TALN 2008, Avignon, 9–13 juin 2008

E-Gen : Profilage automatique de candidatures

Rémy Kessler[1,2] Juan Manuel Torres-Moreno[1] Marc El-Bèze[1]
(1) LIA / Université d'Avignon, 339 chemin des Meinajariès, 84911 Avignon
(2) AKTOR 12, allée Irène Joliot Curie 69800 Saint Priest
{remy.kessler, juan-manuel.torres, marc.elbeze}@univ-avignon.fr

Résumé. La croissance exponentielle de l'Internet a permis le développement de sites d'offres d'emploi en ligne. Le système E-Gen (Traitement automatique d'offres d'emploi) a pour but de permettre l'analyse et la catégorisation d'offres d'emploi ainsi qu'une analyse et classification des réponses des candidats (Lettre de motivation et CV). Nous présentons les travaux réalisés afin de résoudre la seconde partie : on utilise une représentation vectorielle de texte pour effectuer une classification des pièces jointes contenus dans le mail à l'aide de SVM. Par la suite, une évaluation de la candidature est effectuée à l'aide de différents classifieurs (SVM et n-grammes de mots).

Abstract. The exponential growth of the Internet has allowed the development of a market of on-line job search sites. This paper presents the E-Gen system (Automatic Job Offer Processing system for Human Resources). E-Gen will perform two complex tasks : an analysis and categorisation of job postings, which are unstructured text documents, an analysis and a relevance ranking of the candidate answers (cover letter and curriculum vitae). Here we present the work related to the second task : we use vectorial representation before generating a classification with SVM to determine the type of the attachment. In the next step, we try to classify the candidate answers with different classifiers (SVM and ngrams of words).

Mots-clés : Classification de textes, Modèle probabiliste, Ressources humaines, Offres d'emploi.

Keywords: Text Classification, Probabilistic Model, Human Ressources, Job Offer.

1 Introduction

La croissance exponentielle de l'Internet a permis un grand développement de *jobboards* (Bizer & Rainer, 2005; Rafter *et al.*, 2000). Cependant, les réponses des candidats représentent une grande quantité d'information difficile à gérer rapidement et efficacement pour les entreprises (Bourse *et al.*, 2004; Morin, 2004; Rafter *et al.*,). En conséquence, il est nécessaire de la traiter d'une manière automatique ou assistée. Le LIA et Aktor Interactive, agence de communication française spécialisée dans l'e-recruiting, développent le système E-Gen pour résoudre ce problème. Le système E-Gen se compose de deux modules principaux :

1. Un module d'extraction de l'information à partir de corpus des courriels provenant d'offres d'emplois extraites de la base de données d'Aktor.
2. Un module pour analyser et calculer un classement de pertinence du profil du candidat (lettre de motivation et curriculum vitae).

Nos précédents travaux (Kessler *et al.*, 2007; Kessler & El-Bèze, 2008) présentaient le premier module, l'identification des différentes parties d'une offre d'emploi et l'extraction d'informations pertinentes (contrat, salaire, localisation, etc.). Lors de la mise en ligne d'une annonce, Aktor génère une adresse électronique de "réponse" pour chacune de ces offres. Chaque courrier électronique de candidature est par la suite redirigé vers le logiciel de Ressources Humaines, Gestmax[1], afin de pouvoir être traité. Cependant le flux de réponses à une offre d'emploi entraîne un long travail de lecture des candidatures par les recruteurs. Afin de faciliter cette tâche, nous souhaitons mettre en place un système capable de fournir une première évaluation automatisée des candidatures selon divers critères. Nous présentons les premiers travaux concernant le second module du système E-Gen. On présente en section 2 l'architecture globale d'E-Gen et la stratégie pour identifier chaque document de la candidature. Nous présentons en section 3 la méthode utilisé afin d'effectuer un tri entre les pièces jointes avant de présenter en section 4 les travaux concernant l'évaluation d'un curriculum vitae (abrégé CV) d'une candidature avant de détailler les différents résultats obtenus dans la section finale.

2 Vue d'ensemble du système

Nous avons choisi de développer un système répondant aussi rapidement et judicieusement que possible au besoin d'Aktor, et donc aux contraintes du marché de recrutement en ligne. (Kessler & El-Bèze, 2008) détaillent la stratégie mise en place afin de résoudre la tâche 1. Ici, nous présenterons principalement la seconde tâche ainsi que les premiers travaux concernant la tâche 3 (voir figure 1), l'évaluation de la candidature. Lors de la réception d'une candidature par courrier électronique, le système extrait le corps du message, ainsi que les différentes pièces jointes et les transforme au format XML (wvWare[2] traite les documents MS-Word et produit une version texte du document découpé en segments, pdftotext[3]extrait le contenu texte d'un document pdf). Différents processus de filtrage et racinisation permettent au système d'identifier à l'aide de machines à support vectoriel (cf section 3) le contenu de la candidature (composée d'un CV et/ou d'une lettre de motivation présente dans le corps du mail ou dans les pièces jointes). Une fois le CV et la lettre de motivation identifié, le système E-Gen effectuera une première évaluation automatisée de cette candidature selon divers critères tels que la richesse du vocabulaire, le nombre de fautes d'orthographe, la correspondance entre la candidature et l'offre d'emploi (Tâche 1) ainsi qu'une évaluation par des méthodes d'apprentissage de cette candidature(cf section 4). La figure 1 présente une vue d'ensemble du système E-Gen.

2.1 Corpus et exemple de candidatures

Un sous-ensemble de données a été sélectionné à partir de la base de données d'Aktor. Ce corpus regroupe plusieurs missions[4] d'Aktor Sourcing&Selection[5] ainsi que les diverses réponses à ces offres d'emplois classées en différentes *boîtes*[6]. Afin de simplifier le problème nous avons

[1]http ://www.gestmax.fr

[2]http ://wvware.sourceforge.net. La segmentation de textes MS-Word étant difficile, on a opté pour un outil existant. Dans la majorité des cas, il sectionne en paragraphes le document.

[3]http ://www.bluem.net/downloads/pdftotext_en

[4]Mission désigne la pré-sélection effectuée par le cabinet de recrutement pour une offre d'emploi.

[5]http ://www.aktor-selection.fr

[6]Gestmax permet au recruteur de classer une candidature : non lu, oui, non, peut-être, Entretien etc..

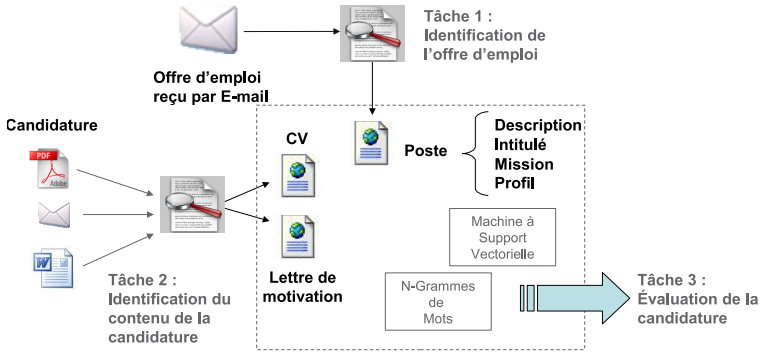


FIG. 1 – Vue d'ensemble du système E-Gen.

ramené chacune de ces boîtes à 3 catégories différentes : OUI, NON ou PEUT-ETRE (PT par la suite). Après consultation du recruteur ayant effectué l'étiquetage des CV, nous avons décidé de regrouper en une seule catégorie les CV appartenant à la classe OUI et à la classe PT. En effet, d'un point de vue ressource humaine, une candidature classée PT est une bonne candidature mais ne correspondant pas forcément à la mission, ou une bonne candidature mais pas la meilleure. Ce regroupement nous a permis d'équilibrer un peu le corpus, celui-ci étant majoritairement composé de candidatures étiquetées NON 2. Les missions peuvent être rédigées en différentes langues, mais notre étude porte sur les offres et les réponses en français (le marché français représente l'activité principale d'Aktor). Ce sous-ensemble, nommé *Corpus de référence* a donc permis d'obtenir un corpus de réponses classées en fonction de la mission ainsi que du jugement d'un recruteur sur les candidatures. Le tableau 1 présente quelques statistiques du corpus.

Nombre de Missions Total	41
Nombre de Mission avec moins de 10 réponses	8
NB Mission avec plus de 10 réponses	13
NB Mission avec plus de 50 réponses	9
NB Mission avec plus de 100 réponses	11
Nombre de candidatures Total	3078

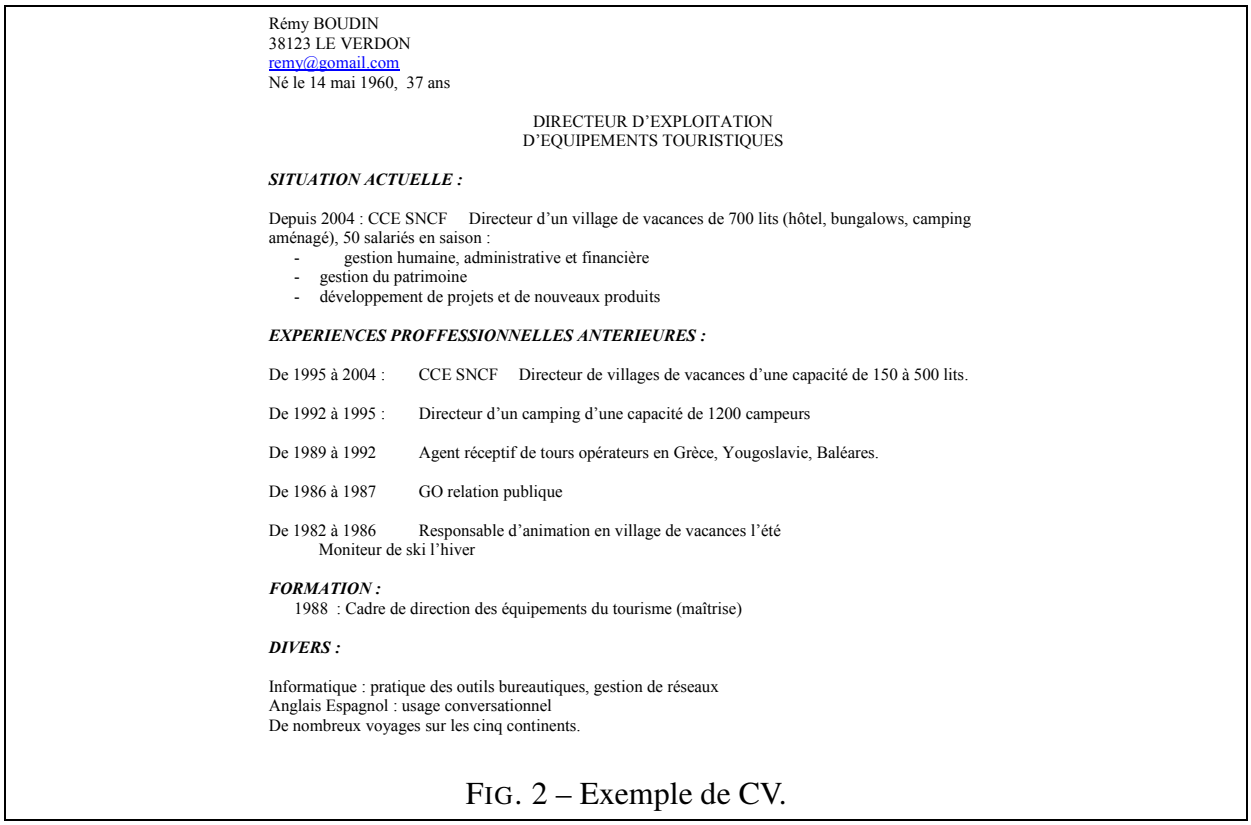
TAB. 1 – Statistisques du *Corpus de référence*.

CV Total	CV noté OUI	CV noté NON	CV noté PT
2755	414	2128	213
LM Total	LM noté OUI	LM noté NON	LM noté PT
2473	407	1882	184

TAB. 2 – Statistisques du *Corpus de référence* en fonction de l'étiquetage.

La figure 2 présente un exemple de CV extrait du *corpus de référence* et la figure 3 un exemple de lettre de motivation (abrégée en LM par la suite). Les documents ont été préalablement anonymisés. De façon évidente le style de chaque document est différent, la lettre de motivation se présentant comme un texte complet alors que le CV résume le parcours professionnel de la personne de façon concise. On observe par ailleurs que les CV, malgré un format libre, présentent des similarités entre eux du point de vue de leur contenu : On retrouve généralement les sections "Expérience professionnelle","Expérience personnelle", "Formation", "Divers" ou encore

"Loisirs") ainsi que certaines collocations pertinentes ("assistante commercial", "baccalauréat scientifique" etc.) comme décrit dans (Roche & Prince, 2008) et d'un point de vue présentation (Texte en gras ou en italique afin de définir chaque partie, indentation etc..), même si les différents outils que nous utilisons afin d'extraire le texte ne nous permettent pas de récupérer la structure du document (cf section 2).



Rémy BOUDIN
38123 LE VERDON
remy@gomail.com
Né le 14 mai 1960, 37 ans

DIRECTEUR D'EXPLOITATION
D'EQUIPEMENTS TOURISTIQUES

SITUATION ACTUELLE :

Depuis 2004 : CCE SNCF Directeur d'un village de vacances de 700 lits (hôtel, bungalows, camping aménagé), 50 salariés en saison :
- gestion humaine, administrative et financière
- gestion du patrimoine
- développement de projets et de nouveaux produits

EXPERIENCES PROFFESSIONNELLES ANTERIEURES :

De 1995 à 2004 : CCE SNCF Directeur de villages de vacances d'une capacité de 150 à 500 lits.

De 1992 à 1995 : Directeur d'un camping d'une capacité de 1200 campeurs

De 1989 à 1992 Agent réceptif de tours opérateurs en Grèce, Yougoslavie, Baléares.

De 1986 à 1987 GO relation publique

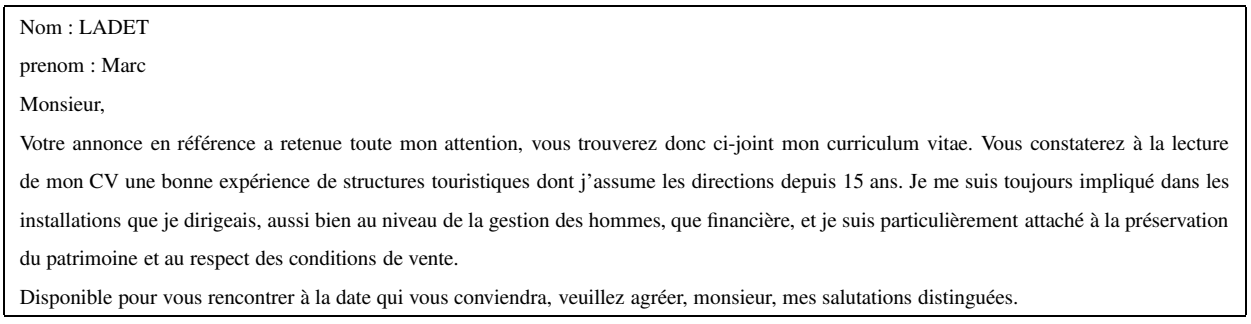
De 1982 à 1986 Responsable d'animation en village de vacances l'été
Moniteur de ski l'hiver

FORMATION :
1988 : Cadre de direction des équipements du tourisme (maîtrise)

DIVERS :

Informatique : pratique des outils bureautiques, gestion de réseaux
Anglais Espagnol : usage conversationnel
De nombreux voyages sur les cinq continents.

FIG. 2 – Exemple de CV.



Nom : LADET

prenom : 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.

FIG. 3 – Exemple de lettre de motivation.

3 Classification de CV/Lettre de motivation par SVM

Nous avons choisi les SVM(Vapnik, 1995) pour cette tâche car cet algorithme d'apprentissage a été utilisé avec succès dans plusieurs tâches de catégorisation de texte auparavant (Joachims, 2008; Pham & Do, 2003). Nous avons tenté évidemment une classification simpliste en se basant uniquement sur les noms des fichiers. Cependant ceci s'est avéré insuffisant[7] en raison de

[7]Le système constituait un corpus tronqué à 1725 CV et 910 LM

la diversité des noms de fichiers[8]. De même, différents tests à base de classifieurs naïfs tels que la longueur moyenne des phrases ou le nombre de mot dans le document ont montré leur limite comme nous le verrons dans la section 5. Nous effectuons donc une première étape de filtrage[9] et de racinisation (Heitz, 2008) [10], nous utilisons la représentation vectorielle de chaque document afin de lui attribuer une étiquette (CV ou LM) à l'aide des SVM. Les SVM permettent de construire un classifieur à valeurs réelles qui découpe le problème de classification en deux sous problèmes : transformation non-linéaire des entrées et choix d'une séparation linéaire optimale. Les données sont projetées dans un espace de grande dimension muni d'un produit scalaire où elles sont linéairement séparables selon une transformation basée sur un noyau linéaire, polynomial ou gaussien. Puis dans cet espace transformé, les classes sont séparées par des classifieurs linéaires qui déterminent un hyperplan séparant correctement toutes les données et maximisant la marge. Elles offrent, en particulier, une bonne approximation du principe de minimisation du risque structurel (c'est-à-dire, trouver une hypothèse h pour laquelle la probabilité que h soit fausse sur un exemple non-vu et extrait aléatoirement du corpus de test soit minimale). Nous utilisons l'implémentation *LibSVM* (Fan *et al.*, 2005) qui a prouvé sa robustesse dans de précédent travaux (Kessler & El-Bèze, 2008).

4 Classification du CV d'une candidature

Nous avons décidé dans un premier temps d'effectuer une classification du CV uniquement, en vue d'un profilage de la candidature (CV, offre d'emploi ainsi que LM) par la suite. Le CV est un document textuel singulier : structure particulière, informations éparses, contenu fortement symbolique, etc. d'où la difficulté de traitement de ces documents (Zighed, 2003). Nous avons pris en considération le genre donné par le *Corpus de référence* (CV ou LM) afin de ne garder que les documents étiquetés comme CV. Après le pré-traitement, nous avons effectué un premier apprentissage par les SVM. Les premiers résultats mitigés (voir section 5.2) nous ont conduit à envisager une classification par n-gramme. Un n-gramme de mots est une séquence de n mots consécutifs. Pour un document donné, on peut générer l'ensemble des n-grammes (n = 1, 2, 3,...) en déplaçant une fenêtre glissante de n cases sur le corpus. À chaque n-gramme, on associe une fréquence. Nos précédents travaux (Kessler & El-Bèze, 2008) ainsi que d'autres dans la littérature (Damashek, 1995; El-Bèze *et al.*, 2005) ont montré l'efficacité de cette approche comme méthode de représentation des textes pour des tâches de classification. Nous avons construit les uni-grammes et les bi-grammes à chaque classe (OUI/NON) avec leur probabilité P puis nous calculons pour obtenir le score $\tilde{t}$ des n-grammes pour un document D :

$$\tilde{t} = ArgMax_t P(t|W) = ArgMax_t \frac{P(W|t)P(t)}{P(W)} = ArgMax_t P(W|t)P(t) \quad (1)$$

Les deux dernières égalités proviennent de l'application du théorème de Bayes. En prenant comme hypothèse, compte tenu de la sous-représentation de la classe OUI :

$$P(t) = 1 \; \forall \, t \quad (2)$$

[8]par exemple PierreDurand.doc, Durand.pdf, Aktor.doc, 13042007.doc, V3.doc etc.

[9]Pour réduire la complexité du texte, différents filtrages du lexique sont effectués : la suppression des verbes et des mots fonctionnels, des expressions courantes, de chiffres (numériques et/ou textuelles) et des symboles.

[10]La racinisation simple trouve la racine des verbes fléchis et à ramène les mots pluriels et/ou féminins au masculin singulier.

on obtient :

$$\tilde{t} \approx ArgMax_t P(W|t) = ArgMax_t \prod_{i=1}^{|D|} P_t(W_i|W_1^{i-1}) \quad (3)$$

avec comme seconde hypothèse, pour obtenir des estimations fiables, malgré la faible taille des corpus disponibles :

$$P_t(W_i|W_1^{i-1}) \approx \lambda P_t(W_i|W_{i-1}) + (1-\lambda)P_t(W_i) \quad (4)$$

Nous travaillons actuellement a l'intégration des tri-grammes dans notre modèle ainsi qu'un lissage des événement non vus (Beaufort *et al.*, 2002) afin de compenser la faible taille de nos corpus ainsi que le manque d'étiquetage grammatical de ceux-ci. Cependant, les résultats très proches de chacun des classifieurs (voir section 5.2) nous ont permis d'envisager une combinaison des deux classifieurs (Grilheres *et al.*, 2004; Plantié M., 2007) sur la base d'un vote simple dans un premier temps afin d'améliorer les performances globales du système.

5 Résultats et discussion

Afin de régler les paramètres et tester nos méthodes, nous avons scindé le *Corpus de référence* en cinq sous-ensembles approximativement de la même taille, respectivement $A1$, $A2$, $A3$, $A4$ et $A5$, avec une répartition aléatoire mais équilibrée des candidatures dans chacun des sous-corpus. Le protocole expérimental a été le suivant : nous avons concaténé quatre des cinq sous-ensembles comme ensemble d'apprentissage et gardé le cinquième pour le test (ex $A2$ a pour ensemble d'apprentissage les sous ensembles 1,3,4,5 et pour validation le sous ensemble 2). Cinq expériences ont été ainsi effectuées à tour de rôle. Nous avons choisi d'effectuer ce découpage afin d'éviter de régler les algorithmes sur un seul ensemble d'apprentissage (et un autre seul de test), ce qui pourrait conduire à deux travers, le biais expérimental et/ou le phénomène de sur-apprentissage (Torres-Moreno *et al.*, 2007). Les algorithmes ont été évalués sur des corpus de test en utilisant la mesure Fscore (5) des documents bien classés, moyennée sur toutes les classes (avec $\beta = 1$ afin de ne privilégier ni la précision ni le rappel)(Goutte & Gaussier, 2005).

$$\text{Fscore}(\beta) = \frac{(\beta^2+1) \times \langle\text{Précision}\rangle \times \langle\text{Rappel}\rangle}{\beta^2 \times \langle\text{Précision}\rangle + \langle\text{Rappel}\rangle} \quad (5)$$

5.1 Classification CV vs. Lettre de motivation

Le tableau 4 présente les différentes statistiques qui ont permis de construire les classifieurs naïfs. Le premier classifieur choisit la classe en fonction de la longueur moyenne des phrases dans le document tandis que le second d'après le nombre de mots rencontrés. On remarque que malgré un nombre de documents identiques et une différence importante dans leur nombre, la moyenne des phrases est a peu près identique entre les deux documents, ce qui explique les résultats du tableau 4 (l'ensemble des documents ont été classés LM pour le classifieur naïf sur la longueur de phrase), ainsi que le peu de "." présent dans un CV, contrairement au LM. Malgré une différence significative entre les moyennes de mots contenus dans chaque document (425 pour les CV et 190 pour les LM), le second classifieur se heurte à l'hétérogénéité des documents dans leur style et leur longueur. Le tableau 5 présente une moyenne de la précision rappel ainsi que le Fscore obtenu pour la tâche de classification de CV/Lettre de motivation

	CV	LM
Nombre de documents	2165	2165
Nombre de phrase total	45655	20658
Longueur moyenne des phrases	17.07	18.97
Nombre de Mots total	922103	412008
Moyenne de mots par documents	425.91	190.30

TAB. 3 – Statitistiques à la base des classifieurs naïfs.

Classifieur	Précision	Rappel	Fscore
Longueur moyenne des phrases	1	0.50	0.66
Nombre de mots	0.35	0.26	0.30

TAB. 4 – Précision, Rappel, Fscore obtenus par les deux classifieurs naïfs.

sur chacun des sous-corpus par les SVM. Le tableau 6 montre la matrice de confusion. Une analyse des CV/Lettre de motivation mal étiquetés montre que les CV mal classés sont de deux types : l'exemple 7 montre le cas d'un mauvais étiquetage dans le *Corpus de référence*, puisque le document contient plus vraisemblablement une lettre de motivation et un lien vers le CV. L'exemple 8, étiqueté LM, est un message généré automatiquement par des sites d'emploi, ceux-ci contenant des versions très courtes du CV avec un lien vers une version complète.

5.2 Classification selon le CV d'une candidature

Afin d'évaluer nos méthodes de classification d'une candidature, et plus particulièrement les CV, nous avons effectué une scission du *Corpus de référence* en plusieurs sous-corpus : un sous-corpus contenant les CV classés en fonction d'une évaluation OUI/NON (désigné comme *Corpus OUI/NON*) ainsi que deux sous-corpus thématiques, afin de tester l'influence du métier sur les caractéristiques de la candidature (c'est à dire, les CV sont-ils indépendants du métier ?). Ceux ci contiennent l'ensemble des CV répondant à des missions de type "commercial" (nommé *corpus commercial*, avec 715 CV) et "comptable" (*corpus comptable*, avec 1546 CV). Le tableau 9 présente les résultats obtenus par les différents noyaux sur le *Corpus OUI/NON* ainsi qu'un test sans racinisation. Le CV étant généralement composé de mots simples et avec peu d'ambiguïté. Le tableau 10 présente les résultats obtenus par les SVM et le calcul de probabilité des n-grammes. Le tableau 11 montre la répartition des erreurs pour chacun des classifieurs. L'observation de ces résultats nous a poussé à envisager un combinaison de classifieurs, les SVM ayant de meilleurs performances sur la classe NON (375 documents bien classés contre 115 pour la méthode probabiliste) alors que les n-grammes classent mieux la classe OUI (107 documents bien classés contre 38 pour les SVM). Le tableau 12 présente le résultat d'un mixage simple entre les SVM et les n-grammes sur la base d'un vote. On observe une très légère amélioration des performances globales (Fscore de 0,66 pour le mixage contre 0,62 pour les SVM et 0,61 pour la méthode probabiliste).

	$A1$	$A2$	$A3$	$A4$	$A5$	Total
Précision	0,98	0,98	0,97	0,98	0,99	0,98
Rappel	0,95	0,95	0,97	0,95	0,97	0,96
Fscore	0,97	0,97	0,97	0,97	0,98	**0,98**

TAB. 5 – Précision, Rappel, Fscore obtenu par les SVM pour la classification de CV/LM.

	Documents type CV	Documents type LM
Documents classés CV	**421**	12
Documents classés LM	6	**428**

TAB. 6 – Matrice de confusion SVM.

6 Conclusion et perspectives

Le traitement des offres d'emploi est une tâche difficile car l'information est en format libre malgré une structure conventionnelle. Ces travaux ont mis en avant le module de traitement des réponses à des offres d'emplois, second module du projet E-Gen, système pour le traitement automatique des offres d'emploi sur Internet. Après différentes étapes de filtrage et de racinisation et de production d'une représentation vectorielle, nous effectuons une classification sur les différentes pièces de la candidature (CV/ Lettre de motivation). Les résultats obtenus sur la tâche de classification de CV/Lettre de motivation par les SVM sont de très bonne qualité (Fscore moyen de 0,98) nous ont permis de commencer les travaux sur la catégorisation du CV d'une candidature. Les résultats mitigés obtenus par les différents classifieurs pour cette tâche nous a fait envisager la mise en place d'une solution mixe utilisant un simple vote. Nous avons observé une très légère amélioration des performances du système mais nous envisageons d'affiner prochainement celui-ci ainsi que de tester d'autres outils tels que boostexter (Schapire & Singer, 2000). Nous prévoyons par ailleurs d'augmenter la taille de notre modèle n-grammes ainsi qu'un lissage pour gérer le problème des événement non vus. De plus, la classification de CV ne représente qu'une partie de l'évaluation globale de la candidature puisque nous souhaitons la coupler avec les informations de la lettre de motivation et du profil du poste. Nous envisageons cependant la mise en place d'un système d'évaluation de CV sur le portail emploi *jobmanager*[11] Les prochaines étapes consisteront donc à évaluer les lettres de motivations et la candidature de façon globale en tenant compte de divers paramètres tels que la richesse du vocabulaire, l'orthographe ainsi que sa correspondance avec l'offre d'emploi (premier module du système).

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

Mr ARVAUX Pierre

45 rue DE CHANTECLAIR 69440 VANNES. Tél 06.06.06.06.06

A la recherche d'un autre emploi, je me permets de vous adresser ma candidature pour le poste de Directeur d'hôtel car je pense correspondre au profil souhaité.En effet j'ai acquis une solide expérience en ma qualité de Responsable de Centre de Profit ainsi que Directeur de Cafétéria. reconnu,homme de terrain, j'ai un sens du commerce trés prononcé, j'ai managé jusqu'à 5O collarateurs.Je vous laisse le soin d'étudier ma candidature et me tiens à votre disposition pour de plus amples renseignements.

Le CV du candidat est consultable à l'adresse suivante :$http : ///CV?code = 3D - 1789031296_1 9543181$

TAB. 7 – 1er exemple de CV mal classé

M. Zidounet Albert 4 rue de la Corniere 42490 Fraisage
akzeddoun@yahoo.fr Portable : 0606060606
Salaire souhaité : 21,000.00 EUR par an
Type d'emploi :Temps Plein Mobile géographiquement : non
Niveau d'études : Maîtrise, IEP, IUP, Bac + 4
Dernière expérience professionelle : 2002 à 2004 : Cabinet d'Expertise Comptable "Cofis" - Assistant en comptabilité
Le CV du candidat est consultable à l'adresse suivante :$http : ///CV?code = 130493543$

TAB. 8 – 2ème exemple de CV mal classé

Noyau SVM	Précision	Rappel	Fscore
Linéaire	0.60	0.61	**0.61**
Polynomial	0.57	0.57	0.57
Radial	0.57	0.55	0.56
Sigmoidal	0.54	0.54	0.55
linéaire sans racinisation	0.57	0.58	0.58

TAB. 9 – Précision, Rappel, Fscore obtenus par les SVM en fonction du noyau.

	Précision		Rappel		Fscore	
Classifieur	SVM	n-grammes	SVM	n-grammes	SVM	n-grammes
Corpus OUI/NON	0,62	0,62	0,62	0,59	0,61	0,61
corpus commercial	**0,66**	0,62	**0,64**	0,57	0,58	0,58
corpus comptable	0,57	0,61	0,59	0,63	**0,64**	**0,64**

TAB. 10 – Précision, Rappel, Fscore obtenu sur chaque corpus.

	Documents de type OUI		Documents de type NON	
	SVM	n-grammes	SVM	n-grammes
Documents classés OUI	38	50	107	309
Documents classés NON	85	375	19	115

TAB. 11 – Matrice de confusion.

Classification par Mixage	(OUI)	(NON)	(Toutes classes)
Précision	0,53	0,83	0,68
Rappel	0,38	0,90	0,64
Fscore	**0,44**	**0,90**	**0,66**

TAB. 12 – Précision, Rappel, Fscore obtenu par mixage des SVM et n-grammes.

Références

BEAUFORT R., DUTOIT T., PAGEL V. & MONS M. (2002). Analyse syntaxique du français. Pondération par trigrammes lissés et classes d'ambiguïté lexicales. *TALN 2002*.

BIZER R. H. & RAINER E. (2005). Impact of Semantic web on the job recruitment Process. *International Conference Wirtschaftsinformatik*.

BOURSE M., LECLÈRE M., MORIN E. & TRICHET F. (2004). Human resource management and semantic web technologies. *ICTTA*.

DAMASHEK M. (1995). A gauging similarity with n-grams : Language independent categorization of text. *Science 267*, p. 843–848.

EL-BÈZE M., TORRES-MORENO J. & BÉCHET F. (2005). Un duel probabiliste pour départager deux Présidents. *RNTI*.

FAN R.-E., CHEN P.-H. & LIN C.-J. (2005). Towards a Hybrid Abstract Generation System, Working set selection using the second order information for training SVM. *NIPS 2005*, p. 1889–1918.

GOUTTE C. & GAUSSIER E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *ECIR 2005*, p. 345–359.

GRILHERES B., BRUNESSAUX S. & LERAY P. (2004). Combining classifiers for harmful document filtering. *RIAO*, p. 173–185.

HEITZ T. (2008). Modélisation du prétraitement des textes . *JADT2006*.

JOACHIMS T. (2008). Text categorization with Support Vector Machines : Learning with many relevant features . *European Conference on Machine Learning*, p. 137–142.

KESSLER R. & EL-BÈZE M. (2008). E-Gen : traitement automatique des offres d'emploi. *JADT2008*, p. 591–601.

KESSLER R., TORRES-MORENO J. M. & EL-BÈZE M. (2007). E-Gen : Automatic Job Offer Processing system for Human Ressources. *MICAI*.

MORIN, EMMANUEL ET LECLÈRE M. E. T. F. (2004). The semantic web in e-recruitment (2004). *The First European Symposium of Semantic Web ESWS*.

PHAM N.-K. & DO T.-N. (2003). Fouille de textes à l'aide de ProximalSVM. *9th national conference in computer science Vietnam*.

PLANTIÉ M., DRAY G. R. M. (2007). Comparaison d'approches pour la classification de textes d'opinion. *3ème défi fouille de textes DEFT 2007*, p. 55–68.

RAFTER R., BRADLEY K. & SMYT B. (2000). Automated Collaborative Filtering Applications for Online Recruitment Services. *RIAO*, p. 363–368.

RAFTER R., SMYTH B. & BRADLEY K. Inferring Relevance Feedback from Server Logs : A Case Study in Online Recruitment.

ROCHE M. & PRINCE V. (2008). Evaluation et détermination de la pertinence pour des syntagmes candidats à la collocation . *JADT2008*, p. 1009–1020.

SCHAPIRE R. E. & SINGER Y. (2000). BoosTexter : A boosting-based system for text categorization. *Machine Learning*, **39**(2/3), 135–168.

TORRES-MORENO J., EL-BÈZE M., BÉCHET F. & N C. (2007). Comment faire pour que l'opinion forgé à la sortie des urnes soient la bonne ? *Actes DEFT2007*.

VAPNIK V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

ZIGHED, D. A. E. C. J. (2003). Data Mining et analyse des CV : une expérience et des perspectives. *Journées sur l'Extraction et la Gestion des Connaissances, Lyon EGC 2003*.