LABORATOIRE DES SCIENCES DU NUMÉRIQUE DE NANTES

LABORATOIRE INFORMATIQUE D'AVIGNON

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

Arthur Amalvy, Vincent Labatut and Richard Dufour

April 2, 2024

AVIGNON UNIVERSITÉ

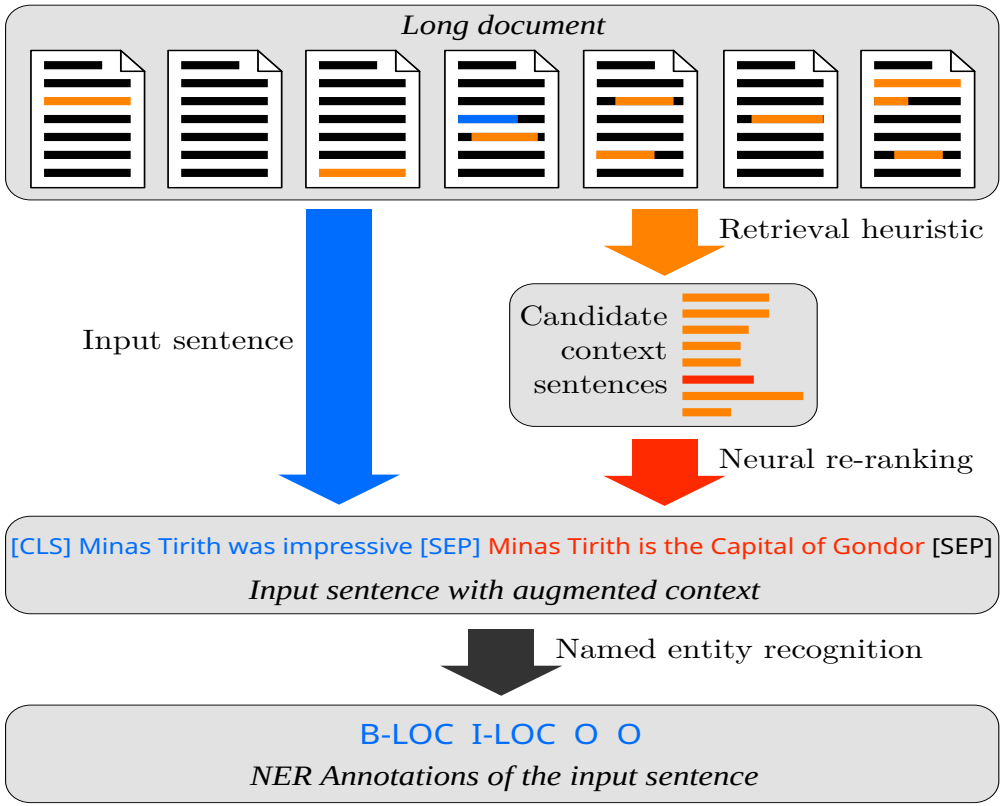
Transformers et longs documents

- Complexité quadratique du mécanisme d'attention : difficulté à traiter de longs documents
  - Sparse transformers
  - Connaissances externes
- Documents traités avec une fenêtre glissante
  - De l'information est perdue au moment de la prédiction

context window

It was a sunny morning.
Raoden stood, and as he did, his eyes fell on **Elantris** again.
He was impressed.
[...]
Elantris is the capital of the Empire.

REN et récupération de contexte



- Contexte à longue portée bénéfique pour la performance [ALD23]
- Comment entraîner un modèle supervisé sans données ?

Jeu de données de REN

- Jeu de données littéraire en anglais annoté par [DKE19] puis amélioré par [ALD23]
- Le premier chapitre de 40 livres
- Documents "suffisamment longs"
- Disponible sur Github !
  `https://github.com/CompNet/conivel`

Générer un jeu de données synthétique

- Comment entraîner un modèle supervisé sans jeu de données ?
- Génération d'un jeu de données en utilisant le grand modèle de langue Alpaca [Tao+23]

Chaque exemple du jeu de données a la forme suivante :

(phrase d'entrée, phrase de contexte, pertinence)

La pertinence vaut 0 ou 1 en fonction du contexte.

Générer des exemples positifs

- Nous déterminons empiriquement quels types de phrases de contexte sont utiles
- Nous écrivons une instruction (prompt) pour chacun d'entre eux

Type d'entité	Contexte supposés utiles
PER	description, action
LOC	description, mouvement vers
ORG	description

Générer des exemples positifs : un exemple

Phrase d'entrée

*"**One-Eye**'s handicap in no way impairs his marvelous insight"*

Phrase de contexte pertinente générée (description)

*"**One-Eye** is a wise and mysterious character with a penchant for coming up with invaluable insights after the fact"*

Pertinence: 1

Générer des exemples négatifs : échantillonage négatif

Phrase d'entrée

"I am afraid that I have been tempted into too great length about the ***Italian Catherine****; but in truth she has been my favourite."*

Phrase de contexte non pertinente prélevée

"said Alice, as she swarm about, trying to find her way out."

Pertinence: 0

Générer des exemples positifs : échange d'exemple positif

Phrase d'entrée

"We left in pretty good time and came after nightfall to ***Klausenburgh****."*

Exemple positif échangé

"Forley was an adventurous and daring individual who was never afraid to take risks."

Pertinence: 0

Expérience : Objectifs

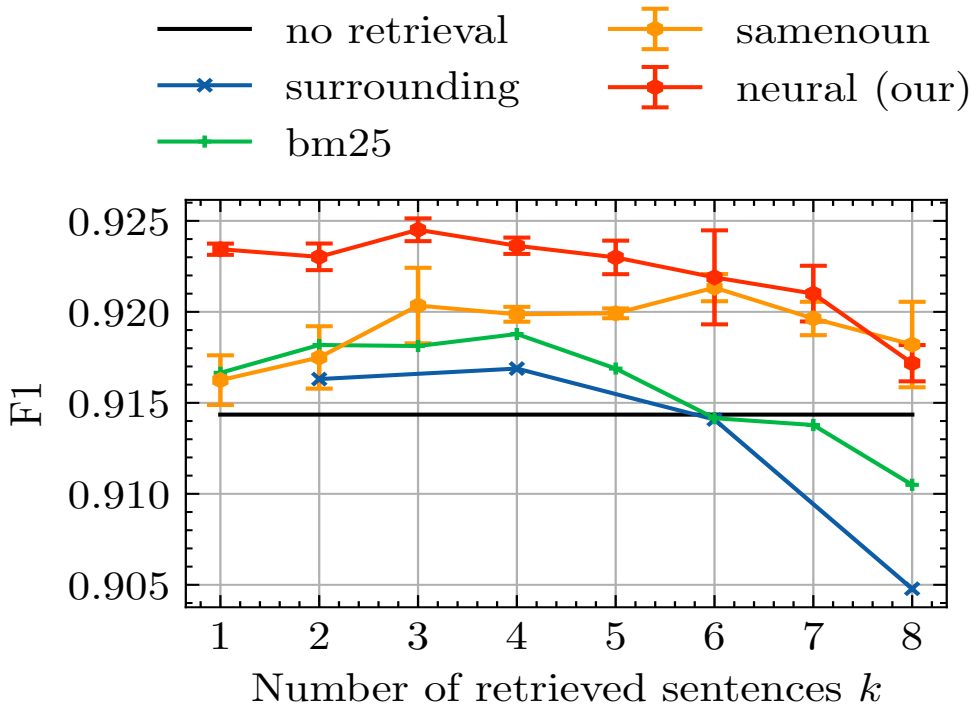
Questions de recherche :

- Est-ce que notre modèle de reclassement peut améliorer les performances d'un modèle de REN ?
- Combien de phrases de contexte est-il utile de récupérer ?
- Est-ce que la taille du grand modèle de langue a une importance ?

Expérience : Protocole

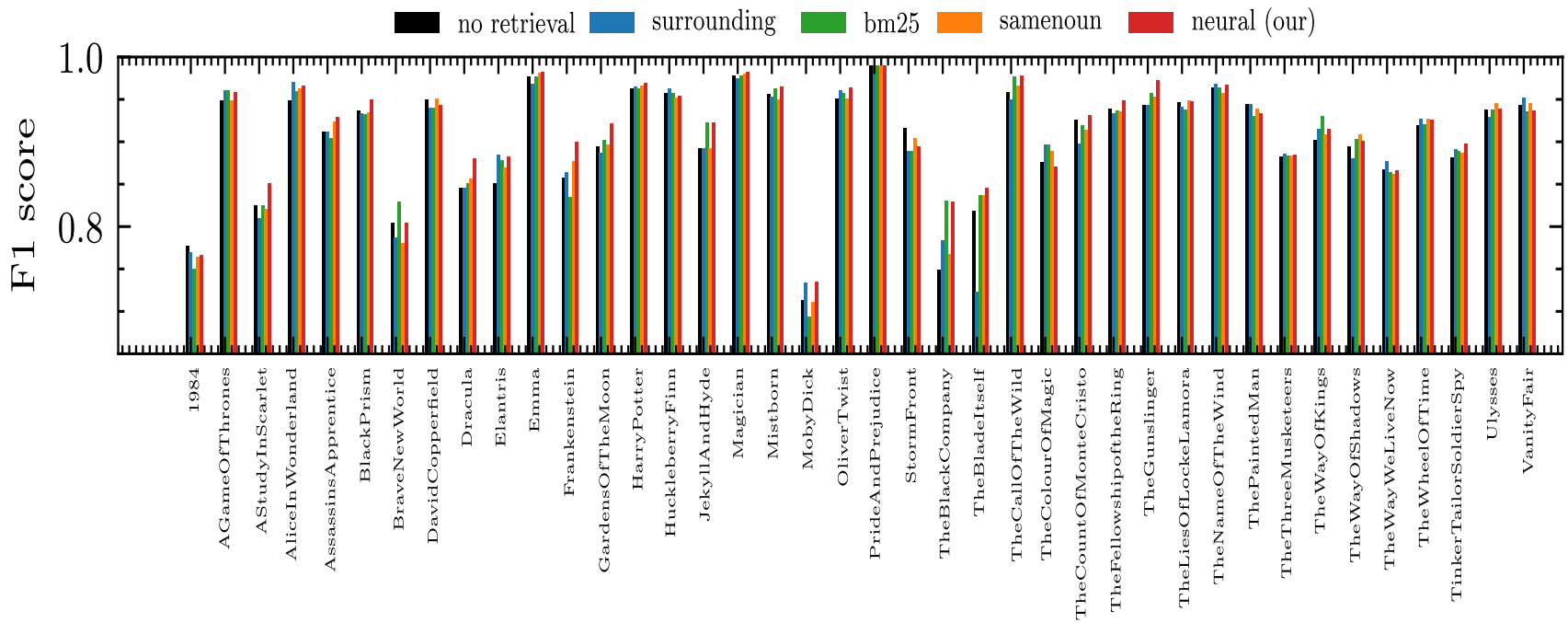
1. Génération d'un jeu de donnée de synthétique de récupération de contexte (~2700 exemples)
2. Entraînement d'un modèle de reclassement basé sur BERT [Dev+19] sur ce jeu de données
3. Entraînement d'un modèle BERT pour la REN sur notre jeu de données de REN
4. Au moment de l'inférence, récupération de contexte grâce à notre modèle de reclassement

Comparaison avec des heuristiques non supervisées



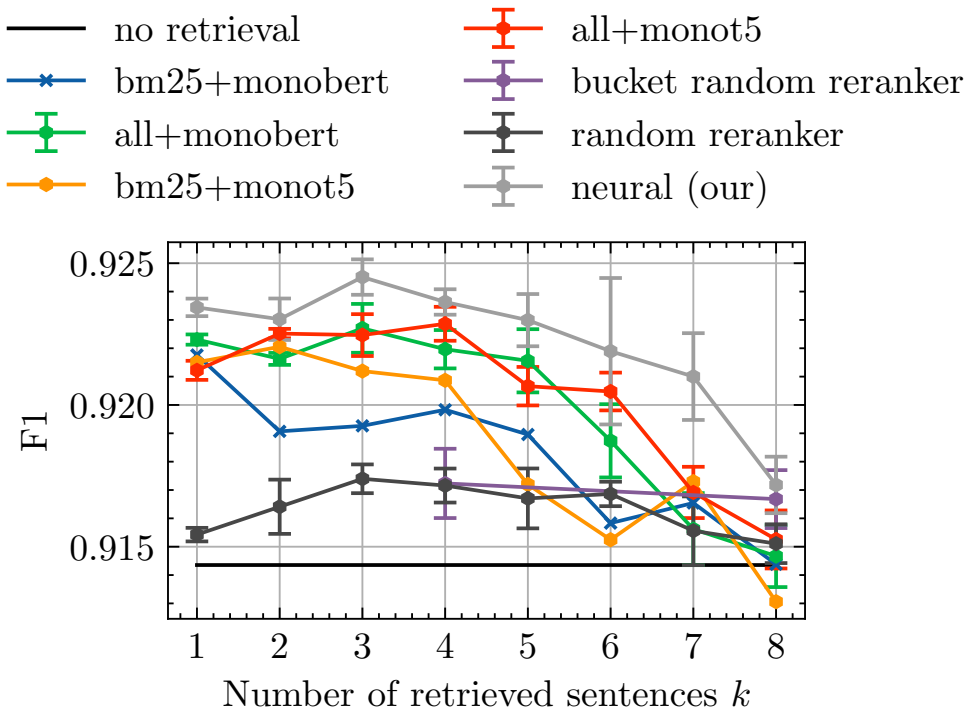
- Notre modèle de reclassement est meilleur que les heuristiques non supervisées
- Le pic de performance est obtenu pour $k = 3$

Résultats par livre



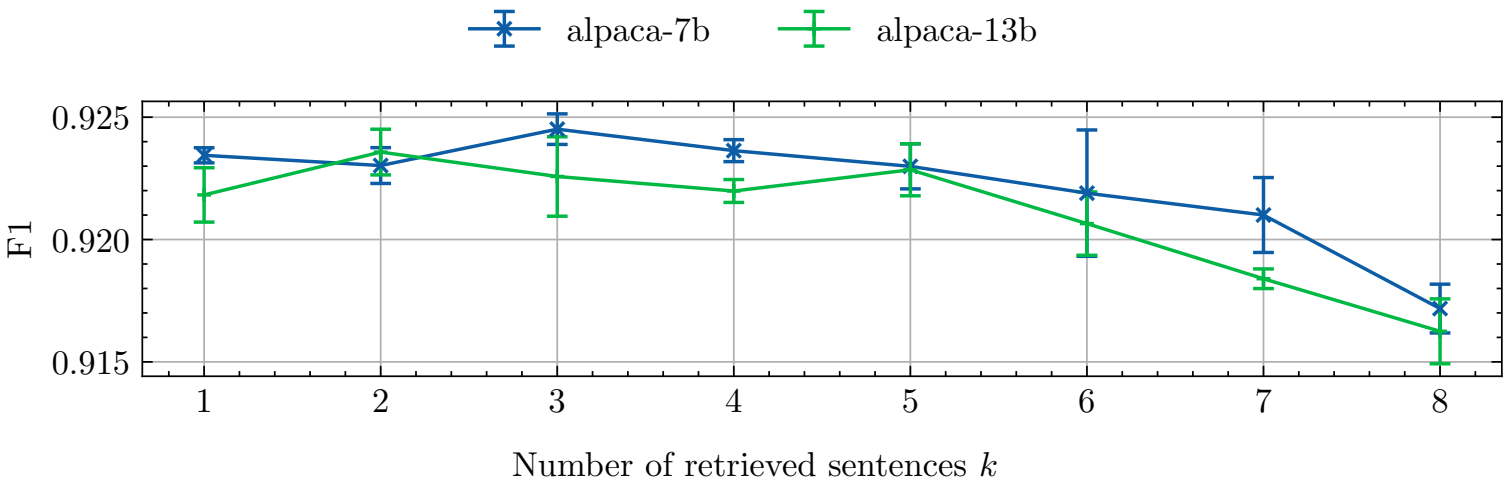
- Les performances varient fortement selon les livres
- Notre méthode de reclassement est la meilleure dans 20 livres sur 40

Comparaison avec des modèles de reclassement existants



- Notre modèle de reclassement est meilleur ou équivalent à des modèles de reclassement existants (MonoBERT [NC20], MonoT5 [Nog+20])

Influence de la taille du modèle de langue



Un modèle de langue avec plus de paramètres permet il d'augmenter les performances ?

- Quantitativement, pas de gain
- Qualitativement, pas de différence observée dans les jeux données

Conclusion

- Nous avons entraîné un modèle de reclassement neuronal pour la REN en utilisant uniquement des données synthétiques
- Pour la REN, la performance de ce modèle est meilleure ou équivalente à des modèles supervisés entraînés sur des jeu de données annotés manuellement
- Augmenter la taille du modèle de langue n'est pas bénéfique

Comment appliquer notre technique à d'autres tâches ?

Références

[ALD23] A. Amalvy, V. Labatut **and** 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.

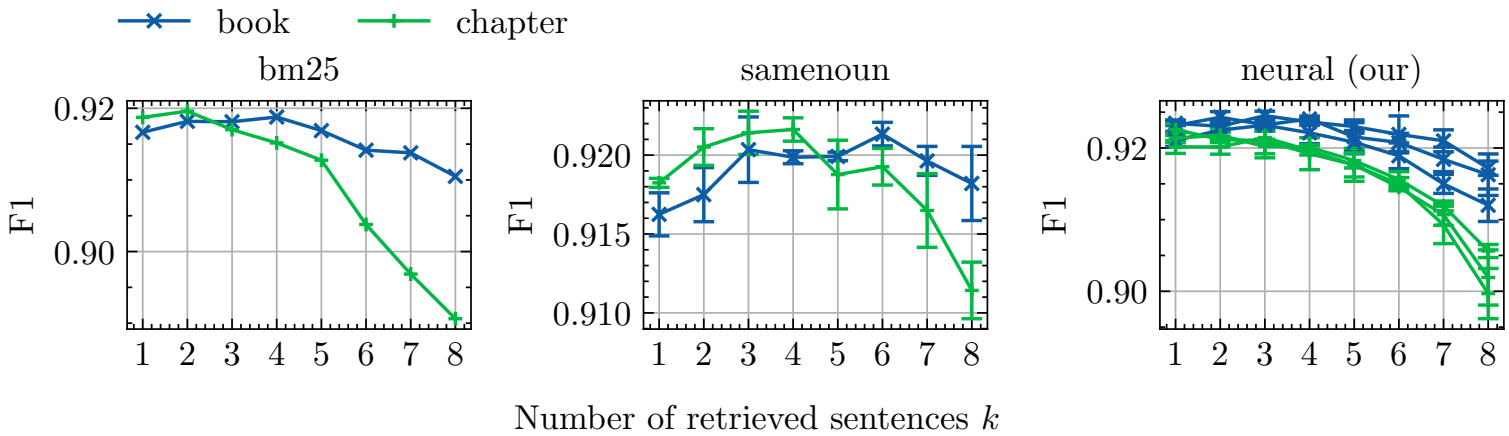
[DKE19] N. Dekker, T. Kuhn **and** M. van Erp. **?**Evaluating named entity recognition tools for extracting social networks from novels**?** **in***PeerJ Computer Science*: 5 (2019), e189. DOI: 10.7717/peerj-cs.189.

[Dev+19] J. Devlin, M. Chang, K. Lee **and** 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, **pages** 4171–4186. DOI: 10.18653/v1/N19-1423.

[NC20] R. Nogueira **and** K. Cho. **?**Passage Re-ranking with BERT**?** **in***arXiv*: cs.IR (2020), **page** 1901.04085. URL: https://arxiv.org/abs/1901.04085.

[Nog+20] R. Nogueira, Z. Jiang, R. Pradeep **and** J. Lin. **?**Document Ranking with a Pretrained Sequence-to-Sequence Model**?** **in***Findings of the Association for Computational Linguistics:*

Annexe 1 : Influence de la fenêtre de contexte



Annexe 2 : Prompts

Description (all classes)	**Prompt template** `'{INPUT SENTENCE}' - In the preceding sentence, {ENTITY} is a character. Invent a one-sentence description for this character, mentioning their name.` **Example input sentence** *[One-Eye]'s handicap in no way impairs his marvelous hindsight.* **Example generated context** *One-Eye is a wise and mysterious character with a penchant for coming up with invaluable insights after the fact.*
Action (PER only)	**Prompt template** `Invent a single sentence depicting the character '{ENTITY}' performing an action, mentioning their name.` **Example input sentence** *"It's my stomach, [Croaker],"* **Example generated context** *Croaker was whistling a jaunty tune as he strolled through the Park.*
Movement (LOC only)	**Prompt template** `Invent a single sentence depicting a character of your invention going to {ENTITY}. You must mention the name of the character.` **Example input sentence** *Lightning from a clear sky smote the [Necropolitan Hill].* **Example generated context** *The gothic vampire Count Necropolis ventured to Necropolitan Hill, his ancient stomping grounds.*

Table 1: Prompts templates and examples of positive context retrieval samples generated by Alpaca-7b.