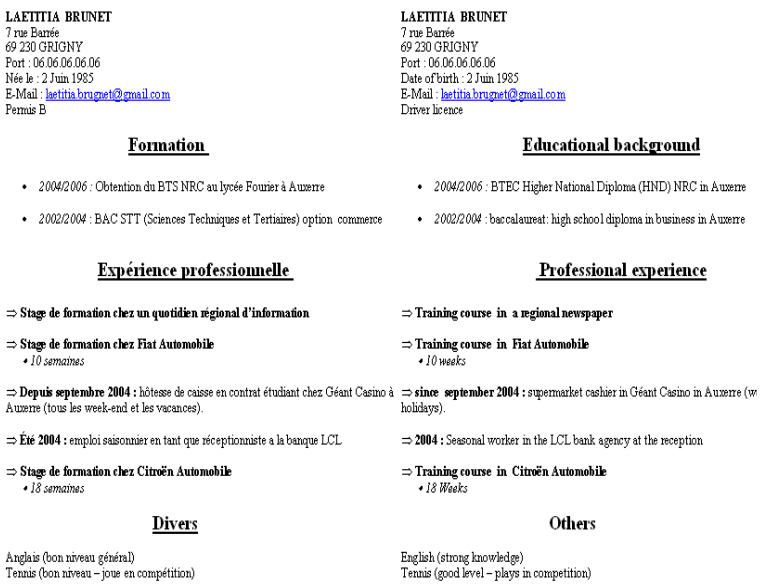


Fig. 2. Job offer segmented.

A job offer exemple with this segmentation is presented in figure 2, translated in English. The content of the job offer is free but we find a rather similar presentation and vocabulary according to every part. This segmentation is used for ranking candidates as we shall see it afterwards in section 4.5.

4.3 Example of candidature

Figure 3 is a example of French curriculum vitae with a translation in English and Figure 4, a French cover letter with a translation in English (documents were



LAETITIA BRUNET
7 rue Barrée
69 230 GRIGNY
Port : 06.06.06.06.06
Née le : 2 Juin 1985
E-Mail : laetitia.brugnet@gmail.com
Permis B

Formation

- *2004/2006* : Obtention du BTS NRC au lycée Fourier à Auxerre
- *2002/2004* : BAC STT (Sciences Techniques et Tertiaires) option commerce

Expérience professionnelle

⇒ **Stage de formation chez un quotidien régional d'information**

⇒ **Stage de formation chez Fiat Automobile**
• 10 semaines

⇒ **Depuis septembre 2004 :** hôtesse de caisse en contrat étudiant chez Géant Casino à Auxerre (tous les week-end et les vacances).

⇒ **Été 2004 :** emploi saisonnier en tant que réceptionniste a la banque LCL

⇒ **Stage de formation chez Citroën Automobile**
• 18 semaines

Divers

Anglais (bon niveau général)
Tennis (bon niveau – joue en compétition)

LAETITIA BRUNET
7 rue Barrée
69 230 GRIGNY
Port : 06.06.06.06.06
Date of birth : 2 Juin 1985
E-Mail : laetitia.brugnet@gmail.com
Driver licence

Educational background

- *2004/2006* : BTEC Higher National Diploma (HND) NRC in Auxerre
- *2002/2004* : baccalaureat: high school diploma in business in Auxerre

Professional experience

⇒ **Training course in a regional newspaper**

⇒ **Training course in Fiat Automobile**
• 10 weeks

⇒ **since september 2004 :** supermarket cashier in Géant Casino in Auxerre (w holidays).

⇒ **2004 :** Seasonal worker in the LCL bank agency at the reception

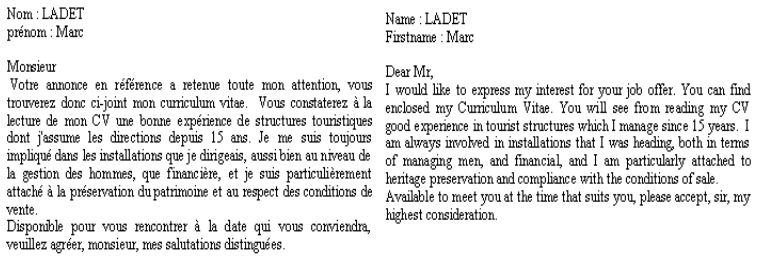
⇒ **Training course in Citroën Automobile**
• 18 Weeks

Others

English (strong knowledge)
Tennis (good level – plays in competition)

Fig. 3. Example of curriculum vitae in french on left and in english on the right.

previously anonymous). A first analysis of CV and cover letter shows that the documents are very different. Cover letter is a complete text, containing references to the job offer while CV summarizes career of candidates. The content of CV are free but we find a rather similar presentation and vocabulary according to each block ("Professional experience", "Educational background","Personal interests",etc.) and some relevant collocations [15] as "assistant director"', "Computer skills", "Driver licence", etc.



Nom : LADET
prénom : Marc

Monsieur
Votre annonce en référence a retenue toute mon attention, vous trouverez donc ci-joint mon curriculum vitae. Vous constaterez à la lecture de mon CV une bonne expérience de structures touristiques dont j'assume les directions depuis 15 ans. Je me suis toujours impliqué dans les installations que je dirigeais, aussi bien au niveau de la gestion des hommes, que financière, et je suis particulièrement attaché à la préservation du patrimoine et au respect des conditions de vente.
Disponible pour vous rencontrer à la date qui vous conviendra, veuillez agréer, monsieur, mes salutations distinguées.

Name : LADET
Firstname : Marc

Dear Mr,
I would like to express my interest for your job offer. You can find enclosed my Curriculum Vitae. You will see from reading my CV good experience in tourist structures which I manage since 15 years. I am always involved in installations that I was heading, both in terms of managing men, and financial, and I am particularly attached to heritage preservation and compliance with the conditions of sale.
Available to meet you at the time that suits you, please accept, sir, my highest consideration.

Fig. 4. Example of cover letter in french on left and in english on the right.

4.4 Experimental protocol

We propose to measure the similarity of a job offer and the candidates. We have 25 job offers associated at least 4 candidates. After representing textual data by vectors in a Salton space model [16], we use different similarity measures between job offers and associated candidate answers to obtain the candidate ranking. We apply several similarity measures presented in section 3.2: Enertex, cosine, Minkowski, Manhattan, Needleman-Wunsch, and Overlap. Finally, we use the Decision measure which combines the previews ones. These measures are described precisely in section 3.2. We use the ROC Curves to evaluate the quality of obtained ranking with tagging defined in 4.1. Initially the ROC curves, detailed in [17], come from the field of signal treatment. ROC curves are often used in the field of medicine to evaluate the validity of diagnostic tests. The ROC curves show in X-coordinate the rate of false positive (in our case, rate of irrelevant candidate answers) and in Y-coordinate the rate of true positive (rate of relevant candidate answers). The surface under the ROC curve (*AUC - Area Under the Curve*), can be seen as the effectiveness of a measurement of interest. The criterion relating to the surface under the curve is equivalent to the statistical test of Wilcoxon-Mann-Whitney, see work of [18]. In the case

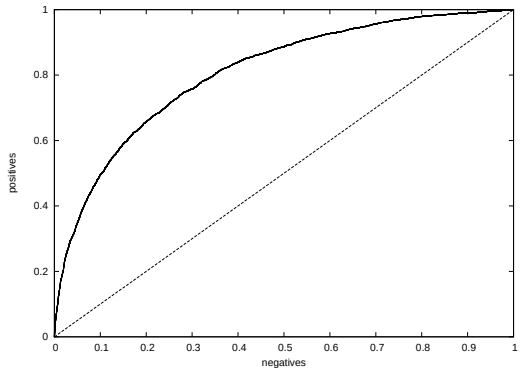


Fig. 5. Example of ROC Curve

of candidate answers ranking, a perfect ROC curve corresponds to obtain all relevant candidate answers at the beginning of the list and all irrelevant at the end. This situation corresponds to $AUC = 1$. The diagonal corresponds to the performance of a random system, progress of the rate of true positive being accompanied by an equivalent degradation of the rate of false positive. This situation corresponds to $AUC = 0.5$. The figure 5 is an example of a ROC Curve with in diagonal a random system distribution. If the candidate answers are ranked by decreasing interest (*i.e.* all relevant candidate answers are after the irrelevant ones) then $AUC = 0$. An effective measurement of interest to order candidate answers consists in obtaining an AUC the highest possible value. This is strictly equivalent to minimizing the sum of the rank of the positive examples.

The advantage of the ROC curves comes from its resistance to imbalance (for example, an imbalance in number of positive and negative examples). For one job offer, we evaluate quality of lists of ranked candidates obtained from the different similarity measures (Enertex, cosine, etc.) by performing an AUC. Then we compute an average of the 25 job offers AUC.

4.5 Results

	AUC	Enertex	Cosine	Minkowski	Manhattan	NW	Overlap	Okapi	Decision
DTMP	CL	0,524	0,567	0,561	0,591	0,481	0,573	0,521	**0,596**
	CV	0,524	**0,604**	0,510	0,503	0,532	0,543	0,541	0,562
	CL+CV	0,523	**0,621**	0,539	0,532	0,509	0,522	0,523	0,571
TMP	CL	0,524	0,560	0,559	0,580	0,473	0,562	0,513	**0,591**
	CV	0,523	**0,622**	0,508	0,501	0,544	0,538	0,542	0,561
	CL+CV	0,523	**0,642**	0,538	0,528	0,526	0,531	0,532	0,592

Table 2. AUC obtained with our different filtering

First we propose to study the structure of data (job offer and candidate answer).
A job offer is composed by a description, a title, a mission, and a profile as described in 4.2. We use two combines of a job offer content:
- conserving only Title, Mission, Profile (called TMP)
- conserving all information of a job offer (called DTMP)
A candidate answer is composed by a CV (Curriculum Vitae) and a CL (Cover Letter).
Table 2 presents the AUC results by studying the impact of the CV and CL in candidate answers with DTMP and TMP. We use the different similarity measures presented in 4.4 to compute ranking. The best AUC is obtained with the cosine measure, using the CV and CL of candidate answers and the TMP job offer. The cosine measure gives best results for all approaches, except by

AUC	Enertex	Cosine	Minkowski	Manhattan	NW	Overlap	Okapi	Decision
CV_1/3	0,525	**0,589**	0,497	0,505	0,533	0,539	0,569	0,579
CV_2/3	0,524	**0,600**	0,524	0,520	0,515	0,577	0,560	0,580
CV_3/3	0,526	**0,526**	0,497	0,503	0,510	0,479	0,506	0,501
CL_1/3	0,527	0,573	0,561	**0,588**	0,480	0,571	0,528	0,580
CL_2/3	0,533	0,565	0,570	**0,578**	0,481	0,578	0,543	0,570
CL_3/3	0,516	0,447	0,528	**0,538**	0,416	0,446	0,439	0,470

Table 3. AUC obtained with different parts of CV and CL

considering CL only (with DTMP and TMP), which is better with the Decision similarity measure. We assume that the results obtained by the decision of algorithm are noisy by the poor performance of certain measures (Overlap or Needleman-Wunsch). The Enertex measure has strangely very similar results for all kind of data, we work to determine the reasons of this results. We observe

that the use of CV is more relevant that CL. These results confirm our intuition, the CV is the main document and contains the most important information of a candidature. Globally, AUC scores are not very relevant. We can explain by the nature of used data and the quality of the human expertise this fact. Finally, we propose to split CV and CL in three parts to identify parts which contain relevant knowledge. Table 3 presents AUC of splited CV and CL. Figure 6 presents graphical results of table 3 by using best similarity measures (Cosine, Minkowski, Manhattan, Overlap, and Decision). We obtain irrelevant scores with the last

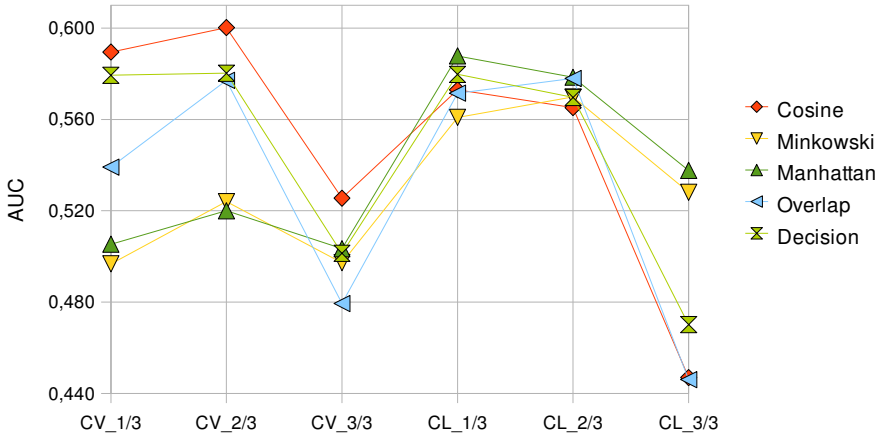


Fig. 6. Best similarity measures comparison with parts of CV and CL

part of CV and CL. We conclude that the relevant knowledge to determine a candidate adapted to an offer, is contained in the 1/3 and 2/3 of the CV and CL. Actually, the last part of CV is generally "Hobbies" or "other" which is rarely a crucial information. In the same way, the last part of CL is generally as "Yours faithfully", "Yours sincerely", "best regards", which are irrelevant informations. We are currently working a finer segmentation for each document.

5 Conclusion and Future work

Processing job information is a difficult task because the information flow is still strongly unstructured. In this paper we show the ranking module, the last component of E-GEN, a modular system to analyse the candidate answers. We tested different measures of similarity and different segmentation of the job offer, curriculum vitae and cover letter. The first results obtained are interesting and we are considering best segmentation to improve performance. We are currently testing our system by using a part-of-speech tagger to improve performance of preprocessing. The first and second module of E-GEN are currently in test on

Aktor's server and allows a considerable saving of time in the daily treatment of job offers. E-GEN is database independent and portable, because it is a modular system with e-mail in input and XML documents as output. We are also setting up a system for evaluating CV on employment portal *jobmanager*[6] allowing internaut to find the jobs offer the more interesting with his profile.

References

1. Bizer, R.H., Rainer, E.: Impact of Semantic web on the job recruitment Process. International Conference Wirtschaftsinformatik (2005)
2. Rafter, R., Bradley, K., Smyt, B.: Automated Collaborative Filtering Applications for Online Recruitment Services. (2000) 363–368
3. Bourse, M., Leclère, M., Morin, E., Trichet, F.: Human resource management and semantic web technologies. In: 1st International Conference on Information & Communication Technologies(ICTTA). (2004)
4. Morin, E., Leclère, M., Trichet, F.: The semantic web in e-recruitment (2004). In: The First European Symposium of Semantic Web (ESWS'2004). (2004)
5. Kessler, R., Torres-Moreno, J.M., El-Bèze, M.: E-Gen: Automatic Job Offer Processing system for Human Ressources. MICAI (2007)
6. Kessler, R., Torres-Moreno, J.M., El-Bèze, M.: E-Gen: Profilage automatique de candidatures. Traitement Automatique de la Langue Naturelle (TALN 2008), Avignon, France (2008)
7. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag (1995)
8. Bellman, R.: Adaptive Control Processes. Princeton University Press (1961)
9. Manning, D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press (2002)
10. Silvia, F., Eric, S., Juan Manuel, T.M.: Textual Energy of Associative Memories: performants applications of ENERTEX algorithm in text summarization and topic segmentation. In: MICAI. (2007)
11. Needleman, S., Wunsch, C.: A general method applicable to the search for similarities in the amino acid sequences of proteins or nucleotides. (1970)
12. Bernstein, A., Kaufmann, E., Kiefer, C., Bürki, C.: Simpack: A generic java library for similarity measures in ontologies. Technical report (August 2005)
13. Bellot, P., El-Bèze, M.: Classification et segmentation de textes par arbres de décision. In: Technique et Science Informatiques (TSI). Volume 20. Hermès (2001)
14. Boudin, F., Moreno, J.M.T.: Neo-cortex: A performant user-oriented multi-document summarization system. In: CICLing. (2007) 551–562
15. Roche, M., Prince, V.: Evaluation et détermination de la pertinence pour des syntagmes candidats à la collocation . JADT2008 (2008) 1009–1020
16. Salton, G.: Developments in automatic text retrieval. Science 253 (1991) 974–979
17. Ferri, C., Flach, P., Hernandez-Orallo, J.: Learning decision trees using the area under the ROC curve. In: Proceedings of ICML'02. (2002) 139–146
18. Yan, L., Dodier, R., Mozer, M., Wolniewicz, R.: Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In: Proceedings of ICML'03. (2003) 848–855

[6] http://www.jobmanager.fr