Answering Complex Open-domain Questions Through Iterative Query Generation

aka Golden Retriever

Qi, Lin, Mehr, Wang & Manning - Stanford - EMNLP

https://arxiv.org/abs/1910.07000

Open-domain QA

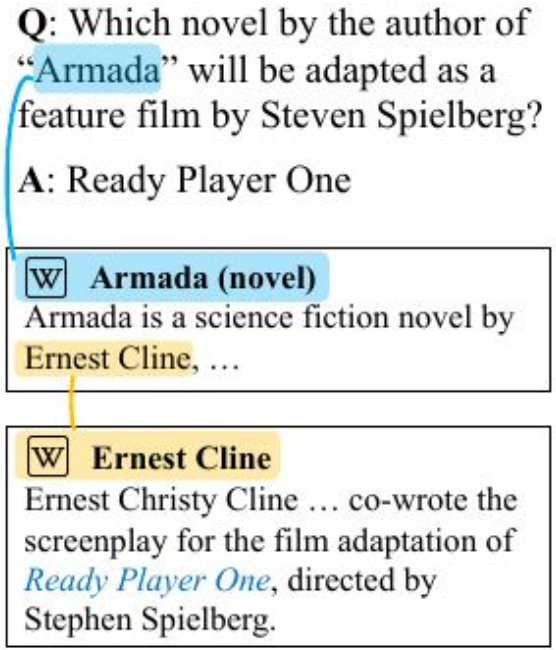
Q: Which novel by the author of "Armada" will be adapted as a feature film by Steven Spielberg?

A: Ready Player One

Setup : multi-hop

Trouver la réponse demande d'exploiter **plusieurs** documents différents (2 pour HotpotQA)

Les documents cibles sont ici connus



Approche usuelle

2 étapes : retrieve & read

- retrieve : utiliser une méthode d'IR pour réduire le corpus a k documents = **contexte**
- read : utiliser un NN (DocReader, BiDAF, BERT…) pour **extraire** la réponse des k documents

→ pas de la génération ! (généralement)

→ il faut prédire les indices de début et fin de la réponse

Approche usuelle

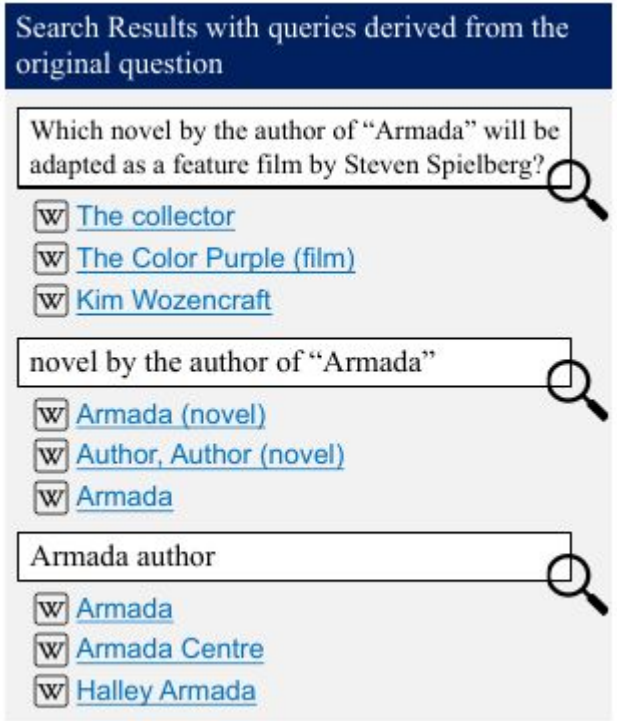
2 étapes : retrieve & read

- retrieve : utiliser une méthode d'IR pour réduire le corpus a k documents = **contexte**
- read : utiliser un NN (DocReader, BiDAF, BERT…) pour **extraire** la réponse des k documents

→ pas de la génération ! (généralement)

Hypothèse : souvent la partie retrieval est pas top → goulet d'étranglement de la performance

Une seule solution : la reformulation



Leur idée : à partir de la requête et des documents déjà retrieved, formuler une *nouvelle requête*

PAS de la génération seq-to-seq

Modélisation comme un pb de QA, i.e. **extraction** d'une requête parmi le contexte.

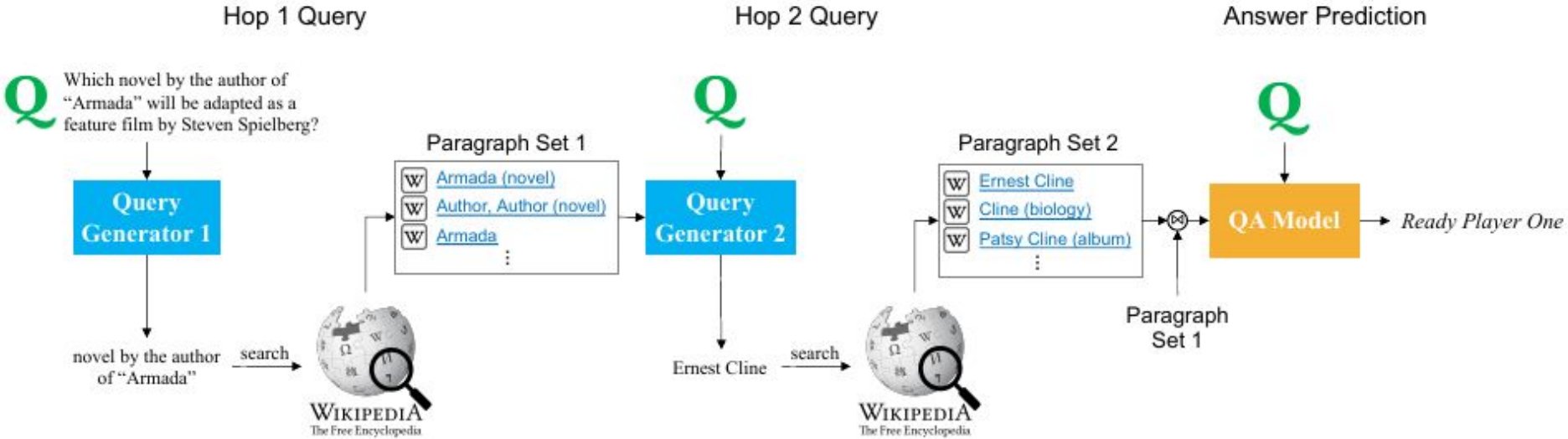


Figure 2: Model overview of GOLDEN Retriever. Given an open-domain multi-hop question, the model iteratively retrieves more context documents, and concatenates all retrieved context for a QA model to answer from.

Modèle & apprentissage

Modele : composante DocReader de DrQA

Apprentissage : **supervisé** grâce à des questions Oracle créées par heuristiques

→ en gros intersection entre le contexte actuel et le document cible

→ puis apprentissage par NLL (comme en QA)

Modèle & apprentissage

Modele : composante DocReader de DrQA

Apprentissage : **supervisé** grâce à des questions Oracle créées par heuristiques

→ en gros intersection entre le contexte actuel et le document cible

→ puis apprentissage par NLL (comme en QA)

Question	Hop 1 Oracle	Hop 2 Oracle
What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?	Corliss Archer in the film Kiss and Tell	Shirley Temple
Scott Parkin has been a vocal critic of Exxonmobil and another corporation that has operations in how many countries?	Scott Parkin	Halliburton
Are Giuseppe Verdi and Ambroise Thomas both Opera composers?	Giuseppe Verdi	Ambroise Thomas

Table 1: Example oracle queries on the HOTPOTQA dev set.

Modèle & apprentissage

Modele : composante DocReader de DrQA

Apprentissage : **supervisé** grâce à des questions Oracle créées par heuristiques

→ en gros intersection entre le contexte actuel et le document cible

→ puis apprentissage par NLL (comme en QA)

Pour le reste :

- BM25 pour le retrieval (classique)
- BiDAF++ pour le QA (classique)

Résultats

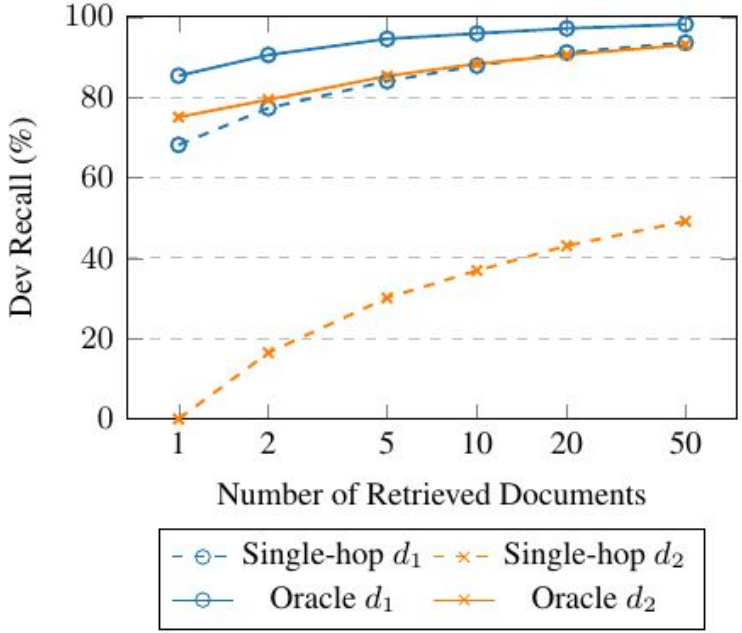


Figure 3: Recall comparison between single-hop queries and GOLDEN Retriever oracle queries for both supporting paragraphs on the HOTPOTQA dev set.

On vérifie que les oracles sont des bons oracles...

Résultats : retrieval performance

System	Answer		Sup Fact		Joint	
	EM	F_1	EM	F_1	EM	F_1
Baseline (Yang et al., 2018)	25.23	34.40	5.07	40.69	2.63	17.85
GRN + BERT	29.87	39.14	13.16	49.67	8.26	25.84
MUPPET (Feldman and El-Yaniv, 2019)	30.61	40.26	16.65	47.33	10.85	27.01
CogQA (Ding et al., 2019)	37.12	48.87	22.82	57.69	12.42	34.92
PR-Bert	43.33	53.79	21.90	59.63	14.50	39.11
Entity-centric BERT Pipeline	41.82	53.09	26.26	57.29	17.01	39.18
BERT pip. (contemporaneous)	45.32	57.34	38.67	70.83	25.14	47.60
GOLDEN Retriever	37.92	48.58	30.69	64.24	18.04	39.13

Table 2: End-to-end QA performance of baselines and our GOLDEN Retriever model on the HOTPOTQA fullwiki test set. Among systems that were not published at the time of submission of this paper, "BERT pip." was submitted to the official HOTPOTQA leaderboard on May 15th (thus contemporaneous), while "Entity-centric BERT Pipeline" and "PR-Bert" were submitted after the paper submission deadline.

Résultats : joint performance

System	Answer		Sup Fact		Joint	
	EM	F_1	EM	F_1	EM	F_1
Baseline (Yang et al., 2018)	25.23	34.40	5.07	40.69	2.63	17.85
GRN + BERT	29.87	39.14	13.16	49.67	8.26	25.84
MUPPET (Feldman and El-Yaniv, 2019)	30.61	40.26	16.65	47.33	10.85	27.01
CogQA (Ding et al., 2019)	37.12	48.87	22.82	57.69	12.42	34.92
PR-Bert	43.33	53.79	21.90	59.63	14.50	39.11
Entity-centric BERT Pipeline	41.82	53.09	26.26	57.29	17.01	39.18
BERT pip. (contemporaneous)	45.32	57.34	38.67	70.83	25.14	47.60
GOLDEN Retriever	37.92	48.58	30.69	64.24	18.04	39.13

59 F1 maintenant

Table 2: End-to-end QA performance of baselines and our GOLDEN Retriever model on the HOTPOTQA fullwiki test set. Among systems that were not published at the time of submission of this paper, "BERT pip." was submitted to the official HOTPOTQA leaderboard on May 15th (thus contemporaneous), while "Entity-centric BERT Pipeline" and "PR-Bert" were submitted after the paper submission deadline.

Quelques samples

	Question	Predicted q_1	Predicted q_2
(1)	What video game character did the voice actress in the animated film Alpha and Omega voice?	voice actress in the animated film Alpha and Omega *(animated film Alpha and Omega voice)*	Hayden Panettiere
(2)	What song was created by the group consisting of Jeffrey Jey, Maurizio Lobina and Gabry Ponte and released on 15 January 1999?	Jeffrey Jey *(group consisting of Jeffrey Jey, Maurizio Lobina and Gabry Ponte)*	Gabry Ponte and released on 15 January 1999 *("Blue (Da Ba Dee)")*
(3)	Yau Ma Tei North is a district of a city with how many citizens?	Yau Ma Tei North	Yau Tsim Mong District of Hong Kong *(Hong Kong)*
(4)	What company started the urban complex development that included the highrise building, The Harmon?	highrise building, The Harmon	CityCenter

Table 5: Examples of predicted queries from the query generators on the HOTPOTQA dev set. The oracle query is displayed in blue in parentheses if it differs from the predicted one.