Classification d'offres d'emploi

Annette Casagrande[*,] — Fabrizio Gotti[*] — Guy Lapalme[*]**

[*] *RALI - Département d'informatique et de recherche opérationnelle*
Université de Montréal
C.P. 6128, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3J7
[**] *CNRS - PACTE UMR Pacte Institut d'études politiques BP 48 38040 Grenoble cedex 9, France*
annette.casagrande@univ-grenoble-alpes.fr {gottif,lapalme}@iro.umontreal.ca

RÉSUMÉ. *Les ressources humaines utilisent de plus en plus les données intelligentes et les techniques du big data pour faciliter le recrutement. Ainsi, grâce aux profils des réseaux sociaux, les recruteurs peuvent identifier des candidats potentiels qui ne sont pas actifs en termes de recherche d'emploi mais qui pourraient être quand même intéressés par une opportunité. Leur intérêt pour une offre non sollicitée est d'autant plus grand lorsque cette dernière correspond bien à leur profil et à leur secteur d'activité. Afin d'améliorer les résultats d'un tel système de recommandation appariant offres d'emploi et profils suivant les compétences et les expériences requises, nous proposons de détecter automatiquement le secteur d'activités des offres à l'aide de techniques d'apprentissage supervisé.*

ABSTRACT. *The field of Human Resources is increasingly using smart data and big data techniques to facilitate recruitment. For instance, on social media, recruiters can identify potential candidates who are not actively engaged in seeking a new job, but may nonetheless be interested in an opportunity. Their interest in such an opportunity is naturally greater if it fits their profile and industry sector. To improve the results of such a recommendation system matching job offers with profiles according to the skills and experiences required, we propose to automatically detect the industry sector in job offers using supervised machine learning techniques.*

MOTS-CLÉS : *Systèmes de recommandation, Classification automatique, E-recrutement.*

KEYWORDS: *Recommendation systems, Automatic classification, E-recruitment.*

1. Introduction

L'utilisation des données intelligentes, du web sémantique ainsi que des techniques du big data se développe dans de nombreux domaines, incluant les ressources humaines. Les progiciels de recrutement, apparus dans les années 1990, s'adaptent aux besoins des entreprises en termes de gestion des offres et des candidatures (Fondeur *et al.*, 2011), et tirent profit notamment de ces données.

Dans cette lignée, le projet Butterfly Predictive Project (BPP), réalisé en partenariat avec LittleBigJob, a pour objectif le développement d'une plate-forme pour améliorer le processus de recrutement de professionnels. Cette plate-forme doit permettre :

– d'augmenter la capacité à récolter et à combiner l'information de multiples sources (dont les données publiques sur les médias sociaux) ;

– d'améliorer la mise en correspondance entre des candidats et des offres d'emploi ;

– de prédire le succès d'un candidat dans un nouveau poste.

LittleBigJob souhaite plus particulièrement proposer à ses clients un moteur qui permette l'identification et le recrutement de candidats dits *passifs*. Ces candidats passifs sont déjà en poste, ne sont pas en recherche active d'emploi, mais pourraient néanmoins être intéressés par une nouvelle opportunité. Ces profils sont très convoités par les recruteurs (Charrière *et al.*, 2014). Cependant, ces derniers rencontrent souvent des difficultés lors de leur recherche de candidats passifs. Le fait de devoir contacter et convaincre le candidat de considérer leur offre contraint en effet les recruteurs à adopter une posture commerciale qui, d'après notre partenaire, n'est pas naturelle. Ce n'est qu'une fois que le candidat a répondu que ce dernier devient *actif* et que les stratégies de recrutement habituelles peuvent être utilisées. Le projet BPP vise notamment à simplifier et à automatiser ce processus.

Les candidats passifs peuvent être repérés à l'aide de leur profil sur les réseaux sociaux car ils y publient des informations professionnelles (voir Figure 1) :

– formations : diplômes, établissements fréquentés, dates de fréquentation, etc.

– expériences : emplois occupés, employeurs, tâches effectuées, etc.

– compétences : savoir-faire et savoir-être, ou compétences spécialisées et générales, etc.

Le projet BPP propose d'apparier les profils issus de réseaux sociaux avec des offres d'emplois. Dans la littérature, les travaux proposant un appariement entre offres et candidats proposent des solutions de type *système de recommandation* (Gu *et al.*, 2016 ; Diaby et Viennet, 2014). Les systèmes de recommandations aident les utilisateurs à gérer la surcharge d'informations en leur proposant des items adaptés à leur besoin. Pour cela, il existe trois familles d'algorithmes : ceux basés sur le *filtrage collaboratif* (Resnick *et al.*, 1994 ; Sarwar *et al.*, 2001), ceux sur le *filtrage sur le contenu* (Pazzani et Billsus, 2007) et les hybrides (Burke, 2007). Les systèmes de recommandation collaboratifs identifient les personnes présentant des goûts similaires et pro-

Formation	2015-2015 2003-2005	ESGI Infosup
Expérience	2008-à ce jour	Responsable informatique : Cequami - Certivea
Compétences	Microsoft Office, Anglais, Microsoft Windows, Cloud Computing, Virtualisation, Cisco Technologies, ITIL, Linux, Microsoft Exchange, VoIP	

Figure 1. *Extrait d'un profil de candidat.*

posent aux personnes des produits en fonction des évaluations laissées par leurs plus proches voisins. Le filtrage basé sur le contenu s'appuie sur les informations données par l'utilisateur (profil par exemple) pour lui proposer des items pertinents. La principale critique du filtrage sur le contenu est l'effet "entonnoir" (Poirier *et al.*, 2010) : l'utilisateur finit par recevoir seulement les recommandations en lien avec son profil. Pour la proposition d'offre, cet effet est très utile : ne présenter que des offres adaptées au profil est très important pour ne pas contacter inutilement une personne. Quant aux systèmes de recommandation hybrides, ils combinent les deux approches : utilisation des évaluations et des informations données par l'utilisateur (Burke, 2002).

Les recherches proposant un appariement entre les exigences contenues dans une offre et un candidat se sont concentrées sur les compétences et les expériences compatibles entre elles (Colucci *et al.*, 2013 ; Trichet *et al.*, 2004 ; Yahiaoui *et al.*, 2006). Dans cette voie et pour le projet BPP, (Dieng, 2016) propose WordMatcher et SkillFinder. Ces deux algorithmes mettent en correspondance une offre d'emploi et une liste de candidats : pour cela, les informations pertinentes d'une offre d'emploi sont extraites et sont intégrées à une requête pour rechercher les profils adéquats. SkillFinder a été intégré dans un système de recommandation, nommé Cerebra, proposant pour une offre donnée une liste de candidats (cf. figure 2). Néanmoins, parmi les limites de ses travaux, (Dieng, 2016) cite l'importance de proposer des candidats issus du même secteur d'activité que l'offre. Cette préférence pour un secteur s'explique en partie par le fait qu'un recrutement a un coût (Florea *et al.*, 2013) et par conséquent les recruteurs cherchent à limiter leur incertitude et ainsi éviter un mauvais recrutement : d'après notre partenaire, un candidat issu du même secteur que la compagnie qui recrute rassure le recruteur tant sur ses compétences que sur son intérêt à obtenir le poste et à le conserver.

Pour pallier ce problème, notre partenaire a proposé une typologie des secteurs d'activité (décrite à la section 2.2) ; chacun des 48 éléments de cette typologie est appelé *univers* . Lors de la proposition de candidats pour une offre, un *career manager* (recruteur) trie les profils en ne conservant que ceux appartenant au même univers que l'offre. Cette opération s'avère cependant longue, fastidieuse et sujette aux erreurs. Notre objectif est d'optimiser cette opération en déterminant automatiquement l'univers associé à une offre ou à un profil. Une offre d'emploi n'appartient qu'à un et un seul univers tandis qu'un profil peut être associé à différents univers en fonction de

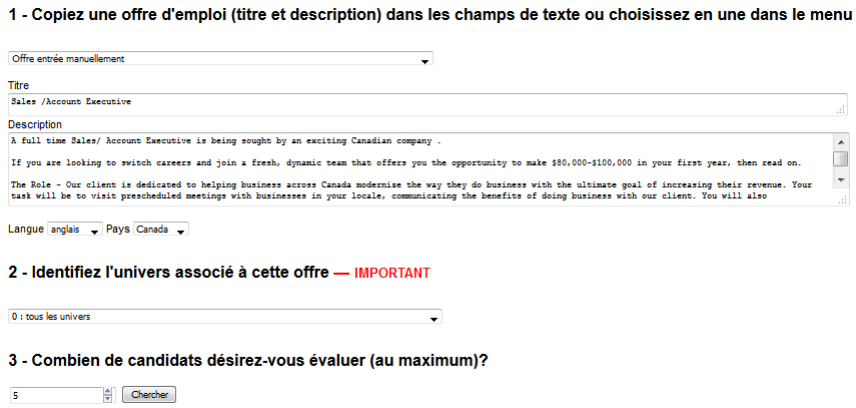

Figure 2. *Système de recommandation utilisant SkillFinder*

ses expériences professionnelles. Dans cette communication, nous nous intéressons à la détection de l'univers des offres d'emplois.

Nous expliquons d'abord la méthodologie de notre recherche (section 2). Puis nous présentons les résultats de nos expériences (section 3). Nous présentons ensuite l'intégration de nos travaux dans notre système de recommandation. Enfin nous concluons et proposons différentes pistes de recherches.

2. Méthodologie

2.1. *Catégorisation des offres*

L'affectation automatique d'un univers à une offre est un problème de catégorisation de documents. Il s'agit d'affecter automatiquement une classe (un univers) à un document (une offre d'emploi). Nos classes ayant été définies par notre partenaire, nous utiliserons des techniques liées à l'apprentissage machine supervisé. L'apprentissage supervisé, à partir des textes déjà classés, cherche à apprendre des règles pour classifier de nouveaux textes. Les étapes principales pour réaliser un apprentissage supervisé sont :

1) définition des classes ;

2) création d'un corpus d'apprentissage ;

3) apprentissage d'un modèle de classification à l'aide du corpus d'apprentissage ;

4) évaluation du modèle.

2.2. *Définition des classes*

Bien qu'il existe de nombreuses typologies des secteurs d'activités (ESCO (Le Vrang *et al.*, 2014), codes ROME[1], CNP-NOC[2], etc.), notre partenaire industriel a défini une typologie adaptée à ses besoins comportant 48 secteurs d'activités baptisés *univers*. Le tableau 1 donne quelques exemples.

Id	Nom	Univers apparentés
2	Aéronautique et aérospatiale	9 Industries automobiles 18 Digital, e-commerce, big data, jeux électroniques 17 Défense et armement, police, sécurité, transport de fonds 20 Énergie, eau, nucléaire, pétrole, gaz 42 Télécoms, hébergement, internet
13	Conseil et services informatiques, édition de logiciels	18 Digital, e-commerce, big data, jeux électroniques 42 Télécoms, hébergement, internet 10 Banque, finance, capital risque, fonds privés 22 Équipements électriques et électroniques, composants, matériel informatique 8 Audiovisuel, cinéma, spectacles, média, publicité, événementiel, divertissement, communication
18	Digital, e-commerce, big data, jeux électroniques	42 Télécoms, hébergement, internet 13 Conseil et services informatiques, édition de logiciels 10 Banque, finance, capital risque, fonds privés 8 Audiovisuel, cinéma, spectacles, média, publicité, événementiel, divertissement, communication 22 Equipements électriques et électroniques, composants, matériel informatique
40	Services aux particuliers (cours, ménage...)	37 Retail, grande distribution, distribution généraliste et spécialisée 10 Banque, finance, capital risque, fonds privés 18 Digital, e-commerce, big data, jeux électroniques 42 Télécoms, hébergement, internet 8 Audiovisuel, cinéma, spectacles, média, publicité, événementiel, divertissement, communication

Tableau 1. *Quelques-uns des 48 univers extraits de la typologie de LittleBigJob, ainsi que leurs univers apparentés selon les transitions entre univers observées dans les profils de candidats.*

Les univers ne sont pas indépendants. En effet, certaines offres se voient comblées par des candidats issus de plusieurs univers différents, et certains candidats passent d'un univers à un autre durant leur parcours professionnel. Ce dernier élément est par-

1. Répertoire Organisationnel des Métiers et Emplois (France)
2. Classification Nationale des Professions - National Occupational Code (Canada)

ticulièrement utile dans cette étude. En analysant les *flux migratoires* entre univers, il nous est possible de mesurer indirectement le degré de parenté entre les univers. Ainsi, les passages relativement nombreux depuis l'univers 13 *Conseil et services Informatiques, édition de logiciels* vers l'univers 18 *Digital, e-commerce, big data, jeux électroniques* indiquent une relation relativement étroite entre les deux. À l'inverse, passer de l'univers 40 *Services à la personne* à l'univers 2 *Aéronautique et aérospatiale* est plus rare, et montre la distance entre ces univers.

Pour faire cette analyse de transitions entre univers, nous avons procédé à une étude de 10 M de profils de candidats à notre disposition. Pour chacun d'eux, nous avons repéré les expériences professionnelles contenant une date de début, une date de fin et un univers. L'univers d'une expérience est l'univers de la société dans laquelle elle s'est déroulée. On utilise pour ceci une correspondance entre société et univers telle que décrite à la section 2.3.2.2. Une fois fait, on peut trier les expériences en ordre chronologique et compter les transitions d'un univers à l'autre au cours de la vie professionnelle de chaque candidat. Nous avons recensé ainsi 2M de transitions suffisamment renseignées pour faire partie de nos statistiques. À partir de ceci, nous avons choisi de considérer, pour chaque univers u, les 5 univers *ancêtres* qui comptent le plus de transitions vers u. Nous les appelons *univers apparentés*.

2.3. *Corpus d'apprentissage*

2.3.1. *Description des données*

Pour développer notre classifieur, nous disposons de 268k offres dont 139k en anglais et 123k en français, collectées sur les sites internet d'Indeed, de Cadremploi et de l'APEC. Chaque offre comporte les informations suivantes :

– un identifiant unique ;

– le poste à pourvoir : statut du poste, profil du candidat (champ *profile*), description de l'offre (champ *description*), type de contrat ;

– la société procédant au recrutement ;

– le lieu d'exercice ;

– la date de mise à jour de l'offre ;

– la personne à contacter et l'URL de l'offre ;

– d'autres champs tels que le secteur d'activité ou le descriptif de la société.

Selon les offres, le descriptif de l'offre peut être écrit soit dans le champ *description*, soit dans le champ *profile* soit dans les deux. Dans la suite de cette communication, nous utiliserons description pour la concaténation des deux champs (même si l'un des deux est vide).

2.3.2. *Corpus d'apprentissage*

Suite aux travaux de (Dieng, 2016), pour vérifier si les résultats de notre système

{ **"profile"** : "Vous justifiez de 5 ans d'expérience dans les métiers de l'eau où vous avez pu développer vos capacités managériales, techniques et commerciales. Rigoureux(se), autonome, vous maîtrisez le pack office et êtes titulaire du permis B.",
"title" : "Chef de Secteur H/F",
"company_name" : "Grand Sud Emploi",
"company_desc" : "Une entreprise moderne au plus proche des territoires. L'activité EAU FRANCE, cœur du métier de Saur et regroupant près de 7000 collaborateurs, assure, pour le compte des collectivités de toutes tailles, les activités de production, distribution et traitement d'eau, l'assainissement et l'épuration des eaux usées ainsi que la construction d'ouvrages. Maillant le territoire au travers de ses Directions Régionales, SAUR est une entreprise à l'écoute de ses clients et proche de ses collaborateurs. Guidée par des valeurs d'entreprise fortes, innovante dans sa gestion des métiers de l'eau, nous affichons une ambition de développement. Nous recrutons un : Chef de Secteur (H/F). Voir toutes les offres de l'entreprise",
"description" : "Rattaché au Chef d'Agence, vous êtes le pilote de la performance commerciale et technique par la qualité de votre management de votre équipe et votre proximité avec les collectivités. Vous êtes responsable de la bonne exécution des contrats, de la prévention et la sécurité des collaborateurs, du développement des activités dans votre secteur et des dépenses placées sous votre initiative. Poste à pourvoir en CDI en Charente (16)." }

Tableau 2. *Extrait d'une offre d'emploi.*

de recommandation Cerebra étaient améliorés en ajoutant l'univers aux critères de recherche, une liste déroulante permettant le choix de l'univers de l'offre a été ajoutée. Cette liste propose les 48 univers de notre typologie. L'analyse des offres soumises au système par les recruteurs humains a permis de dégager deux stratégies qu'utilisent les recruteurs pour déterminer l'univers de l'offre à pourvoir : soit l'univers est déterminé à l'aide du titre d'emploi soit il l'est à l'aide de l'univers de la compagnie. Par exemple, un analyste financier recruté par Bombardier sera classé :

– dans l'univers 2 *Aéronautique et aérospatiale* si on considère l'univers de la compagnie Bombardier ou

– dans l'univers 10 *Banque, finance, capital risque, fonds privés* si nous nous référons à notre référentiel liant titre d'emploi et univers (cf. section 2.3.2.1).

Nous expérimentons avec ces deux stratégies dans ce qui suit. Notons par ailleurs que, pour chaque corpus d'apprentissage que nous constituons, nous avons décidé de ne garder que les offres des univers qui comptent plus de 10 offres.

2.3.2.1. Univers selon le titre d'emploi

Dans le cadre du projet BPP, (Kessler *et al.*, 2016) ont développé le système Agorha, qui construit automatiquement une ontologie des compétences pour un certain nombre d'emplois répartis par univers (cf. figure 3). 296 titres d'emplois ont été affectés à un univers ; pour 173, nous avons la version en anglais et en français et pour 128 nous avons en français la version masculine et la version féminine. Ces emplois recensés ne couvrent que 27 univers. Ainsi, à l'aide des liens entre des titres d'emploi

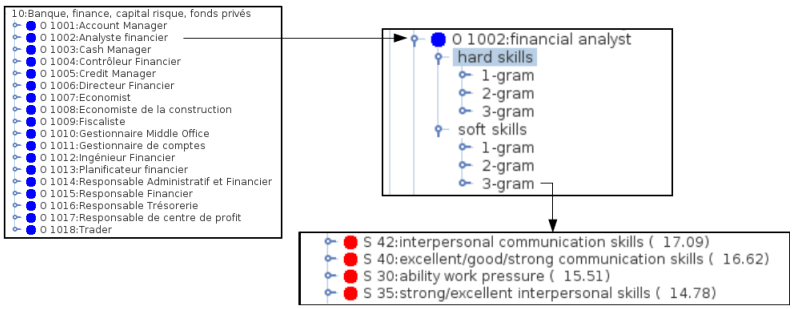

Figure 3. *Ontologie des compétences par métiers et par univers*

et des univers, nous avons pu créer deux corpus d'apprentissage : un en anglais composé de 28k offres (corpus ATA) et un en français avec 17k offres (corpus ATF). Notre corpus en anglais couvre 23 univers et celui en français couvre 24 univers.

2.3.2.2. Univers selon le nom de la compagnie

Pour les besoins du projet, nous avons construit une correspondance entre le nom d'une société et l'univers principal dans lequel la société évolue (par exemple BASF se voit affecter l'univers 12 *Chimie, caoutchouc, plastique*). Pour construire cette ressource, nous sommes partis de 6 M de profils de sociétés présents sur les sites de LinkedIn et Viadeo, et mis à notre disposition par notre partenaire commercial. Ces descriptions de sociétés contiennent, dans 6% des cas, un champ *secteur d'activité* renseigné (par exemple *Chemicals*). Il suffit dès lors de faire le passage du secteur d'activité vers l'univers à l'aide d'un mappage manuellement établi, depuis les 269 secteurs recensés vers nos 48 univers. La correspondance finale permet de trouver l'univers de 300k sociétés, petites et grandes, et réparties partout dans le monde.

Cette ressource a ses limites, notamment lorsque plusieurs sociétés partagent le même nom. Pour le moment, on se contente de choisir arbitrairement un des univers possibles parmi les possibilités offertes par nos ressources.

Les offres d'emploi comportent un champ contenant le nom de la compagnie : pour les offres en anglais, ce champ est renseigné à 98% et pour le français à 95%. A partir des correspondances entre compagnies et univers, nous avons constitué deux corpus d'apprentissage : un en anglais composé de 56k offres (corpus ACA) et un en français

avec 31k offres (corpus ACF). Nos corpus couvrent 45 univers. Nous n'avons pas retenu les offres pour lesquelles la compagnie reconnue était un cabinet de recrutement (univers 45 *Ressources humaines*) : les cabinets de recrutement proposent la plupart du temps des offres d'emplois pour des compagnies clientes (qui n'appartiennent pas à l'univers 45).

2.3.3. *Corpus de test*

Pour évaluer nos modèles d'apprentissage, nous avons également créé des corpus de tests (pour le français) :

– un corpus basé sur le titre d'emploi : 182 offres étiquetées manuellement à partir du titre d'emploi (corpus TTF)

– deux corpus basés sur le nom de la compagnie :

- nous avons retiré 200 offres de notre corpus étiqueté à l'aide du nom de l'entreprise (corpus TCF1)

- 100 offres étiquetées manuellement à partir du nom de la compagnie (offres pour lesquelles le nom de la compagnie n'a pas été trouvé dans notre fichier de correspondance) (corpus TCF2)

Langue	Description	Nom	Nb. offres
Anglais	Corpus entier	Corpus EA	139k
	Corpus d'apprentissage titre d'emploi	Corpus ATA	28k
	Corpus d'apprentissage compagnie	Corpus ACA	56k
Français	Corpus entier	Corpus EF	123k
	Corpus d'apprentissage titre d'emploi	Corpus ATF	17k
	Corpus d'apprentissage compagnie	Corpus ACF	31k
	Corpus de test titre d'emploi	Corpus TTF	182
	Corpus de test compagnie	Corpus TCF1	200
	Corpus de test compagnie	Corpus TCF2	100

Tableau 3. *Description des corpus utilisés dans cette étude.*

2.4. *Apprentissage supervisé*

2.4.1. *Modèles retenus*

Nous avons utilisé Weka[3] pour faire l'apprentissage machine supervisé d'un modèle permettant la catégorisation automatique des offres d'emploi. Après exploration, nous avons retenu quatre algorithmes parmi les plus couramment utilisés :

C4.5 (Quinlan, 2014) est une méthode d'apprentissage qui produit un arbre de décision et qui se base sur la mesure d'entropie.

3. `http://www.cs.waikato.ac.nz/ml/weka/`

K-plus-proches voisins (KNN) classe une instance de test à l'aide des instances qui lui sont le plus similaires. L'algorithme calcule pour un individu sa distance à tous les autres individus de la base et identifie la classe majoritaire parmi les k individus ayant la plus petite distance. Nous avons choisi de prendre $k = 1$.

Forêt d'arbres de décision (RF) (Breiman, 2001) : plusieurs arbres de décision sont entraînés sur des sous-ensembles de données légèrement différents. On procède ensuite à un vote majoritaire pour obtenir la classe prédite. Nous avons choisi de considérer 100 arbres de décision.

SMO (sequential minimal optimization) (Platt, 1999) est un algorithme pour résoudre le problème d'optimisation quadratique qui se produit lors de l'entraînement des machines à vecteurs de support (SVM). Les SVM cherchent à déterminer l'*hyperplan de marge optimale* qui sépare correctement les données de classes distinctes.

2.4.2. *Choix des caractéristiques textuelles*

Ayant observé que les career managers de notre partenaire choisissaient l'univers en fonction soit du titre d'emploi soit du nom de la compagnie, nous avons émis deux hypothèses :

1) Les descriptifs des offres d'emplois dont les titres d'emplois appartiennent au même univers partagent un vocabulaire commun.

2) Les descriptifs des offres d'emplois émises par des compagnies issues du même univers partagent un vocabulaire commun.

Les algorithmes retenus fonctionnent sur des vecteurs où chaque composante correspond à une valeur calculée pour une caractéristique donnée. A partir de nos hypothèses, nous avons choisi de retenir les 1 000 et 1 500 mots les plus discriminants de nos offres. Pour cela, nous avons utilisé des techniques classiques du traitement automatique de la langue. Nous avons procédé de la manière suivante :

1) la description de chaque offre est récupérée.

2) les balises html (par exemple : `<br>` `<p>`) sont supprimées.

3) les offres sont lemmatisées à l'aide de TreeTagger[4].

4) les mots outils sont supprimés à l'aide d'un antidictionnaire.

5) le tf-idf est calculé et ne sont retenus que les 1 000 et 1 500 lemmes ayant les scores les plus élevés. Le tableau 4 présente un extrait des 1 000 lemmes retenus pour l'anglais et le français.

6) nous construisons un vecteur pour chaque offre contenant les valeurs du tf-idf pour chaque lemme retenu.

4. `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

Anglais	ability, able, aboriginal, academic, accept, access, accessibility, accommodate, accommodation, accord, accordance, account, accountability, accountable, accounting, accuracy, accurate, ...
Français	accompagner, accomplir, accès, achat, acquisition, acquérir, acteur, actif, action, activement, activité, actuel, actuellement, adaptation, adapter, administratif, administration, ...

Tableau 4. *Extraits des 1 000 lemmes retenus pour l'anglais et pour le français*

2.4.3. *Évaluation des modèles*

Pour évaluer la qualité de nos modèles, nous avons choisi :

– une validation croisée en 10 strates ;

– une validation à l'aide des corpus de test (pour le français) déjà présentés plus haut.

3. Résultats de nos expériences

3.1. *Expérience à partir du titre d'emploi*

Le tableau 5 présente les résultats obtenus par validation croisée sur nos données d'apprentissage (corpus ATA et ATF) et les résultats sur le corpus de test TTF.

Nous obtenons de bons résultats en validation croisée comme sur notre corpus de test TTF, avec des taux de précision atteignant 74% en anglais et 84% en français. La SVM (SMO) montre les meilleurs résultats, presque systématiquement. Utiliser 1500 mots plutôt que 1000 s'avère bénéfique, ce qui est attendu. Il faut noter cependant que notre corpus ne couvre pas tous les univers de notre typologie. Il serait donc nécessaire d'enrichir nos correspondances entre titre d'emploi et univers. On observe aussi un gain modeste lorsque l'on tient compte des univers apparentés lors du calcul de l'erreur, ce qui est également cohérent.

3.2. *Expérience à partir du nom la compagnie*

Le tableau 6 présente les résultats obtenus par validation croisée sur nos données d'apprentissage (corpus ACA et ACF) et et les résultats sur les deux corpus de test basés sur le titre de la compagnie (TCF1 et TCF2). Les résultats en validation croisée atteignent 77% pour l'anglais et 74% pour le français, soit respectivement un gain de 3% et une perte de 10% par rapport aux résultats de la section précédente. Ici, c'est la forêt d'arbres de décision qui l'emporte, encore avec 1500 mots.

Anglais	1000 mots			1500 mots		
Algo	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
C4.5	66%	66%	66%	68%	68%	68%
KNN	54%	54%	53%	54%	54%	53%
RF	**72%**	70%	68%	**72%**	70%	68%
SMO	**72%**	**72%**	**72%**	**74%**	**74%**	**74%**
Français						
C4.5	76%	76%	76%	79%	79%	79%
KNN	64%	64%	63%	65%	63%	63%
RF	80%	76%	75%	79%	76%	74%
SMO	**81%**	**81%**	**81%**	84%	**83%**	**83%**

Titre : corpus TTF				
	1000 mots		1500 mots	
	Univ. correct	Univ. apparenté	Univ. correct	Univ. apparenté
C4.5	76%	84%	**84%**	87%
KNN	69%	74%	67%	71%
RF	79%	85%	78%	83%
SMO	**81%**	**87%**	**84%**	**88%**

Tableau 5. *Résultats de la classification pour l'étiquetage à partir du titre d'emploi. Le pourcentage en gras indique le résultat maximal, par colonne.*
univ correct : Pourcentage d'offres étiquetées avec le bon univers.
univ. apparenté : Pourcentage d'offres étiquetées avec le bon univers ou un univers apparenté au bon univers, selon les univers apparentés présentés à la section 2.2

Il faut noter le contraste frappant entre les prédictions acceptables faites sur le corpus de test TCF1 (68%) et celles, très faibles, sur le corpus de test TCF2 (23%). Nous avons cherché à comprendre pourquoi nous avions de tels écarts entre le corpus étiqueté automatiquement à partir des noms de compagnies (TCF1) et celui étiqueté de façon entièrement manuelle (TCF2).

Nous avons commencé par vérifier la qualité de l'étiquetage de notre corpus de test TCF1. Nous avons pris les 100 premières offres de notre corpus TCF1 et avons vérifié si l'univers donné était correct. Dans 80% des cas, l'univers proposé par l'étiquetage automatique nous paraissait correct et dans 5% des cas, il s'agissait d'un univers apparenté.

Nous avons ensuite pensé que le fait de connaître l'univers proposé automatiquement avait pu influencer la vérification décrite au paragraphe précédent. Nous avons donc pris les 100 offres restantes du corpus TCF1 et nous leur avons affecté un univers manuellement à l'aveugle, c'est-à-dire sans consulter l'étiquetage automatique à partir

Validation croisée						
Anglais	1000 mots			1500 mots		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
C4.5	54%	53%	53%	56%	55%	55%
KNN	74%	**72%**	**72%**	73%	**71%**	**72%**
RF	**77%**	69%	68%	**77%**	69%	69%
SMO	67%	67%	67%	70%	70%	70%
Français						
C4.5	48%	47%	47%	49%	49%	49%
KNN	66%	**64%**	**65%**	66%	**63%**	**64%**
RF	**73%**	62%	62%	**74%**	63%	63%
SMO	58%	58%	58%	61%	61%	61%

Corpus de test TCF1				
	1000 mots		1500 mots	
	Univ. correct	Univ. apparenté	Univ. correct	Univ. apparenté
C4.5	47%	64%	54%	69%
KNN	**68%**	78%	**68%**	75%
RF	66%	**84%**	65%	**85%**
SMO	61%	74%	67%	77%

Corpus de test TCF2				
	1000 mots		1500 mots	
	Univ. correct	Univ. apparenté	Univ. correct	Univ. apparenté
C4.5	14%	36%	10%	34%
KNN	7%	24%	7%	24%
RF	15%	**48%**	15%	**49%**
SMO	**18%**	38%	**23%**	40%

Tableau 6. *Résultats de la classification pour l'étiquetage à partir du nom de la compagnie.*
Le pourcentage en gras indique le résultat maximal, par colonne.
univ correct : Pourcentage d'offres étiquetées avec le bon univers.
univ. apparenté : Pourcentage d'offres étiquetées avec le bon univers ou un univers apparenté au bon univers, selon les univers apparentés présentés à la section 2.2

du nom de la société. Un fois fait, nous avons comparé l'étiquetage manuel et celui automatique. Nous étions d'accord pour 61 offres.

Une partie de la réponse sur les piètres performances sur TCF2 tient donc à l'ambiguïté de la tâche d'attribution d'un univers. Nous avons deux perspectives pour répondre à la question "quel univers va avec quelle entreprise ?" : celle de nos données

de correspondances et celle de l'étiqueteur humain. Nos données de correspondance sont une vision fragmentaire car nous n'avons pas toutes les entreprises existantes. Des erreurs d'étiquetage ont été observées : par exemple, la société *Advitam* est considérée comme faisant partie de l'univers 40 *Services aux particuliers* mais nous avons une offre qui émane d'une autre société nommée également *Advitam* issue de l'univers 3 *Agriculture*. La perspective de l'étiqueteur manuel présente aussi des risques d'erreur. Son étiquetage va dépendre de son niveau d'expertise pour chacun des univers : par exemple, une société de services informatique dépend-elle de l'univers 13 *Conseil et services informatiques, édition de logiciels* ou de l'univers 18 *Digital, e-commerce, big data, jeux électroniques* ?

Cela dit, les résultats issus du corpus d'apprentissage ACF et ceux issus du corpus de test TCF1 sont cohérents et nous laissent penser que la démarche par apprentissage supervisé est une voie pertinente pour déterminer l'univers d'une offre d'emploi. L'écart entre les résultats de TCF1 et de TCF2 nous incite à améliorer notre méthode de reconnaissance de la compagnie qui émet l'offre dans l'emploi. D'une part, il nous faut enrichir nos correspondances entre compagnie et univers pour se rapprocher de ce qui est attendu par notre partenaire. Nous devons également déterminer le bon univers lorsque nous avons des entreprises homonymes en utilisant les informations à notre disposition dans l'offre pour trancher.

4. Intégration à Cerebra

Nous présentons à la figure 4 comment la détection de l'univers sera intégrée dans notre système de recommandation Cerebra (cf. figure 2). La liste déroulante de l'interface présentera un choix de six univers au lieu des 48 ; le premier univers de la liste sera le(s) univers détecté ou prédit(s) suivi des univers apparentés. La priorité est donnée à l'univers de la compagnie car selon notre partenaire LittleBigJob, les recruteurs cherchent d'abord des candidats issus de l'univers de leur société.

5. Conclusion

Dans cette étude, nous avons présenté les travaux réalisés pour améliorer le processus de sélection de candidats pour une offre d'emploi donnée. Nous nous sommes penchés sur le développement d'un système de détection de l'univers (le secteur d'activité) de l'offre d'emploi afin de proposer des candidats issus du même secteur. Dans le meilleur des cas, il est possible de détecter correctement l'univers avec une précision de 85% en utilisant des algorithmes classiques d'apprentissage supervisé.

La constitution des corpus d'entraînement et de test est relativement délicate, cependant. Pour étiqueter les milliers d'offres nécessaires à l'apprentissage machine, les stratégies proposées ici sont perfectibles. Se fier au titre du poste ou à la société qui recrute est entaché d'erreur, que les algorithmes d'apprentissage sont condamnés à ré-

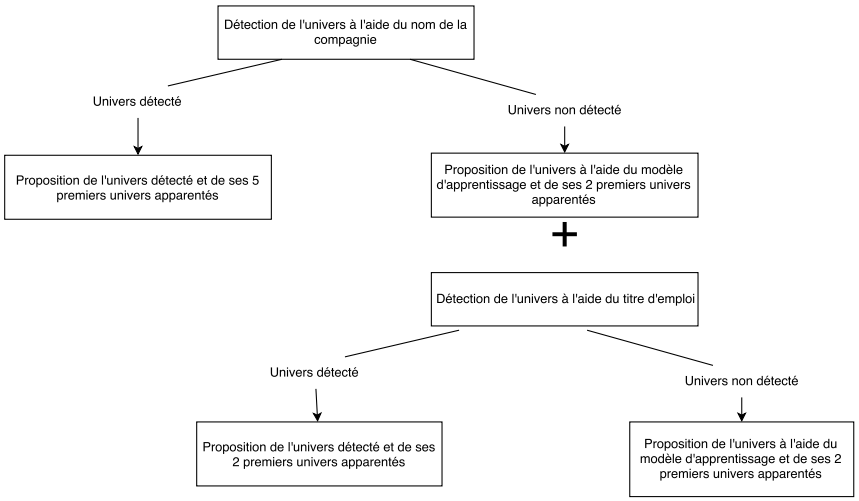

Figure 4. *Proposition de l'univers*

péter. Une façon différente de pré-étiqueter les corpus serait sans doute utile, peut-être en combinant différentes sources d'information afin d'augmenter la précision de ce pré-étiquetage. Dans la foulée, ce processus permettrait peut-être d'identifier d'autres traits ("features") non lexicaux qui pourraient être discriminants. L'offre et les profils de candidats sont en effet très riches en information structurée ou moins structurée.

Cela dit, les performances présentées ici sont forcément de type "intrinsèque", c'est-à-dire qu'elles mesurent exclusivement l'efficacité de classification. Une mesure plus juste du mérite de notre approche consisterait à mesurer l'utilité "extrinsèque" de la classification en univers au sein d'un système complet de recommandation de candidats. Si, comme on l'espère, la composante de classification présentée ici mène à une liste de candidats plus adaptés au poste à pourvoir, alors elle révélera toute son utilité. Une campagne d'évaluation plus complète s'impose donc pour avoir le fin mot sur la qualité de la classification proposée ici.

6. Bibliographie

Breiman L., « Random forests », *Machine learning*, vol. 45, n° 1, p. 5-32, 2001.

Burke R., « Hybrid recommender systems : Survey and experiments », *User modeling and user-adapted interaction*, vol. 12, n° 4, p. 331-370, 2002.

Burke R., « Hybrid web recommender systems », *The adaptive web*, Springer, p. 377-408, 2007.

Charrière V., Dejoux C., Dupuich F., « L'impact des réseaux sociaux et des compétences émotionnelles dans la recherche d'emploi : étude exploratoire », *Management & Avenir*, n° 2, p. 137-163, 2014.

Colucci S., Tinelli E., Giannini S., Di Sciascio E., Donini F. M., « Knowledge compilation for core competence extraction in organizations », *International Conference on Business Information Systems*, Springer, p. 163-174, 2013.

Diaby M., Viennet E., « Développement d'une application de recommandation d'offres d'emploi aux utilisateurs de Facebook et LinkedIn », *Atelier Fouille de Données Complexes de la 14e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'14), Rennes*, 2014.

Dieng M. A., *« Développement d'un système d'appariement pour l'e-recrutement »*, Recherche, Université de Montréal, Montréal, 05/2016, 2016.

Florea N. V. *et al.*, « Cost/Benefit Analysis–A Tool To Improve Recruitment, Selection And Employment In Organizations », *Management and Marketing Journal*, vol. 11, n° 2, p. 274-290, 2013.

Fondeur Y., Larquier G. d., Lhermitte F., « Quand l'informatique outille le recrutement », *Connaissance de l'emploi*, vol. 76, p. 1-4, 2011.

Gu Y., Zhao B., Hardtke D., Sun Y., « Learning Global Term Weights for Content-based Recommender Systems », *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, p. 391-400, 2016.

Kessler R., Tondo E., Lapalme G., « Génération automatique d'une ontologie dans le domaine des ressources humaines. », *CORIA-CIFED*, p. 637-652, 2016.

Le Vrang M., Papantoniou A., Pauwels E., Fannes P., Vandensteen D., De Smedt J., « ESCO : Boosting Job Matching in Europe with Semantic Interoperability », *Computer*, vol. 10, n° 47, p. 57-64, 2014.

Pazzani M. J., Billsus D., « Content-based recommendation systems », *The adaptive web*, Springer, p. 325-341, 2007.

Platt J. C., « 12 fast training of support vector machines using sequential minimal optimization », *Advances in kernel methods*p. 185-208, 1999.

Poirier D., Fessant F., Tellier I., « De la Classification d'Opinion à la Recommandation : l'Apport des Textes Communautaires », *Traitement Automatique des Langues*, vol. 51, n° 3, p. 19-46, 2010.

Quinlan J. R., *C4. 5 : programs for machine learning*, Elsevier, 2014.

Resnick P., Iacovou N., Suchak M., Bergstrom P., Riedl J., « GroupLens : an open architecture for collaborative filtering of netnews », *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, ACM, p. 175-186, 1994.

Sarwar B., Karypis G., Konstan J., Riedl J., « Item-based collaborative filtering recommendation algorithms », *Proceedings of the 10th international conference on World Wide Web*, ACM, p. 285-295, 2001.

Trichet F., Bourse M., Leclère M., Morin E., « Human resource management and semantic web technologies », *Information and Communication Technologies : From Theory to Applications, 2004. Proceedings. 2004 International Conference on*, IEEE, p. 641-642, 2004.

Yahiaoui L., Boufaida Z., Prié Y., « Automatisation du e-recrutement dans le cadre du Web sémantique », *IC-17èmes Journées francophones d'Ingénierie des Connaissances*, p. 51-60, 2006.