Évaluation de modèles de pondération pour la recherche d'information sur TREC AP 88-90

Cours DIC-9345 TALN Présenté à Ngoc Tan Lê

Dominique S. Loyer

Département d'informatique, UQAM

10 avril 2025

Plan

Introduction

- **Recherche d'Information (RI) :** Trouver l'information pertinente dans la vaste collection TREC AP 88-90.
- **Défi :** 243k documents compressés individuellement en GunZip->.gz dans un fichier->.tar les englobants. Traitement séquentiel long vs. en parralèle. Le type d'encodage du fichier XML n'est pas pris en charge par Python -> conversion en .JSON
- **Objectif Principal :** Implémenter et évaluer un système RI monolingue (Anglais) sur la collection.
- **Pertinence Cours :** Application pratique des concepts de TALN vus dans le cours DIC-9345 tels que : (modèles RI, indexation, évaluation, amélioration).

Contexte et Motivations

- Comprendre l'impact de choix fondamentaux en RI :
  - Prétraitement : Aucun (Baseline) vs Racination (Racination Porter). *(Lemmatisation aussi testée mais très peu performante)*.
  - Longueur Requêtes : Courtes (‘<title>‘) vs Longues (‘<title>‘+‘<desc>‘).
  - Modèles Pondération : BM25 vs QLD (Query Likelihood). *(TF-IDF abandonné pour instabilité)*. [1]
- Évaluer une technique d'amélioration : Pseudo-Retour de Pertinence (RM3).
- Utiliser de ces outils technologiques : Pyserini/Lucene, Python, Colab.
- Processus expérimental TREC.

1. Moins utilisé aujourd'hui pour la recherche au profit de modèles comme BM25

État de l'Art (Concepts Clés)

Modèles de Pondération :

- TF-IDF : Classique Salton and Buckley (1988). [a]
- BM25 : Probabiliste robuste Robertson et al. (1995).
- QLD : Modèle de langue + lissage Manning et al. (2008).

État de l'Art (Concepts Clés)

Modèles de Pondération :

- TF-IDF : Classique Salton and Buckley (1988). [a]
- BM25 : Probabiliste robuste Robertson et al. (1995).
- QLD : Modèle de langue + lissage Manning et al. (2008).

Prétraitement :

- Normalisation (tokens, tout en min., mots vides).
- Racination vs Lemmatisation : Compromis Précision/Rappel Manning et al. (2008).

Amélioration Requête :

- PRF (RM3) : Expansion de requête via top docs Abdul-Jaleel et al. (2004).
- Autres : WordNet Miller (1995), Embeddings Mikolov et al. (2013), Grands modèles de langues Nogueira and Lin (2019); Ma and Liu (2023).

a. Difficulté à faire fonctionner ClassicSimilarity (TF-IDF Lucene) avec Colab/Java 21/Pyserini.

État de l'Art (Concepts Clés)

Modèles de Pondération :

- TF-IDF : Classique Salton and Buckley (1988). [a]
- BM25 : Probabiliste robuste Robertson et al. (1995).
- QLD : Modèle de langue + lissage Manning et al. (2008).

Prétraitement :

- Normalisation (tokens, tout en min., mots vides).
- Racination vs Lemmatisation : Compromis Précision/Rappel Manning et al. (2008).

Amélioration Requête :

- PRF (RM3) : Expansion de requête via top docs Abdul-Jaleel et al. (2004).
- Autres : WordNet Miller (1995), Embeddings Mikolov et al. (2013), Grands modèles de langues Nogueira and Lin (2019); Ma and Liu (2023).

Outils/Collections :

- TREC AP 88-90 : Réf. Har (1995).
- Pyserini : Toolkit Python/Lucene Lin et al. (2021).

a. Difficulté à faire fonctionner ClassicSimilarity (TF-IDF Lucene) avec Colab/Java 21/Pyserini.

Pipeline Général

1. **Corpus :** TREC AP 88-90 (Extraction + Décompression Gzip) $\rightarrow$ JSONL.
2. **Prétraitement :** Baseline (Lucene std) vs Racination (NLTK Porter).
3. **Indexation :** 2 Index Lucene (Baseline 825M, Stemmed 1.6G) via Pyserini.
4. **Recherche :** Req. Courtes/Longues ; Modèles BM25/QLD ; Top 1000 $\rightarrow$ 8 runs base.
5. **Amélioration :** RM3 sur meilleur run base (‘baseline long bm25‘) $\rightarrow$ 1 run RM3.
6. **Évaluation :** ‘pytrec eval‘ (MAP, P@10) sur 51 req. jugées.

Données et Outils

- **Collection :** TREC AP 88-90 (242k docs).
- **Requêtes :** 150 Topics TREC (Courtes/Longues).
- **Jugements :** Qrels TREC AP (51 requêtes jugées utilisées).
- **Outils :** Python 3.11, Colab, Pyserini [V.2021], Lucene, Java 21, NLTK (PorterStemmer), Pytrec_eval, Pandas.

Métriques d'Évaluation

- **MAP@1000 :** Qualité globale du classement (Précision + Rappel).
- **P@10 :** Pertinence des 10 premiers résultats (Précision initiale).

Résultats MAP@1000 (Baseline vs Racination)

Requête	Poids	MAP @ 1000	
		Baseline	**Preprocessed (Stem)**
Courte	BM25	0.1677	0.1296
	QLD	0.1481	0.1148
Longue	BM25	**0.2205**	0.1778
	QLD	0.2171	0.1739

Table – Comparaison du MAP (Racination vs Baseline).

Constats : Baseline > Racination ; Longues > Courtes ; BM25 > QLD.

Résultats P@10 (Baseline vs Racination)

Requête	Poids	P@10	
		Baseline	**Preprocessed (Stem)**
Courte	BM25	0.3490	0.2863
	QLD	0.3235	0.2529
Longue	BM25	0.4765	0.3961
	QLD	**0.4804**	0.4059

Table – Comparaison du P@10 (Racination vs Baseline).

Constats : Baseline > Racination ; Longues > Courtes ; QLD $\approx$ BM25 (Long).

Analyse : Base vs Racination

- **Prétraitement (Racination) Négatif :** La racination Porter réduit les performances (MAP et P@10) vs Baseline. Moins pire que lemmatisation, mais contre-productif ici.
  - Hypothèse : Trop agressif ? Perte de nuances ?
- **Longueur Requête Positive :** Requêtes longues $\gg$ Requêtes courtes.
- **Modèle BM25 Robuste :** Meilleur MAP global (0.2205).
- **Meilleure Base :** ‘baseline long bm25‘.

Résultats : Amélioration RM3 (sur Baseline)

RM3 appliqué sur ‘baseline long bm25‘.

Configuration	MAP	P@10
Baseline + Long + BM25 (Base)	0.2205	0.4765
Baseline + Long + BM25 + RM3	**0.2948**	**0.5078**
Amélioration Relative	+33.7%	+6.6%

Table – Impact de RM3 sur la meilleure configuration baseline.

- Gain MAP très important (+34%).
- Gain P@10 notable (+6.6%). [2]
- MAP final 0.295 : niveau respectable.

2. Bon, car parfois le PRF peut introduire du bruit qui nuit aux tout premiers résultats

Limitations

- Performances absolues finales (0.30 MAP) correctes mais améliorables.
- Prétraitement personnalisé (Racination/Lemma) inefficace ici.
- Pas d'optimisation des paramètres (BM25, RM3).
- Une seule méthode d'amélioration (RM3) testée.
- Analyse qualitative des erreurs à approfondir.

Conclusion

- Pipeline RI implémenté avec Pyserini sur TREC AP.
- Meilleure approche : Baseline + Req. Longues + BM25 + RM3 (MAP final 0.2948).
- Prétraitement personnalisé (racination/lemmatisation) a nui aux performances.
- RM3 a montré un gain très significatif sur la meilleure base.
  - 34/100 montre que l'ajout de termes issus des documents pseudo-pertinents a permis de reformuler les requêtes de manière beaucoup plus efficace pour trouver les documents pertinents sur l'ensemble de la liste classée

Travaux Futurs

- Analyser/Corriger le Prétraitement (autre stemmer comme Snowball ? Une liste de mots vides ?).
- Optimiser les Paramètres (BM25, RM3).
- Autres Améliorations (Embeddings->Word2Vec, Re-ranking transformer avec BERT).
- Analyse d'Erreurs Approfondie.
- Intégration potentielle des Entités Nommées.
- Utiliser BabelNet au lieu de WordNet

Références I

1995. Overview of the third text retrieval conference (trec-3). In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, Gaithersburg, USA. NIST Special Publication 500-225. Contexte pour la collection AP.

Naveen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Díaz, Leah S. Larkey, Xiaoyan Li, Donald Metzler, Mark D. Smucker, Trevor Strohman, Howard Turtle, and Courtney Vantonder. 2004. UMass at TREC 2004 : Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*. NIST. Contextualise les modèles de pertinence utilisés dans TREC, dont RM3 est un exemple.

Jimmy Lin, Xueguang Ma, Craig Macdonald, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, pages 2356–2360.

Guangyuan Ma and Pengfei Liu. 2023. Query generation for information retrieval using large language models: A survey. *arXiv preprint arXiv :2310.11496*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11) :39–41.

Rodrigo Nogueira and Jimmy Lin. 2019. Passage re-ranking with bert. *arXiv preprint arXiv :1901.04085*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 109–126, Gaithersburg, USA. NIST Special Publication 500-225.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5) :513–523.

Références II

ACL Anthology Organization. ACL anthology GitHub repository. https ://github.com/acl-org/acl-anthology/ 2025.
ACM Special Interest Group on Information Retrieval (SIGIR). ACM
SIGIR conference on research and development in information retrieval. https ://sigir.org 2025.
Anusha. Natural language processing (NLP) with python – tutorial for beginners.
https ://pub.towardsai.net/natural-language-processing-nlp-with-python-tutorial-for-beginners-1f54e610a1a0 2021.
Association for Computational Linguistics. ACL anthology. https ://aclanthology.org 2025.
DataCamp. Stemming and lemmatization in python. https ://www.datacamp.com/tutorial/stemming-lemmatization-python 2023.
Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2018). BERT : Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv :1810.04805.
Jurafsky, D. et Martin, J. H. (2023). Speech and language processing (3rd (Draft)). Online.
https ://web.stanford.edu/ jurafsky/slp3/
Lin, J., Ma, X., Macdonald, C. et Nogueira, R. (2021). Pyserini : a python toolkit for reproducible information retrieval research with sparse and dense representations. Dans Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval (SIGIR '21) (p. 2356 2360). https ://doi.org/10.1145/3404835.3463238
Cambridge University Press. https ://nlp.stanford.edu/IR-book/ National Institute of Standards and Technology (NIST). Text retrieval conference (TREC). https ://trec.nist.gov 2025.
Nie, J.-Y. Information retrieval : Implementing and evaluating search engines (online material).
https ://www.iro.umontreal.ca/ nie/IR-book/Preface.html 2010.
Nogueira, R. et Lin, J. (2019). Passage re-ranking with BERT. arXiv preprint arXiv :1901.04085. https ://arxiv.org/abs/1901.04085
Stanford University. CS224n : Natural language processing with deep learning - syllabus.
https ://web.stanford.edu/class/archive/cs/cs224n/cs224n.1162/syllabus.shtml 2016.

Discussion

Merci de votre attention !