Approche hybride d'extraction compétences dans les CVs en format PDF

Florentin Flambeau Jiechieu Kameni, Norbert Tsopze

Approche hybride d'extraction compétences dans les CVs en format PDF

Florentin Flambeau Jiechieu Kameni
Norbert Tsopze

IRD UMI 209 UMMISCO, Université de Yaoundé I, Yaoundé, Cameroun
Département d'Informatique
jkflorex@gmail.com
tsopze.norbert@gmail.com

RÉSUMÉ. L'objectif de ce travail est d'utiliser une approche hybride pour extraire les compétences dans les CVs. L'approche d'extraction des compétences proposée s'effectue en deux grandes phases : une phase de segmentation du CV en sections classées suivant leurs contenus et à partir desquelles des compétences de base où les termes représentant les compétences sont extraits du CV; et une phase de prédiction qui consiste à partir des caractéristiques extraites précédemment, à prédire un ensemble de compétences qu'un expert aurait déduit et qui ne seraient pas nécessairement mentionnées dans le CV de l'expert. Les principales contributions de ce travail sont : l'utilisation de l'approche hiérarchique de segmentation du CV en sections et classification supervisée des sections, l'utilisation d'un modèle d'apprentissage multi-label basée sur les SVMs afin de prédire parmi un ensemble de compétences, celles dont on peut déduire à la lecture du CV. Les expérimentations effectuées sur un jeu de CVs collectés sur Internet ont montré une amélioration de plus de 10% dans l'identification des blocs par rapport à un modèle de l'état de l'art. Le modèle de prédiction multi-label des compétences, permet de retrouver la liste des compétences avec une précision et un rappel respectivement de l'ordre de 90,5% et 92,3%.

ABSTRACT. The aim of this work is to use a hybrid approach to extract CVs' competences. The extraction approach for competences is made of two phases: a segmentation into sections phase within which the terms representing the competences are extracted from a CV; and a prediction phase that consists from the features previously extracted, to foretell a set of competences that would have been deduced and that would not have been necessary to mention in the resume of that expert. The main contributions of the work are two folds : the use of the approach of the hierarchical clustering of a résume in section before extracting the competences; the use of the multi-label learning model based on SVMs so as to foretell among a set of skills, those that we deduce during the reading of a CV. Experimentation carried out on a set of CVs collected from an internet source have shown that, more than 10% improvement in the identification of blocs compared to a model of the start of the art. The multi-label competences model of prediction allows finding the list of competences with a precision and a reminder respectively in an order of 90.5 % and 92.3 %. .

MOTS-CLÉS : Skill Gap, CV, Extraction des Compétences, segmentation

KEYWORDS : Skill Gap, Resume, Skills Extraction, clustering

1. Introduction

L'écart entre les compétences détenues par la ressource humaine d'une entreprise et celles qu'elle a besoin pour son développement est connu sous le nom de *skill-gap* [6, 2, 21]. Pour combler cet écart, les entreprises organisent généralement les recrutements en sélectionnant les candidats parmi ceux qui ont déposé un dossier physique, ou un dossier électronique sur leur site internet (ou par mail) ou encore sur les sites d'offres d'emploi. Ces dossiers sont composés entre autres des demandes d'emploi, des CVs et des lettres de motivations. Une des tâches de la direction des ressources humaines serait alors de retrouver ces compétences détenues des candidats en lisant leurs CVs. L'extraction des compétences dans les CVs est un problème bien connu dans la littérature [21, 5, 17]. Il s'agit d'une activité qui trouve son application dans les processus de recrutement automatique dont l'objectif est d'effectuer un matching automatique entre les compétences extraites dans les CVs et les compétences requises par l'organisation pour effectuer une tâche ou occuper un poste [10]. L'extraction automatique des compétences à partir des CVs permet aussi de réduire la subjectivité liée à l'Homme dans le processus de recrutement.

Plusieurs travaux dans la littérature ont abordé le problème de *skill-gap* allant de la médélisation ontologique [17] au matching entre les profils des candidats et les offres d'emplois [10] en passant par les modèles d'évolution des compétences et l'extraction des compétences [5, 8]. La compétence est une notion abstraite et est généralement représentée par des phrases, des mots, ou groupes de mots. (Ex : *Java, computer programming, big data, troubleshoot network infrastructure*, ...). L'extraction automatique des compétences est une étape très importante dans le matching automatique entre CVs et offre d'emplois. Parmi les méthodes d'extraction, l'approche hiérarchique [5] considérant un CV comme un ensemble de blocs (*Education, Publications, Expérience professionnelle*,...) permet de conserver la structure hiérarchique facilitant ainsi l'opération de spécialisation au cas où l'on s'intéresse au contenu des sections.

Cet article est une extension de [8]. En plus de la classification des blocs (ou sections) extraits des CVs à partir des caractéristiques (comme dans [8]), le présent article propose un modèle pour classer les CVs en fonction des compétences du candidat. Ces compétences peuvent être explicites (clairement mentionnées dans le CV) ou implicites (non mentionnées dans le CV, mais pouvant être déduites des autres compétences). L'objectif de ce travail est donc d'adapter le modèle proposé par Zhen Chen et al. [5] qui se base sur l'approche hiérarchique consistant à segmenter le CV en ses différentes sections, afin d'appliquer une méthode d'extraction de caractéristiques appropriée dans chaque section. En fait, l'algorithme de segmentation hiérarchique utilisé dans l'approche de Zhen Chen etal [5] est basé sur la densité des mots dans la section. Et pour cela d'après les auteurs, cet algorithme ne produit pas les bonnées performances. L'idée ici étant d'identifier d'abord les sections avant d'extraire les caractéristiques. Cette proposition est baseé sur le fait qu'un CV est organisé en sections ayant chacun un titre ; un CV est donc traité comme un ensemble de sections (*identification, Education, Language*,...) ; et aussi que les caractéristiques morphologiques utilisées pour écrire les titres sont généralement différentes de celles utilisées pour écrire les détails des sections. Pour faciliter l'identification des titres de sections, un dictionnaire comportant les mots et expressions les plus utilisés est incorporé à cette étape. En plus, après l'extraction des caractéristiques de chaque bloc, un modèle de classification multi-label [9] est proposé pour prédire les compétences des candidats. Le format de CV traité est le format *PDF* car est encore très utilisé dans les

processus de recrutement en ligne. De plus, les autres formats comme le *HMTL* ne posent pas de problème de réel problème de segmentation car en général bien structuré.

Le reste de cet article est organisé de la manière suivante : la prochaine section présente un ensemble de travaux présents dans la littérature sur le traitement du *skill-gap* en général et sur l'extration des compétences en particulier. La section 3 sera consacrée à notre proposition, le modèle général et les détails sur les différentes étapes y seront présentés. Les expérimentations et les résultats obtenus feront l'objet de la quatrième section. Nous terminerons enfin par la conslusion.

2. Etat de l'art

Le *skill-gap* est un problème traité par plusieurs communautés de chercheurs. Plusieurs aspects du problème et plusieurs approches sont présentés dans la littérature : établir les correspondances entre le CVs et les offres d'emplois, faire des systèmes de recommandations d'offre d'emplois aux candidats, modélisation à travers les ontologies des compétences, recherche des compétences dans les CVs. L'objet de cette section est de présenter quelques travaux traitant ce sujet.

Benson etal. présentent dans [3] une discussion sur un ensemble de compétences permettant de profiter des opportunités présentées par les médias sociaux. Il s'agit pour eux de comprendre comment les professionnels exploitent les réseaux sociaux et de proposer les extensions de la formation dans l'exploitation les opportunités offertes par les réseaux sociaux. Ils recommandent au gouvernement britanique de se pencher sur cette question afin de créer des compétences se basant sur la manière avec laquelle les professionnels exploitent les réseaux sociaux. Dans [2], l'auteur propose un modèle d'évolution des compétences intégrant l'évolution technologique. Partant du constat de la disparution des certaines compétences aux Etats-Unis et de la difficulté à avoir les travailleurs qualifiés pour certaines tâches, il propose un modèle d'évolution des compténces qui tient compte du caractère (du travail) variant avec le temps et de la liason avec les caractéristiques de la population.

Miranda etal. dans [17] proposent une ontologie pour représenter aussi bien les compétences que les offres d'emplois dant le but d'améliorer la stratégie d'employabilité. L'ontologie est construite à partir de la littérature spécialisée sur les compétences et les offres d'emplois ; et aussi sur le recruitement dans un projet de R&D. Cette ontologie est donc utile dans l'interopérabilité entre les outils de gestion des ressources humaines et la détection des influences entre les compétences. A travers cette ontologie, il est alors possible de faire des inférences et déduire des compétences nouvelles. D'autres aspects importants de l'ontologie concernent la planification et la gestion des changements dans l'entreprise, l'évaluation des performances et la programmation des formations. D'autres recherches comme celles présentées par Shaha dans [1] ont proposé de traiter le *skill-gap* comme un problème de recommandation bidirectionnelle : recommander un emploi à un candidat ou recommander un candidat à un recruteur. Dans la même lancée, Guo et al. dans [10] propose *RésuMatcher*, un système capable d'extraire les qualifications et l'expérience professionnelle des candidats, d'extraire également les caractéristiques sur les offres d'emplois et d'utiliser ces informations (caractéristiques du candidat et caractéristiques de l'offre d'emploi) pour calculer un score permettant de déterminer si l'offre correspond au candidat.

Au sujet de l'extraction des compétences dans les textes, des approches basées sur la reconnaissance des entités nommées [21, 12] et les approches basées sur l'utilisation des

graphes [13] ont aussi été proposées. Les approches du premier groupe construisent par apprentissage un modèle d'étiquettage des entités à partir d'un corpus, et utilise ce modèle pour détecter les compétences. Leurs performances sont fortement liées à la qualité de l'algorithme d'apprentissage et à la qualité du jeu d'apprentissage.

La plupart des modèles d'extraction des compétences dans les CVs traitent ces derniers comme des textes bruts ne possédant aucune structure particulière [21, 14]. Ces modèles ne font pas d'hypothèse concernant la *semi-structuration* (organisation en sections) du CV. Or Jun Yu et al. [20] ont montré que la prise en compte de la propriété *semi-structurée* du CV permet de faire une extraction d'information de meilleure qualité si une technique d'extraction appropriée était appliquée à chacune de ses sections. Maheshwari et al.[14] ont proposé un modèle hiérarchique de traitement des CVs dans lequel ils supposent que chaque section du CV (pour un candidat donné) détient une information spéciale permettant de le distinguer. Le modèle recherche d'abord cette information spéciale et l'utilise pour classer les candidats dans le processus de sélection.

L'une des modèles hiérarchiques d'extraction d'information dans les CVs les plus récents a été proposé en 2016 par Chen et al. [5] ; il s'agit d'une amélioration du modèle hybride en cascade proposé par Jun Yu [20] pour l'extraction des informations dans les CVs au format PDF. La principale limite de ce modèle est qu'il n'extrait pas les compétences ; mais plutôt les informations d'état civil (*nom, prénom, âge,...*) ainsi que les informations sur l'éducation (*diplômes, années d'obtention, etc.*). De plus, pendant la phase de segmentation, les caractéristiques des titres de section sont définies de manière empirique pour tous les CVs [5]. Pourtant, en pratique ces caractéristiques varient d'un CV à un autre.

Le modèle développé dans ce travail apporte une amélioration à celui présenté dans [5] en s'intéressant au contenu des sections et en utilisant une autre approche de segmentation. Ce travail aborde aussi la recherche des compétences dites de haut niveau des candidats. Les compétences de haut niveau sont celles que le candidat utilise pour résumer son CV. Les exemples peuvent être *Software engineer*, *Data scientist*,... Certaines de ces compétences sont explicitement mentionnées dans le CV (compétence explicite) et d'autres pas (compétence implicite).

3. Méthodologie

Le modèle d'identification des compétences dans les CVs développé dans cet article est un modèle hybride. Le processus général procède en des étapes suivantes :

– extraction des caractéristiques traduisant la compétence dans les CVs (termes et expressions exprimant les compétences) ;

– utilisation des caractéristiques extraites à l'étape précédente pour prédire des compétences de haut niveau que possède le candidat.

3.1. Extraction des caractéristiques par approche hiérarchique

Les approches hiérarchiques d'extraction des informations dans les CVs procèdent généralement en trois principales étapes telles que schématisées dans la figure 1 :

1) Segmentation du CV en blocs de texte représentant des sections potentielles ;

2) Classification des blocs ;

3) Extraction des informations détaillées dans chaque bloc.

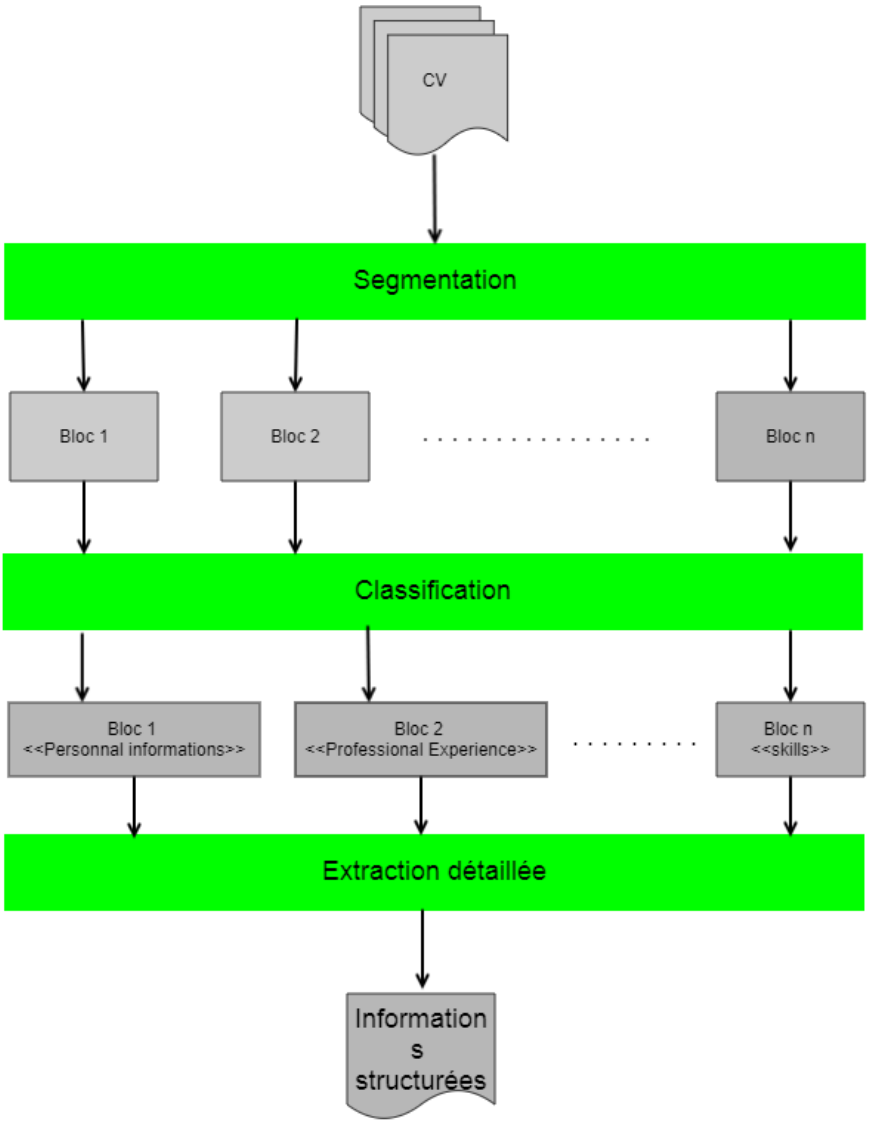


Figure 1. *Schéma d'extraction hiérarchique des informations dans un CV*

La première étape consiste à décomposer le CV en des blocs de texte, chacun représentant à priori une section ou la partie d'une section dans un CV. La deuxième étape quant à elle, a pour but d'affecter à chaque bloc identifié à la première étape une étiquette représentant une classe sémantique d'information généralement contenue dans le CV. Une classe sémantique peut être *EDUCATION*, *SKILLS*, etc,... Enfin la dernière étape va consister à extraire les informations détaillées dans chaque section en fonction de la nature des informations qui y sont contenues.

Une section dans le CV est un groupe d'informations appartenant à une même classe sémantique. Elle est généralement composée d'un titre et d'un contenu qui détaille les informations du candidat relatives à cette section. Par exemple la section "*SKILLS*" contient généralement les compétences de l'expert ; alors que la section "*EDUCATION*" contient les informations relatives à l'éduction.

3.1.1. Segmentation des blocs

Comme des paragraphes dans un texte qui se distiguent visuellement par le retour à la ligne à la fin et l'indentation en début, les blocs dans les CVs sont des ensembles consécutifs de textes proches les uns des autres qui au niveau visuel se délimitent les uns des autres par la largeur de l'espacement vertical ou horizontal (CV organisé en colonnes) entre ces blocs ou par un titre de section. Autrement dit, un bloc dans un CV est graphiquement une région dense en terme de distribution des mots dans l'espace [5]. Et l'espacement entre deux blocs est une zone peu dense toujours en terme de distribution de mots. L'espacement entre le contenu d'une section et le titre de la section suivante est plus grand que celui entre les lignes décrivant la section.

Définitions :

Définition 1 *Deux mots m_1 et m_2 sont proches si la distance entre les deux mots est inférieure à un seuil s.*

La figure 2 illustre de manière schématique comment la proximité entre deux mots est évaluée. Cette distance est évaluée comme étant la largeur de l'espacement horizontal (dh) (si ces deux mots sont sur la même ligne) ou verticale (dv) (si les deux mots appartiennent à des lignes consécutives) qui existent entre ces deux mots. Si les deux mots m_1 et m_2 sont sur la même ligne, la distance entre eux est mesurée comme étant le nombre de pixels entre la position x de son dernier caractère de m_1 et la position x du premier caractère du mot m_2. Dans l'autre cas où les deux mots sont sur des lignes consécutives, la distance est évaluée comme étant l'écart vertical en pixels entre les deux mots (dh sur la figure 2). Dans la pratique, le seuil de proximité horizontal (s_h) est différent du seuil de proximité vertical (s_v).

Définition 2 *Le s-voisinage d'un mot m noté $V(m)$ est l'ensemble des mots qui sont proches de m au regard d'un seuil des proximités s_h et s_v.*

Définition 3 *La fermeture transitive d'un mot m est l'ensemble des mots qu'on peut atteindre verticalement ou horizontalement à partir de m par pas de longueur inférieure à s_h si le déplace est fait horizontalement ou à s_v s'il s'est fait verticalement.*

Plus formellement, la fermeture transitive $F(m)$ d'un mot m dans le CV peut être défini comme suit :

- $m \in F(m)$
- $V(m) \subset F(m)$

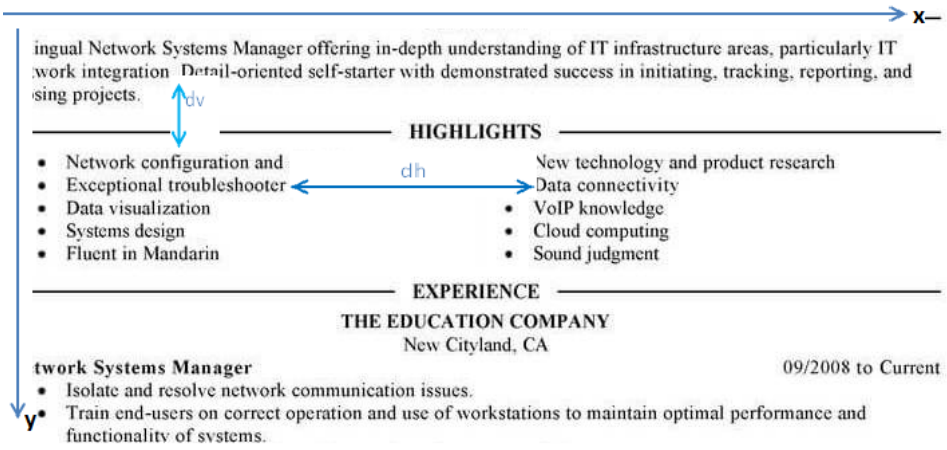


Figure 2. *Evaluation de la distance entre mots*

– $m_{k,(k\geq 2)}$ appartient à $F(m)$ si et seulement si il existe une suite de mots $m_1, m_2, ..., m_{k-1}$, telle que $\forall_{i=1,2,...k-1}\ d(m_i, m_i + 1) \leq s$.

Chaque mot dans un bloc a donc pour fermeture transitive le bloc lui-même. Le principe de l'algorithme de segmentation des blocs consiste donc à identifier toutes les fermetures transitives des mots du bloc. La rencontre d'un titre de section est également considérée comme étant la séparation entre deux blocs.

L'algorithme de segmentation est semblable à celui décrit dans l'article [5] et procède comme suit :

1) Définir un critère d'indentification des titres de section ;

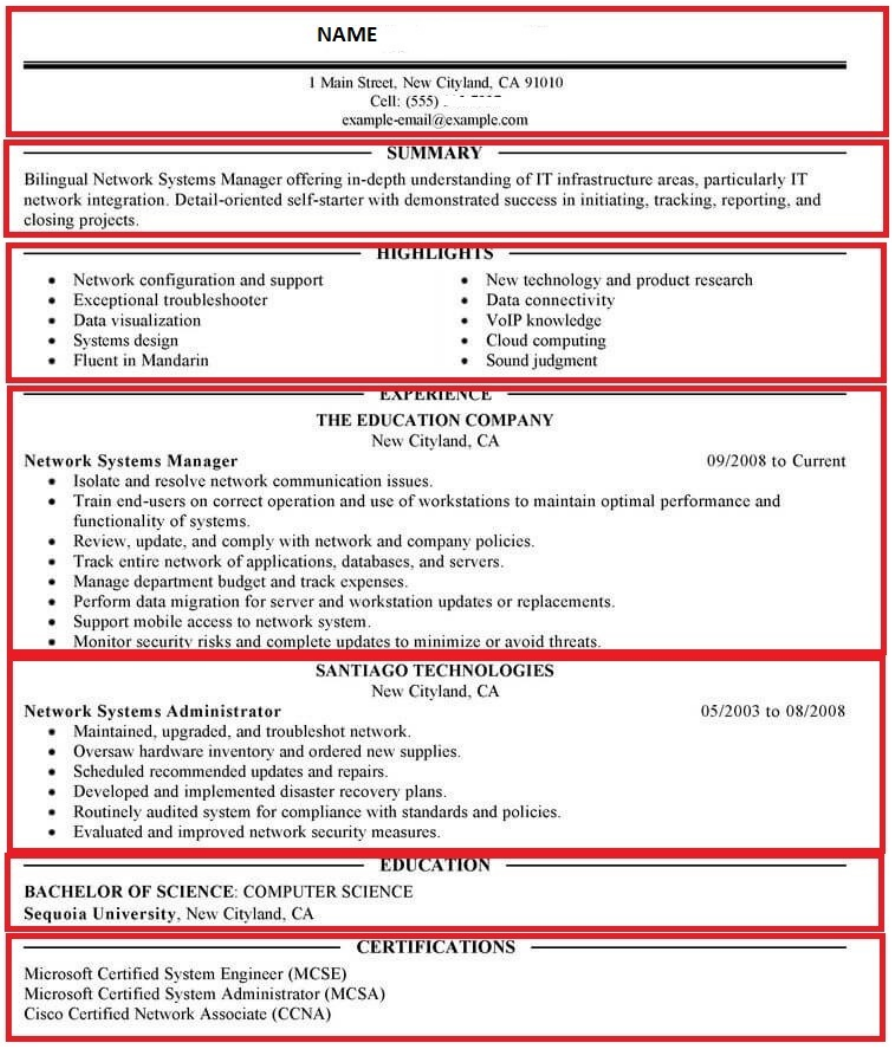
2) Initialement, on considère que chaque mot est un bloc. Les fermetures de ces mots sont construites en rajoutant progressivement les mots voisins, les voisins des voisins et ainsi de suite jusqu'à ce qu'on ne puisse plus rajouter un mot qui respecte le critère de proximité ou alors lorsqu'on recontre un titre de section.

3) S'il y a chevauchement entre des blocs alors ces blocs sont fusionnés.

Un exemple de résultat de segmentation appliquée à un CV anonyme est présenté en figure 3. Les différents blocs identifiés par l'algorithme sont représentés par les rectangles.

3.1.2. Identification des titres des différents blocs.

La rencontre d'un titre de section est un indicateur de la fin d'un bloc et du début d'un autre bloc. Identifier les titres de section dans un CV permet de délimiter les sections. A la différence de la méthode utilisée dans l'article [5] qui définit les caractéristiques des titres de section de manière empirique (taille de la police de caractère, épaisseur de la police de caractère, couleur) avec les mêmes valeurs peu importe le CV, l'approche proposée dans cet article pour identifier les titres se base sur l'hypothèse selon laquelle, les titres sont rédigés généralement avec le même style dans un CV et leurs caractéristiques se distinguent visuellement du reste du texte. Cette considération peut trouver une illustration dans le CV présenté en exemple à la figure 3 où les titres de section (*SUMMARY, HIGHLIGHTS, EXPERIENCE, THE EDUCATION COMPANY, SANTIAGO TECHNOLOGY, EDUCATION, CERTIFICATIONS*) ont des caractéristiques morphologiques (*majuscule, taille de la police de caractère, epaisseur de la police de caractère*) qui se distinguent clairement du reste du texte. Mais les caractéristiques qui font différer les titres du reste du texte sont propres à chaque CV et ne sauraient être généralisées à tous les CVs comme



NAME

1 Main Street, New Cityland, CA 91010
Cell: (555)
example-email@example.com

SUMMARY

Bilingual Network Systems Manager offering in-depth understanding of IT infrastructure areas, particularly IT network integration. Detail-oriented self-starter with demonstrated success in initiating, tracking, reporting, and closing projects.

HIGHLIGHTS

- Network configuration and support
- Exceptional troubleshooter
- Data visualization
- Systems design
- Fluent in Mandarin
- New technology and product research
- Data connectivity
- VoIP knowledge
- Cloud computing
- Sound judgment

EXPERIENCE

THE EDUCATION COMPANY
New Cityland, CA

Network Systems Manager 09/2008 to Current

- Isolate and resolve network communication issues.
- Train end-users on correct operation and use of workstations to maintain optimal performance and functionality of systems.
- Review, update, and comply with network and company policies.
- Track entire network of applications, databases, and servers.
- Manage department budget and track expenses.
- Perform data migration for server and workstation updates or replacements.
- Support mobile access to network system.
- Monitor security risks and complete updates to minimize or avoid threats.

SANTIAGO TECHNOLOGIES
New Cityland, CA

Network Systems Administrator 05/2003 to 08/2008

- Maintained, upgraded, and troubleshot network.
- Oversaw hardware inventory and ordered new supplies.
- Scheduled recommended updates and repairs.
- Developed and implemented disaster recovery plans.
- Routinely audited system for compliance with standards and policies.
- Evaluated and improved network security measures.

EDUCATION

BACHELOR OF SCIENCE: COMPUTER SCIENCE
Sequoia University, New Cityland, CA

CERTIFICATIONS

Microsoft Certified System Engineer (MCSE)
Microsoft Certified System Administrator (MCSA)
Cisco Certified Network Associate (CCNA)

Figure 3. *Example de CV segmenté en blocs*

dans [5]. Car les choix de ces caractéristiques sont propores à chaque candidat. Pour certains, l'epaisseur de la police de caractère fait la distinction, pour d'autres, la couleur pourrait aussi faire la différence, etc. C'est pour cette raison que l'approche proposée se base sur l'identification des caractéristiques distinctives propres au CV en procédant par la reconnaissance des titres bien connus.

Il existe des formulations de titres qui apparaissent très fréquemment dans les CVs. Ces formulations bien connues sont par exemple : Langues (*Languages*), Compétences (*Skills*), Formation académique (*Education*), Expériences professionnelles (*Work Experience*), ... Un dictionnaire de ces différentes formulations connues des titres de CV rédigé en langue anglaise a été ainsi constitué et contient 284 manières (pour nos expérimentations) de formuler les différents titres de sections du CV. Le dictionnaire va servir à reconnaitre des titres fréquemment utilisés. Ces caractéristiques vont permettre de retrouver les autres titres du CV qui ne se retrouveraient pas dans le dictionnaire. Les étapes d'identification les titres de sections du CV sont donc les suivantes :

1) Identifier dans le CV des lignes de textes qui correspondent aux titres renseignés dans le dictionnaire.

2) Extraire les caractéristiques morphologiques communes de ces lignes de texte.

3) Enfin, comparer les caractéristiques des titres-exemples identifiés à la première étape avec celles des autres lignes de texte ; en cas de similarité des caractéristiques d'une ligne avec un titre-exemple (titre inclus dans le dictionnaire), la ligne texte est donc considérée comme étant un titre.

Sur une echelle de 0 à 1, le seuil de similarité doit-être suffisamment grand (>0.95 utilisé pendant les expérimentations) pour éviter de détecter un titre qui n'en est pas un (faux positif).

Les caractéristiques utilisées pour évaluer la similarité entre les titres sont :

– La taille de la police de caractère ;

– Un indicateur positionné à 1 si la première lettre du titre est majuscule et le reste en minuscule et 0 sinon ;

– Un indicateur positionné à 1 si tout le titre est en majuscule et 0 sinon ;

– La couleur du titre (niveau de rouge de gris et de bleu) ;

– L'épaisseur de la police de caractère ;

La mesure de similarité utilisée est la similarité cosinus [11]. A chaque caractéristique est associé un poids qui traduit son degré de discrimination entre les titres et les portions de textes qui ne sont pas des titres. Il est possible dans de rares cas qu'une ligne ne représentant pas un titre dans le CV, mais apparaîssant dans le dictionnaire soit identifié comme titre. Etant donné que les titres de section ont des caractéristiques similaires qui se distinguent des autres lignes de texte du CV, un filtrage de ces lignes est fait en comparant entre elles les titres indentifiés et en ne retenant que ceux ayant des caractéristiques telles que (la police de caractère élevée, l'épaisseur, ou encore une couleur différente, tous les caractère en majuscule) similaires.

3.1.3. Classification des blocs

Une fois les blocs délimités, il faut les classer en utilisant un modèle de classification construit à partir d'un algorithme d'apprentissage. Le modèle de classification utilisé est le même que celui décrit dans l'article [5]. Les différentes classes que nous avons considérées pour les expérimentations sont les suivantes : *Work Experience*, *Education*, *Skills*, *Languages*, *Certification*, *Personnal Information*, *Hobbies*. Le modèle d'appren-

tissage utilisé est le SVM [19]. L'apprentissage est fait avec les exemples constitués des caractéristiques extraites des blocs. Ces caractéristiques extraites dans des blocs pour la classification sont celles listées dans [5] à savoir :

– La présence des séparateurs (",", ";" ou " :")

– La présence des dates

– La présence de nom de personnes

– La présence d'adresse mail

– Les mots clés apparaissant dans les titres de chaque bloc

A ces caractéristiques sont ajoutées la présence des noms des organisations et verbes d'action (*design, implement, troubleshoot*, etc.). Ces verbes sont très souvent utilisés dans les sections relatives à l'*expérience professionnelle* dans un CV. Une bonne identification des titres de section a également un grand impact dans la classification des blocs car le titre est l'élément qui contient l'information la plus importante qui permet de classer les blocs. Les termes extraits aussi bien dans les titres que dans les blocs sont lemmatizés en utilisant le *lemmatizer* de *wordnet* (WordnetLemmatizer) [7] à l'effet d'associer les termes qui partagent en commun un même lemme et traduisent à la base une information similaire. Ces informations sont utilisées pour représenter chaque bloc sous forme d'un vecteur de numérique en y associant sa classe qui n'est rien d'autre que le titre de la section (*EDUCATION, SKILLS*,...). Le jeu de données ainsi consitué est utilisé pour entrainer un SVM à l'effet de prédire pour un nouveau bloc son étiquette étant donnée les caractéristiques associées au bloc.

Après avoir étiqueté les différents blocs, la prochaine étape consiste à extraire les informations détaillées dans chacun des blocs étiquettés en fonction de la nature et de la structure des informations qui y sont contenues.

3.2. Extraction des informations détaillées dans chaque bloc

L'article [5] ne s'intéresse qu'à l'extraction des informations détaillées dans les blocs informations personnelles et éducation. Ce travail l'étend aux autres blocs qui renferment les caractéristiques permettant d'apprécier la compétence du candidat à savoir :

– Les blocs étiquetés « Compétences » et « Certifications »

– Les blocs étiquetés « Expérience professionnelle »

3.2.1. Extraction des compétences dans les sections *compétences* et Certifications

Généralement, dans les CVs, les éléments de la section « Compétence » (*SKILLS, COMPETENCIES, TECHNOLOGIES, etc.*) sont constitués de liste des compétences délimitées par un séparateur qui peut être (la virgule, le point-virgule, ou le caractère de retour à la ligne). Ainsi, pour cette section, les compétences sont extraites en scindant les éléments du texte avec comme séparateur la virgule, le point-virgule, l'espace ou le caractère de retour à la ligne. Pour chaque terme, une taxonomie du domaine est utilisée pour valider s'il s'agit d'une compétence du domaine ou pas. La taxonomie utilisée a été construite sur la base des compétences extraites de dice.com [1] qui contient plus de 5000 termes du domaine de l'informatique.

1. consulté le 21 septembre 2018

3.2.2. Extraction des compétences dans les blocs *Expérience professionnelle*

Plusieurs titres de section dans un CV correspondent à l'expérience professionnelle. Il s'agit des titres tels que : (*Work experience, Project experience, Publications, etc.*) Dans ces sections, les compétences sont généralement exprimées sous forme de phrases commençant par un verbe d'action (par exemple *identify cyber threat signature*) ou sous sa forme nominale du verbe suivie par la préposition « of » (par exemple *Indentification of cyber threat signatures*). Un dictionnaire de tels verbes d'action et de leur forme nominale a été constitué. Une règle est appliquée pour identifier ce style de proposition dans les textes. La règle qui a été définie pour extraire ces clauses est la suivante :

<VRB|NN of><complément>[<IN><complément>]. (R)

avec la nomemclature proposée par [4] où *VRB* pour *verbe* (Verb), *IN* pour *dans* (in), *NN* pour *nom* (Noun), etc. Avant d'appliquer cette règle, la simplification de phrase est utilisée pour décomposer toutes les phrases composées (ayant plusieurs propositions reliées par les conjonctions) en plusieurs sous phrases ayant un sens unique. Il a été montré dans [18] que la simplification des phrases est un prétraitement qui permet aux outils de Traitement Automatique de la Langue Naturelle (TALN) d'avoir de meilleurs résultats.

Exemple 1 *Le résultat de la simplification syntaxique l'expression "detection and identification of cyber-attack signatures" est constitué des propositions simples suivantes :*

– detection of cyber-attack signatures.

– identification of cyber-attack signatures.

Après cette étape de simplification, l'algorithme d'extraction peut facilement extraire les savoir-faire en appliquant la règle précédemment définie.

L'outil *stanford dependency parser* [4] est utilisé pour étiquetter le texte c'est-à-dire attribuer une étiquette syntaxique à chaque mot. Cet outil identifie aussi les dépendances entre les mots en déterminant les relations entre les mots dans l'expression (par exemple *subj* : sujet du verbe, *cobj* : complément d'objet, *comp* : mot composé à, etc.). A partir de ces étiquettes et des relations entre les mots, un parcours du graphe de dépendance est effectué à l'effet d'extraire les expressions caractérisant des compétences : celles qui respectent la règle (R).

L'algorithme d'extraction utilisant la règle (R) appliquée à l'exemple 1 va générer les savoir-faire ci-après :

– detection of cyber-attack signatures.

– identification of cyber-attack signatures.

Les compétences exprimées sous forme de savoirs (technologies, outils, sujets maîtrisés) sont également extraites dans cette section en recherchant les *n-grams* (de 1 à 5) et en se servant d'une taxonomie du domaine pour déterminer si chaque *gram* extrait de la section *Experience professionnelle* correspond à une compétence. Chaque *gram* est lemmatisé avant d'être comparé au lemme du nom de l'entité dans la taxonomie. Les mots vides (*stop words*) sont également éliminés.

Exemple 2 *La procédure d'extraction des compétences appliquées à l'exemple 1, le résultat sur les savoirs suivants avec leur classe est :*

– cyber threat signatures : Topic

– cyber attack : Topic

3.3. Prédiction des compétences de haut niveau à partir des caractéristiques extraites

Une compétence est dite de haut niveau si elle est obtenue par regroupement d'un certain ensemble de connaissances élémentaires. Elle peut être explicite ou implicite. Dans cette partie, il s'agit d'utiliser les caractéristiques extraites dans les étapes précédentes, pour déduire des compétences de haut niveau.

Exemple 3 *Des connaissances élémentaires html, css, javascript, php peuvent se déduire la connaissance de haut niveau développeur web. De même des connaissances android, ionic, ios, java, cordova peut se déduire developpeur mobile ; ou encore : de Switch, routeur, paquet, réseau peut également se déduire administrateur réseau.*

Pour cette tâche, une approche suppervisée où la liste des compétences est connue à l'avance pour chaque CV. Il est donc question de prédire la liste des compétences explicites (mentionnées dans le CV) et implicites (non mentionnées sur le CV) détenues par un expert à partir de l'ensemble des caractéristiques extraites de son CV. Généralement dans les processus de recrutement, l'ensemble des compétences attendues des candidats est connues à l'avance. En suivant ce principe, le modèle dévelopé dans cette section se base sur une liste de compétences mentionnées dans les CVs et ces compétences feront office de classe. Ainsi étant donné un CV, il est question de savoir lesquelles parmi ces compétences (explicites et implicites) il possède. Pour nos expérimentations la liste de compétence retenue est la suivante :

project manager	oracle database administrator	business analyst
technical lead	java developer	software ingeneer
consultant	analyst	software developer
network administrator	web developer	IT analyst
IT consultant	manager	

Le problème de prédiction des compétences d'un expert peut être vu comme un problème de classification multi-label [9] où pour chaque expert, dans un CV, plusieurs compétences peuvent être mentionnées. Par exemple, il est fréquent de rencontrer *analyst* et *software developer* dans les CVs. Formellement un problème de classification multi-label est défini par la définition 4.

Définition 4 *Etant donné un ensemble de données X et un ensemble de classe Y , la classification multi-label consiste à associer, à un élément de X, un sous ensemble de Y*

Pour les expérimentations, l'ensemble Y est constitué des quatorze compétences citées plus haut.

La figure 4 illustre l'architecture globale de construction et de fonctionnement du modèle de prédiction des compétences de haut niveau des CVs. Partant d'un jeu de données constitués à partir d'un ensemble de CVs, chacun étiquetté par l'ensemble des compétences qui y apparaissent le processus se fait en plusieurs étapes décrites de la manière suivante :

– Représentation des données ;

– transformation du jeux de données en n jeux chacun centré sur une compétence ;

– Apprentissage d'un modèle par jeu pour prédire la compétence associée au jeu ;

– Aggregation des résultats des prédicteurs de base

3.3.1. Représentation des CV

Chaque CV est initialement représenté par un sac de mots, puis transformé en une représentation vectorielle en utilisant la technique *word2vec* [16]. Les mots dans le sac sont ceux qui renseignent sur les compétences. Le modèle de représentation considéré est le modèle *skip gram* [15] avec comme taille de la fenêtre 5 et comme dimension de 300. La représentation *skip gram* permet de prendre en considération le lien sémantique entre le terme dans la construction du modèle d'apprentissage. Deux termes (*programmer* et *developer*) par exemple auront des représentations vectorielles similaires car ils sont employés dans des contextes similaires.

3.3.2. Construction du jeu de données

A cette étape, le processus de classification *multi-label* est le *Binary Relevance* [9] qui consiste à transformer le problème de classification multi-label en autant de problèmes de classification binaire qu'il y a de classes. Le jeu de données initial est utilisé n fois (où n représente le nombre de classes/compétences), chaque fois pour une compétence particulière. Pour chaque compétence, le vecteur des caractéristiques du CV obtenu à l'étape de représentation reste inchangé. Seule la classe change. Si le jeu i est fait pour la compétence C_i, alors la classe de chaque exemple représentera la présence (1) ou l'absence (0) de la compétence C_i dans l'exemple considéré. Ainsi donc, le modèle construit à partir du jeu de données i permettra de prédire si un exemple possède (1) ou non (0) la compétence C_i.

3.3.3. Apprentissage du modèle de classification

Une fois le jeu de données initial transformé en n jeux centrés chacun sur une compétence unique, chaque jeu de données est ensuite utilisé pour entrainer un prédicteur binaire dont le rôle est de prédire si l'exemple placé en entrée possède la compétence. Le modèle de classification utilisé pour se faire est le SVM qui est bien adapté pour les problèmes de classification binaire.

3.3.4. Aggregation des résultats

A la fin du processus, les prédictions de chaque modèle de base sont aggrégées en faisant l'union des classes prédites par chacun. Ainsi donc, nous avons la liste des compétences prédites de l'expert parmi les compétences d'intérêt. A la fin, l'ensemble des compétences de l'expert est donc l'union des compétences prédites par chaque prédicteur de base.

4. Expérimentations

4.1. Données

Un "*web scraper*" a été développé pour collecter automatiquement les CVs sur les sites Internet. Ces CVs sont structurés au format PDF. Un jeu de 800 CVs rédigés en anglais a été constitué pour les expérimentations. Le jeu données a été divisé en 4 sous-ensembles. l'expérimentation s'effectue en quatre étapes et à chaque étape, un sous-ensemble distinct (1/4 du jeu de donnée) est utilisé pour l'expériementation et les autres sous-ensembles (3/4 du jeux de données) sont utilisés pour la validation. Les résultats finaux consignés dans des tableaux sont obtenus en faisant la moyenne des résultats sur les quatres exépriences. Chaque aspect important du modèle a été évalué individuellement. Une mesure d'évaluation appropriée a été utilisée pour chacun de ces aspects.

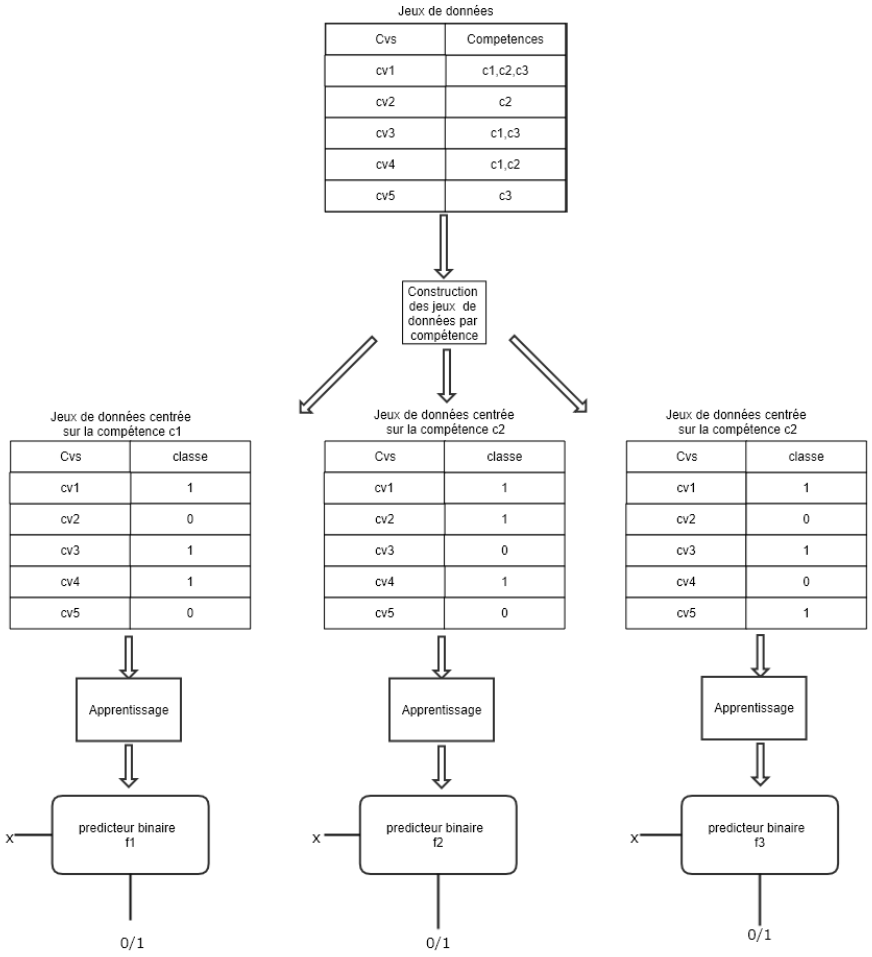


Figure 4. *Modèle de prédiction multi-label*

4.2. Evaluation de l'algorithme de division du CV en blocs

Mesure d'évaluation

La mesure d'évaluation utilisée pour mesurer l'acuité de l'algorithme de division du CV en blocs est la *précision* que nous nous proposons de la calculer avec la formule 1.

$$Precision = \frac{NBC}{NT}. \qquad [1]$$

Tableau 1. *Résultat de la division du CV en bloc.*

Format du CV	Tabulaire	Linéaire
Précision	84,32%	89,86%

Dans la formule 8, NBC représente le nombre de blocs correctement délimités et NT le nombre total de blocs. Un bloc est considéré comme correctement délimité s'il s'agit

d'une section de premier niveau (bloc non imbriqué dans un autre) ou d'une sous section (bloc situé au premier niveau d'imbrication). Cette considération évite de constituer des blocs de très petites tailles. Car les blocs de très petites tailles ne permettent pas dans bien des cas, d'avoir suffisamment d'informations pour déduire la classe du bloc. L'idéal aurait été que les blocs correspondent exactement aux sections du CV. Néanmoins, après la phase de classification, les blocs ayant la même étiquette sont fusionnés pour reconstituer la section initiale correspondante à l'étiquette en question. La précision varie en fonction de la structure du CV. La détection des blocs opère très bien (98.32%) sur les CVs rédigés dans une structure linéaire-verticale (où les sections sont empilées les unes sur les autres 3). En effet, dans les CVs rédigés sous ce dernier format, les erreurs constatées dans la division des blocs sont principalement dues aux erreurs de détection des titres de section (le style d'écriture du titre de la section se distingue très peu du style d'écriture du reste du texte). Par contre, lorsque le CV présente une structure complexe (à l'exemple des CVs disposés sous forme de tableau), l'algorithme de détection des blocs ne produit pas toujours de très bons résultats comme dans le cas linéaire.

4.3. Evaluation de la classification des blocs

Cette sous section présente les résultats du modèle de classification des blocs.

Mesure d'évaluation

Pour une évaluation détaillée de l'algorithme de classification des blocs, une matrice de confusion (C) (tableau 4.3) est calculée. Dans cette matrice, les lignes représentent les classes réelles et les colonnes représentent les classes prédites. Chaque élément C_{ij} de la matrice représente le nombre d'éléments donc la classe réelle est i et qui ont été prédits par le modèle comme appartenant à la classe j. Par la suite, trois mesures synthétiques sont utilisées pour évaluer l'acuité du modèle pour chaque classe d'une part, et l'acuité générale du modèle en faisant la moyenne des valeurs obtenues pour chaque classe d'autre part. Ces mesures sont :

– Le Rappel (R_i) pour une classe i, mesure la proportion d'exemples qui ont été correctement assignés à la classe i parmi tous les exemples dont la classe réelle est i. La formule est donnée par :

$$R_i = \frac{C_{ii}}{\sum_j C_{ij}}. \quad [2]$$

– La précision (P_i) qui mesure la proportion des bonnes prédictions sur le nombre total de prédictions, et donc la formule est donnée par l'équation 3 :

$$P_i = \frac{C_{ii}}{\sum_j C_{ji}}. \quad [3]$$

– La precision moyenne peut être obtenue en faisant simplement la moyenne des précisions des différentes classes :

$$Precision = \frac{\sum_i P_i}{N}. \quad [4]$$

– Le rappel moyen est obtenu en calculant la moyenne des rappels des différentes classes :

$$Rappel = \frac{\sum_i R_i}{N}. \quad [5]$$

		Classes prédites							
		Education	Work Experience	Skills	Certifications	Personal Infos	Languages	Hobbies	Rappel
Classes réelles	Education	180	12	0	4	4	0	0	90%
	Work Experience	7	178	13	2	0	0	0	89%
	Skills	2	3	178	10	10	6	0	89%
	Certifications	16	0	7	171	6	0	0	85,5%
	Personal infos	0	7	0	0	193	0	0	96,5%
	Languages	2	0	7	0	0	191	0	95,5%
	Hobbies	0	0	5	0	0	0	195	97,5%
	Précision	86,95%	90,35%	84,76%	91,44%	92,34%	96,95%	100%	

La matrice de confusion représentée dans le tableau 4.3, nous montre que le modèle prédit certaines classes avec une précision plus importante que d'autres. Par exemple, le modèle prédit un bloc d'étiquette réelle ***Languages*** avec une précision de 96,95% alors qu'un bloc d'étiquette réelle *Education* est prédite avec une précision de 86,95%. La valeur élevée de la précision pour la section ***Languages*** est probablement dûe au fait que cette section se distingue avec peu d'ambiguité des autres sections par la présence des *language*, des termes clés d'appréciation du niveau de la langue, et par la taille relativement courte du bloc. Les erreurs dans les autres blocs sont dûes au ressemblance de contenu dans les blocs. Comme on pourrait le constater dans la matrice de confusion. Par exemple, sur un total de 178 exemples appartenant à la classe "*Work Experience*", 7 ont été prédits comme étant de la classe *Education*, et 13 ont été prédit s comme étant de la classe *Skills*. En effet, les même caractéristiques qu'on peut retrouver dans la section "*Work Expérience*" peuvent se retrouver dans la classe "*Education* par exemple" Ce qui peut justifier les confusions effectuées par le modèle. Le modèle de classification des blocs opère généralement avec une précision moyenne évaluée à 91.83% et un rappel global évalué à 91,85% (tableau 2).

Tableau 2. *Classification des blocs : Evaluation globale*

	Précision moyenne	Rappel moyen	F-mesure
Modèle	91,82%	91,85%	91,39%

4.4. Evaluation du modèle de prédiction d'extraction des compétences explicites

Mesure d'évaluation

L'idée de l'évaluation est de mesurer l'aptitude du système à extraire tous les termes caractérisants la compétence et uniquement ceux là. Pour un CV i donné, notons Y_i est l'ensemble des termes clés extraits par le modèle ; Z_i est l'ensemble des termes clé extraits par l'expert du domaine et n le nombre de CV soumis à l'évaluation. Les mesures utilisées pour évaluer la classification des compétences par le modèle sont les suivantes :

– L'exactitude qui mesure pour un CV i donné, le pourcentage des termes correctement identifiés par l'algorithme sur l'ensemble des termes (ceux prédits par l'algorithme et ceux identifiés par l'expert). L'exactitude moyenne est obtenue en faisant la moyenne des exactitudes calculées sur tous les CV. Elle se calcule par la formule 6 :

$$Exactitude = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i \cap Z_i}{Y_i \cup Z_i}\right|. \quad [6]$$

– Le Rappel qui mesure pour un CV i donné, la proportion des termes identifiés par l'expert qui ont été correctement identifiés par l'algorithme. Le rappel moyen est caculé en faisant la moyenne des rappels obtenus sur tous les instances de CVs. Il est donné par la formule 7 :

$$Rappel = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i \cap Z_i}{Z_i}\right|. \quad [7]$$

– La précision pour un CV i mesure la proportion des termes clés qui ont été correctement prédites parmi celles que le modèle a prédite. La précision moyenne (formule 8) est déterminée en faisant la moyenne des précision calculée sur tous les instances de CVs utilisés pour le test.

$$Precision = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i \cap Z_i}{Y_i}\right|. \quad [8]$$

– La *F-mesure* qui est une mesure qui agrège les deux mesures précédentes, de trouver un compromis entre la précision et le rappel.

$$F - mesure = \frac{2 * Rappel * Precision}{Rappel + Precision}. \quad [9]$$

Résultats

Les résultats d'expérimentations sont capturés dans le tableau 3. La précision de 96.2% signifie 96.2% des CVs indentifiés par le modèle sont corrects. Les erreurs sont principalement dues au fait que le modèle identifie des termes homonymes des termes du domaine mais dont le contexte d'emploi n'en fait pas une caractéristique du domaine ; l'exactitude de 81,4% traduit qu'en moyenne, le pourcentage de correspondance entre les compétences prédites et celles attendues. D'autre part, le rappel de 84.32% signifie que le modèle réussi à retrouver 84.32% des caractéristiques qu'on souhaitait extraire du modèle et qui ont été identifiée au préalable par un expert du domaine informatique. Le gap au niveau du rappel est du au fait que la règle (R) n'est pas une règle exhaustive et ne permet pas d'extraitre tous les compétences de type savoir faire. D'autres part, les erreurs liées à la lemmatisation induisent le modèle à ne pas reconnaitre des termes qui ont une même racine lexicale.

Tableau 3. *Résultat de l'extraction des compétences explicites.*

	Exactitude	Précision	Rappel	F-mesure
Modèle	81,4%	96,2%	84,32%	89,86%

4.5. Evaluation du modèle de prédiction des compétences implicites

Les CVs ont été étiquettés en y introduisant les compétences que les candidats n'ont pas explicitement mentionnés. Ces nouvelles étiquettes sont des compétences implicites du candidat.

Mesure d'évaluation

Les mêmes mesures d'évaluation utilisées au niveau de l'extraction des compétences explicites sont considérées dans cette partie . La semantique de ces mesures dans le contexte de la prédiction des compétences implicites est la suivante :

– L'exactitude (formule 6) qui mesure la proportion des compétences prédites correctement parmi l'ensemble des compétences (prédites et celles qu'on devait prédire).

– La précision (formule 8) qui mesure la proportion des compétences correctement prédites parmi les compétences qui ont été prédite par la modèle.

– Le rappel (formule 7) qui mesure la proportion des compétences correctement prédites parmi les compétences qui devaient être prédites.

Le résultat de la prédiction des compétences implicites est représenté dans le tableau 4. Dans ce tableau, l'exatitude traduit à priori en moyenne que la probabilité que le résultat prédit par le modèle corresponde au résultat correct est de 88.6%. La précision de 92,3% signifie qu'en moyenne 92,3% des compétences prédites sont correctes. Le rappel de 90,5% signifie que le modèle prédit en moyenne 90,5% des compétences consituant les étiquettes de chaque exemple. Ces résultats montrent la capacité du modèle à proposer aux récruteurs les autres compétences (non mentionnées) des candidats.

Tableau 4. *Résultat de la prédiction des compétences implicites.*

	Exactitude	Précision	Rappel	F-mesure
Modèle	88,6%	92,3%	90,5%	91,39%

4.6. Contribution de l'approche de détection des titres sur la division des blocs

Le tableau 5 présente les résultats des approches de segmentation.

Tableau 5. *Comparaison des approches de segmentation.*

Approches	Précision
Approche basée sur les titres-exemples	82.00%
Approche caractérisant les titres de façon empirique	68.82%

Résultats

Comme le montre le tableau 5 l'approche modifiée pour effectuer la segmentation est plus précise que celle décrite dans l'article [5]. Ceci peut s'expliquer par le fait que dans un cas, les caractéristiques des titres sont définies de manière empirique ; or dans l'autre cas, ces caractéristiques sont extraites des titres identifiés dans le CV et donc sont spécifiques au CV en question.

5. Conclusion et Perspectives

Cet article décrit l'utilisation du modèle hybride pour l'extraction des compétences dans les CVs. Le CV est segmenté en sections en utilisant une version modifiée de l'algorithme proposé dans [5]. La classification des sections a aussi été adaptée en intégrant le titre de la section ainsi que les verbes d'action communément utilisés dans la rédaction des CVs comme caractéristiques pour la classification. Après l'étiquettage des blocs, les caractéristiques ont été extraites et utilisées pour construire un modèle de claissification multi-label permettant de classer les CVs suivant les compétences. L'algorithme de segmentation modifié offre une meilleure précision dans la segmentation que la méthode de base. Le modèle de classification multi-label a offert des résultats satisfaisants (de l'ordre de 90%).

Comme travail futur, il sera question d'utiliser les caractéristiques explicites extraites du CV pour déduire des compétences implicites. Il est fréquent que les candidats ne mentionnent pas toutes leurs compétences dans les CVs, mais qu'en lisant ce CV, un Homme est capable de déduire ces compétences non mentionnées.

6. Bibliographie

Shaha Alotaibi. A survey of job recommender systems. *International Journal of the Physical Sciences*, 7, 07 2012.

Ziemowit Bednarek. Skills gap : The timing of technical change. *Journal of Economics and Business*, 74 :57 – 64, 2014.

Vladlena Benson, Stephanie Morgan, and Fragkiskos Filippaios. Social career management : Social media and employability skills gap. *Comput. Hum. Behav.*, 30 :519–525, January 2014.

Marie catherine De Marneffe and Christopher D. Manning. Stanford typed dependencies manual, 2008.

Jiaze Chen, Liangcai Gao, and Zhi Tang. Information extraction from resume documents in pdf format. *Electronic Imaging*, 2016(17) :1–8, 2016.

ASSOCIATION FOR TALENT DEVELOPMENT. Bridging the skills gap : workforce development is everyone's business n2, 2015.

Christiane Fellbaum, editor. *WordNet : an electronic lexical database*. MIT Press, 1998.

Jiechieu Kameni Florentin Flambeau and Norbert Tsopze. Approche hiérarchique pour extraire les compétences dans les cvs en format pdf. In *Proceedings de la conférence sur la recherche en Informatique CRI'17*, ENSP, Yaoundé, 2017.

Eva Gibaja and Sebastián Ventura. Multi-label learning : A review of the state of the art and ongoing research. *Wiley Int. Rev. Data Min. and Knowl. Disc.*, 4(6) :411–444, November 2014.

Shiqiang Guo, Folami Alamudun, and Tracy Hammond. Résumatcher : A personalized résumé-job matching system. *Expert Systems with Applications*, 60 :169 – 182, 2016.

Anna Huang. Similarity measures for text document clustering. In *Proceedings of the 6th New Zealand Computer Science Research Student Conference*, 01 2008.

Thimma Reddy Kalva. *All Graduate Plan B and other Reports*, chapter Skill Finder : Automated Job-Resume Matching System. 2013.

Ilkka Kivimäki, Alexander Panchenko, Adrien Dessy, Dries Verdegem, Pascal Francq, Hugues Bersini, and Marco Saerens. A graph-based approach to skill extraction from text. In *Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing*, pages 79–87.

Association for Computational Linguistics, 2013.

Sumit Maheshwari, Abhishek Sainani, and P. Krishna Reddy. An approach to extract special skills to improve the performance of resume selection. In Shinji Kikuchi, Shelly Sachdeva, and Subhash Bhalla, editors, *Databases in Networked Information Systems*, pages 256–273, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

Sergio Miranda, Francesco Orciuoli, Vincenzo Loia, and Demetrios Sampson. An ontology-based model for competence management. *Data Knowl. Eng.*, 107(C) :51–66, 2017.

Yifan Peng, Catalina O Tudor, Manabu Torii, Cathy H Wu, and K Vijay-Shanker. isimp : A sentence simplification system for biomedicail text. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Kun Yu, Gang Guan, and Ming Zhou. Resume information extraction with cascaded hybrid model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 499–506, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

Meng Zhao, Faizan Javed, Ferosh Jacob, and Matt McNair. SKILL : A system for skill identification and normalization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 4012–4018, 2015.