Peuplement de base de connaissances, évaluation et système end-to-end

Maxime Prieur[*,**], Cédric du Mouza[*], Guillaume Gadek[**], Bruno Grilheres[**]

[*] Laboratoire Cédric, Conservatoire National des Arts et Métiers
maxime.prieur.auditeur@lecnam.net, dumouza@cnam.fr
[**] Airbus Defence and Space, Élancourt France
guillaume.gadek, bruno.grilheres@airbus.com

Résumé. Les bases de connaissances (KB) sont utilisées dans de nombreux domaines, comme l'intelligence économique ou l'assistance aux utilisateurs. Elles regroupent des connaissances pouvant être exploitées par l'ordinateur. Cependant, leur création reste complexe pour un expert devant extraire et lier chaque nouvelle information. Dans cet article, nous décrivons une méthode d'évaluation de système d'enrichissement de base de connaissances. Cette évaluation est mise à l'épreuve par ELROND un système complet conçu autour d'une chaîne de traitement composée de 4 modules.

1 Introduction

Les bases de connaissances sont des structures de données régies par des ontologies prédéfinies et sont très utiles pour agréger l'information dans le but de simplifier sa visualisation et son analyse. Ces bases sont utilisées dans de multiples domaines tels que les métiers de l'intelligence économique (Shue et al., 2009) afin de faciliter la prise de décision ou bien pour extraire les éléments liant des publications scientifiques (Luan et al., 2018) par exemple.

Cependant, la construction et la mise à jour manuelles de ces bases sont coûteuses puisque souvent les domaines d'utilisation exploitent des informations en constante évolution. L'alternative, un enrichissement automatique, permettrait d'extraire et d'ajouter les éléments souhaités depuis des sources d'intérêts vers la structure de données. Néanmoins, les solutions existantes sont encore limitées et il n'existe à notre connaissance aucun protocole d'évaluation couvrant l'entièreté de la tâche de peuplement (Min et al., 2018; Mesquita et al., 2019).

Dans cet article, nous tentons de combler ce manque en proposant les contributions suivantes :

— Nous formalisons une méthode d'évaluation des systèmes complets d'enrichissement de base de connaissances depuis des textes.
— Nous introduisons ELROND, un système de peuplement end-to-end en 4 étapes sur lequel nous testons notre méthode d'évaluation.
— Nous évaluons et comparons le modèle proposé en appliquant le mode d'évaluation présenté au dataset DWIE (Zaporojets et al., 2021).

La suite de cet article est structurée comme suit : après avoir présenté les travaux récents sur les systèmes d'enrichissement de base de connaissances (Knowledge Base Population, KBP)

et leur évaluation (section 2), nous formaliserons et détaillerons notre protocole d'évaluation (section 3). Nous décrirons ELROND (section 4) puis présenterons l'expérience menée et les résultats obtenus (section 5) avant de discuter des pistes à explorer dans des travaux futurs.

2 État de l'art

Enrichissement de base de connaissances. Le peuplement d'une base de connaissances depuis des textes consiste à extraire les éléments et leurs relations afin de compléter des informations déjà connues et structurées. Cette tâche englobe en général plusieurs étapes : reconnaissance d'entités nommées (REN), résolution de coréférence, extraction de relations et liage des entités à la base. TinkerBell (Al-Badrashiny et al., 2017), l'un des premiers systèmes complets, combine deux Bi-LSTM (Graves et al., 2013) utilisant le texte et les caractéristiques linguistiques pour assigner des tags aux mots du document. Le liage d'entité somme des scores de popularité, de similarité et de cohérence. KnowledgeNet (Mesquita et al., 2019) résout l'extraction de relations par un modèle Bi-LSTM utilisant des caractéristiques linguistiques et les représentations obtenues par un modèle BERT (Devlin et al., 2019), mais ne détecte pas de relations au niveau supra-phrastiques (Yao et al., 2019). Ces approches reposent sur Wikipédia et ne sont pas adaptées lorsque les sources présentent une grande proportion d'entités non répertoriées dans l'encyclopédie. KBPearl (Lin et al.) propose d'utiliser des frameworks d'extraction d'information ouverte (OIE). Le système extrait et lie les connaissances du texte par une méthode de densification de graphe appliquée au graphe sémantique des connaissances du texte. En plus d'une profusion d'informations potentiellement inintéressantes pour l'utilisateur produite par les frameworks OIE, la sélection des candidats est réalisée par correspondance d'alias, peu adaptée lors de l'apparition de nouvelles mentions.

Évaluation d'un système de KBP. S'il existe des benchmarks et des métriques pour l'évaluation des sous-tâches (F1-score pour le REN et l'extraction de relations, Hit@k pour le liage, etc), peu de propositions existent pour les systèmes complets, construisant des bases de connaissances depuis des textes. Les workshops TAC KBP évaluent un système en calculant la précision sur des requêtes 1-hop, « Que porte Frodon au Mordor ? », et 2-hop, « Qui a créé ce que porte Frodon au Mordor ? ». Le numéro étant le nombre de relations séparant les deux entités. Le coût d'une évaluation manuelle de systèmes retournant un grand nombre de réponses (Ellis et al., 2015) oblige les évaluateurs à se concentrer sur un nombre restreint de requêtes et n'évaluent donc pas la base dans son entièreté. Min et al. (2018) rendent possible l'évaluation automatique en mesurant l'alignement de triplets *(sujet, relation, objet)* entre la référence et ce qui est prédit. Le liage d'une entité créée avec une entité de référence est fait si l'entité produite partage plus de 50% des mentions avec l'entité de référence. Cet alignement pose les questions du cas où le système n'extrairait qu'un nombre faible de mentions et du choix arbitraire du seuil à 50%. KnowledgeNet (Mesquita et al., 2019) mesure le $F1$ score sur l'extraction de triplets annotés dans des phrases et le liage du couple d'entités sujet et objet à leur page Wikidata. Les phrases du jeu de données n'étant annotées que pour un couple et une relation à chaque phrase, il est impossible d'évaluer correctement la précision puisque des résultats pourraient être considérés à tort comme de faux positifs. Comme Min et al. (2018), l'évaluation se fait au niveau textuel et écarte la construction d'une base. Ces méthodes d'évaluation incomplètes soulignent la nécessité d'un protocole évaluant les performances d'un système KBP.

3 Modèle de données et définitions

3.1 L'enrichissement de bases de connaissances

Une KB se compose d'éléments (des entités, des attributs et de relations entre ces derniers) suivant une ontologie définie. Elle peut ainsi être modélisée par un graphe dans lequel les nœuds sont les différents éléments et les arêtes traduisent l'existence d'une relation entre ces éléments. On définit ainsi une KB comme suit :

Definition 3.1 (Base de Connaissances) *Une base de connaissance est une structure de données modélisable par un graphe $G = (V, E, \Phi, \Psi)$ où V est l'ensemble des nœuds du graphe, E celui des arcs entre deux nœuds de V, $\Phi : V \rightarrow \mathcal{A}$ est une fonction qui pour tout nœud v_i de V associe un ensemble d'attributs $A_i \in \mathcal{A}$ représentés par des tuples (clé, valeur) et $\Psi : E \rightarrow \mathcal{E}$ une seconde fonction qui associe à chaque arc $e_i \in E$ un type d'arc $E_i \in \mathcal{E}$, avec $\mathcal{A}$ et $\mathcal{E}$ désignant respectivement l'ensemble des attributs et l'ensemble des types de relation.*

Enrichir une KB avec du contenu textuel consiste donc à y ajouter les éléments extraits de textes en respectant une ontologie. Pour lier les informations d'une même entité rencontrées dans plusieurs textes, les entités doivent posséder un identifiant unique (URI). Ceci permet d'obtenir pour un ensemble de k textes, une base de référence, $G_k = (V_k, E_k, \Phi_k, \Psi_k)$ et de mesurer la proportion d'information correctement extraite par un système ayant construit une base $G'_k = (V'_k, E'_k, \Phi'_k, \Psi'_k)$.

Exemple de workflow pour la tâche de KBP. Un système d'enrichissement de base de connaissances s'articule autour de composants ou de modules constituant la chaîne de traitement pour résoudre la tâche de KBP. Le premier est chargé de reconnaître les entités nommées (REN) et les autres éléments d'intérêt dans le texte (attributs, entités non nommées) tout en leur attribuant un type en s'aidant du document. Par exemple, pour le passage *« Joe Biden, le président américain, s'est montré virulent envers Trump »*, l'étape permet d'obtenir les mentions [*(Joe Biden, Per), (le président américain, Per), (président, Rôle), (américain, Nationalité), (Trump, Per)*]. La deuxième brique de traitement, regroupe les éléments textuels qui coréférent. Les mentions du REN appartenant à un cluster composé de mentions de même type sont gardées en coréférence et les mentions restantes sont considérées comme des singletons. En reprenant l'exemple précèdent, on obtient deux clusters [*(1, PER, Joe Biden, le président américain), (2, PER, Trump)*]. On détermine ensuite, par une troisième brique, les relations liant les éléments [*(1, VS, 2), (1, Rôle, Président), (1, Nationalité, Américain)*]. Ces relations en plus d'être une partie de l'information à extraire constituent un support pour la dernière étape, le liage d'entité. Chaque entité du texte, lorsque c'est possible, est associée à une entité de la base. *« Trump »* doit être rattachée à l'entité *« Donald Trump »* et non *« Fred Trump »*. Au final, l'ensemble des informations extraites du texte viennent enrichir les informations de la base en complétant celles des entités déjà connues ou en ajoutant de nouvelles entités.

3.2 Mesurer la qualité de l'enrichissement

Les entités se caractérisent donc par des attributs et des relations les liant à d'autres entités de la base. Lors de la comparaison d'une entité de référence et d'une entité construite, il faut vérifier qu'à la fois les attributs et les relations correspondent.

Definition 3.2 (Similarité d'attributs et de relations) *Nous adoptons pour les attributs et les relations les définitions de similarité suivantes :*

- ***similarité d'attributs** : les attributs correspondent s'ils possèdent le même type, valeur et texte d'inférence (dans lequel l'attribut apparaît). Bien qu'inclure le texte d'inférence crée une multiplication de l'information, cela permet de vérifier que le système extrait correctement l'information à chaque fois qu'elle est mentionnée.*
- ***similarité de relations** : les relations sont semblables si elles impliquent le même type de relation, texte d'inférence et que l'ensemble des mentions de l'entité objet de la relation construite sont inclues dans les mentions de l'entité objet de la base de référence.*

Pour vérifier si une entité a été correctement extraite, nous comparons les caractéristiques extraites avec celles associée à l'entité de référence. Afin de mesurer la qualité de l'extraction, nous proposons d'adapter la précision, le rappel et la mesure F1 comme suit :

$$
\begin{aligned}
P_{v_i,v_j,k} &= \frac{\alpha|TP_{rel}| + \beta|TP_{attr}|}{\alpha(|TP_{rel}| + |FP_{rel}|) + \beta(|TP_{attr}| + |FP_{attr}|)} \\
R_{v_i,v_j,k} &= \frac{\alpha|TP_{rel}| + \beta|TP_{attr}|}{\alpha(|TP_{rel}| + |FN_{rel}|) + \beta(|TP_{attr}| + |FN_{attr}|)} \\
F1_{v_i,v_j,k} &= 2\frac{P_{v_i,v_j,k} \times R_{v_i,v_j,k}}{P_{v_i,v_j,k} + R_{v_i,v_j,k}}
\end{aligned}
\tag{1}
$$

Avec TP, FP et FN pour respectivement vrais positifs, faux positifs et faux négatifs, v_i et v_j des nœuds appartenant respectivement à G_k et G'_k, et $0 \leq \alpha, \beta \leq 1$ des poids tels que $\alpha + \beta = 1$ permettant de donner une importance différente aux attributs et aux relations.

Alignement des entités. Nous alignons chaque entité de la base de référence avec une entité de la base construite en utilisant le score $F1$ proposé ci-dessus et l'algorithme Hongrois (Kuhn, 1955). L'alignement est possible pour des paires avec un score strictement positif. Les entités sans correspondance de G_k et G'_k sont respectivement considérées comme des faux négatifs et des faux positifs. En warm-start, les entités initialement présentes reste alignées et leur F1-score ne prend en compte que les nouvelles informations. Cette phase d'appariement permet de construire Ω_k un ensemble de paires (v_{i,G_k}, v_{j,G'_k}).

Scores finaux pour la qualité de l'extraction. La comparaison entre les bases après k textes s'effectue en agrégeant les scores de similarités de paires d'entités formées. Pour mesurer la proportion d'information correctement extraite, deux scores $F1$, un $F1_{micro}$ et un $F1_{macro}$, peuvent être obtenus :

$$
\begin{aligned}
P_{micro,k} &= \frac{\alpha \sum_{(v_i,v_j)\in\Omega_k} |e_{v_j} = e_{v_i}| + \beta \sum_{(v_i,v_j)\in\Omega_k} |a_{v_j} = a_{v_i}|}{\alpha|E_{G'_k}| + \beta|A_{G'_k}|} \\
R_{micro,k} &= \frac{\alpha \sum_{(v_i,v_j)\in\Omega_k} |e_{v_j} = e_{v_i}| + \beta \sum_{(v_i,v_j)\in\Omega_k} |a_{v_j} = a_{v_i}|}{\alpha|E_{G_k}| + \beta|A_{G_k}|} \\
F1_{micro,k} &= 2\frac{P_{micro,k} \times R_{micro,k}}{P_{micro,k} + R_{micro,k}} \qquad F1_{macro,k} = \frac{\sum_{(v_i,v_j)\in\Omega_k} F1_{v_i,v_j,k}}{|\Omega_k| + |FN| + |FP|}
\end{aligned}
\tag{2}
$$

Le $F1_{macro}$ est une moyenne des scores de similarité des entités alignées, à l'inverse du $F1_{micro}$, il ne prends pas en compte la différence de distribution (nombre et type de relations ou d'attributs) entre les entités. Le $F1_{micro}$ est pondéré et calculé en fonction des éléments identiques entre les entités alignées.

Avantages du mode d'évaluation. Le protocole proposé mesure à différents intervalles la distance entre les bases, ce qui permet de montrer la résilience d'un système aux erreurs pouvant être commises et risquant de polluer la base construite. L'impact d'un module sur la chaîne complète est mesurable en utilisant les vrais résultats sur le reste de cette chaîne. De plus, remplacer la correspondance exacte d'entités par la proportion d'informations identiques dans une paire pour calculer les scores de similarité, apporte plus de souplesse et une meilleure représentativité des performances des systèmes.

4 ELROND : un système de KBP complet

Cette section présente, ELROND (Entity Linking and Relation extraction On News Documents). Un système implémentant les étapes explicitées en partie 3. Ce système fait office de baseline pour la tâche de KBP et permet de montrer l'intérêt de la méthode d'évaluation détaillée en partie 3. Les composants principaux et leurs interactions, sont illustrés en figure 1.

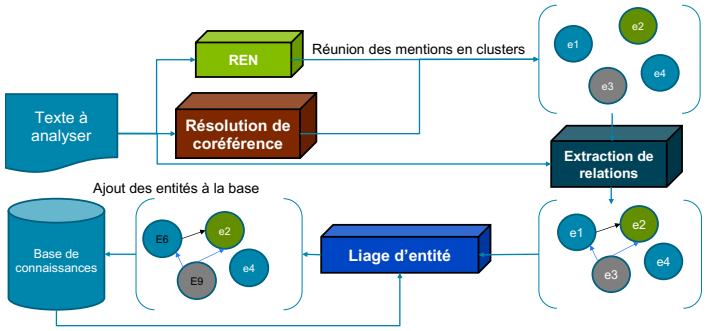


FIG. 1 – Schéma du workflow des composants dans ELROND.

Reconnaissance d'entités nommées. La brique de REN utilise un modèle RoBERTa (Liu et al., 2019), pré-entraîné et fine-tuné. Le choix de RoBERTa a été motivé par ses performances actuelles sur les différents benchmarks pour la reconnaissance d'entités nommées

Résolution de coréférence. Pour cette étape, nous utilisons le modèle pré-entraîné, *Word-level Coreference Resolution* (Dobrovolskii, 2021). Celui-ci forme des groupes de mots en coréférence (entités nommées, non nommées, pronoms, etc). La représentation d'un token est obtenue en pondérant les vecteurs de ses sous-tokens produits par RoBERTa. Les poids résultent d'une fonction softmax appliquée à la projection des vecteurs par une matrice d'attention. Le modèle prédit une coréférence lorsque la somme d'une projection bilinéaire entre

deux tokens et la sortie d'un réseau de neurones est positive. Nous avons choisi d'intégrer *Word-level Coreference Resolution* à ELROND en raison de ses performances pour cette tâche.

Extraction de relations. Les relations sont obtenues à l'aide du modèle ATLOP (Zhou et al., 2021) qui représente chacune des entités en appliquant une fonction de pooling sur les vecteurs des mentions obtenus par un PLM. Pour chaque couple d'entités, on obtient un coefficient d'attention utilisé dans une fonction bi-linéaire calculant la plausibilité d'un type de relation. Si le score est supérieur au type « Null », on prédit cette relation comme existante.

Liage des entités. Le dernier composant applique une recherche par mention. Pour chaque entité du texte, on retourne celles de la base, de même type et partageant au moins une mention avec l'entité du texte. Si aucune entité n'est retournée, une nouvelle est créée. Si plusieurs correspondances sont trouvées, une sélection par popularité, similaire à Al-Badrashiny et al. (2017) est appliquée. L'entité avec le plus de mentions en commun est sélectionnée.

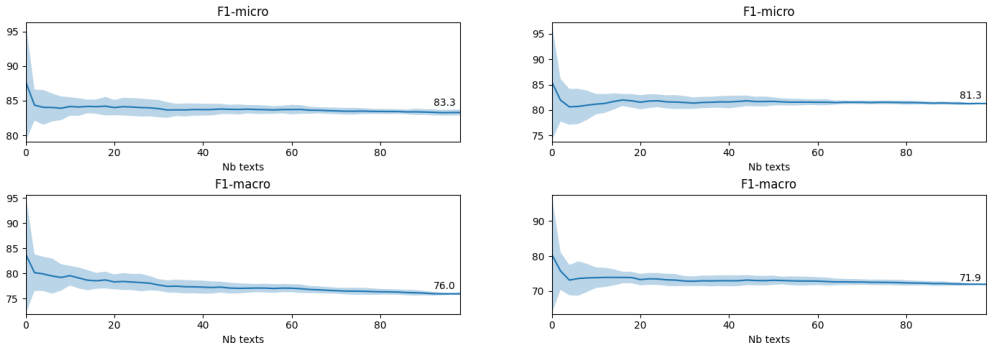
5 Expériences

Détails d'implémentation. L'ensemble des approches proposées sont implémentées en Python et utilisent la librairie *Pytorch* [1]. Le modèle REN est entraîné à l'aide du framework Flair (Akbik et al., 2019). Le code est disponible via un répertoire git [2].

Jeu de données Pour une mesure complète des systèmes, l'exhaustivité de l'annotation du jeu de données sur toutes les dimensions de l'information à extraire est nécessaire. Nous avons donc utilisé DWIE (Zaporojets et al., 2021), le seul dataset gratuit à notre connaissance respectant cette contrainte. Dans les 800 articles de presse en anglais (700 textes d'entraînement et 100 textes de test) de DWIE, les entités y sont annotées suivant une quinzaine de classes et de relations. Pour la tâche de KBP, les types et les alias font office d'attributs. Étant donné que la présence de certains éléments dans une base, et donc l'ordre dans lequel ces éléments apparaissent, peut favoriser ou non le fonctionnement des systèmes, les performances sont mesurées et moyennées sur 10 ordonnancements différents du jeu de test.

Résultats et discussion La figure 2 montrent les score moyens obtenus par ELROND pour la tâche de KBP au fil des 100 textes de test pour 10 ordonnancements différents. Dans le cas du warm-start, la base initiale est constituée des informations contenues dans les 700 textes d'entraînement. On observe que la distance entre les bases est plus grande en warm-start en raison de la difficulté à lier les nouvelles informations avec celles possédées initialement. Dans les deux cas, les scores diminuent au fil des textes, attestant d'une accumulation d'erreurs au cours de l'enrichissement. Le score $F1_{micro}$ est plus élevé que le $F1_{macro}$ dans les deux cas et semble plus stable à la fin de l'évaluation. Ceci s'explique par le fait que les entités populaires telles que les pays et les villes, ont plus de poids dans la base finale et sont plus faciles à reconnaître. Ceci induit que le $F1_{macro}$, qui lisse la différence de distribution, sera plus faible.

1. `https://pytorch.org/`
2. `https://github.com/Todaime/KBP.git`



(a) Cold-start (b) Warm-start

FIG. 2 – Performances d'ELROND pour l'enrichissement de base avec les textes de DWIE.

Nous avons également évalué les performances d'ELROND et du modèle introduit dans DWIE à l'aide du protocole décrit en section 3. Le modèle de DWIE ne possédant pas de solution pour le liage d'entité, nous avons utilisé celle adoptée dans ELROND. En s'intéressant aux résultats finaux des modèles (tableau 1), on observe qu'ELROND apporte une amélioration allant jusqu'à 2.1% pour le $F1_{macro}$ du scénario warm-start.

	Cold-start		Warm-start	
Model	F1-Micro	F1-Macro	F1-Micro	F1-Macro
ELROND	**83.3**	**76.0**	**81.3**	**72.0**
DWIE	82.8	75.6	80.3	69.9

TAB. 1 – Scores finaux de ELROND et DWIE pour la tâche de KBP en Cold et Warm start.

6 Conclusion

Nous avons formalisé un mode d'évaluation automatique, complet et modulable pour la tâche de KBP depuis des textes. Celui-ci permet de comparer des méthodes dans des scenarios warm-start et cold-start. Ce protocole a été utilisé pour mesurer les performances d'ELROND, un premier système servant de baseline pour de futures améliorations. De futurs travaux visent à étudier l'apport des NER multi-classes pour favoriser l'apprentissage et déterminer la place de l'ontologie dans la tâche de KBP. Nous travaillons également sur la production d'un dataset français pour entraîner et évaluer les systèmes de KBP.

Références

Akbik, A., T. Bergmann, D. Blythe, K. Rasul, S. Schweter, et R. Vollgraf (2019). FLAIR : An easy-to-use framework for state-of-the-art NLP. In *NAACL*, pp. 54–59.

Al-Badrashiny, M., J. Bolton, A. T. Chaganty, K. Clark, C. Harman, L. Huang, M. Lamm, J. Lei, D. Lu, X. Pan, et al. (2017). Tinkerbell : Cross-lingual cold-start knowledge base construction. In *TAC*.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186.

Dobrovolskii, V. (2021). Word-level coreference resolution. In *EMNLP*, pp. 7670–7675.

Ellis, J., J. Getman, D. Fore, N. Kuster, Z. Song, A. Bies, et S. M. Strassel (2015). Overview of linguistic resources for the tac kbp 2015 evaluations : Methodologies and results. In *TAC*.

Graves, A., N. Jaitly, et A.-r. Mohamed (2013). Hybrid speech recognition with deep bidirectional lstm. In *IEEE Work. ASRU*, pp. 273–278.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly 2*(1-2), 83–97.

Lin, X., H. Li, H. Xin, Z. Li, et L. Chen. Kbpearl : A knowledge base population system supported by joint entity and relation linking. *VLDB Endowment 13*(7).

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, et V. Stoyanov (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

Luan, Y., L. He, M. Ostendorf, et H. Hajishirzi (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP*.

Mesquita, F., M. Cannaviccio, J. Schmidek, P. Mirza, et D. Barbosa (2019). Knowledgenet : A benchmark dataset for knowledge base population. In *EMNLP-IJCNLP*, pp. 749–758.

Min, B., M. Freedman, R. Bock, et R. M. Weischedel (2018). When ace met kbp : End-to-end evaluation of knowledge base population with component-level annotation. In *LREC*.

Shue, L.-Y., C.-W. Chen, et W. Shiue (2009). The development of an ontology-based expert system for corporate financial rating. *Expert Systems with Applications 36*(2), 2130–2142.

Yao, Y., D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, et M. Sun (2019). Docred : A large-scale document-level relation extraction dataset. In *CoRR*.

Zaporojets, K., J. Deleu, C. Develder, et T. Demeester (2021). Dwie : An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management 58*(4), 102563.

Zhou, W., K. Huang, T. Ma, et J. Huang (2021). Document-level relation extraction with adaptive thresholding and localized context pooling. In *AAAI*, pp. 14612–14620.

Summary

Knowledge Bases (KB) are used in many fields and gather knowledge that can be exploited by the computer. However, their creation remains complex for an expert who has to extract and link each new information. In this paper, we describe a method for evaluating a Knowledge Base Population system. This evaluation is put to the test with ELROND, a complete system designed around a processing chain composed of 4 modules.