Soutenance de projet de fin d'études

La réponse aux questions à partir des rapports sur le prix et la qualité du service (RPQS)

Préparée et présentée par M. Youssef ASSIS

Membres du Jury

M. Rachid BENMANSOUR
Mme. Kaoutar EL HARI
M. Abdeslam KADRANI

Sous la direction de

M. Abdeslam KADRANI
M. Amir NAFI
M. Ahmed SAMET

Année universitaire : 2019 - 2020

Plan de présentation

Introduction

Contexte du projet

- Laboratoire **ICube**[1] à Strasbourg.
- Équipe de la **science des données et de connaissances (SDC)**
  - Recherche en intelligence artificielle.
- Équipe de la **conception, système d'information et processus inventifs (CSIP)**
  - Étude, la compréhension et le développement de nouveaux produits et systèmes.

1. Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (ICube)

Introduction

Problématique

- Les services d'eau et d'assainissement en France archivent leurs activités sous forme de documents appelés rapports RPQS [2].
- Les rapports RPQS facilitent l'accès à l'information en exploitant un certain nombre d'indicateurs de performance :
  - **Techniques** : ouvrages, conditions sanitaires de l'eau, qualité de l'eau, composants de l'eau, etc.
  - **Financiers** : tarifs de l'eau, investissements, dettes, remboursements, etc.

2. Rapport annuel sur le prix et la qualité des services (RPQS)

Introduction

Problématique

- Processus d'extraction de données à partir des rapports RPQS est :
  - Effectué par des spécialistes du domaine de l'eau de d'assainissement.
  - Manuel, chronophage et très gourmand en énergie.
  - Nécessite l'attribution d'un budget de la part de la direction des services.
  - Résultats ne sont pas toujours satisfaisants.

Introduction

Solution proposée

Système de questions-réponses

- Extraire les réponses pertinentes aux questions posées en français.
- Application aux rapports RPQS.
- Adaptation aux spécificités des connaissances contenues dans ces rapports.

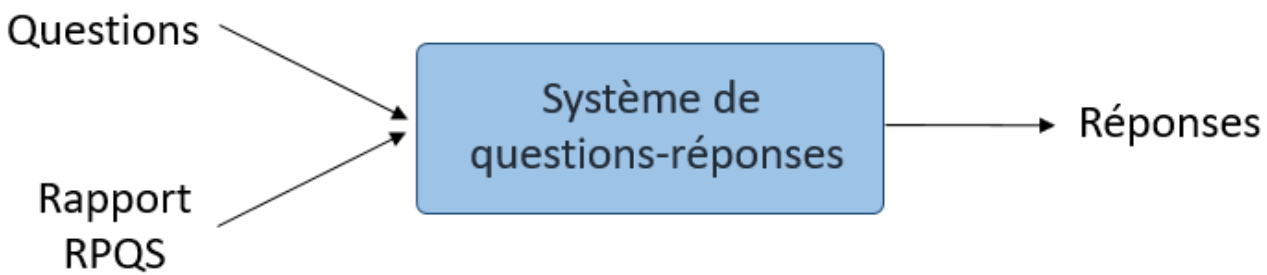


FIGURE – Système de questions-réponses

Introduction

Complexités

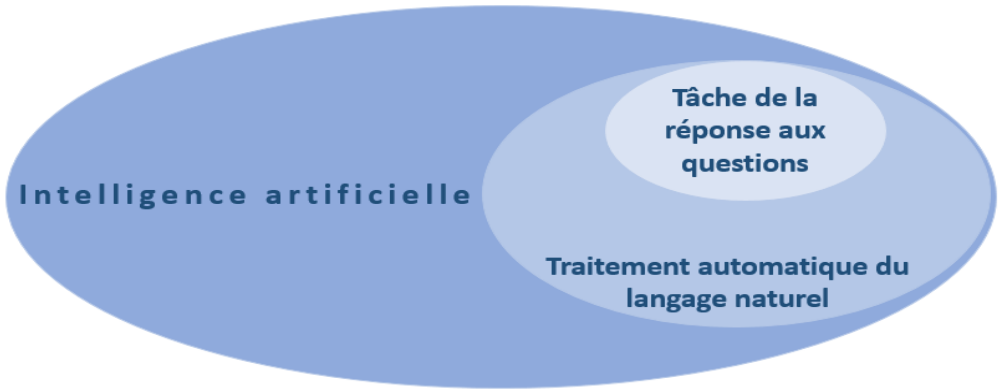
Rapports RPQS

- Rapports très longs, rédigés en français.
- Variété et complexité des structures de données contenues dans ces rapports : paragraphes, tables, images, schémas, etc.
- Diversité de formats des rapports : pdf et word.

État de l'art

Tâche de la réponse aux questions

- Adopte les techniques utilisées dans le domaine du traitement automatique du langage naturel, pour comprendre le langage naturel.
- Tente de répondre à des questions posées en langage naturel, par une recherche dans un ensemble de documents d'entrée suivie d'une extraction de réponses.



État de l'art

Approches existantes

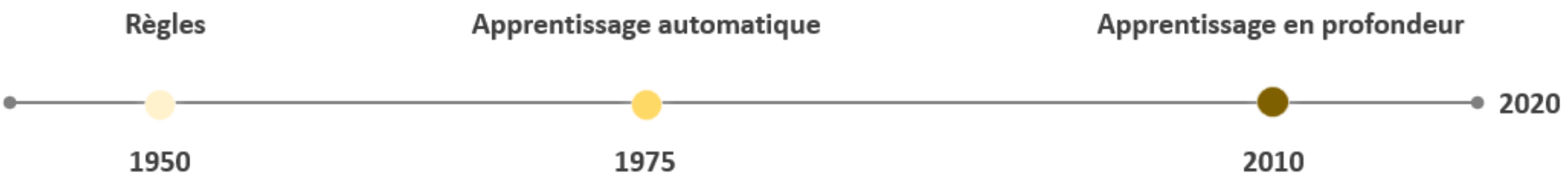


FIGURE – Approches adoptées dans les systèmes de questions-réponses

Approches basées sur les règles

- Fondées sur des représentations logiques d'arbres de décisions :
  - **Racines** représentent les questions ;
  - **Feuilles** représentent les réponses ;
  - **Arcs** sont des règles grammaticales qui définissent les chemins d'accès aux réponses.
- Lourds et inadaptées aux problèmes avec de larges données.

Approches basées sur l'apprentissage automatique

- Apprend les fonctionnalités linguistiques du texte, sans être explicitement programmées.
- Traite de très grandes quantités de données.
- Cartographe et représente les questions et les réponses candidates sous forme de représentations vectorielles et par la suite mesurer leur similitude.

Approches basées sur l'apprentissage profond

- Utilise les **réseaux de neurones**.
- Différentes architectures de réseaux sont utilisées :
  - Réseau neuronal convolutif (CNN) ;
  - Réseau neuronal récurrent (RNN) ;
  - Réseau Transformers ;
  - Modèle BERT [3].

3. Bidirectional Encoder Representations from Transformers (BERT).

État de l'art

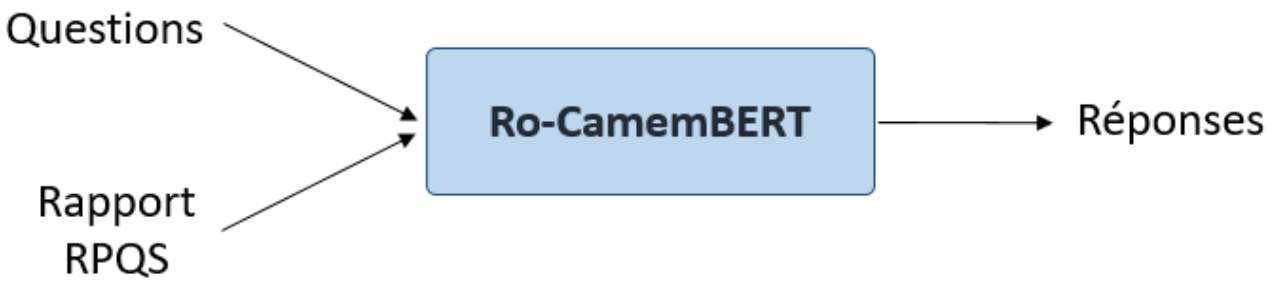
Synthèse

- Les approches basées sur l'apprentissage profond sont les plus efficaces.
- Le modèle BERT montre de meilleures performances pour la tâche de la réponse aux questions.
  - Produit des représentations sémantiques de haute qualité ;
  - Présente une faiblesse dans le traitement des séquences qui dépassent environ 300 à 500 mots.

Ro-CamemBERT

Solution proposée

- Nommée Ro-CamemBERT [4].
- Basée sur le modèle pré-entraîné de représentation du langage CamemBERT, dédié pour la langue française.
- Intègre et exécute séquentiellement trois modules dont chacun traite un volet spécifique de la problématique.



4. Recurrence over CamemBERT (Ro-CamemBERT)

Ro-CamemBERT

Solution proposée

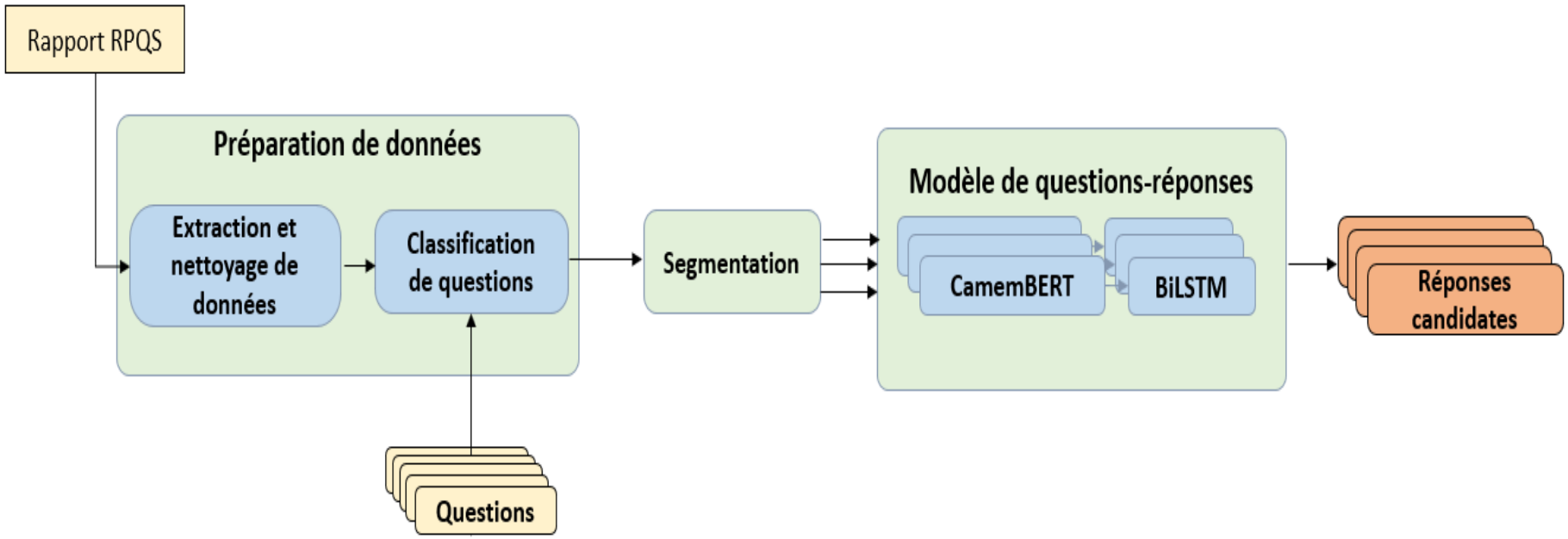


FIGURE – Ro-CamemBERT

Ro-CamemBERT

Préparation de données

- **Extraction et nettoyage de données** :
  - Transforme les connaissances représentées dans les différentes structures de données (paragraphes, images, tables, etc.) du rapport RPQS en format textuel ;
  - Supprime toute information bruyante et inutile.
- **Classification de questions** :
  - Décompose le rapports en sections ;
  - Retourne les sections du rapport concernées par chaque question d'entrée ;

Ro-CamemBERT

Segmentation

- Dépasser la limitation des modèles pré-entraînés du représentation du langage dans le traitement des longues séquences de texte.
- Segmenter les sections retournées en segments de tailles identiques.

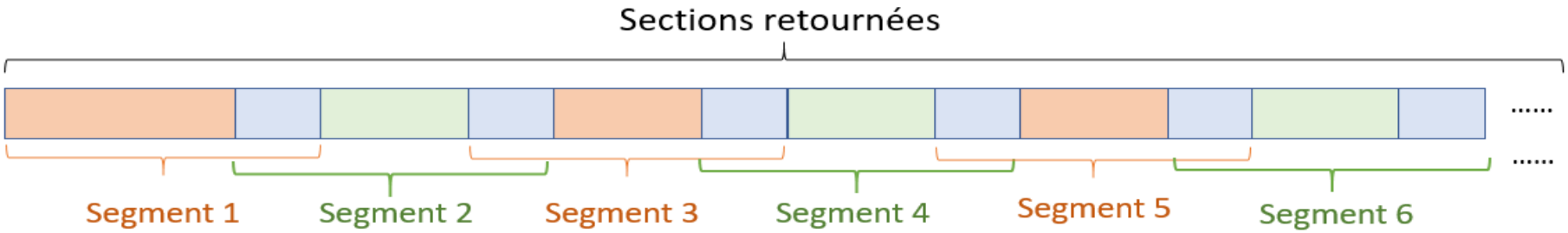


FIGURE – Module de segmentation

Ro-CamemBERT

Modèle de questions-réponses

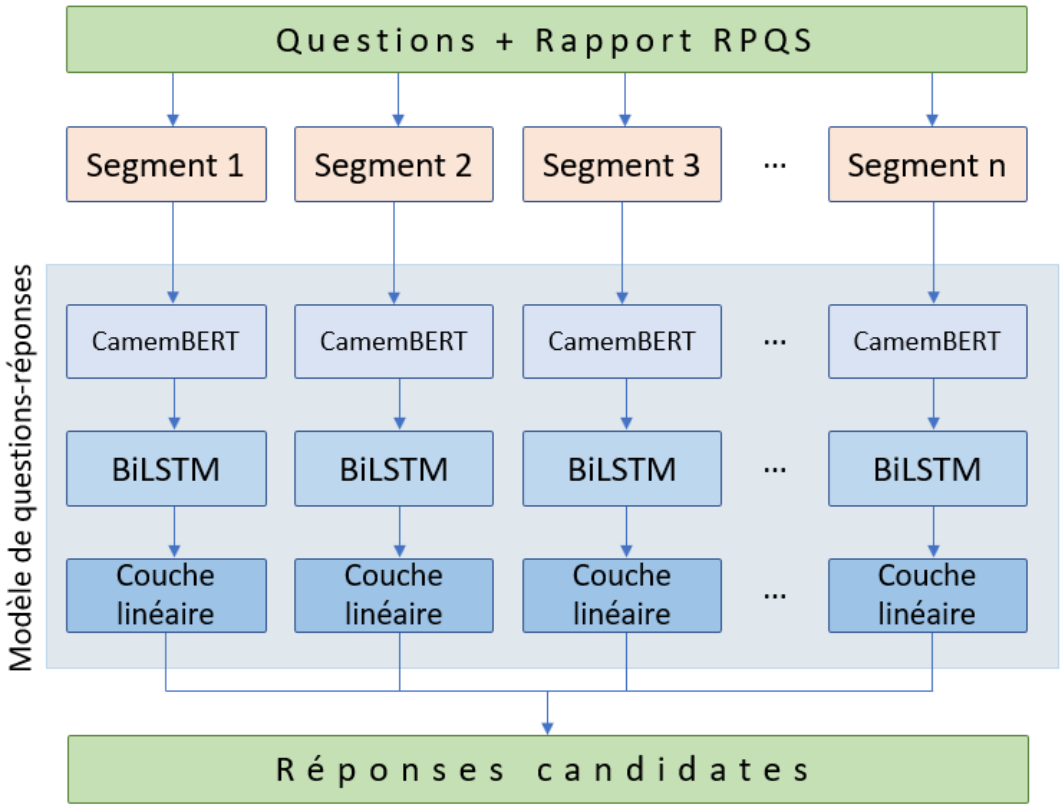


FIGURE – Modèle de questions-réponses

Réalisation

Étapes

1. Choix du modèle de questions-réponses le plus adéquat.
2. Ajout d'un réseau BiLSTM [5] au modèle.
3. Intégration de la tâche de classification de questions.
4. Création du jeu de données RPQS_QA : fine-tuning du modèle.

5. Bidirectional long short-term memory (BiLSTM)

Réalisation

Logiciel avec interface graphique

Conclusion

Résumé

- Ro-CamemBERT est conceptuellement simple mais empiriquement performant.
- Le premier système de questions-réponses spécialisé dans le domaine de l'eau et l'environnement en France.
- Peut être facilement adaptée et utilisée pour d'autres domaines.
- Contribuer dans les efforts de la communauté de la science des données française, par le jeu de données de questions-réponses RPQS_QA.

Conclusion

Perspectives

- Unification des architectures des rapports RPQS entre les différents services.
- Intégration d'un processus automatisé pour l'alimentation d'un entrepôt de données
- Analyse de l'évolution des indicateurs de performances dans chaque service d'eau et assainissement.

Conclusion

Merci pour votre aimable attention !