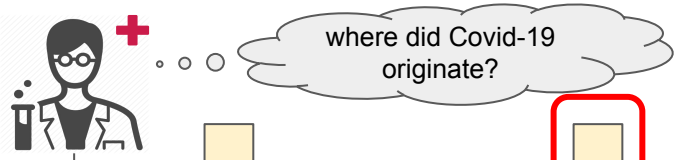
A Neural Model for Joint Document and Snippet Ranking in Question Answering for Large Document Collections

*Dimitris Pappas, Ion Androutsopoulos

Retrieval Pipeline



BM25

Document Re-ranking

We only examine the top k documents in snippet re-ranking

Snippet Re-ranking does not give feedback to Document Re-ranking (i.e. the yellow bot cannot talk back to the red bot).

Snippet Re-ranking

Snippets are sentences

PMC3901466: Only cell lines originating from goats and camels showed efficient replication of MERS-CoV.

PMC7118693: Originated in **bats**, 2019-nCoV/ severe acute respiratory syndrome coronavirus (SARS-CoV)-2 likely experienced ...

PMC7929414: Data science in unveiling COVID-19 pathogenesis and diagnosis: evolutionary origin to drug repurposing.

PMC7263917: … genetic evidence suggests that **bats** are a reservoir host of the virus, and pangolins are a probable intermediate.

Exact Answer Extraction

We only examine the top N snippets in exact answer extraction

Answer: “Bats”

Relevant Docs

Irrelevant Docs

Relevant Sentences

Irrelevant Sentences

Pappas and Androutsopoulos, “Joint Document and Snippet Ranking”, ACL-IJCNLP 2021

Retrieval Pipeline

where did Covid-19 originate?

BM25

this is a good snippet

this is a good document

Document & Snippet Re-ranking

Documents with no relevant snippets should be ranked to the bottom of the list

Snippet Re-r… by hand…
- rele…
- relevant … ippets

Pipeline methods missed article PMC7113610 in document retrieval so the sentence retrieval module did not have the chance to examine the snippet

PMC7113610: Genomic analysis revealed that SARS-CoV-2 ..., therefore **bats** could be the possible primary reservoir.
PMC7118693: Originated in **bats**, 2019-nCoV/ severe acute respiratory syndrome coronavirus (SARS-CoV)-2 likely experienced ...
PMC7929414: Data science in unveiling COVID-19 pathogenesis and diagnosis: evolutionary origin to drug repurposing.
PMC7263917: … genetic evidence suggests that **bats** are a reservoir host of the virus, and pangolins are a probable intermediate.

We only examine the top N snippets in exact answer extraction

Exact Answer Extraction

Answer: “Bats”

Relevant Docs | Irrelevant Docs | Relevant Sentences | Irrelevant Sentences

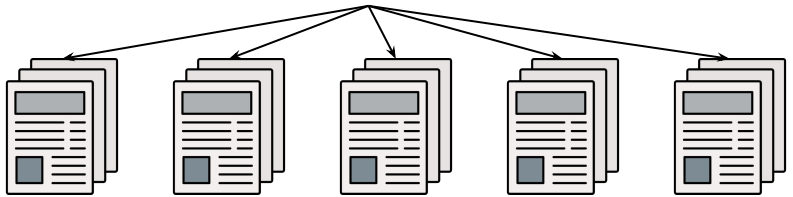
Pappas and Androutsopoulos, “Joint Document and Snippet Ranking”, ACL-IJCNLP 2021

Data collection and Indexing

BioASQ 2,747 questions

Which disease is caused by de novo VPS4A mutations?

PubMed



32.6M biomedical articles
27.8M biomedical articles in English

Data collection and preprocessing

21.8M articles with
Title and Abstract

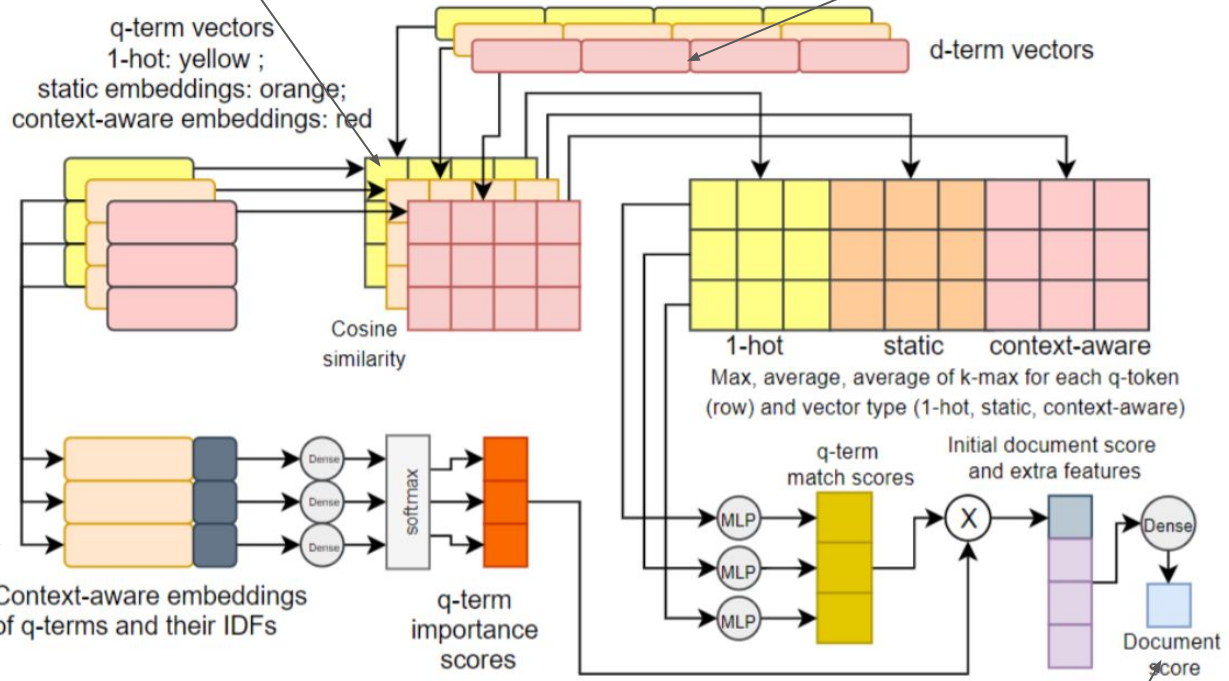
BM25

Pappas and Androutsopoulos, "Joint Document and Snippet Ranking", ACL-IJCNLP 2021

PDRMM

And computes three cosine similarity matrices across query and document terms

PDRRM uses three kinds of word representations:
- one hot for exact match
- pretrained w2v vectors
- cnn contextual representations



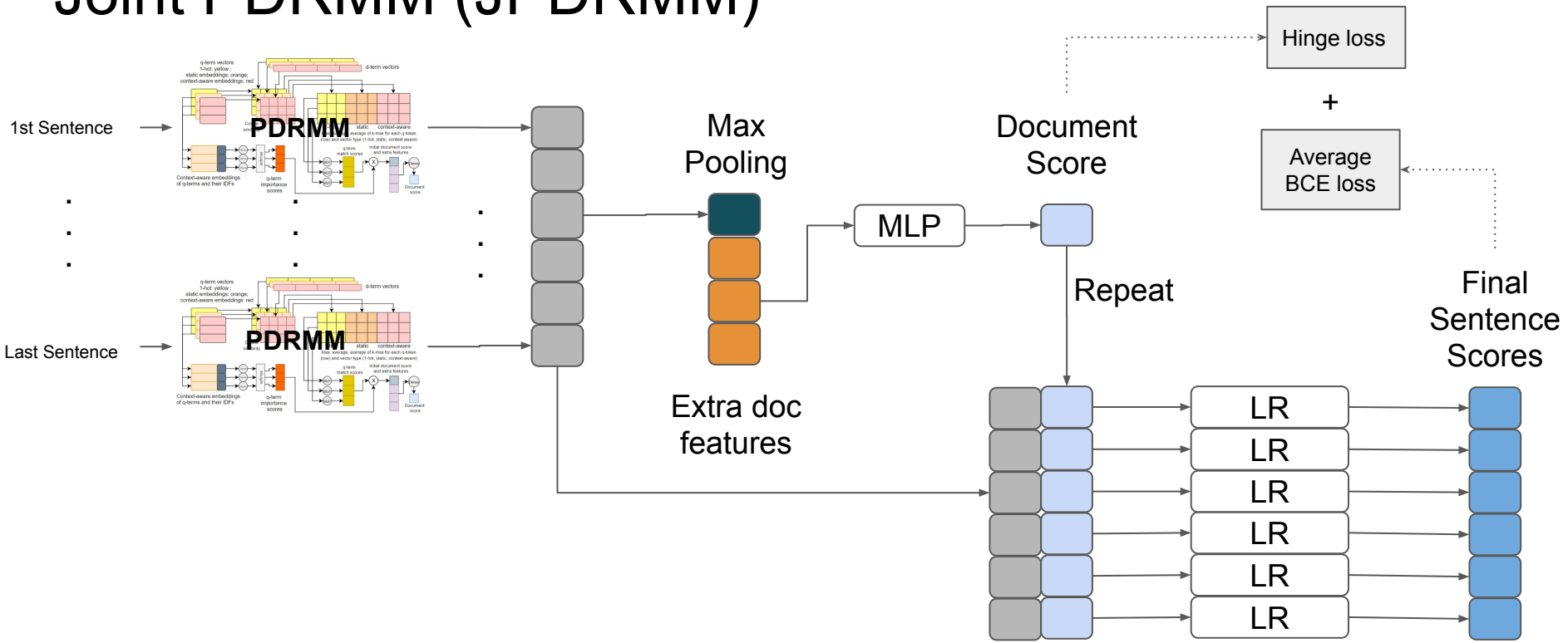
PDRRM uses three kinds of word representations:
- one hot for exact match
- pretrained w2v vectors
- cnn contextual representations

as well as an importance score for each query term

to compute a relevance score for the entire document

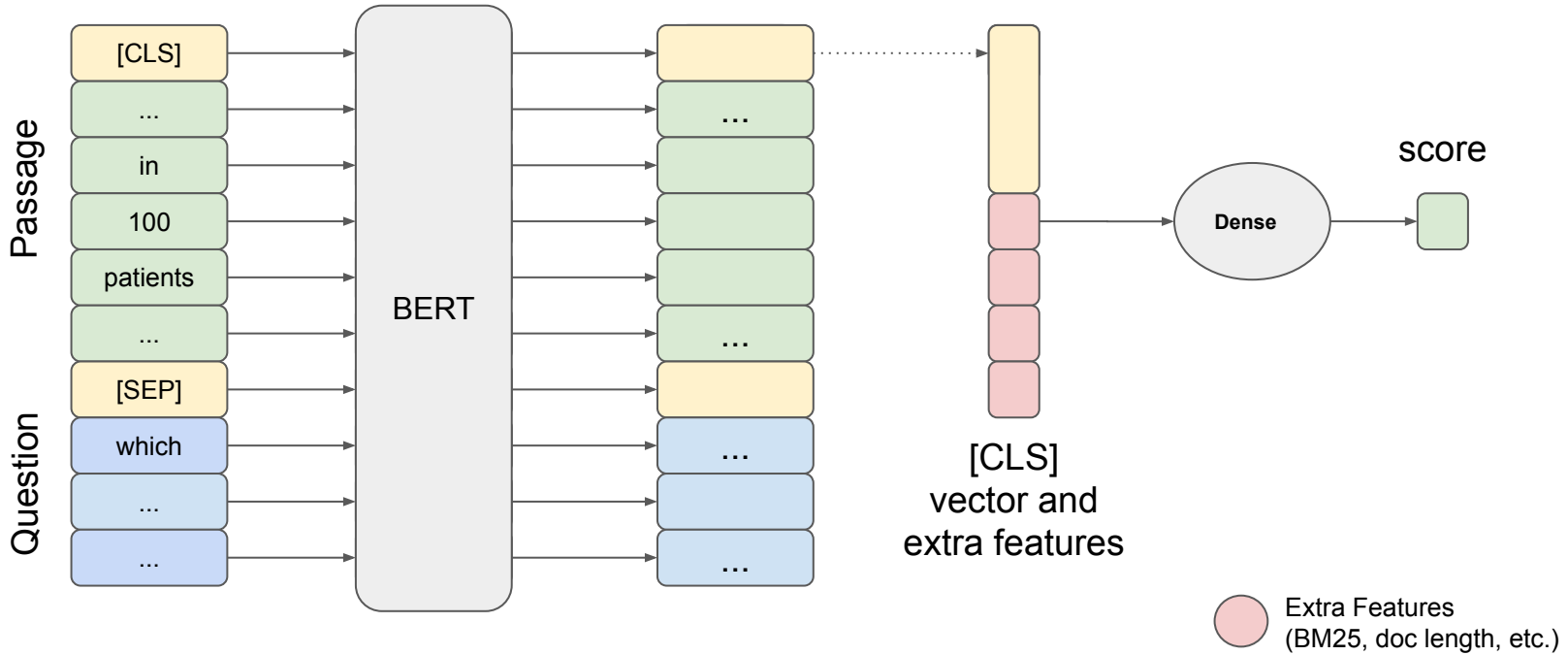
Pappas and Androutsopoulos, "Joint Document and Snippet Ranking", ACL-IJCNLP 2021

Joint PDRMM (JPDRMM)



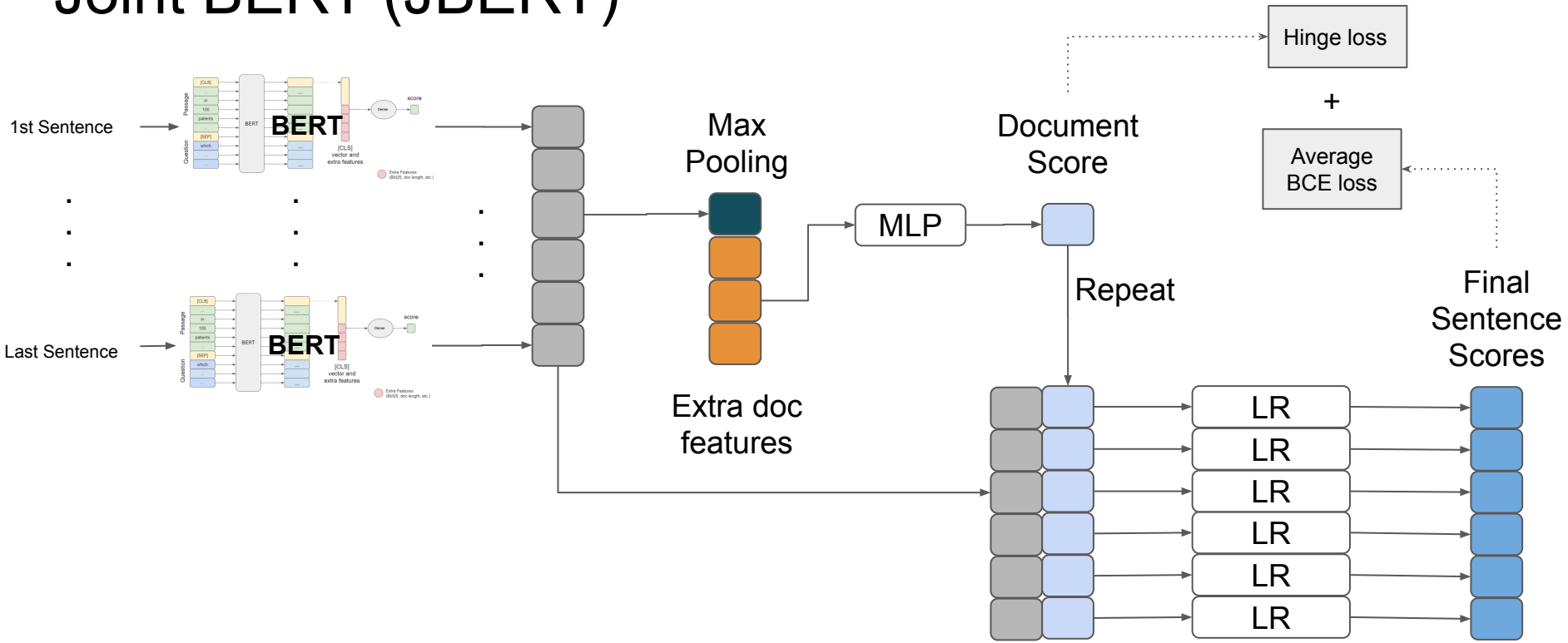
Pappas and Androutsopoulos, "Joint Document and Snippet Ranking", ACL-IJCNLP 2021

BERT



Pappas and Androutsopoulos, "Joint Document and Snippet Ranking", ACL-IJCNLP 2021

Joint BERT (JBERT)



Pappas and Androutsopoulos, "Joint Document and Snippet Ranking", ACL-IJCNLP 2021

BioASQ 7 Results (Best of 14 methods)

Models perform better when Bert weights are not fine-tuned for the retrieval task

Method	Method Type	Params	Doc. MAP	Snip. MAP
BERT+PDRMM	Pipeline	109.5M	8.79	9.63
JPDRMM	Joint	5.79k	6.69	15.72
BJPDRMM-ADAPT-NF	Joint	3.5M	7.42	17.35
JBERT-ADAPT-NF	Joint	6.3K	7.84	16.53
Oracle	n/a	0	19.24	25.18

In ADAPT models we use a linear combination of all embeddings that Bert produces throughout its layers

JPDRMM remains competitive in snippet retrieval despite using orders of magnitude fewer parameters.

Pappas and Androutsopoulos, "Joint Document and Snippet Ranking", ACL-IJCNLP 2021

BioASQ 7 Results (post-contest expe

Human annotators examine all results submitted by contestants to identify unseen gold data

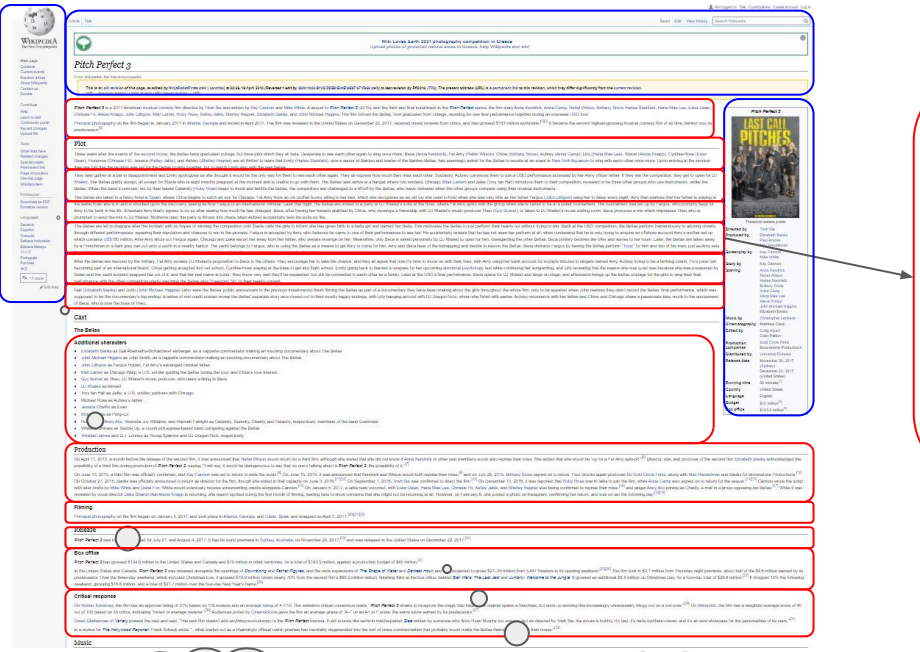
			Before expert inspection		After expert inspection	
Method	**Method Type**	**Parameters**	**Doc. MAP**	**Snip. MAP**	**Doc. MAP**	**Snip. MAP**
BERT+PDRMM	Pipeline	109.5M	7.29	7.58	14.86	15.61
JPDRMM	Joint	5.79K	5.16	12.45	16.55	21.98
BJPDRMM-NF	Joint	3.5M	6.18	13.89	14.65	23.96
Best BIOASQ 7 competitor	n/a	n/a	n/a	n/a	13.18	14.98

BIOASQ 7 **test batches 4 and 5** before and after post-contest

Pappas and Androutsopoulos, "Joint Document and Snippet Ranking", ACL-IJCNLP 2021

Original Natural Questions Dataset

T. Kwiatkowski et al. 2019



Boilerplate and tables (blue color)

Paragraphs (red color)

Question: where do the bellas go in pitch perfect 3

HTML content: <h2>Plot</h2> The Bellas are taken to a fancy hotel in Spain, where Chloe begins to catch an eye for Chicago. Fat Amy finds an old stuffed bunny sitting in her bed, which she recognizes as an old toy she used to hold when she was very little as her father Fergus (John Lithgow) sang her to sleep every night. Amy then realizes that her father is staying in the same hotel she is in and is shocked upon the discovery, seeing as how Fergus is an international criminal. Later that night, the Bellas are invited to a party at DJ Khaled's suite at the hotel, where Fat Amy splits with the group when she is called to be at a poker tournament. The tournament was set up by Fergus, who promptly begs for Amy to be back in his life. A hesitant Amy finally agrees to do so after seeing how much he has changed. Beca, after having her breasts grabbed by Chloe, who develops a friendship with DJ Khaled's music producer Theo (Guy Burnet), is taken to DJ Khaled's music editing room. Beca produces a mix which impresses Theo who is prompted to send the mix to DJ Khaled. Moments later, the party is thrown into chaos when Aubrey accidentally sets the suite on fire.

Answer: Spain

URL: https://en.wikipedia.org/?title=Pitch_Perfect_3&oldid=837158847

We handle each paragraph of the entire webpage as a document which we index in our database

Pappas and Androutsopoulos, "Joint Document and Snippet Ranking", ACL-IJCNLP 2021

Natural Questions Modification

Question: where do the bellas go in pitch perfect 3

HTML content: <h2>Plot</h2> The Bellas are taken to a fancy hotel in Spain, where Chloe begins to catch an eye for Chicago. Fat Amy finds an old stuffed bunny sitting in her bed, which she recognizes as an old toy she used to hold when she was very little as her father Fergus (John Lithgow) sang her to sleep every night. Amy then realizes that her father is staying in the same hotel she is in and is shocked upon the discovery, seeing as how Fergus is an international criminal. Later that night, the Bellas are invited to a party at DJ Khaled's suite at the hotel, where Fat Amy splits with the group when she is called to be at a poker tournament. The tournament was set up by Fergus, who promptly begs for Amy to be back in his life. A hesitant Amy finally agrees to do so after seeing how much he has changed. Beca, after having her breasts grabbed by Chloe, who develops a friendship with DJ Khaled's music producer Theo (Guy Burnet), is taken to DJ Khaled's music editing room. Beca produces a mix which impresses Theo who is prompted to send the mix to DJ Khaled. Moments later, the party is thrown into chaos when Aubrey accidentally sets the suite on fire.

We handle each paragraph of the entire webpage as a document which we index in our database

Plain text extraction

Question: where do the bellas go in pitch perfect 3

Sentences:

The Bellas are taken to a fancy hotel in **Spain**, where Chloe begins to catch an eye for Chicago.
Fat Amy finds an old stuffed bunny sitting in her bed, which she recognizes as an old toy she used to hold when she was very little as her father Fergus (John Lithgow) sang her to sleep every night.
Amy then realizes that her father is staying in the same hotel she is in and is shocked upon the discovery, seeing as how Fergus is an international criminal.
Later that night, the Bellas are invited to a party at DJ Khaled's suite at the hotel, where Fat Amy splits with the group when she is called to be at a poker tournament.
The tournament was set up by Fergus, who promptly begs for Amy to be back in his life.
A hesitant Amy finally agrees to do so after seeing how much he has changed.
Beca, after having her breasts grabbed by Chloe, who develops a friendship with DJ Khaled's music producer Theo (Guy Burnet), is taken to DJ Khaled's music editing room.
Beca produces a mix which impresses Theo who is prompted to send the mix to DJ Khaled.
Moments later, the party is thrown into chaos when Aubrey accidentally sets the suite on fire.

"Spain" is the answer therefore the yellow sentence is annotated as relevant.

Pappas and Androutsopoulos, "Joint Document and Snippet Ranking", ACL-IJCNLP 2021

Modified Natural Questions Results

	Document Retrieval (paragraph retrieval)			Snippet Retrieval (sentence retrieval)		
Method	**MRR**	**Recall@1**	**Recall@2**	**MRR**	**Recall@1**	**Recall@2**
Pipeline: BM25+BM25	30.18	16.50	29.75	8.19	3.75	7.13
Pipeline: PDRMM+PDRMM	40.33	28.25	38.50	22.86	13.75	22.75
Joint: JPDRMM	36.50	24.50	36.00	26.92	19.00	25.25

We use Recall@1 and Recall@2 because there are at most two relevant documents and two relevant snippets


joint model does not outperform pipelines in document retrieval


Joint model outperforms pipelines in snippet retrieval


Both models outperform BM25

Pappas and Androutsopoulos, "Joint Document and Snippet Ranking", ACL-IJCNLP 2021

Conclusions

- All the neural pipelines and joint models we considered improve the traditional BM25 ranking on both datasets.
- Joint models vastly outperform the corresponding pipelines in snippet retrieval.
- The Joint PDRMM model that does not use BERT is competitive with BERT-based models.
- We surpassed all competitive systems in BioASQ 7.

Future Work

- Other pre-trained models on biomedical data (e.g. BioBERT) or QA data (e.g. SQuAD)
- Extend our joint models to cover selecting exact answers too (Document Retrieval, Snippet extraction, Factoid QA)
- Use external knowledge (e.g. graph embeddings)

http://nlp.cs.aueb.gr/publications.html

@dvpappas, @AUEBNLPGroup

pappasd@aueb.gr