Un algorithme d'apprentissage profond et semi-supervisé basé sur la représentation de graphes pour la classification des CV

Wissem Inoubli * ***, Armelle Brun **

* Université d'Artois, CRIL, Lens, France
** Université de Lorraine, LORIA, Nancy, France
***Keep In Touch, Strasbourg, France

Résumé. Les demandeurs d'emploi et les départements des ressources humaines sont confrontés à une abondance d'offres d'emploi et de CV, rendant impossible une évaluation manuelle complète de chaque CV et de chaque offre. Pour résoudre ce défi, les systèmes de recommandation sont utilisés pour suggérer aux demandeurs d'emploi des offres pertinentes et aux recruteurs des CV correspondants. Dans ce travail, nous proposons de représenter les données (les CV) sous forme de graphes et d'aborder ce problème de recommandation comme un problème de classification. Nous présentons DGL4C, un modèle d'apprentissage profond semi-supervisé à base de graphes. Les expériences menées sur un jeu de données publique de CV anonymisés montrent que DGL4C améliore significativement la précision d'un modèle traditionnel d'apprentissage profond.

1 Introduction

Le recrutement peut être vu comme le processus qui vise à faire correspondre des offres d'emploi et des CV. Cette correspondance est effectuée à la fois par les demandeurs d'emploi (manuellement) et par les services des ressources humaines (RH) (grâce à l'utilisation de systèmes de gestion des candidaI'm ts (ATS), par exemple JobSCAN [1]). En raison de l'énorme volume de CV et d'offres, cette tâche ne peut plus être effectuée manuellement. Les algorithmes de recherche d'information (RI) sont traditionnellement utilisés pour effectuer cette tâche. Par ailleurs, la classification des CV est désormais utilisées comme un outil commun conçu pour recommander aux RH les CV qui correspondent à une offre d'emploi donnée, ou à un demandeur d'emploi les offres pertinentes selon son profil Giabelli et al. (2021).

La classification des CVs a traditionnellement reposé sur le pré-traitement des documents, avec des modèles tels que TF-IDF, LDA et word2vec pour l'extraction de caractéristiques et la représentation des textes. Les méthodes classiques de classification, comme les forêts aléatoires, les arbres de décision et les machines à vecteurs de support, étaient ensuite utilisées pour effectuer la classification à l'aide de ces représentations. Cependant, il est devenu évident que la qualité des représentations textuelles joue un rôle crucial dans les performances des classifieurs. Récemment, l'apprentissage profond a été exploré comme une approche plus

1. https ://www.jobscan.co/applicant-tracking-systems

avancée. Les réseaux de neurones profonds, tels que les LSTM, les RNN et les CNN, ont été étudiés et ont montré des améliorations significatives par rapport aux méthodes d'apprentissage automatique traditionnelles. Les structures de graphes sont traditionnellement adoptées pour gérer des données riches et structurées. Récemment, des modèles d'apprentissage profond de graphes Yao et al. (2019), qui permettent d'apprendre un espace non-euclidien de données, ont émergé. De manière surprenante, a notre connaissance ils n'ont pas été étudiés dans le contexte des RH, en particulier pour la classification de CV. Dans ce travail, nous proposons DGL4C, pour *Deep Graph Representation Learning for Classification*, un nouveau modèle de classification basé sur l'apprentissage profond des graphes. DGL4C est un modèle semi-supervisé, conçu pour la classification de CV, qui gère à la fois des données étiquetées (CV) et non étiquetées (éléments de CV).

Concrètement, nous proposons deux variantes de DGL4C. DGL4C-GCN, est un réseau neuronal convolutif de graphe de bout en bout, qui apprend toutes les étapes entre la phase initiale d'entrée et le résultat final de sortie (classification du résumé). DGL4C-GRL est composé de deux étapes : (i) la représentation du texte (CV) par une architecture GCN, et (ii) un classifieur basé sur l'apprentissage automatique.

La suite de ce document est organisée comme suit. La section 2 présente la littérature relative à la classification des CV. Dans la section 3, nous présentons DGL4C et ses deux variantes DGL4C-GCN et DGL4C-GRL. Ensuite, dans la section 4, les résultats expérimentaux sont décrits et analysés. Enfin, dans la section 5, nous concluons et proposons des perspectives.

2 État de l'art

Dans cette section, nous présentons des travaux relatifs à la classification de CV dans le domaine des ressources humaines.

La littérature a proposé plusieurs approches, que nous choisissons de diviser en trois catégories : (i) les modèles basés sur les ontologies, (ii) les modèles d'apprentissage automatique et (iii) les modèles d'apprentissage profond. Considérons tout d'abord les modèles basés sur les ontologies. Une ontologie est un méta-modèle conceptuel qui représente une connaissance du domaine Fazel-Zarandi et Fox (2009). Après une étape d'extraction de caractéristiques, ces modèles utilisent des ontologies pour effectuer la classification. Quelques bases de connaissances internationales et nationales en matière de RH ont été publiées. Les plus connues sont le DISCO [2], la CITP [3] et l'ESCO [4] de Groot et al. (2021). Ces bases de connaissances représentent des groupes de professions à différents niveaux de granularité. En ce qui concerne les modèles d'apprentissage automatique, largement utilisés, ils reposent sur des données d'entraînement et nécessitent une étape de pré-traitement dédiée à l'extraction de caractéristiques des éléments à classer. Les modèles d'apprentissage automatique, tels que les forêts aléatoires, les arbres de décision et les machines à vecteurs de support, etc. ont montré une efficacité et des performances élevées pour la tâche de classification de CV Fareri et al. (2021).

Dans ces modèles, la qualité de l'étape d'extraction des caractéristiques a un impact important sur les performances de classification. À l'opposé des modèles basés sur les ontologies et de l'apprentissage automatique, les modèles d'apprentissage profond considèrent à la fois

2. Dictionnaire européen des aptitudes et des compétences
3. Classification internationale type des professions
4. European Skills, Competences, Qualifications and Occupations

l'extraction de caractéristiques et la classification en une seule étape, ce qui réduit la potentielle perte d'informations dans l'étape d'extraction de caractéristiques. Les modèles d'apprentissage profond sont très populaires et ont montré une amélioration significative des performances. Plusieurs travaux ont été proposés pour la classification des offres d'emploi Jiechieu et Tsopze (2021); Giabelli et al. (2021); Sajid et al. (2022); Abdollahnejad et al. (2021) où les architectures de réseau neuronal convolutif 1-D (CNN) et de réseau neuronal récurrent (RNN) ont été adaptées dans le contexte des RH. Les techniques d'apprentissage de représentation de graphes sont apparues récemment et sont utilisées dans nombreuses applications. La structure de graphe est un moyen traditionnel de représenter les données, mais les modèles qui s'appuient sur une telle représentation souffrent de la rareté des données et d'un manque de robustesse au bruit, ce qui diminue la performance des modèles prédictifs. Pour surmonter ces limites, l'apprentissage par représentation de graphes a été conçu pour transformer les données dans un nouvel espace de dimension réduite. Il a montré son efficacité pour les données non structurées telles que les images, les textes et les graphes. L'apprentissage de la représentation de graphes peut être classé en trois familles : les modèles basés sur la factorisation matricielle, les modèles basés sur la marche aléatoire et les modèles basés sur les réseaux de neurones de graphes (GNNs - Graph Neural Networks) (Kipf et Welling (2016)). À notre connaissance, l'architecture GNN n'a pas été étudiée pour la modélisation de CV ou d'offres d'emploi, et nous faisons l'hypothèse qu'ils pourraient améliorer les performances de classification, ce que nous proposons d'étudier ci-dessous.

3 DGL4C : un algorithme de classification basé sur les réseaux convolutifs de graphes

Partant du constat que la représentation des données est une étape essentielle pour les algorithmes de classification, nous proposons ici une nouvelle approche pour la représentation des CV, inspirée de Yao et al. (2019), et basée à la fois sur une structure de graphe et sur des informations contextuelles (les caractéristiques des noeuds). Concrètement, nous proposons DGL4C, un modèle d'apprentissage semi-supervisé de la représentation de CV, basé sur une architecture GNN. La principale motivation pour le choix d'un apprentissage de représentation de graphes, et plus particulièrement d'une architecture GNN, est motivé par le fait que ces architectures exploitent des informations de voisinage (des voisins) lors de l'étape d'apprentissage, et qui permettent une représentation enrichie du graphe.

La quantité de données de CV étant généralement limitée, nous optons pour un algorithme d'apprentissage semi-supervisé, c'est-à-dire qui exploite à la fois des exemples étiquetés et non étiquetés ; le processus d'entraînement repose donc sur ces deux types de données Kipf et Welling (2016). DGL4C vise à apprendre une nouvelle représentation latente de qualité des CV, qui sera utilisée ensuite dans une étape de classification. Dans les sous-sections suivantes, nous présentons la façon dont nous proposons de construire un graphe de CV, puis l'idée centrale de l'apprentissage de représentation de graphes, et la manière dont nous développons DGL4C, conçu pour encoder un jeu de données de CV dans un espace vectoriel latent.

3.1 Construction du graphe

Avant l'étape d'apprentissage d'une représentation de données, la première phase consiste à construire le graphe des CV. Nous proposons de nous inspirer du travail mené dans Yao et al. (2019), notamment l'étape de construction de graphe.

Soit $D = (R, L)$ un jeu de données. R est l'ensemble des CV, $R_i \in R$ est un CV avec $R_i = \{wd_{i1}, wd_{i2}, ..., wd_{ij}\}$ est l'ensemble des j mots du CV R_i. $\mathbb{W} = \bigcup_i R_i$ est le vocabulaire de D, c'est-à-dire l'ensemble des mots distincts dans R.

Y représente l'ensemble complet des étiquettes possibles(les classes), c'est-à-dire les classes auxquelles les CV peuvent appartenir. Chaque CV R_i est associé à une étiquette Y_i qui fait partie de cet ensemble Y. La cardinalité de Y est notée $|Y|$.

Definition 3.1 (R-graphe)**.** Soit G un graphe hétérogène, avec attributs et non pondéré construit à partir de D. G = (V, E) avec V et E représentant respectivement les nœuds et les arêtes de G. L'ensemble de nœuds $V = \mathbb{R} \cup \mathbb{W}$ est l'union de l'ensemble des CV et de l'ensemble des mots uniques des CV. Ainsi, V est composé de deux types de nœuds : les nœuds de type CV et les nœuds de type mots. L'ensemble des arêtes E est également divisé en deux types, les arêtes entre mot et CV et les arêtes mot à mot. En d'autres termes, une arête entre deux mots représente leur co-occurrence dans le même CV, tandis qu'une arête entre un mot et un CV indique la présence de ce mot dans ce CV.

De manière similaire à Yao et al. (2019), une arête entre deux nœuds existe si la similarité entre ces nœuds est positive. La façon dont cette similarité est évaluée dépend du type d'arête. Les arêtes mot à un cv sont évaluées à l'aide du traditionnel TF-IDF, et les arêtes mot à mot sont évaluées par l'information mutuelle spécifique (PMI). Pour plus de détails voir Inoubli et Brun (2022).

3.2 Apprentissage de représentation d'un graphe

Après avoir construit le graphe G, l'apprentissage de la représentation de graphe représente la seconde étape de DGL4C.

Un GCN est un réseau de neurones convolutifs sur les graphes qui effectue des opérations similaires à celles du CNN, à l'exception qu'il applique une convolution sur un graphe au lieu d'une convolution sur un tableau 2-D Kipf et Welling (2016). Un GCN apprend une représentation latente en propageant l'information des voisins directs dans le graphe et applique une transformation linéaire.

Dans Yao et al. (2019), les auteurs ont utilisé les caractéristiques initiales des nœuds comme une matrice d'identité $X = I_{|V|}$. Dans DGL4C, la matrice des caractéristiques des nœuds X est le vecteur de caractéristiques de chaque nœud de G. Le modèle pré-entraîné bien connu sBERT (Reimers et Gurevych (2019)) est utilisé pour encoder à la fois les mots et les CV pour construire X.

Nous proposons deux variantes de DGL4C, qui diffèrent par le nombre d'étapes qui les composent : (i) DGL4C-GCN est un modèle compact (de bout en bout) avec une étape unique qui apprend à la fois la représentation et la classification et (ii) DGL4C-GRL est un modèle composé de deux étapes : la représentation du texte (résumé) puis la classification.

4 Expérimentations

Nous nous intéressons maintenant à l'évaluation de DGL4C, et en particulier des deux modèles DGL4C-GCN et DGL4C-GRL. Nous choisissons d'évaluer les modèles au travers de la précision, que nous comparons à celle d'algorithmes de l'état de l'art.

4.1 Protocole expérimental

Le jeu de données utilisé est un corpus de 2484 CV librement disponibles et anonymisés [5]. Chaque CV est associé à une étiquette, qui représente le profil du CV. 24 profils (classes) sont disponibles. Chaque CV est écrit en langage naturel et contient des informations personnelles, la formation, des expériences, etc. Dans les expérimentations menées, nous nous intéressons en particulier à l'impact du nombre de classes (profils) sur la précision des modèles. Ainsi, nous formons cinq ensembles de données qui varient par le nombre de classes qu'il contient. Les ensembles de données D1, D2, D3, D4 et D5 ont respectivement des tailles différentes en termes de nombres de CVs, avec 500, 1000, 1500, 2000 et 2484 CVs chacun. De plus, le nombre de classes varie pour ces ensembles, étant de 5, 10, 16, 20 et 24 classes respectivement.

DGL4C-GCN et DGL4C-GRL ont été implémentés en utilisant le framework DGL [6] avec deux couches de convolution de l'architecture GraphSage Hamilton et al. (2017) pour permettre le passage de messages entre les nœuds, et l'agrégation moyenne. D'un point de vue architectural, nous avons fixé la taille d'intégration de la première couche de convolution à 500. Nous avons réglé d'autres paramètres et fixé le taux d'apprentissage à 0,001, le *dropout* à 0,2. Pour chaque jeu de données, nous utilisons aléatoirement 80 % des CV pour l'entraînement et le reste pour le test.

4.2 Résultats expérimentaux

Pour évaluer l'efficacité de DGL4C-GCN et DGL4C-GRL, nous comparons leurs performances avec plusieurs modèles de la littérature, qui diffèrent soit par la représentation du texte, soit par l'étape du classification. Chaque modèle consiste en deux étapes, représentation de texte et un classifieur, à l'exception du modèle de bout en bout DGL4C-GCN. Les modèles de représentation de texte les plus populaires dans la littérature, et mentionnés dans la section 2, sont utilisés.

Impact de la représentation du texte

Nous nous concentrons tout d'abord sur l'évaluation de l'impact de la représentation du texte, en fixant le classifieur. Nous choisissons ainsi d'utiliser l'algorithme populaire de la forêt aléatoire (RF). Le tableau 1 présente la précision de ces modèles, y compris DGL4C, en fonction du jeu de données utilisé.

Étudions tout d'abord la précision des modèles sur le jeu de données complet (D5). Comme attendu, TF-IDF+RF est le modèle le moins performant (précision=30,09), TF-IDF étant une représentation simple. Les représentations basées sur l'apprentissage profond : Word2Vec+RF et sBERT+RF sont plus performantes, avec une précision de 49,65% et 60,23% respectivement. sBERT+RF est plus performant que Word2Vec+RF, ce qui est conforme à la littérature. En ce

5. https ://www.kaggle.com/datasets/snehaanbhawal/resume-dataset
6. https ://www.dgl.ai/

Modèlejeux de données	D1	D2	D3	D4	D5
TF-IDF+RF	70,50	42,43	35,65	34,65	30,09
Word2Vec+RF	78,43	63,16	51,76	52,24	50,64
sBERT+RF	92,23	73,15	68,14	66,97	61,65
DGL4C-GCN	93,54	**75,23**	**74,65**	73,66	62,43
DGL4C-GRL+RF	**94,38**	73,65	73,76	**75,76**	**67,87**

TAB. 1 – *Précision moyenne des modèles étudiés.*

qui concerne les modèles de représentation basés sur les graphes, que nous proposons, DGL4C-GCN, modèle de bout en bout, est légèrement plus performant que sBERT+RF, mais cette augmentation n'est pas statistiquement significative. Quant à DGL4C-GRL+RF, il est plus performant que DGL4C-GCN et significativement plus performant que sBERT+RF. Nous pouvons conclure que l'information de voisinage (avec la représentation de graphe) dans l'apprentissage de la représentation, qui combine à la fois l'information sémantique et structurelle permet d'améliorer la performance. Intéressons-nous maintenant à l'impact du nombre de classes sur les performances des modèles, en étudiant les performances sur les jeux de données D1 à D5, c'est-à-dire de entre 5 et 24 classes. Comme attendu, les performances de chaque modèle sont négativement impactées par l'augmentation du nombre de classes. Par exemple, la précision de DGL4C-GRL+RF est de 94,38% avec 5 classes et diminue à 67,87% avec 24 classes. Cependant, cette performance ne diminue pas linéairement avec le nombre de classes. En particulier, les performances entre 11 et 20 classes restent stables. Une diminution significative se produit entre 20 et 24 classes, de 75,76% à 67,87%. Une diminution similaire se produit également pour l'autre modèle à base de graphe DGL4C-GCN. Cependant, ce n'est pas le cas pour les modèles basés sur l'apprentissage profond (Word2Vec), ni pour le TF-IDF. Il est par ailleurs important de noter que, quel que soit le nombre de classes, les modèles à base de graphe sont toujours ceux qui ont la précision la plus élevée.

Impact du classifieur

Nous nous concentrons maintenant sur l'évaluation de l'impact du classifieur sur les performances de DGL4C-GRL. Nous évaluons plusieurs classifieurs populaires dans la littérature, à savoir la classification par vecteurs support (SVC), le perceptron multicouche (MLP), la régression logistique (RL), que nous comparons à la forêt aléatoire (RF) précédemment étudiée. La précision moyenne de ces modèles est présentée dans le tableau 2, qui rappelle également les performances du meilleur modèle d'apprentissage profond sBERT+RF et du modèle DGL4C-GCN bout en bout basé sur les représentation de graphes. Tout d'abord, nous pouvons constater

Modèle/jeux de données	D1	D2	D3	D4	D5
DGL4C-GRL+LR	**95,67**	73,99	74,23	74,85	67,10
DGL4C-GRL+SVC	94,20	72,03	74,60	75,65	**69.04**
DGL4C-GRL+MLP	95,08	72,25	72,96	73,76	65,77
DGL4C-GRL+RF	94,38	73,65	73,76	**75,76**	67,87
DGL4C-GCN	93,54	**75,23**	**74,65**	73,66	62,43
sBERT+RF	92,23	73,15	68,14	66,97	61,65

TAB. 2 – *Précision moyenne de DGL4C-GRL avec différents classifieurs*

que quel que soit le classifieur utilisé, DGL4C-GRL est plus performant que sBERT+RF sur la plupart des versions des jeux de données. Si nous nous concentrons sur l'impact du classifieur sur les performances de DGL4C-GRL, LR et SVC sont les deux classifieurs les plus performants, qui surpassent légèrement les performances de RF. Cependant, cette amélioration n'est pas statistiquement significative. Nous pouvons donc conclure que la nature du classifieur n'a pas d'impact significatif sur la performance du modèle. Au contraire, la représentation de graphe semble être l'étape la plus influente, ce qui confirme les résultats de la littérature. En ce qui concerne DGL4C-GCN, il s'agit du modèle le plus performant pour deux jeux de données (D2 et D3). Cependant, DGL4C-GCN a une performance significativement plus faible sur D5 (62,43% précision) par rapport au modèle le plus performant DGL4C-GRL+SVC (69,04% précision). Cela peut s'expliquer par le fait qu'un modèle de bout en bout a une fonction d'optimisation qui optimise en même temps l'apprentissage de la représentation et la classification, alors que DGL4C-GRL a deux fonctions d'optimisation utilisées séparément, ce qui rend le classifieur plus flexible.

Impact de la représentation du texte
Les hyper-paramètres jouent un rôle important dans l'apprentissage de la représentation des graphes, car ils déterminent la manière dont les plongements de nœuds seront générés. Dans cette expérience, nous essayons d'analyser l'impact de la dimension de plongement. Le tableau 3 montre le comportement du modèle par rapport à la variation de la taille de plongement.

Jeux de données\Taille de plongement	64	128	256	512	1024
D1	92,1	93,15	**94,23**	93,85	92,12
D2	70,20	72,03	72,43	**76,15**	72,05
D3	70,08	72,25	73,60	**76,26**	72,77
D4	71,23	73,65	73,56	**74,34**	71,87
D5	58,54	59,23	61,57	**63,69**	61,13

TAB. 3 – *Impact de taille de plongement sur la précision du modèle*

Dans cette expérimentation nous avons fait varier la dimension de plongement de 64 à 1024 pour toutes les données étudiées. Les résultats expérimentaux (voir Tableau 3) montrent clairement que le modèle obtient de meilleures performance avec une dimension de 1024 pour les jeux de données D2, D3, D4 et D5 tandis que la meilleure précision de D1 est obtenue avec une dimension de 512. Cela peut être expliqué par le fait que l'augmentation du nombre de dimensions d'un système de coordonnées permet de repérer d'une manière efficace un point dans l'espace, et qui permet d'avoir une meilleure représentation de données. Mais lorsque la taille de plongement dépasse un certain seuil, la performance diminue lentement.

5 Conclusion

Dans cet article, nous avons proposé DGL4C, un modèle d'apprentissage profond semi-supervisé pour la classification des CV basé sur la représentation de graphes. DGL4C s'appuie sur une approche d'apprentissage de représentation profonde et adapte l'architecture GCN à partir de données textuelles. Les expériences menées démontrent les très bonnes performances des deux variantes de DGL4C : DGL4C-GCN et DGL4C-GRL. Dans nos travaux futurs, nous

prévoyons d'adopter un apprentissage non supervisé, et d'évaluer notre modèle avec des jeux de données plus larges.

Références

Abdollahnejad, E., M. Kalman, et B. H. Far (2021). A deep learning bert-based approach to person-job fit in talent recruitment. In *CSCI*. IEEE.

de Groot, M., J. Schutte, et D. Graus (2021). Job posting-enriched knowledge graph for skills-based matching. *arXiv preprint arXiv :2109.02554*.

Fareri, S., N. Melluso, F. Chiarello, et G. Fantoni (2021). Skillner : Mining and mapping soft skills from any text. *Expert Systems with Applications*.

Fazel-Zarandi, M. et M. S. Fox (2009). Semantic matchmaking for job recruitment : an ontology-based hybrid approach. In *ISWC*, Volume 525, pp. 2009.

Giabelli, A., L. Malandri, F. Mercorio, M. Mezzanzanica, et A. Seveso (2021). Skills2job : A recommender system that encodes job offer embeddings on graph databases. *Applied Soft Computing 101*, 107049.

Hamilton, W., Z. Ying, et J. Leskovec (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems 30*.

Inoubli, W. et A. Brun (2022). Dgl4c : a deep semi-supervised graph representation learning model for resume classification.

Jiechieu, K. et N. Tsopze (2021). Skills prediction based on multi-label resume classification using cnn with model predictions explanation. *Neural Computing and Applications*.

Kipf, T. N. et M. Welling (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv :1609.02907*.

Reimers, N. et I. Gurevych (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv :1908.10084*.

Sajid, H., J. Kanwal, S. U. R. Bhatti, S. A. Qureshi, A. Basharat, S. Hussain, et K. U. Khan (2022). Resume parsing framework for e-recruitment. In *IMCOM*, pp. 1–8. IEEE.

Yao, L., C. Mao, et Y. Luo (2019). Graph convolutional networks for text classification. In *AAAI*, Volume 33, pp. 7370–7377.

Summary

Job seekers aim to find job offers that align with their qualifications, as do human resource departments in seeking candidates whose resumes match their expectations. In this work, the proposal suggests using graph representation for data and treating the problem as a classification task. The proposed DGL4C model, a semi-supervised graph deep learning approach, learns representations from graphs and trains a classifier on this latent data. Experiments conducted on an anonymous resume dataset demonstrate that DGL4C notably enhances precision and accuracy compared to traditional deep learning models.