Apprendre à classer le contexte pour la reconnaissance d'entités nommées en utilisant un jeu de données synthétique

Arthur Amalvy[1], Vincent Labatut[1] and Richard Dufour[2]

[1]*Laboratoire Informatique d'Avignon*
[2]*Laboratoire des Sciences du Numérique de Nantes*

Abstract
Même si les modèles récents à base de transformeurs obtiennent de bons résultats dans de nombreuses tâches de traitement du langage, leur portée limitée est un problème lorsqu'il s'agit de traiter de longs documents comme des romans. Il est possible de récupérer du contexte au niveau du document de manière non supervisée pour améliorer les performances d'une tâche, mais entraîner un modèle à récupérer ce contexte de manière supervisée est difficile à cause du manque de données. Pour la tâche de reconnaissance des entités nommées (REN), nous proposons de générer un jeu de données synthétique de récupération de contexte afin d'entraîner un modèle de reclassement. Nous montrons qu'un tel modèle peut surpasser des approches non supervisées lorsqu'il s'agit d'améliorer la performance d'un modèle de REN sur un corpus littéraire.

Keywords
Reconnaissance d'Entités Nommées, Textes Littéraires, Apprentissage Profond

1. Introduction

La reconnaissance d'entités nommées (REN) est une tâche fondamentale pour le traitement du langage naturel. Si les transformeurs pré-entraînés tels que BERT [1] ou LUKE [2] obtiennent de très bonnes performances, les textes longs restent problématiques à cause de la complexité quadratique de leur mécanisme d'attention en fonction du nombre de tokens d'entrée. De tels documents sont donc souvent traités par blocs, ce qui interdit l'accès au contexte global du document et peut résulter en une perte de performance [3]. Des méthodes de recherche d'information (RI) peuvent partiellement contrecarrer ce problème en récupérant du contexte pertinent dans le document. Malheureusement, le manque de jeux de données de recherche de contexte adaptés à la tâche de REN ne permet pas d'entraîner un modèle de recherche supervisé.

Afin de résoudre ce problème, nous proposons dans l'article *"Learning to Rank Context for Named Entity Recognition Using a Synthetic Dataset"* [4], que nous résumons ici, de générer un jeu de données de recherche de contexte à l'aide du grand modèle de langue Alpaca [5]. Grâce à ce jeu de données, nous entraînons ensuite un modèle de reclassement de contexte adapté à la tâche de REN. Nous présentons ci-dessous la méthode de génération de ce jeu de données synthétique, ainsi que l'amélioration apportée par notre modèle de reclassement entraîné sur un tel corpus pour la tâche de REN, notamment en le comparant avec des modèles de recherche d'information non supervisés et supervisés existants.

2. Méthode

Nous réalisons toutes nos expériences sur un jeu de données littéraire comportant 40 chapitres de romans en anglais, initialement proposé par Dekker et al. [6] puis amélioré par nos soins [3]. Ce jeu de données contient trois classes d'entités : *personnes* (PER), *lieux* (LOC) et *organisations* (ORG).

. *Coria 2024*
. arthur.amalvy@univ-avignon.fr (A. Amalvy); vincent.labatut@univ-avignon.fr (V. Labatut); richard.dufour@univ-nantes.fr (R. Dufour)
. 0000-0003-4629-0923 (A. Amalvy); 0000-0002-2619-2835 (V. Labatut); 0000-0002-9421-8566 (R. Dufour)

2.1. Génération du jeu de données

Nous formulons la tâche de recherche de contexte comme une tâche de classification binaire sur une paire "*texte d'entrée / contexte*" : l'exemple est positif si le contexte est pertinent pour le texte d'entrée sur lequel réaliser la prédiction, négatif sinon.

Afin de générer des exemples positifs, nous supposons qu'un contexte pertinent pour la tâche de REN contient des informations permettant de préciser la classe des entités du texte d'entrée. Comme ces informations dépendent de la classe des entités, nous déterminons par observation pour chacune de ces classes un certain nombre de types de phrases permettant d'améliorer la prédiction. Pour les *personnes*, nous considérons par exemple que les phrases décrivant ces personnes ou les dépeignant en train d'effectuer des actions sont pertinentes. Nous utilisons le grand modèle de langue Alpaca [5] afin de générer des phrases de ces types pour des passages donnés de notre jeu de données d'entraînement de REN.

Générer des exemples négatifs est plus simple. Nous utilisons deux méthodes différentes, le *prélèvement négatif* et l'*échange d'exemples positifs*. Le *prélèvement négatif* consiste à récupérer au hasard une phrase de contexte dans un roman différent du texte d'entrée. Pour ce qui est de l'*échange d'exemples positifs*, nous utilisons une phrase de contexte provenant d'un contexte positif généré et l'appairons avec un texte d'entrée contenant une entité différente.

2.2. Reconnaissance d'entités nommées

Pour la reconnaissance d'entités nommées, nous utilisons un modèle BERT-base [1] suivi d'une tête de classification. Nous entraînons notre modèle de recherche de contexte, également basé sur BERT-base, sur le jeu de données synthétique décrit plus haut. À l'inférence, nous récupérons d'abord un certain nombre de contextes candidats à l'aide de plusieurs heuristiques non supervisées comme BM25 [7], puis nous utilisons notre modèle pour garder les contextes les plus pertinents parmi ceux-ci. Nous les concaténons ensuite à l'entrée.

3. Résultats

Nous entraînons un seul modèle de REN, et étudions les résultats sur cette tâche en utilisant différentes méthodes de recherche de contexte. Nous comparons notre modèle de recherche à des heuristiques non supervisées comme BM25 [7], ainsi qu'à des modèles supervisés existants (MonoBERT [8] et MonoT5 [9]) entraînés sur le jeu de données MSMarco [10].

Nos résultats sur la REN montrent que notre modèle surpasse les heuristiques non supervisées, et permet de gagner en moyenne environ un point de micro F1 sur notre corpus d'évaluation. Cette amélioration dépend fortement du roman étudié, et peut aller jusqu'à 8,1 points de micro F1. En outre, la comparaison avec les modèles supervisés existants montre que notre modèle de reclassement est capable d'égaler les modèles MonoBERT et MonoT5, alors même que ceux-ci sont entraînés sur des jeux de données annotés manuellement et que leur nombre de paramètres surpasse celui de notre modèle.

4. Conclusion

Nous avons proposé de générer un jeu de données synthétique à l'aide d'un grand modèle de langue afin d'entraîner un modèle de reclassement de contexte pour la REN [4]. Ce modèle permet d'augmenter la performance d'un modèle de REN qui doit traiter un long texte en le découpant par blocs.

Notre méthode de génération d'exemples de contextes positifs demande cependant de faire des suppositions sur les exemples utiles à la tâche, suppositions qui sont intrinsèquement liées à la tâche que nous traitons (la REN). Or, la REN n'est pas la seule tâche qui pâtit des problèmes de portée des modèles transformeurs. De futurs travaux pourraient donc s'intéresser à la généralisation de notre méthode de génération, soit en proposant des méthodes de génération d'exemples de contexte pour d'autres tâches, ou en proposant une méthode s'affranchissant des spécificités de la tâche traitée.

Références

[1] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT : Pre-training of deep bidirectional transformers for language understanding, in : Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, volume 1, 2019, pp. 4171–4186. doi :10.18653/v1/N19-1423.

[2] I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Matsumoto, LUKE : Deep contextualized entity representations with entity-aware self-attention, in : Conference on Empirical Methods in Natural Language Processing, 2020, pp. 6442–6454. doi :10.18653/v1/2020.emnlp-main.523.

[3] A. Amalvy, V. Labatut, R. Dufour, The role of global and local context in named entity recognition, in : 61st Annual Meeting of the Association for Computational Linguistics, 2023. doi :10.18653/v1/2023.acl-short.62.

[4] A. Amalvy, V. Labatut, R. Dufour, Learning to rank context for named entity recognition using a synthetic dataset, in : 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 10372–10382. doi :10.18653/v1/2023.emnlp-main.642.

[5] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca : An instruction-following llama model, 2023. URL : https://github.com/tatsu-lab/stanford_alpaca.

[6] N. Dekker, T. Kuhn, M. van Erp, Evaluating named entity recognition tools for extracting social networks from novels, PeerJ Computer Science 5 (2019) e189. doi :10.7717/peerj-cs.189.

[7] S. E. W. Robertson, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in : B. W. Croft, C. J. van Rijsbergen (Eds.), SIGIR '94, 1994, pp. 232–241.

[8] R. Nogueira, K. Cho, Passage re-ranking with bert, arXiv cs.IR (2020) 1901.04085. URL : https://arxiv.org/abs/1901.04085.

[9] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document ranking with a pretrained sequence-to-sequence model, in : Findings of the Association for Computational Linguistics : EMNLP 2020, 2020, pp. 708–718. doi :10.18653/v1/2020.findings-emnlp.63.

[10] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, Ms marco : A human generated machine reading comprehension dataset, arXiv cs.CL (2018) 1611.09268. URL : https://arxiv.org/abs/1611.09268.