Récupération de passages basée sur un graphe d'attention amélioré par des entités

Lucas Albarede[1, 2] Lorraine Goeuriot[1] Philippe Mulhem[1] Claude Le Pape-Gardeux[2] Sylvain Marié[2] Trinidad Chardin-Segui[2]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP[*], LIG, Grenoble, France
(2) Schneider Electric Industries SAS
lucas.albarede@protonmail.com, lorraine.goeuriot@imag.fr, philippe.mulhem@imag.fr

RÉSUMÉ
La recherche de passages est cruciale dans les domaines spécialisés où les documents sont longs et complexes, tels que les brevets, les documents juridiques, les rapports scientifiques, etc. Nous explorons dans cet article l'intégration d'entités et de passages dans des modèles de graphes d'attention hétérogènes dédiés à la recherche de passages. Nous utilisons les deux architectures de recherche de passages basées sur le reclassement proposées dans (Albarede *et al.*, 2022). Nous expérimentons notre proposition sur la tâche de recherche de passages TREC CAR Y3. Les résultats obtenus montrent une amélioration par rapport aux techniques de pointe et prouvent l'efficacité de l'approche. Nos expériences montrent l'importance d'utiliser des paramètres adéquats pour une telle approche.

MOTS-CLÉS : Graph Attention Networks, Conceptual Representation.

KEYWORDS: Recherche de passages, Représentation conceptuelle.

1 Introduction

Cet article est un résumé de (Albarede *et al.*, 2023), dans lequel nous proposons d'utiliser l'apprentissage par graphe hétérogène avec des bases de connaissances pour tirer parti de la recherche de passages. Il étend un travail précédent (Albarede *et al.*, 2022), en utilisant explicitement des représentations d'entités dans le processus de recherche de passages. Les réseaux neuronaux graphes basés sur l'attention, ou *GNNs*, peuvent aider à calculer les représentations sémantiques contextualisées des passages en tenant compte des interactions entre les différentes parties d'un document, et ces représentations peuvent être utilisées efficacement pour effectuer la recherche de passages.

Dans cette contribution, nous considérons les documents comme des graphes composés de passages, de sections et d'entités, et nous étudions l'utilisation de *GNNs* pour calculer des représentations contextualisées de passages améliorées par les entités, en particulier pour la tâche de recherche de passages. Les passages et les sections sont des éléments structurels définis en fonction de la structure physique du document et dont le contenu peut couvrir plusieurs sujets, tandis que les entités sont des références sémantiques qui représentent généralement un seul concept du domaine. Ces deux éléments étant de nature différente, nous étudions comment ils pourraient interagir efficacement au sein des

*. Institute of Engineering Univ. Grenoble Alpes.

GNNs. Dans ce qui suit, nous décrivons d'abord les éléments importants des Heterogeneous Graph Attention Networks, appelés HGATs, de (Albarede *et al.*, 2022) sur lesquels nous basons ce travail. Nous nous concentrons ensuite sur la représentation des graphes de documents en tenant compte à la fois des passages et des entités. Ensuite, nous étendons les HGATs avec des représentations de passage renforcées par des entités en introduisant l'**E**ntity-**E**nhanced **HGAT** (*EE-HGAT*). Enfin, nous intégrons les représentations de passages dans un modèle de classement neuronal et réalisons des expériences sur la tâche TREC CAR Y3 Passage Retrieval.

2 Vue d'Ensemble des HGATs

Les GAT (Graph Attention Networks) sont des réseaux neuronaux multicouches qui calculent les représentations sémantiques des nœuds d'un graphe en tenant compte des informations fournies par leurs voisins (Bahdanau *et al.*, 2014). Les HGATs (Albarede *et al.*, 2022) s'appuient sur les GAT en intégrant différents types d'arêtes. Chacune de leurs couches contient : (1) Un *composant d'échantillonnage* qui définit le voisinage N_i du nœud i, incluant i lui-même ; (2) un *composant de propagation* qui calcule une représentation pour le nœud i en agrégeant la représentation de chaque nœud dans son voisinage N_i en fonction du type d'arête qui les relie. Afin de les adapter à la recherche de passages, nous intégrons les modifications suivantes : *No Out* (NO) : les liens sortants sont ignorés lors du calcul de voisinage pendant le processus d'échantillonnage ; *No Self* (NS) : les boucles sont ignorées lors du calcul de voisinage pendant le processus d'échantillonnage ; *Lambda Separator* (LB) : un coefficient spécifique λ est ajouté aux boucles et $1 - \lambda$ à toutes les autres arêtes.

3 Intégration des entités dans les HGATs

Un document est représenté sous la forme d'un graphe hétérogène où les nœuds représentent les éléments structurels (passages, sections) et les entités, et les arêtes représentent leurs relations. Nous étendons le graphique de contextualisation présenté dans (Albarede *et al.*, 2022), en ajoutant le type d'*entité* du nœud et le type d'*entity_rel* de l'arête. Nous définissons la relation entre une entité et un passage comme la mention de l'entité dans le texte du passage.

Un encodeur neuronal calcule les représentations sémantiques pour chaque nœud du graphe. Les entités sont représentées soit par leur **étiquette** (label), soit par leur **description** (desc). Nos propositions pour la *composante d'échantillonnage* de l'*EE-HGAT* visent à étudier la manière dont les entités interagissent avec les passages. *Une entité doit servir de lien entre les passages* (Yu *et al.*, 2021). Sur la base de cette hypothèse, nous étudions le *composant d'échantillonnage* de **liens**. Ce composant définit N_i le voisinage du nœud d'*entité* i tel qu'il contient chaque nœud j où $A_{ji} \neq \emptyset$ ainsi que i lui-même (A étant la matrice d'adjacence associée au graphe hétérogène). Cela permet à l'information de circuler à travers les nœuds d'entité (c'est-à-dire que deux nœuds connectés à un nœud d'entité peuvent recevoir des informations l'un de l'autre). *Une entité doit injecter ses informations dans des éléments structurels* (Dong *et al.*, 2022; Ju *et al.*, 2022; Xiong *et al.*, 2017; Das *et al.*, 2019). Sur la base de cette hypothèse, nous étudions le *composant d'échantillonnage* **injecté**. Ce composant définit N_i le voisinage du nœud d'*entité* i comme l'ensemble vide : les nœuds d'*entités* propagent leurs informations et les empêche de recevoir des informations d'autres nœuds. Les représentations des nœuds *entités* ne sont pas modifiées.

Nous présentons dans le tableau 1 trois modifications de l'architecture des *HGATs* dans nos *EE-HGATs* (cf. Section 2) : i) $BASE$ pour étudier nos propositions sans aucune modification ; ii) NO et LB, d'autres expériences (soumission en cours d'examen, non rapportée ici) ont montré de bonnes performances dans les *HGATs* ; iii) NO, NS et LB, car d'autres expériences (soumission en cours d'examen, non rapportée ici) ont montré les meilleures performances moyennes dans les *HGATs* à travers plusieurs tâches de recherche de Passages.

Représentation des entités		**Composant d'échantillonnage**		**Modifications de l'architecture**			**identifiant *EE-HGAT***
étiquette	description	lien	injecter	NO	NS	LB	
✓		✓					$label\_link_{BASE}$
✓		✓		✓		✓	$label\_link_{NO\_LB}$
✓		✓		✓	✓	✓	$label\_link_{NO\_NS\_LB}$
	✓	✓					$desc\_link_{BASE}$
	✓	✓		✓		✓	$desc\_link_{NO\_LB}$
	✓	✓		✓	✓	✓	$desc\_link_{NO\_NS\_LB}$
✓			✓				$label\_inject_{BASE}$
✓			✓	✓		✓	$label\_inject_{NO\_LB}$
✓			✓	✓	✓	✓	$label\_inject_{NO\_NS\_LB}$
	✓		✓				$desc\_inject_{BASE}$
	✓		✓	✓		✓	$desc\_inject_{NO\_LB}$
	✓		✓	✓	✓	✓	$desc\_inject_{NO\_NS\_LB}$

TABLE 1 – Les modèles *EE-HGAT* utilisés.

Afin d'effectuer la recherche de passages, nous combinons nos *EE-HGATs* avec des cadres de classement de passages. Sur la base des travaux réalisés dans (Albarede *et al.*, 2022), nous considérons deux cadres : i) Le cadre standard, dans lequel la pertinence d'un passage est estimée à l'aide d'une représentation unique contenant à la fois des informations sur le contenu et le contexte ; et ii) Le cadre d'injection tardive, dans lequel la pertinence d'un passage est estimée à l'aide d'une représentation contenant des informations sur le contenu et d'une représentation contenant uniquement des informations sur le contexte. Ces cadres utilisent un encodeur qui apprend à différencier les requêtes et les documents à l'aide d'un token spécial "[Q]" (query) et d'un token spéciel "[D]" (document) ajoutés au texte. Nous introduisons un token spécial "[E]" qui est ajouté à la forme textuelle d'une entité. Conformément à la section 2, nous associons les modèles *EE-HGAT* en utilisant les combinaisons de modifications $BASE$ et NO et LB, avec le cadre standard et avec le cadre de l'injection tardive de (Albarede *et al.*, 2022) : par exemple $desc\_inject_{NO\_NS\_LB}$ utilisé avec l'injection tardive est noté $desc\_inject_{late\_NO\_NS\_LB}$.

4 Expérimentations et résultats

Nous expérimentons nos approches sur la tâche de Recherche de Passages TREC CAR Y3 (Dietz & Foley, 2019). Un **label** d'entité est le titre du document Wikipédia correspondant, et une **description** d'entité est le premier passage du document Wikipedia correspondant. Pour représenter les documents sous forme de graphe, un nœud est créé pour chaque entité et relié aux nœuds des passages la mentionnant. Nous définissons 3 types de nœuds ($passage$, $section$, $entity$) et 6 types d'arêtes : $horizontal$, pour les $passage$ voisins et $horizontal_i$ son symétrique ; $hierarchical$ pour les liens entre un nœud $section$ et un nœud $passage$ ou entre deux nœuds $section$ et $hierarchical_i$ son symétrique ; $entity\_rel$ pour la relation entre un nœud $entity$ et un nœud $passage$ et $entity\_rel_i$ son symétrique.

Nous réalisons nos expériences en utilisant le système de RI `Pyterrier` (Macdonald & Tonellotto, 2020) et le framework `Pytorch` (Paszke *et al.*, 2019). Nous avons fixé une longueur maximale de représentation de passage de 180 tokens et une longueur maximale de représentation de requête de

Modèle de classement des passages	P@10	nDCG@20	MAP
$label\_link_{std\_BASE}$	0.156 ± 0.018	0.216 ± 0.014	0.110 ± 0.021
$label\_link_{std\_NO\_LB}$	$0.196^{il} \pm 0.017$	$0.266^{il} \pm 0.010$	$0.150^{il} \pm 0.022$
$label\_link_{late\_NO\_NS\_LB}$	0.185 ± 0.038	0.255 ± 0.051	0.151 ± 0.046
$desc\_link_{std\_BASE}$	0.176 ± 0.017	0.231 ± 0.015	0.134 ± 0.023
$desc\_link_{std\_NO\_LB}$	$0.206^{il} \pm 0.017$	$0.291^{ijkl} \pm 0.005$	$0.174^{ikl} \pm 0.019$
$desc\_link_{late\_NO\_NS\_LB}$	$0.215^{i} \pm 0.046$	$0.301^{ijlk} \pm 0.044$	$0.181^{il} \pm 0.037$
$label\_inject_{std\_BASE}$	$0.194^{il} \pm 0.019$	$0.287^{ijkl} \pm 0.011$	$0.198^{ijklm} \pm 0.021$
$label\_inject_{std\_NO\_LB}$	$0.250^{ijklmnor} \pm 0.010$	$0.377^{ijklmnor} \pm 0.009$	$0.241^{ijklmnor} \pm 0.015$
$label\_inject_{late\_NO\_NS\_LB}$	$0.255^{ijklmno} \pm 0.028$	$0.381^{ijklmnor} \pm 0.037$	$0.254^{ijklmnor} \pm 0.041$
$desc\_inject_{std\_BASE}$	$0.233^{ijklmo} \pm 0.020$	$0.316^{ijklmo} \pm 0.016$	$0.211^{ijklmno} \pm 0.021$
$desc\_inject_{std\_NO\_LB}$	$0.261^{ijklmnopqr} \pm 0.012$	$0.386^{ijklmnopqr} \pm 0.004$	$0.256^{ijklmnopr} \pm 0.017$
$desc\_inject_{late\_NO\_NS\_LB}$	$\mathbf{0.265}^{ijklmnoqr} \pm 0.027$	$\mathbf{0.401}^{ijklmnor} \pm 0.041$	$\mathbf{0.262}^{ijklmnor} \pm 0.036$

TABLE 2 – Résultats (mean $\pm$ stdev). $i, ..., r$: significativité stat. pour chaque modèle dans l'ordre décroissant (test U de Mann-Whitney, p-value =0.05). En gras : meilleur résultat par colonne.

120 tokens. Nos *EE-HGAT* sont composés de 3 couches, chaque couche ayant plusieurs fonctions d'attention avec 8 têtes avec un *dropout* égal à 0.7. La recherche de passage est effectuée en 2 phases : nous utilisons *BM25* (Robertson *et al.*, 1995) pour récupérer les 100 premiers documents et nous classons les passages de ces documents à l'aide de notre proposition. Plus de détails sur l'implémentation peuvent être trouvés dans (Albarede *et al.*, 2023).

Nous présentons les résultats de nos approches dans le tableau 2 sous la forme $mean \pm st.dev$. Nous constatons que les approches exploitant la composante d'**injection** donnent globalement de meilleurs résultats que celles exploitant la composante de **lien**. Plus précisément, nous constatons que l'utilisation d'un composant d'échantillonnage par injection améliore toujours les performances de manière significative (par ex. $label\_inject_{std\_NO\_LB}$ améliore sensiblement les performances par rapport à $label\_link_{std\_NO\_LB}$ de $+41,7\%$, $+60,6\%$ et $+27.5\%$ pour les mesures nDCG@20, MAP et P@10, respectivement). Une explication potentielle de cet effet est que le **lien** permet aux passages de recevoir des informations d'autres passages par le biais d'une relation d'entité partagée, mais des passages provenant de documents très différents peuvent mentionner la même entité. Un passage peut alors recevoir des informations trop larges ou trop éloignées de son contenu intrinsèque, entravant le calcul de sa représentation contextualisée. En nous concentrant sur le type de représentation des entités, nous constatons que les approches exploitant la représentation des **desc** donnent globalement de meilleurs résultats que celles exploitant la représentation des **étiquettes**. Ainsi, les représentations d'entités plus riches sont bénéfiques pour le calcul de la représentation des passages.

5 Conclusion

Cet article présente une proposition détaillée de définition de l'extension de l'entité de HGATs pour la recherche de passages. Il décrit également comment ces HGATs peuvent être intégrés dans une stratégie de reclassement des passages. Les expériences réalisées sur le corpus CAR Y3 montrent que les architectures et les choix d'apprentissage de graphes adéquats peuvent être stables.

Remerciements

Ce travail a été partiellement soutenu par le MIAI@Grenoble Alpes (ANR-19-P3IA-0003), ainsi que par l'Association Nationale de la Recherche et de la Technologie (ANRT).

Références

ALBAREDE L., GOEURIOT L., MULHEM P., PAPE-GARDEUX C. L., MARIÉ S. & CHARDIN-SEGUI T. (2023). Entity enhanced attention graph-based passages retrieval. In *2nd Workshop on Augmented Intelligence for Technology-Assisted Reviews Systems : Evaluation Metrics and Protocols for eDiscovery and Systematic Review Systems - ECIR Workshop*.

ALBAREDE L., MULHEM P., GOEURIOT L., LE PAPE-GARDEUX C., MARIÉ S. & CHARDIN-SEGUI T. (2022). Passage retrieval on structured documents using graph attention networks. In *Proceedings of ECIR 2022*, p. 13–21, Stavanger, Norway.

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate.

DAS R., GODBOLE A., KAVARTHAPU D., GONG Z., SINGHAL A., YU M., GUO X., GAO T., ZAMANI H., ZAHEER M. & MCCALLUM A. (2019). Multi-step entity-centric information retrieval for multi-hop question answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, p. 113–118, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5816.

DIETZ L. & FOLEY J. (2019). Trec car y3 : Complex answer retrieval overview. In *Proceedings of Text REtrieval Conference (TREC)*.

DONG Q., LIU Y., CHENG S., WANG S., CHENG Z., NIU S. & YIN D. (2022). Incorporating explicit knowledge in pre-trained language models for passage re-ranking. DOI : 10.48550/ARXIV.2204.11673.

JU M., YU W., ZHAO T., ZHANG C. & YE Y. (2022). Grape : Knowledge graph enhanced passage reader for open-domain question answering. DOI : 10.48550/ARXIV.2210.02933.

MACDONALD C. & TONELLOTTO N. (2020). Declarative experimentation in information retrieval using pyterrier. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, p. 4526–4533 : ACM.

PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KOPF A., YANG E., DEVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J. & CHINTALA S. (2019). Pytorch : An imperative style, high-performance deep learning library. In H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ-BUC, E. FOX & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 32*, p. 8024–8035. Curran Associates, Inc.

ROBERTSON S., WALKER S., JONES S., HANCOCK-BEAULIEU M. M. & GATFORD M. (1995). Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, p. 109–126 : Gaithersburg, MD : NIST.

XIONG C., CALLAN J. & LIU T.-Y. (2017). Word-entity duet representations for document ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, p. 763–772, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3077136.3080768.

YU D., ZHU C., FANG Y., YU W., WANG S., XU Y., REN X., YANG Y. & ZENG M. (2021). Kg-fid : Infusing knowledge graph in fusion-in-decoder for open-domain question answering. *CoRR*, **abs/2110.04330**.