



**HAL**  
open science

## Approche hiérarchique d'extraction des compétences dans des CVs en format PDF

Florentin Flambeau Jiechieu Kameni, Norbert Tsopze

► **To cite this version:**

Florentin Flambeau Jiechieu Kameni, Norbert Tsopze. Approche hiérarchique d'extraction des compétences dans des CVs en format PDF. 2019. hal-01898913v2

**HAL Id: hal-01898913**

**<https://hal.science/hal-01898913v2>**

Preprint submitted on 18 Jun 2019 (v2), last revised 1 Oct 2019 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# Approche hiérarchique d'extraction des compétences dans des 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 l'approche hiérarchique d'extraction des informations dans le CV pour en extraire les compétences. 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 les termes représentant les compétences (compétences de bases) sont extraits; et une phase de prédiction qui consiste à partir des caractéristiques extraites précédemment, à prédire un ensemble de compétences qu'un expert aurait déduites, et que ces compétences ne seraient pas nécessairement mentionnées dans le CV (compétences implicites). Les principales contributions de ce travail sont : l'utilisation de l'approche hiérarchique de segmentation du CV en sections pour extraire les compétences dans le CV; l'amélioration de la l'approche de segmentation des CV; enfin, l'utilisation de l'approche binary relevance de classification multi-label pour prédire les compétences implicites du CV. Les expérimentations effectuées sur un jeu de CVs collectés sur Internet ont montré une amélioration de la précision de l'identification des blocs de plus de 10% comparé à un modèle de l'état de l'art. Aussi, 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 résumé in section before extracting the competences; the use of the multi-label learning model based on Support Vector Machine (SVM) 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, classification multi-label

**KEYWORDS :** Skill Gap, Resume, Skills Extraction, multi-label classification



---

## 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* [2, 21, 12]. Pour combler cet écart, les entreprises organisent généralement des recrutements en sélectionnant les candidats parmi ceux qui ont déposé un dossier physique, ou électronique à travers le site internet de l'entreprise, la messagerie électronique, ou encore un site d'offres d'emploi. Ces dossiers sont composés entre autres des demandes d'emploi, des CVs et des lettres de motivation. Une des tâches de la direction des ressources humaines est alors d'identifier les compétences détenues par les candidats à la lecture de leur CV. L'extraction des compétences dans les CVs est un problème bien connu dans la littérature [12, 4, 14]. Il s'agit d'une activité qui trouve son application dans les processus de recrutement automatique et dont la finalité est d'effectuer un *matching* automatique entre les compétences extraites du CV et celles requises par l'organisation pour effectuer certaines tâches [16]. L'extraction automatique des compétences à partir des CVs permet aussi de réduire l'erreur liée à la subjectivité humaine dans le processus de recrutement.

Plusieurs travaux dans la littérature ont abordé le problème de *skill-gap* allant de la modélisation ontologique [14] au matching entre les profils des candidats et les offres d'emplois [16] en passant par les modèles d'évolution et d'extraction des compétences [4, 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 offres d'emploi. L'approche hiérarchique d'extraction des informations dans le CV[4] considère un CV comme un ensemble de blocs ou sections (*Education, Publications, Expérience professionnelle,...*) 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 segmentation du CV en sections, ainsi que de l'extraction des compétences de base particulières à chaque section (comme dans [8]), le présent article propose un modèle pour prédire des compétences implicites (compétences qui peuvent être déduites de la lecture du CV sans pour autant être mentionnées dans le CV) à partir des compétences de base extraites dans les différentes sections. L'objectif de ce travail est donc d'adapter le modèle proposé par Zhen Chen et al. [4] qui se base sur l'approche hiérarchique consistant à segmenter le CV en sections pour extraire les caractéristiques appropriées dans chaque section. En fait, l'algorithme de segmentation hiérarchique utilisé dans l'approche de Zhen Chen et al. [4] se base sur la densité des mots et considère les titres de section comme délimitateurs des sections. Les auteurs estiment que cet algorithme peut être amélioré, car les performances n'étant pas très satisfaisantes ; ceci peut s'expliquer entre autres par le fait que les caractéristiques utilisées par les auteurs pour identifier les titres de section soient définies de manière empirique, et les mêmes caractéristiques sont utilisées pour tous les CVs. Or en pratique ces caractéristiques peuvent varier d'un CV à l'autre. Pour améliorer la détection des titres de section et par ricochet la segmentation du CV en sections, nous avons proposé une approche basée sur l'identification des titres connues (*titres-exemples*) à l'aide d'un dictionnaire, et l'utilisation de leurs caractéristiques morphologiques dans le CV pour retrouver les autres titres de section. Après l'extraction des compétences de base de chaque section, un modèle de classification *multi-label* [6] est proposé pour prédire les compétences des candidats : ce modèle a l'avantage de pouvoir prédire des compétences qui ne sont pas explicitement

mentionnées dans le CV. 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 réel problème de segmentation car, en général bien structurés.

Le reste de cet article est organisé de la manière suivante : la prochaine section présente un ensemble de travaux traitant différents aspects du problème de *skill-gap* y compris l'extraction des compétences. 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 par la conclusion.

---

## 2. Etat de l'art

Le *skill-gap* est un problème traité par plusieurs communautés de chercheurs. Plusieurs aspects du problème et plusieurs approches de résolution sont présentés dans la littérature parmi lesquels : la correspondance ou *matching* entre les CVs et les offres d'emploi ; la mise en place des systèmes de recommandation d'offre d'emploi aux candidats ; la construction des modèles de compétence en utilisant les ontologies ; la recherche des compétences dans les CVs, etc.

Benson et al. 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 des opportunités offertes par ces réseaux. Ils recommandent au gouvernement britannique de se pencher sur cette question afin de créer des compétences se basant sur la manière avec laquelle les professionnels exploitent les réseaux sociaux. Partant du constat de la disparition des certaines compétences aux Etats-Unis et de la difficulté à avoir les travailleurs qualifiés pour certaines tâches, Bednarek propose dans [21] un modèle d'évolution des compétences qui tient compte du caractère (du travail) variant avec le temps et de la liaison avec les caractéristiques de la population.

Miranda et al. dans [14] proposent une ontologie pour représenter aussi bien les compétences que les offres d'emplois dans 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'emploi ; et aussi sur le recrutement 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 [16] proposent *RésuméMatcher*, un système capable d'extraire les qualifications et l'expérience professionnelle des candidats, d'extraire également les caractéristiques sur les offres d'emploi 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. A la différence de cet outil (*RésuméMatcher*) qui fait le *matching* entre les CVs et les offres d'emploi, le travail présenté dans cet article s'intéresse à l'extraction des compétences contenues dans les CVs.

Au sujet de l'extraction des compétences dans les textes, des approches basées sur la reconnaissance des entités nommées [12, 17] et les approches basées sur l'utilisation des graphes [9] ont aussi été proposées. Les approches du premier groupe construisent par apprentissage un modèle d'étiquetage des entités à partir d'un corpus, et utilisent 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 [12, 10]. 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. Mahe-shwari et al.[10] 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'un des modèles hiérarchiques d'extraction d'information dans les CVs les plus récents a été proposé en 2016 par Chen et al. [4]; 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 [4]. 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 [4] 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 (compétences explicites) et d'autres pas (compétences implicites).

---

### 3. Méthodologie

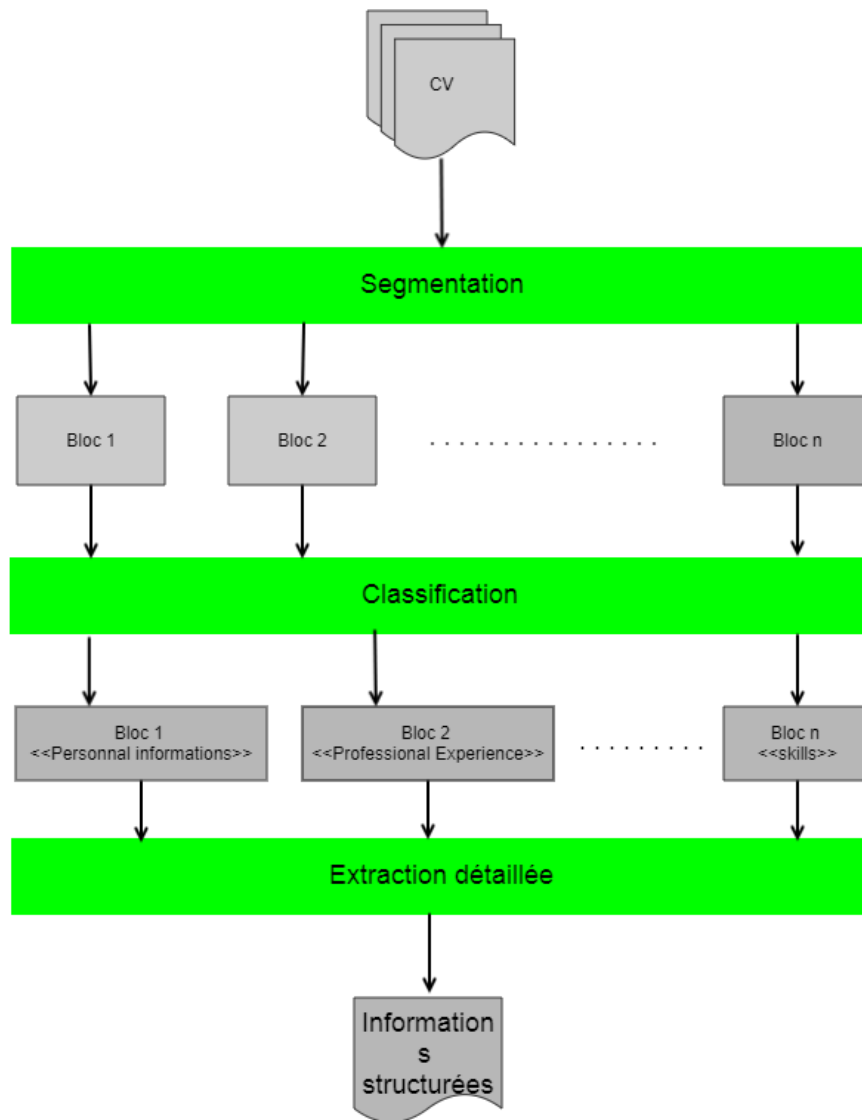
Le modèle d'identification des compétences dans les CVs développé dans cet article est un modèle basé sur l'approche hiérarchique d'extraction des informations dans les CVs. Le processus général procède en des étapes suivantes :

- division du CV en sections et extraction dans chaque section, des termes traduisant la compétence que l'on appellera aussi compétences de base;
- 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 l'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;



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

- 2) Classification des blocs ;
- 3) Extraction des informations détaillées dans chaque bloc.

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. La 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 son éducation.

### 3.1.1. Segmentation des blocs

Comme des paragraphes dans un texte qui se distinguent 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 relativement proches les uns des autres qui au niveau visuel se délimitent par la largeur de l'espace 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 [4]. Et l'espace entre deux blocs est une zone peu dense. L'espace entre le contenu d'une section et le titre de la section suivante est plus grand que celui entre les lignes décrivant la section.

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

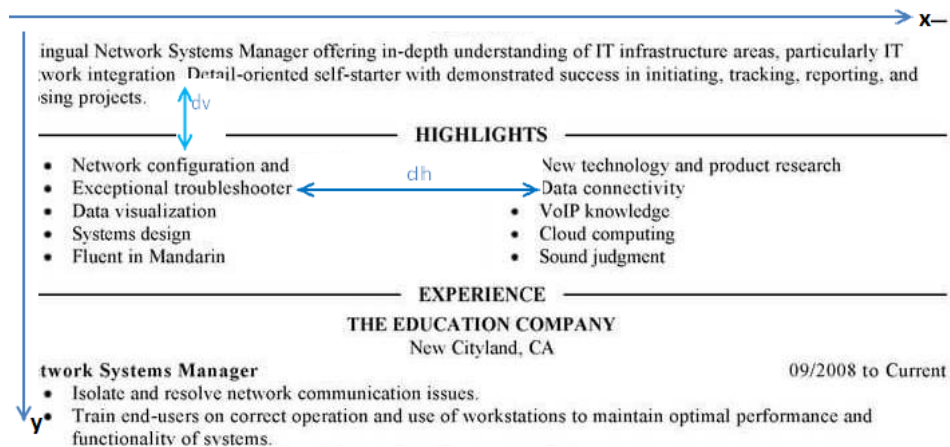
La figure 2 illustre de manière schématique comment la proximité entre deux mots est évaluée. Cette distance est évaluée comme étant la largeur de l'espace horizontal ( $dh$ ) (si ces deux mots sont sur la même ligne) ou verticale ( $dv$ ) (si les deux mots appartiennent à des lignes consécutives) qui existe entre ces deux mots. Si les deux mots  $m_1$  et  $m_2$  sont sur la même ligne, alors, la distance entre eux est définie comme étant le nombre de pixels entre la position  $x$  du dernier caractère du mot  $m_1$  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 défini comme étant l'ensemble des mots qui sont proches de  $m$  au regard des seuils de proximité  $s_h$  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 déplacement est fait horizontalement ou à  $s_v$  s'il 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  $m_1, m_2, \dots, m_{k-1}$ , telle que  $\forall_{i=1,2,\dots,k-1} d(m_i, m_{i+1}) \leq s$ .

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 considérée comme étant la séparation entre deux blocs.

L’algorithme de segmentation est semblable à celui décrit dans l’article [4] 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 rencontre 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 [4] 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 généralement rédigés 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, épaisseur 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 dans [4]. Car



<b>NAME</b>
<hr/> 1 Main Street, New Cityland, CA 91010 Cell: (555) 555-5555 example-email@example.com
<b>SUMMARY</b>
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.
<b>HIGHLIGHTS</b>
<ul style="list-style-type: none"> <li>• Network configuration and support</li> <li>• Exceptional troubleshooter</li> <li>• Data visualization</li> <li>• Systems design</li> <li>• Fluent in Mandarin</li> <li>• New technology and product research</li> <li>• Data connectivity</li> <li>• VoIP knowledge</li> <li>• Cloud computing</li> <li>• Sound judgment</li> </ul>
<b>EXPERIENCE</b>
<b>THE EDUCATION COMPANY</b> New Cityland, CA
<b>Network Systems Manager</b> <span style="float: right;">09/2008 to Current</span> <ul style="list-style-type: none"> <li>• Isolate and resolve network communication issues.</li> <li>• Train end-users on correct operation and use of workstations to maintain optimal performance and functionality of systems.</li> <li>• Review, update, and comply with network and company policies.</li> <li>• Track entire network of applications, databases, and servers.</li> <li>• Manage department budget and track expenses.</li> <li>• Perform data migration for server and workstation updates or replacements.</li> <li>• Support mobile access to network system.</li> <li>• Monitor security risks and complete updates to minimize or avoid threats.</li> </ul>
<b>SANTIAGO TECHNOLOGIES</b> New Cityland, CA
<b>Network Systems Administrator</b> <span style="float: right;">05/2003 to 08/2008</span> <ul style="list-style-type: none"> <li>• Maintained, upgraded, and troubleshoot network.</li> <li>• Oversaw hardware inventory and ordered new supplies.</li> <li>• Scheduled recommended updates and repairs.</li> <li>• Developed and implemented disaster recovery plans.</li> <li>• Routinely audited system for compliance with standards and policies.</li> <li>• Evaluated and improved network security measures.</li> </ul>
<b>EDUCATION</b>
<b>BACHELOR OF SCIENCE: COMPUTER SCIENCE</b> Sequoia University, New Cityland, CA
<b>CERTIFICATIONS</b>
Microsoft Certified System Engineer (MCSE) Microsoft Certified System Administrator (MCSA) Cisco Certified Network Associate (CCNA)

Figure 3. Example de CV segmenté en blocs

les choix de ces caractéristiques sont propres à chaque candidat. Pour certains, l'épaisseur 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 stocker des titres fréquemment utilisés. Les caractéristiques des titres reconnus vont permettre de retrouver les autres titres du CV qui ne se retrouvent pas dans le dictionnaire. Les étapes d'identification des 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 de texte est considérée comme étant un titre.

Sur une échelle 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.

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 [7]. 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 apparaissant 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 eux, les titres identifiés et en ne retenant que ceux ayant des caractéristiques similaires telles que la police de caractère élevée, l'épaisseur, ou encore une couleur différente, la casse.

### 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 [4]. Les différentes classes que nous avons considérées pour les expérimentations sont les suivantes : *Work Experience*, *Education*, *Skills*, *Languages*, *Certification*, *Personal Information*, *Hobbies*. Le modèle d'apprentissage 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 [4] 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 un élément très informatif sur la nature du contenu du bloc. Les termes extraits aussi bien dans les titres que dans les blocs sont lemmatisés en utilisant le *lemmatizer* de *wordnet* (WordnetLemmatizer) [5], à l'effet d'associer les termes qui partagent en commun un même lemme et qui traduisent à la base une information similaire. Ces informations sont utilisées pour représenter chaque bloc sous forme d'un vecteur de numériques en y associant sa classe qui n'est rien d'autre que le titre de la section (*EDUCATION, SKILLS,...*). Le jeu de données ainsi constitué est utilisé pour entraîner un SVM à l'effet de prédire pour un nouveau bloc son étiquette étant données les caractéristiques associées au bloc.

Après avoir étiqueté les différents blocs, la prochaine étape consiste à extraire les informations détaillées dans chacun des blocs étiquetés 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 [4] 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*<sup>1</sup> 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 *Identification 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 :

$\langle VRB \backslash NN \text{ of} \rangle \langle \text{complément} \rangle [ \langle IN \rangle \langle \text{complément} \rangle ] . (R)$

avec la nomenclature proposée par [11] où *VRB* est mis pour *verbe* (Verb), *IN* pour préposition (in), *NN* pour *nom* (Noun), etc. Avant d'appliquer cette règle, la simplification de phrase est utilisée pour décomposer toutes les phrases composées (celles ayant plusieurs propositions reliées par les conjonctions) en plusieurs sous phrases ayant un sens unique. Il a été montré dans [15] 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 de 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* [11] est utilisé pour étiqueter 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, *obj* : 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 *Expérience 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, permet d'obtenir les compétences suivants :*

- *cyber threat signatures*

– cyber attack

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

Une compétence est dite de haut niveau si elle est déduite de la connaissance d'un certain nombre de compétences de base (*skills*). Elle peut être explicite ou implicite. Dans cette partie, il s'agit d'utiliser les compétences de base extraites dans les étapes précédentes, pour déduire des compétences de haut niveau.

**Exemple 3** *Il s'agira par exemple de déduire la compétence développeur web à partir des compétences de base : html, css, javascript, php. Ou encore de déduire la compétence développeur mobile à partir des compétences de base : android, ionic, ios, java, cordova.*

Pour cette tâche, une approche supervisée où la liste des compétences est connue à l'avance pour chaque CV est utilisée. Ainsi, Il est 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 connu à l'avance. En suivant ce principe, le modèle développé dans cette section se base sur une liste de compétences mentionnées utilisées comme classes. Ainsi étant donné un CV, il est question de savoir lesquelles parmi ces compétences il possède. Pour nos expérimentations la liste des compétences retenues comme classes 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	

Il s'agit des compétences qui apparaissent plus fréquemment dans le jeu de CVs utilisés.

Le problème de prédiction des compétences d'un expert peut être vu comme un problème de classification multi-label [6] où chaque CV peut être étiqueté par plusieurs compétences simultanément. Par exemple, un expert peut avoir en même temps les compétences *system administrator* et *network administrator*. 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 étiqueté par un ensemble de compétences, le processus de construction du modèle de prédiction se fait en plusieurs étapes comme suit :

- 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 CVs

Chaque CV est initialement représenté par un sac de mots, puis transformé en une représentation vectorielle en utilisant la technique *word2vec* [13]. Les mots dans le sac sont ceux qui renseignent sur les compétences. Le modèle de représentation utilisé est le modèle *skip gram* [18] avec comme taille de la fenêtre 5 et comme dimension 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 très souvent employés dans des contextes similaires.

### 3.3.2. Construction du jeu de données

La méthode *Binary Relevance* est utilisée pour faire la classification *multi-label*[6] des CVs. Elle 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), et 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 entraîner 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

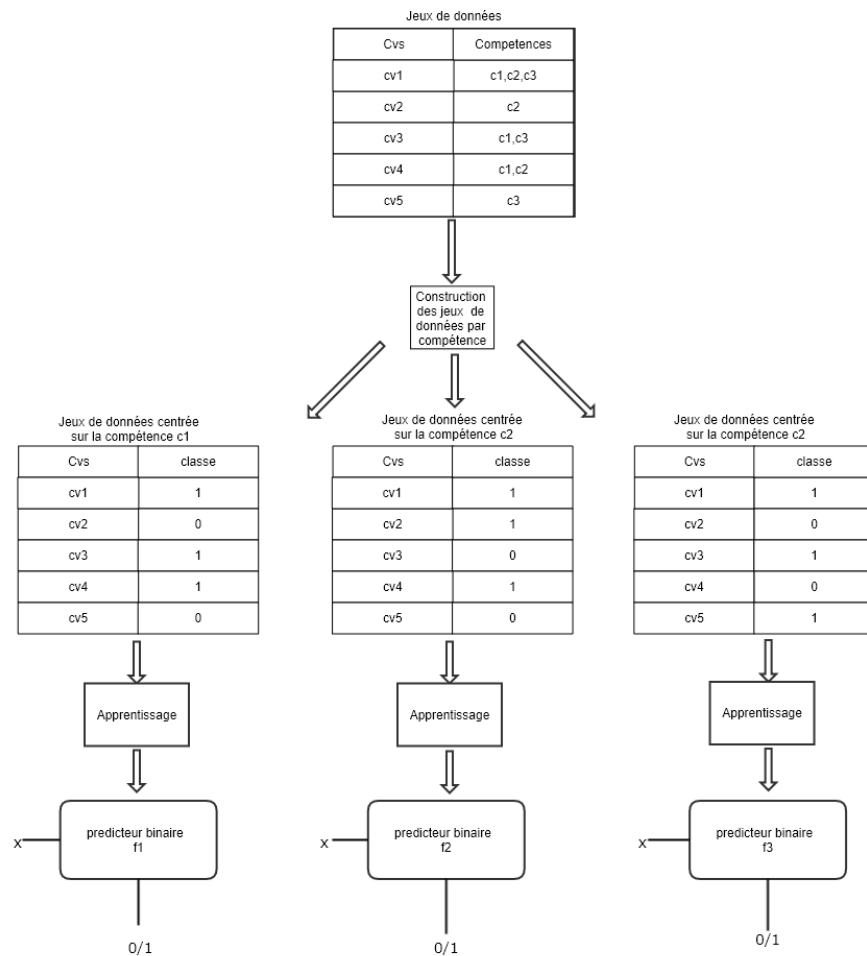
A la fin du processus, les prédictions de chaque modèle de base sont agrégées en faisant l'union des classes prédites par chacun. Ainsi, nous obtenons la liste des compétences prédites de l'expert parmi les compétences d'intérêt. A la fin, l'ensemble des compétences de l'expert est donc l'union des compétences 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 stockés au format PDF. Un jeu de 800 CVs rédigés en anglais a été constitué pour les expérimentations. Le jeu de 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 données) est utilisé pour le test et les autres sous-ensembles (3/4 du jeu de données) sont utilisés pour la construction du modèle. Les résultats finaux consignés dans des tableaux sont obtenus en faisant la moyenne des résultats sur les quatre expériences.



**Figure 4. Modèle de prédiction multi-label**

Chaque aspect important du modèle a été évalué individuellement. Une mesure d'évaluation appropriée a été utilisée pour chacun de ces aspects.

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

### Mesure d'évaluation

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

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

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

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

Dans la formule 1,  $NBC$  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. Le tableau 1 présente la précision obtenue dans la tâche d'extraction des blocs. La précision varie en fonction de la structure du CV. La détection des blocs opère assez bien avec une précision de 89,86% 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 structurés de cette façon, 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 2) 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, quatre mesures synthétiques sont utilisées pour évaluer la qualité du modèle pour chaque classe d'une part, et la qualité 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 précision 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 :

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

**Tableau 2.** Matrice de confusion de la classification des blocs  
Classes prédites

		Classes prédites							Rappel
		Education	Work Experience	Skills	Certifications	Personal Infos	Languages	Hobbies	
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 2, 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 90%. La valeur élevée de la précision pour la section **Languages** est probablement due au fait que cette section se distingue avec peu d'ambiguïté 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 dues 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 prédits comme étant de la classe **Skills**. En effet, les mêmes caractéristiques qu'on peut retrouver dans la section "Work Experience" 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 globalement avec une précision moyenne évaluée à 91,83% et un rappel global évalué à 91,85% (tableau 3).

**Tableau 3.** Classification des blocs : Evaluation globale

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

## 4.4. Evaluation du modèle 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  $Y_i$  l'ensemble des termes clés extraits par le modèle,  $Z_i$  l'ensemble des termes clés extraits par l'expert du domaine, et  $n$  le nombre de CVs 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 calculé en faisant la moyenne des rappels obtenus sur tous les 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 classes qui ont été correctement prédites parmi celles que le modèle a prédites. La précision moyenne (formule 8) est déterminée en faisant la moyenne des précisions calculées sur toutes 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, permet 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 répertoriés dans le tableau 4. La précision de 96.2% signifie 96.2% des termes exprimant les compétences qui ont été 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 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 réussit à retrouver 84.32% des caractéristiques qu'on souhaitait extraire du modèle et qui ont été identifiées au préalable par un expert du domaine informatique. Le gap au niveau du rappel est dû au fait que la règle (R) n'est pas une règle exhaustive et ne permet pas d'extraire tous les compétences de

**Tableau 4.** *Résultat de l'extraction des compétences explicites.*

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

type savoir faire. D'autres part, les erreurs liées à la lemmatisation induisent le modèle à ne pas reconnaître des termes qui ont une même racine lexicale.

#### **4.5. Evaluation du modèle de prédiction des compétences implicites**

Les CVs ont été étiquetés 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 sémantique 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 prédites par le 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 5. Dans ce tableau, l'exactitude traduit à priori qu'en moyenne, 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 constituant les étiquettes de chaque exemple. Ces résultats montrent la capacité du modèle à proposer aux recruteurs les autres compétences non explicitement mentionnées dans les CVs par les candidats.

**Tableau 5.** *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 6 présente les résultats des approches de segmentation.

**Tableau 6.** *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 6 l'approche modifiée pour effectuer la segmentation est plus précise que celle décrite dans l'article [4]. 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 la construction d'un modèle basé sur l'approche hiérarchique d'extraction des informations dans le CV décrite dans [4] pour extraire les compétences. Le CV est segmenté en sections en utilisant une version modifiée de l'algorithme proposé dans [4]. 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'étiquetage des blocs, les caractéristiques ont été extraites et utilisées pour construire un modèle de classification 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%).

En perspective, il sera question de voir dans quel mesure utiliser un réseau de neurones profond en l'occurrence le réseau de neurones convolutionnels pour extraire automatiquement les caractéristiques pertinentes dans le but de faire la prédiction. En effet, les filtres du réseau de neurones convolutionnels peuvent permettre d'identifier automatiquement les termes ou compétences de base qui caractérisent chaque compétence qui constitue la liste des classes.

---

## 6. Bibliographie

- [1] ALOTAIBI, SHAHA, « A survey of job recommender systems », *International Journal of the Physical Sciences*, vol. 7, 2012.
- [2] ASSOCIATION FOR TALENT DEVELOPMENT, « Bridging the skills gap : workforce development is everyone's business », n° 2, 2015.
- [3] BENSON, VLADLENA, FILIPPAIOS, FRAGKISKOS, MORGAN, STEPHANIE, « Social Career Management : Social Media and Employability Skills Gap », *Comput. Hum. Behav.*, vol. 30, 519-525, 2014.
- [4] CHEN, JIAZE, GAO, LIANGCAI, TANG, ZHI, « Information Extraction from Resume Documents in PDF Format », *Electronic Imaging*, n° 17, n° 17, 1-8, 2016.
- [5] FELLBAUM, CHRISTIANE, , « WordNet : an electronic lexical database », 1998.
- [6] GIBAJA, EVA, SEBASTIÁN, , « Multi-label Learning : A Review of the State of the Art and Ongoing Research », *Wiley Int. Rev. Data Min. and Knowl. Disc.*, vol. 4 n° 6, 411-444, 2014.
- [7] HUANG, ANNA, « Similarity measures for text document clustering », *Proceedings of the 6th New Zealand Computer Science Research Student Conference*. 2008.
- [8] JIECHIEU KAMENI FLORENTIN FLAMBEAU, NORBET TSOPZE, « Approche hiérarchique pour extraire les compétences dans les CVs en format pdf », *Proceedings de la conférence sur la recherche en Informatique CRI'17, ENSP Yaoundé*, 2017.

- [9] KIVIMÄKI, ILKKA, PANCHENKO, ALEXANDER, DESSY, ADRIEN, VERDEGEM, DRIES, FRANCO, PASCAL, BERSINI, HUGUES, SAERENS, MARCO, « A Graph-Based Approach to Skill Extraction from Text », *Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing*, 79-87, 2013.
- [10] MAHESHWARI, SUMIT, SAINANI, ABHISHEK, REDDY, P. KRISHNA, « An Approach to Extract Special Skills to Improve the Performance of Resume Selection », *Databases in Networked Information Systems*, 256-273, 2010.
- [11] MARIE-CATHERINE DE MARNEFFE, CHRISTOPHER D. MANNING, « Stanford typed dependencies manual », *Comput. Hum. Behav.*, 2008.
- [12] MENG ZHAO, FAIZAN JAVED, FEROSH JACOB, MATT MCNAI, « SKILL : A System for Skill Identification and Normalization », *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, 4012-4018, 2015.
- [13] MIKOLOV, TOMAS, SUTSKEVER, ILYA, CHEN, KAI, CORRADO, GREG S, DEAN, JEFF, « Distributed Representations of Words and Phrases and their Compositionality », *Advances in Neural Information Processing Systems 26*, 3111-3119, 2013.
- [14] MIRANDA, SERGIO, ORCIUOLI, FRANCESCO, LOIA, VINCENZO, SAMPSON, DEMETRIOS, « An Ontology-based Model for Competence Management », *Data Knowl. Eng.*, vol. 107, n° C, 51-66, January 2017.
- [15] PENG, YIFAN, CATALINA O, TORII, MANABU, WU, CATHY H, VIJAY-SHANKER, K, « iSimp : A sentence simplification system for biomedical text », *Bioinformatics and Biomedicine (BIBM)*, 2012 IEEE International Conference on, 1-6, 2012.
- [16] SHIQIANG GUO, FOLAMI ALAMUDUN, TRACY HAMMOND, « Résumatcher : A personalized résumé-job matching system », *Wiley Int. Rev. Data Min. and Knowl. Disc.*, vol. 60, 169-182, 2016.
- [17] THIMMA REDDY KALVA, « All Graduate Plan B and other Reports », *Skill Finder : Automated Job-Resume Matching System*, 2013.
- [18] TOMAS MIKOLOV, KAI CHEN, GREG CORRADO, JEFFREY DEAN, « Efficient Estimation of Word Representations in Vector Space », *Databases in Networked Information Systems*, 256-273, 2013.
- [19] VAPNIK, VLADIMIR N., « Statistical Learning Theory », n° 1998.
- [20] YU, KUN, GANG GUAN, MING ZHOU, « Resume information extraction with cascaded hybrid model », *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*, 499-506, 2005.
- [21] ZIEMOWIT BEDNAREK, , « Skills gap : The timing of technical changes », *Journal of Economics and Business*, vol. 74, 54-64, 2014.