Connecter les BDDs aux LLMs

Pourquoi le RAG permet de sourcer les affirmations d'un chatbot

Luc Pommeret

CRED - Paris XIV^{e}

3 juin 2025

Contexte de recherche

Problématique : Comment réconcilier les paradigmes de recherche d'information classiques avec les nouvelles approches neuronales ?

Information Retrieval Classique

Index inversé
TF-IDF, BM25

Gap sémantique

Approches Neuronales

Embeddings denses
Transformers

Hypothèse : Les architectures RAG permettent de combiner les avantages des deux approches.

RAG : Un nouveau paradigme

Retrieval-Augmented Generation : évolution majeure des systèmes de question-réponse.

Requête	**Retrieval**	**Augment.**	**Génération**
Langage naturel	Recherche vectorielle	Contexte enrichi	Réponse synthétique

Innovation : Précision de la recherche vectorielle + capacité de synthèse des LLMs.
Applications : Question-answering, résumé automatique, exploration de corpus.

Architecture expérimentale

Corpus de documents (Web scraping)

↓

Pipeline : Chunking + Embeddings

↓

Base vectorielle : PostgreSQL + pgvector

↓

Modèle de langage : Ollama + Llama3

Méthodologie

Domaine : Bibliothèques Paris-Saclay

Corpus : Pages web des bibliothèques universitaires

Pipeline expérimental :

1. Extraction et normalisation via scraping web
2. Segmentation en chunks + génération d'embeddings
3. Indexation vectorielle et optimisation
4. Évaluation qualitative et quantitative

Démonstration technique

Pipeline développé :

1. Collecte de données

```
python -m qa_haystack.src.run scrape \
  --output library_data.json --max-pages 50
```

2. Indexation vectorielle

```
python -m qa_haystack.src.run ingest \
  --input library_data.json
```

3. Requêtage intelligent

```
python -m qa_haystack.src.run query \
  --model llama3:8b --top-k 5
```

Résultat : Requête en langage naturel $\rightarrow$ réponse structurée **avec sources**.

Comparaison paradigmatique

Recherche traditionnelle

Requête par mots-clés exacts.

Correspondance lexicale stricte.

L'utilisateur s'adapte au vocabulaire du système.

Pas de contextualisation.

Approche RAG

Requête en langage naturel.

Similarité vectorielle sémantique.

Le système s'adapte à l'utilisateur.

Réponses synthétiques avec justifications.

Contribution : Système hybride alliant robustesse des BDD et flexibilité des modèles neuronaux.

Défis de recherche

Challenges techniques et scientifiques :

Scalabilité : Optimisation sur gros corpus. Solutions : indexation hiérarchique.

Qualité sémantique : Cohérence lors des mises à jour. Versioning des embeddings.

Hallucinations : Réduction des erreurs factuelles. Validation par retrieval.

Évaluation : Métriques adaptées. Mesures automatiques + évaluation humaine.

Biais : Analyse et mitigation dans modèles et corpus.

Résultats et perspectives

Observations expérimentales :

Amélioration significative de la satisfaction utilisateur.

Réduction des itérations : de 3-4 requêtes à 1-2 en moyenne.

Capacité de synthèse d'informations dispersées.

Perspectives :

Extension à d'autres domaines documentaires.

Intégration d'apprentissage par renforcement.

Architectures multimodales (texte + images + métadonnées).

Contributions scientifiques

Apports de cette recherche :

Théorique : Complémentarité recherche symbolique et neuronale.

Méthodologique : Architecture reproductible et open-source.

Pratique : Validation sur cas d'usage réel (Bibliothèques de Paris-Saclay).

Applications futures :

Aide à la recherche scientifique et découverte de littérature.

Systèmes de recommandation contextuels.

Interfaces conversationnelles pour grandes bases de connaissances.

Discussion et Questions

LISN - Laboratoire Interdisciplinaire des Sciences du Numérique
Équipe Langue Interaction Parole et Signes