Système pour la catégorisation automatique des offres d'emploi en une typologie de fonctions

Julie Séguéla[*,**]

[*]Cédric-CNAM, 292 rue Saint-Martin, 75003 Paris, France
[**]Multiposting.fr, 33 rue Réaumur, 75003 Paris, France
jseguela@multiposting.fr

Résumé. Depuis les deux dernières décennies, l'augmentation du nombre de sites d'emploi sur Internet a accentué la nécessité de proposer des outils d'aide à la décision adaptés aux besoins des recruteurs. Cet article présente un système pour la catégorisation des textes d'offres d'emploi destinées à être diffusées sur Internet. Après un pré-traitement adapté des offres, les termes descripteurs sont choisis en fonction de leur pouvoir discriminant vis-à-vis des différentes classes ce qui permet de réduire leur nombre de manière significative. Les offres sont ensuite représentées par leurs coordonnées dans l'espace factoriel obtenu par analyse des correspondances et la classification réalisée dans un cadre supervisé à l'aide de SVM.

1 Introduction

Contexte. La multiplication rapide du nombre de supports Internet pour la publication des offres d'emploi a rendu indispensable l'évaluation et la comparaison des performances dans le cadre du développement d'outils d'aide à la décision pour les recruteurs. Les indicateurs de performance permettent de mesurer et comparer l'efficacité des sites d'emploi en décrivant les candidatures reçues d'un point de vue aussi bien quantitatif que qualitatif (nombre de retours, coût, etc.). Dans ce contexte, nous proposons un système permettant d'affecter automatiquement une fonction (selon une typologie prédéfinie) à toute offre d'emploi diffusée sur Internet, l'intérêt étant d'obtenir une nomenclature unique pour l'ensemble des annonces diffusées sur les plusieurs centaines de sites d'emploi existant. En effet, chaque site d'emploi dispose de sa propre nomenclature pour caractériser la fonction (métier au sens large), rendant très difficile l'agrégation d'information et l'exploitation des connaissances acquises en termes de performance des annonces suite à leur diffusion. Nous disposons d'une base de données d'annonces diffusées par Multiposting.fr [1] sur un grand nombre de sites d'emploi.

Afin de mettre au point l'algorithme de catégorisation, nous utilisons un corpus d'offres d'emploi rédigées en français et étiquetées manuellement par des recruteurs lors de leur publication sur un site d'emploi généraliste. Les étiquettes attribuées par les recruteurs appartiennent à une typologie prédéfinie par le site d'emploi.

1. Multiposting.fr est une plateforme de multidiffusion d'offres d'emploi sur Internet.

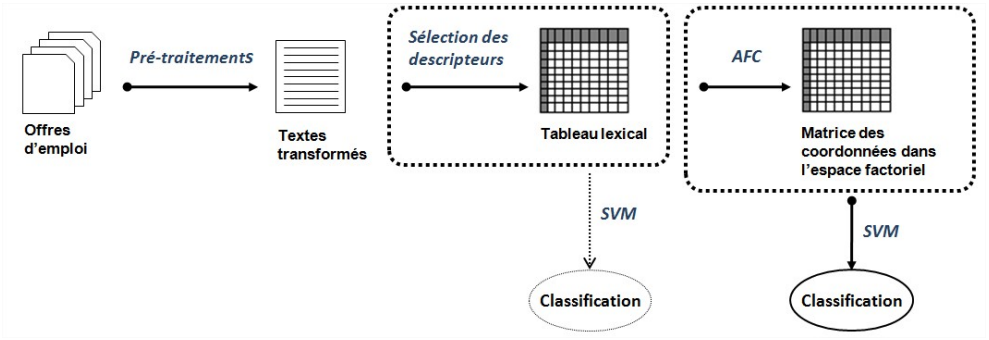

FIG. 1 – *Vue générale du système de catégorisation automatique.*

Présentation générale du système. Le système de catégorisation s'articule en quatre étapes principales, illustrées par la figure 1. Dans un premier temps, le système assure le pré-traitement des textes des offres (lemmatisation et étiquetage morphosyntaxique, filtrage des termes). Il effectue ensuite la sélection de l'ensemble des termes qui vont être utilisés pour décrire les offres d'emploi via une représentation vectorielle des textes, cette représentation étant également appelée « sacs de mots » (Salton et al., 1975). Ensuite, une analyse des correspondances réalisée sur le tableau lexical croisant offres et termes projette les vecteurs des offres dans un nouvel espace, dont les coordonnées sont utilisées pour représenter les textes. Enfin, les machines à vecteurs de support (SVM) sont utilisées pour l'étiquetage supervisé des offres selon notre typologie de fonctions.

Après un rapide état des lieux des travaux concernant le traitement textuel des offres d'emploi (section 2), la méthodologie de représentation et classification des offres d'emploi est ensuite présentée (section 3). Les résultats obtenus dans le cadre d'une application à un corpus réel sont présentés dans la section 4, à travers la comparaison de plusieurs méthodes de représentation des offres. Enfin, la section 5 est consacrée à une discussion des résultats obtenus.

2 Traitement textuel des offres d'emploi : état des lieux et contribution

Dans le contexte du traitement des offres d'emploi, les principales applications de ces dernières années se sont focalisées sur le « matching » offres-CV, visant à détecter de manière automatique les CV les plus qualifiés pour une offre donnée et inversement. Tandis que certains auteurs se basent sur des ontologies pour établir des modèles de compétences (Radevski et Trichet, 2006), d'autres préfèrent adopter la représentation vectorielle des textes. C'est le cas de Kessler (2009), qui obtient les meilleurs résultats avec les représentations : TF (*term frequency*), TF-IDF (*term frequency-inverse document frequency*), pondération selon l'étiquette grammaticale, et une méthode d'enrichissement de l'offre par *relevance feedback* mais qui ne s'avère par pertinente par rapport à notre tâche.

Notre méthode diffère des techniques de recherche d'information classiques car nous les combinons avec des techniques issues de l'analyse statistique des données s'avérant efficaces pour répondre à deux problématiques : éliminer les termes non pertinents pour la description des offres et corriger en partie les problèmes de synonymie et polysémie rencontrés. Etant donné la tâche de catégorisation des textes, nous souhaitons utiliser une statistique permettant de juger du pouvoir discriminant des termes. Différentes techniques existent dans la littérature pour la sélection des descripteurs pertinents : *mutual information*, *information gain*, *term strength* ou encore la statistique du χ^2. Yang et Pedersen (1997) mettent en évidence dans une étude comparative de ces critères les très bons résultats de la statistique du χ^2 pour la tâche de catégorisation. Dans nos travaux, nous souhaitons comparer le χ^2 à des statistiques issues de l'analyse textuelle permettant d'évaluer le sur-emploi d'un mot au sein d'une catégorie : la spécificité lexicale positive et le Z-score.

Nos apports concernant le traitement textuel des offres d'emploi résident donc dans les points suivants. D'abord, nous mettons en évidence l'existence de vocabulaires spécifiques aux types de métiers permettant une réduction considérable en amont du nombre des descripteurs, tout en maintenant l'efficacité de l'algorithme de catégorisation. De plus, nous montrons que la statistique de spécificité lexicale permet de réaliser cette tâche avec une efficacité au moins équivalente à celle du χ^2, tandis que le Z-score est légèrement inférieur. Enfin, notre application se distingue par la méthode de représentation adoptée pour les offres : nous montrons ici qu'il est pertinent de représenter des documents de type offre d'emploi par les coordonnées factorielles issues de l'analyse des correspondances.

3 Méthodologie

3.1 Représentation des textes d'offres d'emploi

3.1.1 Pré-traitement des textes

Dans un premier temps, le système assure le pré-traitement des textes des offres afin de le rendre exploitable et de procéder à une première réduction du nombre de termes :

- Lemmatisation et étiquetage morphosyntaxique des termes. L'algorithme de Schmid (1994) est utilisé pour cette tâche. Nous choisissons la lemmatisation plutôt que la racinisation des termes, technique moins bien adaptée à la langue française ayant un fort taux de flexion (Namer, 2000). En effet, les algorithmes de racinisation suivent une démarche de troncation visant à réduire les différentes formes d'un mot à une racine commune. En revanche, la lemmatisation prend en compte la flexion (variation de la forme d'un mot en fonction de facteurs grammaticaux) afin de ramener au lemme (masculin singulier pour un adjectif, infinitif pour un verbe conjugué, etc.).
- Filtrage de certaines catégories morphosyntaxiques (prépositions, adverbes, symboles, déterminants, etc.). Les offres d'emploi étant des textes relativement courts, composés de phrases courtes et en partie d'énumération de tâches, nous décidons de conserver uniquement les noms, adjectifs et verbes pour représenter les offres (à l'instar de Kessler, 2009).

3.1.2 Sélection des descripteurs

Nous travaillons sur la table des fréquences qui comporte initialement un grand nombre de lignes (offres) et de colonnes (termes). Chaque offre n'utilise qu'une petite partie du vocabulaire complet et par suite, la matrice est principalement constituée de zéros (c'est une matrice *sparse*). Pour réduire le degré de « sparsité » de cette matrice, nous considérons comme descripteurs les termes apparaissant 5 fois ou plus dans la base d'apprentissage. De plus, nous souhaitons limiter le vocabulaire servant à décrire les offres à l'ensemble des termes pertinents pour la discrimination des différentes catégories, afin de réduire la dimension du problème de manière significative (et ainsi réduire les temps d'apprentissage) tout en maintenant le pouvoir de généralisation de l'algorithme. Comme évoqué précédemment, nous nous focalisons sur trois méthodes issues de l'analyse statistique des données : le χ^2, la spécificité lexicale positive et le Z-score.

Statistique du χ^2. La statistique $\chi^2(t, c_i)$ mesure le manque d'indépendance entre un terme *t* et une catégorie c_i. Elle évalue l'importance de l'écart entre la fréquence observée et la fréquence attendue s'il y avait indépendance.

Spécificité lexicale positive. Introduites par Lafon (1980), les spécificités positives sont très utilisées en analyse textuelle pour décrire une partie d'un corpus à travers les formes qui y sont significativement sur-employées par rapport aux autres parties. Nous utilisons la formule de Lafon basée sur le modèle hypergéométrique pour calculer le score de spécificité $spec(t, c_i)$ du terme *t* au sein de la catégorie c_i.

Z-score. Le Z-score $Z(t, c_i)$ est une statistique indiquant le degré d'appartenance d'un terme *t* au vocabulaire spécifique d'une catégorie c_i. Son calcul est basé sur l'hypothèse que le nombre d'occurrences d'un mot au sein d'une catégorie suit une loi binomiale.

Afin de mesurer la pertinence d'un terme pour la discrimination entre les *m* catégories, les scores suivants sont calculés pour chacun des termes :

$$\chi^2_{max}(t) = \max_{i=1}^{m}\{\chi^2(t, c_i)\},\ spec(t, c_i) = \max_{i=1}^{m}\{1 - spec(t, c_i)\},\ Z(t, c_i) = \max_{i=1}^{m}\{Z(t, c_i)\}$$

Par la suite, les termes seront classés selon le score obtenu et leur sélection s'effectue en fonction de la statistique choisie et du nombre de termes retenus.

3.1.3 Analyse des correspondances

La représentation vectorielle des textes est souvent critiquée car elle ne prend pas en compte les relations sémantiques entre les termes ou la structure des phrases. Proche de l'Analyse Sémantique Latente (Deerwester et al., 1990), l'analyse des correspondances (AC) est une technique d'analyse en axes principaux basée sur la décomposition en valeurs singulières qui permet de répondre en partie aux problèmes de polysémie et synonymie. De plus, les structures des données qui nous intéressent ne représentent qu'une faible part de la variabilité totale (Lebart, 2004), et l'AC permet de réduire la sparsité du tableau en éliminant la partie bruitée des données. Nous appliquons l'AC au tableau lexical croisant offres et termes (dont chaque

élément représente la fréquence d'apparition d'un terme donné dans une offre donnée). Les axes principaux sont ici utilisés comme nouveaux descripteurs des textes des offres d'emploi, c'est-à-dire que les textes sont représentés par leurs coordonnées dans le sous-espace engendré par les *k* premiers vecteurs principaux produits par l'AC. Nous étudions les résultats de la classification pour plusieurs valeurs du paramètre *k*.

3.2 Classification

Les machines à vecteurs de support (Vapnik, 1995) se sont montrées efficaces à de nombreuses reprises pour des tâches de catégorisation sur données textuelles (Joachims, 1998). Dans le cadre du traitement textuel des offres d'emploi, les SVM ont été utilisés avec succès par Kessler et al. (2007) pour l'étiquetage des différentes parties de l'annonce. Les SVM reposent sur la projection des données dans un espace de grande dimension par une transformation basée sur un noyau, dont les plus répandus sont les noyaux polynomiaux et gaussiens. Il est également possible d'utiliser un noyau linéaire, ramenant au cas d'un classifieur linéaire, sans changement d'espace. Des premières expérimentations sur le corpus des annonces ont montré que les noyaux les plus complexes n'apportaient pas un gain de performance dans la classification par rapport au noyau linéaire. Étant donné la grande dimension du problème et afin de minimiser les risques de sur-apprentissage, les SVM linéaires sont utilisés pour la classification des annonces. L'implémentation *libsvm* de l'algorithme de Chang et Lin (2001) est appliquée et permet d'assurer une classification multi-classes par une méthode de vote *one-against-one*. L'optimisation du paramètre de coût est effectuée par validation croisée (*10-fold*).

3.3 Critère de performance

Pour mesurer la qualité de l'algorithme de catégorisation, le corpus est divisé en deux parties : 75% du corpus sont dédiés à l'apprentissage, tandis que les 25% restants constituent l'échantillon de test. Les données de cet échantillon sont totalement extérieures à la définition de l'ensemble des descripteurs et à l'apprentissage du modèle. Les échantillons de test et apprentissage sont construits de sorte à préserver la répartition des catégories observée sur l'ensemble du corpus, ceci dans le but de se prémunir d'éventuels biais et d'assurer un apprentissage sur l'ensemble des catégories. Au sein de chaque catégorie, la répartition entre apprentissage et test est faite par un tirage aléatoire. Le processus d'évaluation de l'erreur sur l'échantillon test est illustré par la figure 2.

Un des objectifs du système est de permettre la comparaison des performances des annonces entre les catégories. La mesure privilégiée pour estimer l'efficacité de l'algorithme est ici la précision, car nous souhaitons minimiser le bruit au sein de chaque catégorie obtenue. Toutefois, le rappel est également une mesure d'intérêt dans la mesure où chaque catégorie doit réunir un nombre d'offres « représentatif » de la fonction. La F_β-mesure (Van Rijsbergen, 1979) est un indicateur de synthèse qui permet d'accentuer l'importance de la précision ou du rappel :

$$F_\beta = \frac{(1+\beta^2) \cdot precision \cdot rappel}{\beta^2 \cdot precision + rappel}$$

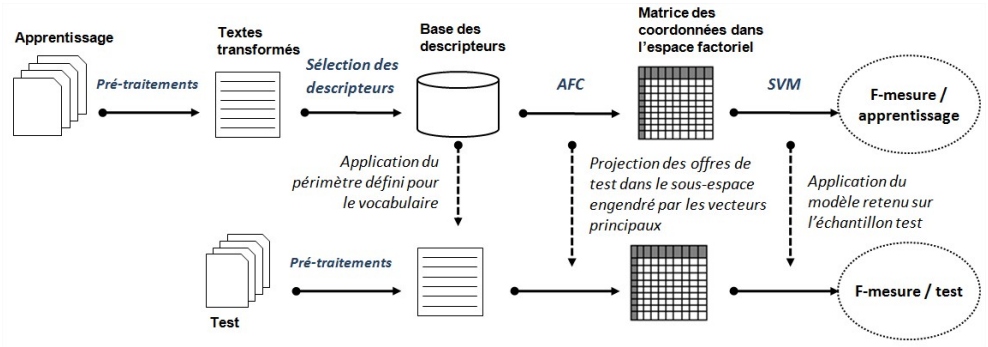

FIG. 2 – *Processus d'évaluation de l'erreur.*

Les indicateurs *precision* et *rappel* sont calculés comme des macro-moyennes sur l'ensemble des *m* catégories :

$$precision = \frac{\sum_{i=1}^{m} p_i}{m} \text{ et } rappel = \frac{\sum_{i=1}^{m} r_i}{m}$$

où p_i et r_i sont respectivement la précision et le rappel associés à la classe i. Etant donné le contexte de nos travaux cité plus haut, nous choisissons la F_β-mesure avec $\beta = 0.5$ comme critère de performance d'un modèle de classification.

4 Application et résultats

4.1 Statistiques du corpus

Le corpus est composé de 4704 offres d'emploi ayant toutes été postées sur le même site d'emploi généraliste au cours des deux dernières années. Les annonces sont étiquetées en 21 catégories (ou fonctions) distinctes. La répartition des offres entre les 21 catégories est visible dans le tableau 1. Certaines catégories présentent un faible effectif pouvant rendre plus difficile la généralisation à de nouvelles annonces. Les 4704 annonces sont associées à un vocabulaire d'environ 28 400 termes distincts (avant pré-traitements).

4.2 Choix des paramètres

Dans un premier temps, nous étudions la qualité de la classification en fonction du nombre de termes retenus pour la représentation vectorielle des offres et de la statistique choisie pour la sélection. Dans un second temps, nous étudions les résultats obtenus après analyse des correspondances sur le tableau lexical. La performance est mesurée en fonction du nombre d'axes retenus, pour des statistiques et un nombre de termes choisis. Nous faisons varier le nombre de termes conservés depuis la totalité du vocabulaire (à savoir 4189 termes) jusqu'à 500 termes (par décrément de 500 termes), pour chacune des statistiques de score calculées. Les F_β-mesures et précisions obtenues sont présentées dans la figure 3.

Catégorie	Effectif	%	Catégorie	Effectif	%
Architecture / Création	55	1.2	Juridique	111	2.4
BTP	229	4.9	Logistique / Transport	302	6.4
Commercial / Vente	527	11.2	Management / Stratégie	142	3.0
Comptabilité / Finance	589	12.5	Marketing / Communication	281	6.0
Edition	49	1.0	Production	263	5.6
Formation / Education	79	1.7	Recherche / Etudes	125	2.7
Hôtellerie / Restauration	100	2.1	Ressources Humaines	196	4.2
Informatique	487	10.4	Santé	106	2.3
Ingénierie	327	6.9	Services administratifs	275	5.8
Inspection / Qualité	117	2.5	Services clientèle	165	3.5
Installation / Maintenance	179	3.8	**Total**	**4704**	**100**

TAB. 1 – *Répartition des annonces entre les catégories.*

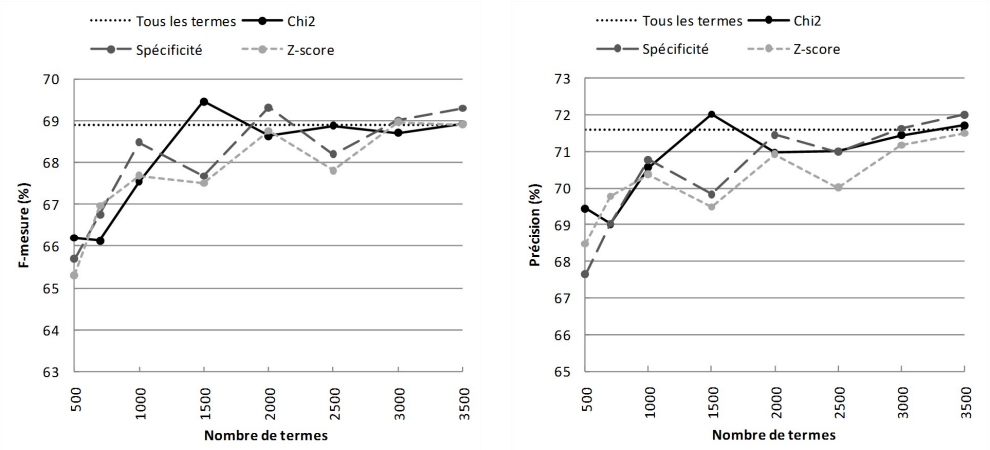

FIG. 3 – *F_β-mesure et précision en fonction du nombre de termes retenus.*

La sélection des termes par la statistique du χ^2 permet une réduction jusqu'à 1500 termes tout en maintenant une performance équivalente à un modèle conservant la totalité du vocabulaire. Les courbes de la statistique du Z-score et du score de spécificité ont une allure proche, bien que le Z-score reste inférieur en termes de performance. Pour les trois statistiques, la F_β-mesure est stable et équivalente jusqu'à 2000 termes à celle d'un modèle conservant tous les termes. La décroissance de la F_β-mesure est fortement marquée à partir de 1000 termes.

Nous nous intéressons maintenant à l'évolution du critère de performance en fonction du nombre d'axes principaux choisi pour représenter les offres après AC. Nous faisons varier le nombre d'axes de 50 à 600 par incrément de 50, puis de 100. Cette évolution est représentée pour plusieurs ensembles de termes descripteurs dans la figure 4. Nous avons choisi de retenir 2000 termes descripteurs (soit environ 48% de l'ensemble de départ) et de comparer les résultats pour les statistiques de spécificité et du χ^2.

La figure 4 permet de constater qu'appliquer l'AC au tableau lexical a amélioré les résultats

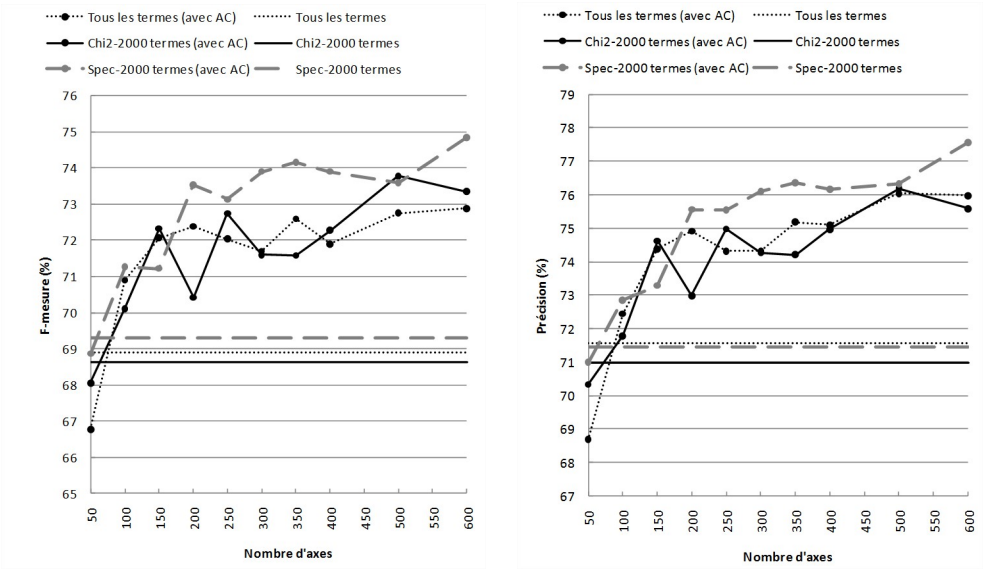

FIG. 4 – *F_β-mesure et précision en fonction du nombre d'axes retenus.*

de la classification, tout en réduisant de manière importante le nombre de variables explicatives fournies aux SVM. En effet, à partir de 100 axes retenus, les résultats sont toujours meilleurs après AC. Cette amélioration est visible quel que soit l'ensemble des termes retenus pour la description, bien qu'une légère supériorité semble apparaître pour la sélection avec score de spécificité selon le nombre d'axes retenus.

4.3 Illustration des résultats

Bien que plusieurs choix soient possibles, nous retenons comme procédé la sélection des termes descripteurs par la statistique de spécificité, puis une analyse des correspondances à 600 axes. La classification par SVM fournit alors les résultats détaillés pour chaque catégorie présentés dans la figure 5.

Nous pouvons constater des disparités assez marquées en termes de performance pour la prédiction des différentes catégories. D'une manière générale, un niveau de précision d'au moins 55% est assuré, et la moitié des classes ont une précision supérieure à 75%. Bien qu'un faible niveau de rappel ait été obtenu pour certaines catégories, il demeure supérieur à 70% pour plus de la moitié des classes. Le choix de la $F_{0.5}$-mesure comme critère de performance a permis de répondre aux objectifs de départ en accordant plus d'importance à la précision tout en tenant compte du niveau de rappel. Des résultats très satisfaisants sont obtenus pour certaines catégories (*Juridique*, *Hôtellerie / Restauration*, *Santé*, *Ressources Humaines* ou encore *Comptabilité / Finance*), ce qui laisse supposer qu'un vocabulaire très spécifique de ce type de fonction est employé lors de la rédaction des annonces. A contrario, certaines catégories

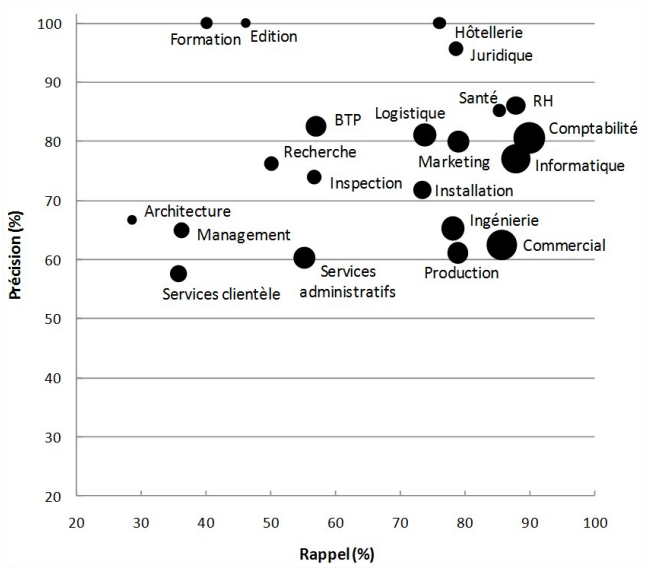

FIG. 5 – *Représentation des 21 catégories de fonctions dans le plan rappel× précision (taille des bulles proportionnelle à l'effectif de la catégorie).*

(*Architecture / Création*, *Management*, *Services clientèle* et *Services administratifs*) semblent plus difficiles à prédire. Plusieurs raisons possibles à cela, notamment :

- un faible effectif de la catégorie au départ, induisant un vocabulaire moins représentatif de l'ensemble de la fonction, et par suite une généralisation à de nouvelles annonces plus difficile ;
- un vocabulaire peu spécifique, commun à plusieurs fonctions, entrainant une discrimination des catégories plus complexe.

Nous avons pu en partie répondre au deuxième type de problème grâce à l'analyse factorielle des correspondances appliquée au tableau lexical des annonces.

5 Discussion et conclusions

Comme nous l'avons évoqué dans la section précédente, la qualité de prédiction reste peu satisfaisante pour certaines catégories. Nous tentons dans cette section d'y apporter quelques éléments d'explication. Les erreurs étant dues au manque de séparation entre certaines catégories, nous présentons dans le tableau 2 la distribution des catégories observées sur les catégories prédites afin de visualiser plus en détail la nature des erreurs.

L'analyse du tableau 2 peut s'accompagner de celle de la figure 5, les valeurs en diagonale n'étant autres que le rappel associé à chaque catégorie. Nous pouvons alors expliquer certains mauvais résultats. En effet, les offres de *Management* semblent principalement confondues avec celles de la fonction *Commercial*, de même que les offres de *Services clientèle*. La fonction *BTP* est prédite également en *Ingénierie* et *Installation*, tandis que la fonction *Services*

	Catégories observées																				
	Architecture	BTP	Commercial	Comptabilité	Edition	Formation	Hôtellerie	Informatique	Ingénierie	Inspection	Installation	Juridique	Logistique	Management	Marketing	Production	Recherche	RH	Santé	Services admin.	Services client.
Architecture	**29**														1						
BTP		**57**							5		4			3							
Commercial	7	7	**86**	2		15	8	4			2		9	33	13			4	4	9	29
Comptabilité			3	**90**							2	11		14				2		17	12
Edition					**46**																
Formation						**40**															
Hôtellerie							**76**														
Informatique	21		2	3		10		**88**	5	7	2		1		3	6	12	2			5
Ingénierie	7	21							**78**	3	9					6	28				
Inspection									2	**57**	2	4				2			4		
Installation		10							2		**73**					3					7
Juridique												**79**									
Logistique			2										**74**			3		2		9	2
Management							12							**36**	1			2		1	
Marketing	21				23			2					1		**79**	2			4	1	2
Production	7	3			8		4	2	5	20	4		7	3	3	**79**	6			3	2
Recherche									1	7			1				**50**			1	
RH				2		10								6				**88**			
Santé						5							1	6					**85**		
Services admin.	7	2	4	2	23	15			1	3		7	3				3			**55**	5
Services client.			2			5				3			3						4	3	**36**
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

TAB. 2 – *Distribution des catégories observées sur les catégories prédites (% en colonne, les valeurs inférieures à 1% n'apparaissent pas).*

administratifs est confondue avec *Comptabilité*. Une certaine logique semble ressortir d'une partie des erreurs de prédiction, dans la mesure où les postes des catégories confondues sont proches sémantiquement ainsi qu'en termes de tâches à accomplir. De plus, il ne faut pas négliger le fait que l'étiquetage humain induit une interprétation plus ou moins subjective du poste à proposer. Ainsi, si le système est en désaccord avec l'étiquetage humain, cela désigne potentiellement un recruteur en désaccord avec les autres ou une erreur d'étiquetage.

Afin de confronter nos résultats à la réalité, le système de catégorisation est maintenant employé pour étiqueter un échantillon de 1306 offres diffusées sur un second site généraliste (ayant sa propre nomenclature). La classification obtenue et la classification d'origine du site sont mises en regard dans le tableau 3.

Des petites catégories comme *Audit* et *Export* sont parfaitement associées à nos fonctions *Comptabilité* et *Commercial*. Certaines catégories de la nouvelle nomenclature semblent également trouver leur équivalent dans notre typologie : *Gestion-Compta-Finance*, *Systèmes d'Information-Télécom*, ou encore *Logistique-Transport*. La fonction *Production-Maintenance-Qualité*, large par construction, se retrouve dans nos fonctions *Production*, *Installation / Maintenance*, *Inspection / Qualité* et *Ingénierie*. A nouveau, nous observons une certaine cohérence dans les résultats obtenus.

L'étude de l'impact du nombre de descripteurs et de leur nature sur la performance du système de catégorisation nous a amenés à plusieurs conclusions. Grâce à la sélection des descripteurs selon leur pouvoir discriminant, leur nombre a pu être réduit de manière importante

	Nouvelle nomenclature																	
	Administration	Audit	Commercial-Vente	Communication-Création	Conseil	Direction générale	Etudes-Recherche	Export	Gestion-Compta-Finance	Internet / e-Commerce	Juridique-Fiscal	Logistique-Transport	Marketing	Production-Maintenance-Qualité	RH-Formation	Santé	Systèmes d'Information-Télécom
Architecture / Création				3													
BTP						3	5							9			
Commercial / Vente	4		**68**	17	15	32	5	**100**	2			3	10	2	7	8	2
Comptabilité / Finance	4	**100**	16	6	35	5			**92**		11	1	6	3			2
Edition				3													
Formation / Education	4														3		
Hôtellerie / Restauration			1			11											
Informatique	7		1	6	30		11			**35**		5	2	6	2		**91**
Ingénierie			1		2		**62**							**21**			
Inspection / Qualité					2									**16**		8	
Installation / Maintenance												1		**16**			
Juridique	4				1						**89**						
Logistique / Transport	4		3			3			1			**83**	1				
Management / Stratégie	4		3		6	**43**	2		2	5		3				17	
Marketing / Communication			3	**67**	5				2	**60**		1	**75**		2		
Production	4					3	2					1	1	**27**		8	
Recherche / Etudes							**12**										
RH	4		2		5				2			1	1		**85**		2
Santé																**58**	
Services administratifs	**64**																
Services clientèle			1										3				
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Effectif	28	26	296	36	66	37	56	12	196	20	28	88	79	94	94	12	138

TAB. 3 – *Comparaison de la classification fournie par le système avec la classification d'un autre site généraliste (% en colonne, les valeurs inférieures à 1% n'apparaissent pas).*

tout en maintenant l'efficacité de l'algorithme de classification. Trois scores basés sur des statistiques de test ont été comparés. La sélection obtenue par la statistique de spécificité s'est avérée au moins aussi efficace que celle obtenue par la statistique du χ^2. De plus, l'AC appliquée au tableau lexical a permis de réduire encore le nombre de descripteurs tout en améliorant sensiblement la qualité de la classification. L'évaluation du système sur des échantillons test d'annonces postées sur le même site et sur un deuxième site généraliste a permis de valider la capacité du système à fournir des catégories homogènes. Par la suite, nous utiliserons ce système dans le cadre de l'analyse et de la compréhension des performances des offres d'emploi.

Ce système présente l'avantage de pouvoir associer une fonction à une offre de manière automatique, à partir du seul descriptif du poste. Aucune information additionnelle n'est nécessaire, toutefois, la prise en compte d'informations complémentaires (secteur de l'entreprise, niveaux de diplôme et d'expérience requis, etc.) pourrait être envisagée afin d'en améliorer la performance. Nous envisageons également d'introduire dans la représentation des textes des bi-grammes de mots pour permettre une meilleure gestion de la sémantique grâce à la prise en compte de mots composés.

Références

Chang, C.-C. et C.-J. Lin (2001). libsvm: a library for support vector machines. Documentation, http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

Deerwester, S., S. Dumais, G. W. Furnas, T. K. Landauer, et R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41*, 391–407.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *ECML 1998 : European Conference on Machine Learning*, pp. 137–142.

Kessler, R. (2009). *Traitement automatique d'informations appliqué aux ressources humaines*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse.

Kessler, R., J.-M. Torres-Moreno, et M. El-Bèze (2007). E-Gen: automatic job offer processing system for Human Resources. In *MICAI 2007 : Mexican International Conference on Artificial Intelligence*, pp. 985–995.

Lafon, P. (1980). Un analyseur flexionnel du français à base de règles. *Mots 1*, 127–165.

Lebart, L. (2004). Validité des visualisations de données textuelles. In *JADT 2004 : Journées internationales d'Analyse statistique des Données Textuelles*, pp. 708–715.

Namer, F. (2000). Un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues 41*, 247–523.

Radevski, V. et F. Trichet (2006). Ontology-based systems dedicated to Human Resources Management : an application in e-recruitment. In R. Meersman, Z. Tari, et P. Herrero (Eds.), *OTM Workshops 2006*, Volume 4278 of *LNCS*, pp. 1068–1077. Springer.

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing. *Communications of the ACM 18*, 613–620.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pp. 44–49.

Van Rijsbergen, K. (1979). *Information Retrieval*. London: Butterworths.

Vapnik, V. (1995). *The nature of statistical learning theory*. Berlin: Springer Verlag.

Yang, Y. et J. P. Pedersen (1997). A comparative study on feature selection in text categorization. In *ICML 1997 : International Conference on Machine Learning*, pp. 412–420.

Summary

The increasing number of online job boards has made crucial the introduction of decision-making tools adapted to recruiter needs. This paper is introducing a system for automatic categorization of online job postings. After an adapted preprocessing, term features are selected according to their discriminatory power in relation with the different classes, which allows to reduce their number significantly. Then, offers are represented by their coordinates in the factorial space resulting from correspondence analysis. Finally, classication task is made in a supervised framework by implementing an SVM algorithm.