LIUM
Laboratoire d'Informatique
Le Mans Université

UMR IRISA

DEFT 2023
-
Qui de DrBERT, Wikipédia ou Flan-T5 s'y connaît le plus en questions médicales ?
-
5 juin 2023

Clément Besnard - Mohamed Ettaleb - Christian Raymond - Nathalie Camelin

Notre approche

Analyser le type de question

- Question de type QCM
- Des énoncés différents pour une recherche différente
- Permet de définir comment appliquer le système qui va répondre à la question

Définir des systèmes pour répondre

- Exploiter une base de connaissances
- Utiliser un modèle de langage génératif
- Définir un Méta-système sur un ensemble de descripteurs variés

Détection du type de question

Bonne(s) réponse(s) recherchée(s) : fausse(s) ou correcte(s) ?

Exemple d'énoncés

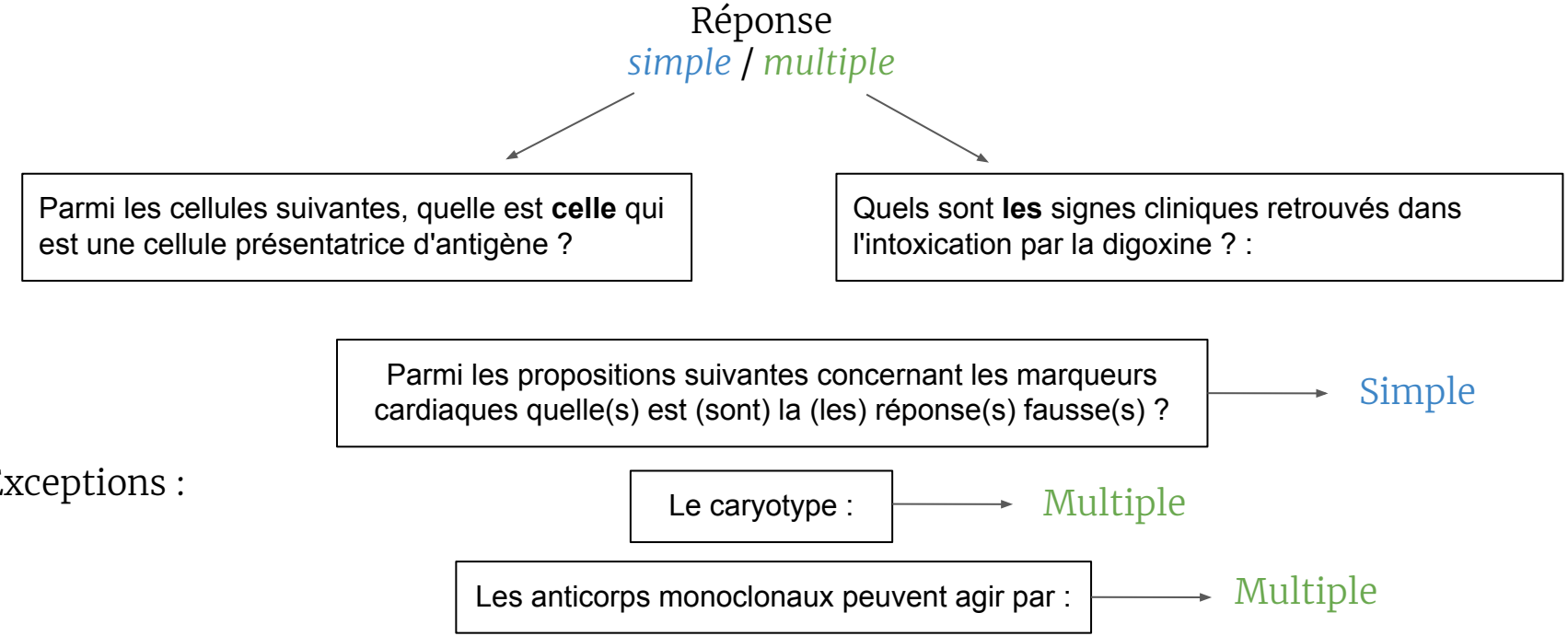
*Parmi les affirmations suivantes, une seule est **fausse**, indiquer laquelle : les particules alpha*

*Parmi les propositions suivantes, une seule est **exacte**. Laquelle ? La sérotonine est le (la) :*

Détection par expression régulière

- sauf
- fausse
- fausses
- ne
- inexacte
- inexactes
- n'

Détection du type de question



Détection : CamemBERT + couche de classification

F1-score : 95,90 (Dev)

FBC-ngram-rule : Fouille dans une base de connaissances

ÉTAPE 1 : Définir une base de connaissances pour chaque question

Mise en œuvre :

- Utilisation des articles de Wikipédia
- Sélection des *n* articles les plus proches de la question

Deux approches :

1. Modèle vectoriel TF.IDF : calcul de similarité entre la question et des titres d'articles
2. API de recherche Wikipédia

Prétraitements :

- liste stopwords Spacy + '%exact%', 'proposition%', 'indique%', 'réponse%', 'fausse%', 'affirmation%', 'propos', 'vraie%', 'coche%', 'donner', 'trouve'
- Suppression de la ponctuation

FBC-ngram-rule : Fouille dans une base de connaissances

ÉTAPE 2 : Attribuer un score à chaque réponse

<u>Mise en œuvre :</u>

- Comptage des uni- et bi-grammes communs entre la base de connaissances et la réponse
- Calcul d'un score :

Score :
(2 * NbBigram + NbUnigram) / NbTokens

Score (avec pondération idf) :
(idf * NbBigram + idf * NbUnigram) / NbTokens

<u>Prétraitements :</u>

- Caractères grecs → Caractère latins | β → beta
- Chiffres romains → Notation arabe | II → 2
- Suppression des sauts de lignes
- Normalisation unicode (caractères spéciaux / accents)
- Lemmatisation (fr_core_news_md - Spacy)
- Suppression de la ponctuation
- Liste de stopwords Spacy \ {'moins', 'plus', 'peu'}

FBC-ngram-rule : Fouille dans une base de connaissances

ÉTAPE 3 : Prédiction à l'aide de règles

Exemple de vecteur de scores :

A	B	C	D	E
145	0	123	66	40

<u>Ensemble de règles appliquées</u>

A	B	C	D	E
145	0	123	66	40

La réponse correcte

A	B	C	D	E
145	0	123	66	40

La réponse fausse

Si scores égaux :
Bonne réponse : 'd', 'c', 'b', 'a', 'e'
Mauvaise réponse : 'd', 'c', 'e', 'b', 'a'

A	B	C	D	E
145	0	123	66	40

Les réponses correctes

A	B	C	D	E
145	0	123	66	40

Les réponses fausses

Si scores nuls :
Retourne la réponse 'bcd'

Flan-T5 : modèle génératif

- Modèle de langage spécialisé sur plus de 1 000 tâches
- Modèle génératif

Chung et al., 2022, Scaling instruction-finetuned language models

<u>Spécialisation à une nouvelle tâche avec de nouvelles instructions :</u>

Choisis la bonne réponse : {question} (A) {réponse A} (B) {réponse B} (C) {réponse C} (D) {réponse D} (E) {réponse E} context : {contexte}

Instructions

Choisis les bonnes réponses : {question} (A) {réponse A} (B) {réponse B} (C) {réponse C} (D) {réponse D} (E) {réponse E} context : {contexte}

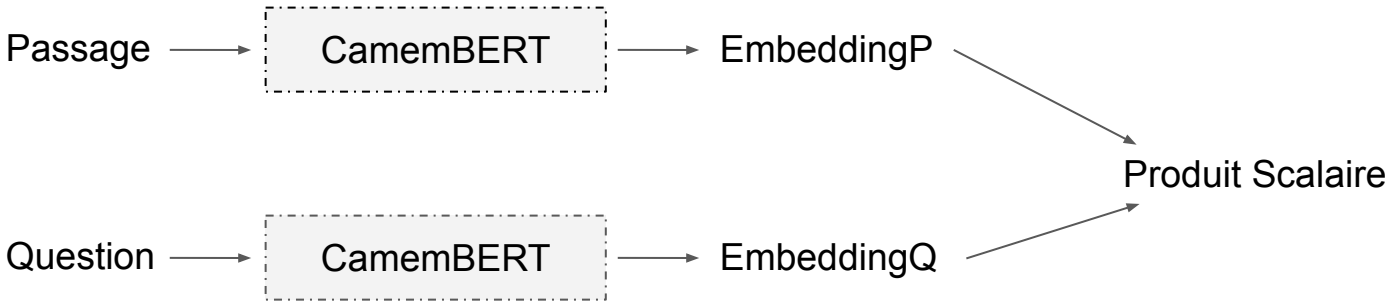
Sortie du modèle → Génération des bonnes réponses

Exemple → A B D

D'où vient le contexte ?

Flan-T5 : modèle génératif

- Associer un contexte à chaque question



Modèle Dense Passage Retriever (DPR)

Karpukhin et al., 2020, Dense passage retrieval for open-domain question answering

- Utilisation de deux encodeurs pré-entraîné sur 90 562 questions (PIAF, FQuAD, SQuAD-FR)
- Passage de 100 mots issus des articles de Wikipédia (recouvrement de 10 mots)
- Contexte → Passage le plus proche (distance euclidienne)

https://huggingface.co/etalab-ia/dpr-ctx_encoder-fr_qa-camembert
https://huggingface.co/etalab-ia/dpr-question_encoder-fr_qa-camembert

Méta-système : décider selon plusieurs descripteurs

Objectif :

- Prendre une décision sur :
  - cette réponse en particulier est-elle valide par rapport à la question posée ?
- En fonction d'un grand ensemble de descripteurs, issus de :
  - de nouveaux indices de similarités/base de connaissances
  - un système expert spécialisé dans le médical
  - les 2 premiers systèmes

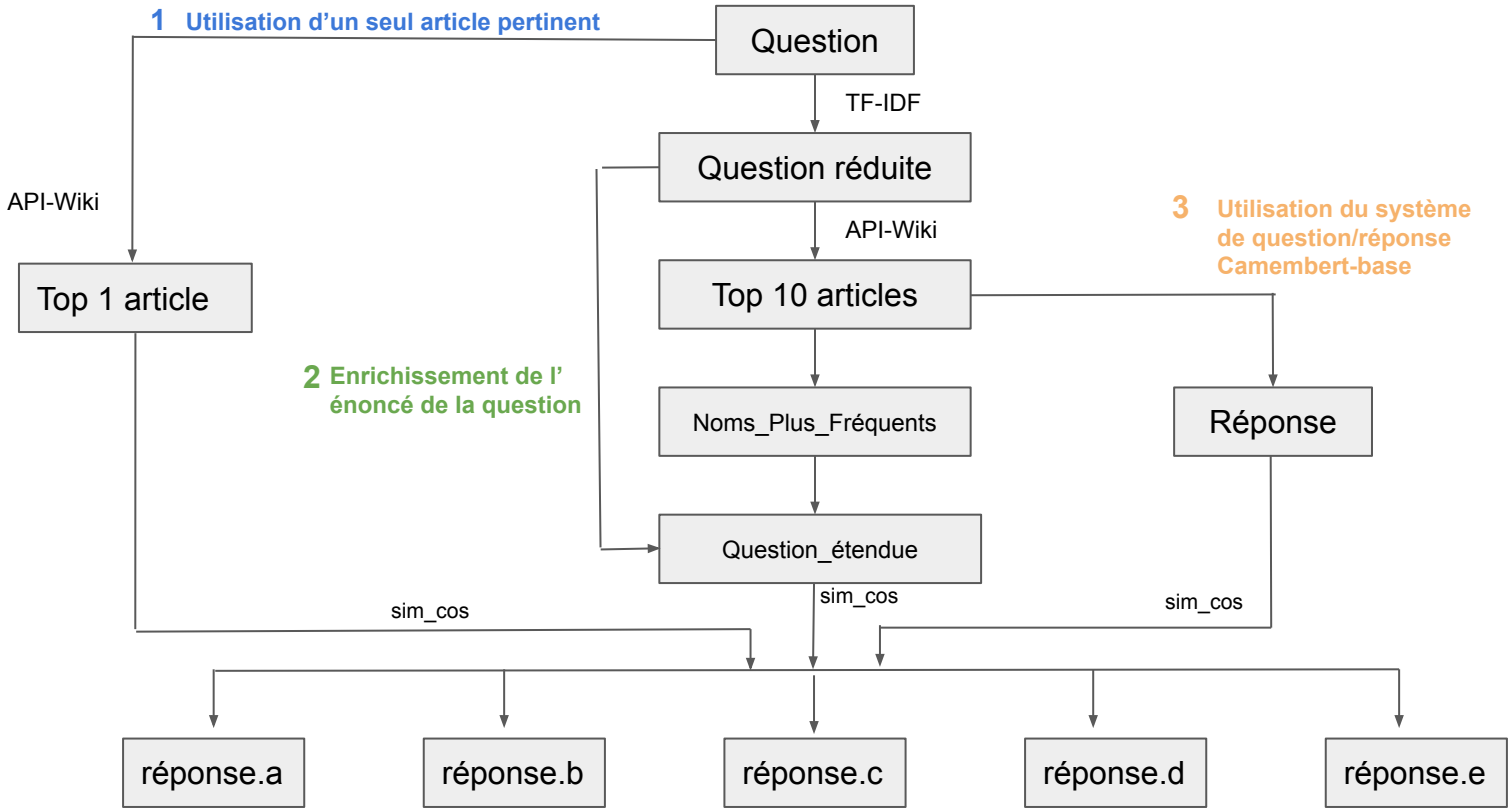
Mise en œuvre :

- Génération du corpus de questions en version binaire :

  ➔ Création de paires question/réponseX associées à la classe ***oui*** si réponseX est bonne, ***non*** sinon.

- Utilisation d'un algorithme d'apprentissage supervisé classique : Gradient Boosting Classifier

Méta-système : de nouveaux descripteurs



Méta-système : autres descripteurs

- **Descripteur biomédical** - DrBERT

  Labrak et al., 2023, A Robust Pre-trained Model in French for Biomedical and Clinical domains

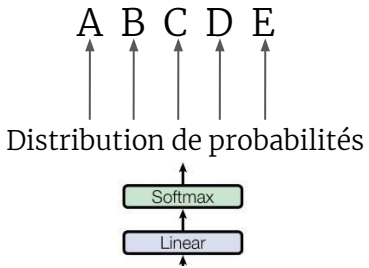
  Similarité cosinus entre couple Question/RéponseX (token [CLS])

- **FBC-ngram-rule**

  Nombre d'unigrammes et de bigrammes avec 1, 5 et 20 articles

- **Flan-T5**

  Génération de la première réponse :



Résultats : FBC-ngram-rule

Hamming → Taux de bonnes réponses parmi l'ensemble des hypothèses et référence

EMR → Taux de bonnes réponses exactes

		1 article		5 articles		20 articles	
		Hamming	**EMR**	**Hamming**	**EMR**	**Hamming**	**EMR**
API Wikipédia	**Dev** *(pond. idf)*	37,32	18,27	38,24	18,59	38,03	17,31
	Dev	36,43	17,31	**38,74**	**18,91**	38,35	17,95
	Test	-	-	36,72	17,85	-	-
Modèle Vectoriel	**Dev**	35,60	17,31	34,78	16,35	-	-

Résultats système 1 : FBC-ngram-rule

Résultats : Flan-T5

	Wiki passages DPR		Sans contexte	
	Hamming	**EMR**	**Hamming**	**EMR**
Dev	44,88	**25,64**	**45,04**	25,32
Test	43,24	22,19	42,75	22,83

Résultats système 2 : Flan-T5

Résultats : Méta-système

	Hamming	EMR
Dev	**44,42**	**24,36**
Test	35,47	18,49

Résultats système 3 : Méta-système

Descripteur	Importance
Flan-T5	**0,187**
Enrichissement de l'énoncé de la question	0,154
Utilisation d'un système de question/réponse	0,137
Utilisation d'un seul article pertinent	0,094
FBC-ngram-rule	0,043
Descripteur biomédical	0,039

Importance des descripteurs dans la prédiction du modèle

GradientBoostingClassifier.feature_importances_() : permet de mesurer l'importance relative des différentes caractéristiques (features) utilisées pour effectuer les prédictions

Conclusion

- 1re participation à la campagne d'évaluation DEFT 2023
- Grande exploitation de Wikipédia comme base de connaissances généraliste
  (système à base de règles, différents descripteurs de similarité cosine)
- Meilleure performance obtenue avec un grand modèle de langage génératif non spécialisé dans le domaine médical
- DrBERT n'a pas contribué de manière significative aux performances