Data Augmentation for Biomedical Factoid Question Answering

*Dimitris Pappas
Makis Malakasiotis
Ion Androutsopoulos

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ATHENA Research & Innovation Information Technologies

ΔΗΜΟΚΡΙΤΟΣ
DEMOKRITOS
NATIONAL CENTRE FOR SCIENTIFIC RESEARCH

BioASQ Factoid QA

Question:

"Orteronel was developed for treatment of which cancer?"

Snippets:

- “Orteronel plus prednisone in patients with chemotherapy-naive metastatic **castration-resistant prostate cancer** (ELM-PC 4): a double-blind, multicentre, phase 3, randomised, placebo-controlled trial”
- "On the basis of these and other data, orteronel is not undergoing further development in metastatic **castration-resistant prostate cancer**."
- "This study examined orteronel in patients with metastatic **castration-resistant prostate cancer** that progressed after docetaxel therapy."
- “The experimental interventions tested in these studies were enzalutamide, ipilimumab, abiraterone acetate, orteronel and cabazitaxel.”

Answer:

“castration-resistant prostate cancer”

Triplets: Question - Snippet - Answer

("Orteronel was developed for treatment of which cancer?", “Orteronel plus prednisone in patients with chemotherapy…” , “castration-resistant prostate cancer”)

("Orteronel was developed for treatment of which cancer?", “On the basis of these and other data, orteronel is not …”, “castration-resistant prostate cancer”)

("Orteronel was developed for treatment of which cancer?", “This study examined orteronel in patients with metastatic …”, “castration-resistant prostate cancer”)

Pre-training using generic-domain QA data

Pre-trained huggingface QA models

Model	**BioASQ corpus**
DISTILBERT(SQUAD)	64.27
BIOBERT(SQUAD-V2)	69.22
ALBERT (SQUAD-V2)	75.05

We do not further train the already pretrained models.
We just evaluate on BioASQ's data

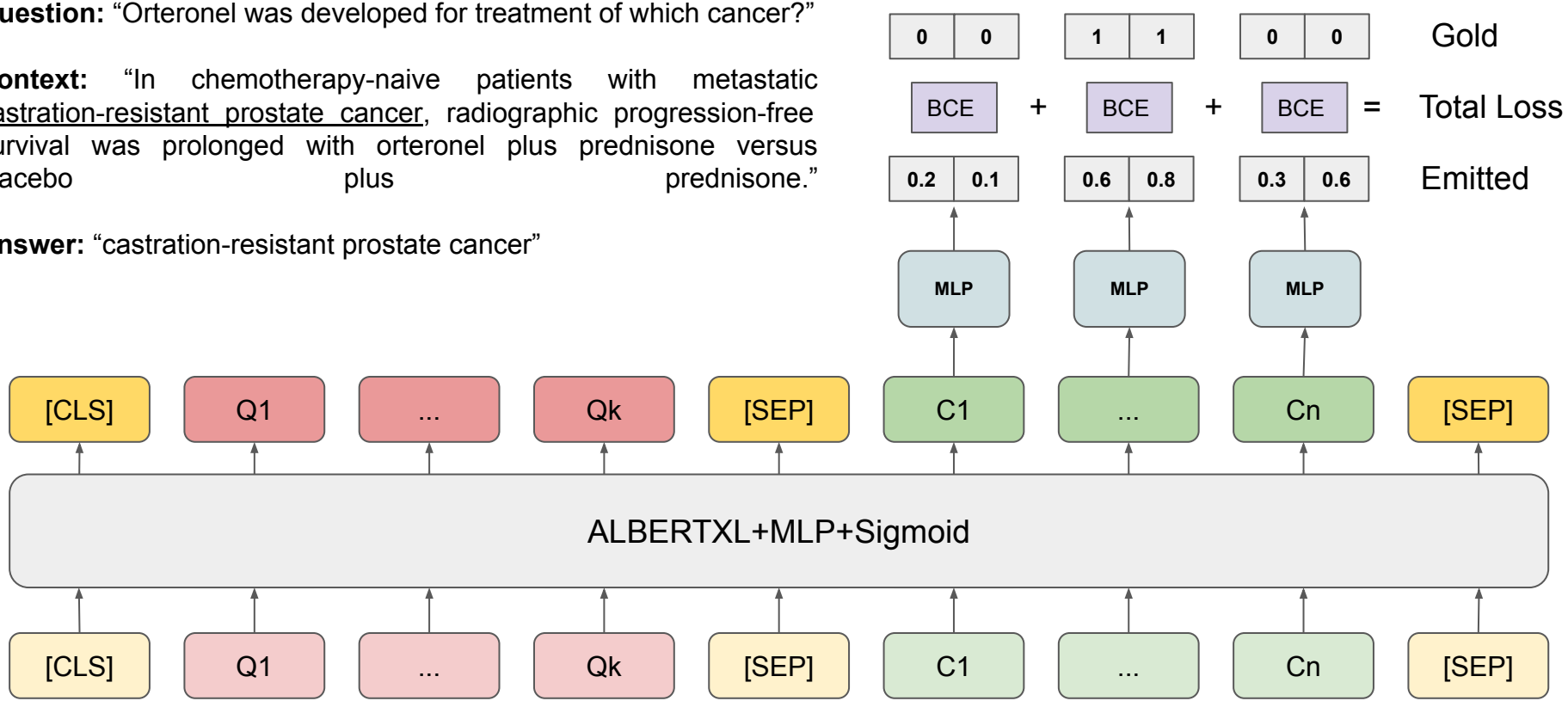
Pappas et al., "Data Augmentation for Biomedical Factoid Question Answering", BIONLP 2022

Factoid QA model

Question: “Orteronel was developed for treatment of which cancer?”

Context: “In chemotherapy-naive patients with metastatic castration-resistant prostate cancer, radiographic progression-free survival was prolonged with orteronel plus prednisone versus placebo plus prednisone.”

Answer: “castration-resistant prostate cancer”



Data Augmentation Techniques

- Back Translation
- Context Increasing
- Information Retrieval
- Word Embedding Substitution
- Bert Masking
- Question Generation
- Machine Reading Comprehension
- Synonym Replacement
- Random Insertion/Deletion/Swap
- QWERTY Error Insertion

Pappas et al., "Data Augmentation for Biomedical Factoid Question Answering", BIONLP 2022

Machine Reading Comprehension (BIOMRC)

Context (Abstract):
OBJECTIVE: The proximal segment of the anterior cerebral artery (A1) is among the most uncommon locations for occurrence of an **@entity438** . These **@entity439** may be missed if small or misinterpreted when they are near the internal cerebral artery bifurcation or Anterior Communicating Artery region. The association with **@entity154** and multiplicity makes them unique. METHODS: Seventeen A1 **@entity439** were diagnosed in sixteen **@entity1** between January 2000 and October 2014 in our institution. A retrospective review of the clinical, radiological, and management (microsurgical and endovascular) details of these **@entity1** was conducted. RESULTS: The incidence of A1 **@entity439** was 1.71% of all **@entity1** harboring **@entity439** and 1.19% of all **@entity439** . Half of these **@entity1** exhibited **@entity510** . Fourteen **@entity439** underwent microsurgical or endovascular intervention. All **@entity1** recovered well, except for one **@entity1** who died in the postoperative period. CONCLUSIONS: A1 **@entity439** are rare, with wide anatomic variations. In this article, we discuss those variations in detail with illustrative cases and pictures. We also discussed the microsurgical and endovascular strategies to encounter them highlighting the technical challenges.

Question (Title): Management of Proximal Anterior Cerebral **XXXX** Anatomical Variations and Technical Nuances.

Choices:
"@entity1:: ('9606', 'Species') :: ['patients', 'patient']",
"@entity154 :: ('MESH:D000013', 'Disease') :: ['congenital vascular anomalies']",
"@entity510 :: ('MESH:D013345', 'Disease') :: ['subarachnoid hemorrhage']",
"@entity439 :: ('MESH:D000783', 'Disease') :: ['aneurysms', 'aneurysm']",
"@entity438 :: ('MESH:D002532', 'Disease') :: ['intracranial aneurysm']"

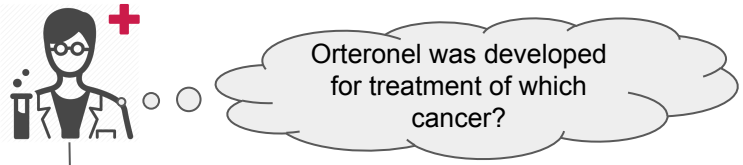
Answer:
@entity438:: (MESH:D002532,Disease) :: ['Artery Aneurysms']

Snippet:
The proximal segment of the anterior cerebral artery (A1) is among the most uncommon locations for occurrence of an Artery Aneurysms .

Question:
Management of Proximal Anterior Cerebral [MASK] Anatomical Variations and Technical Nuances.

Answer:
Artery Aneurysms

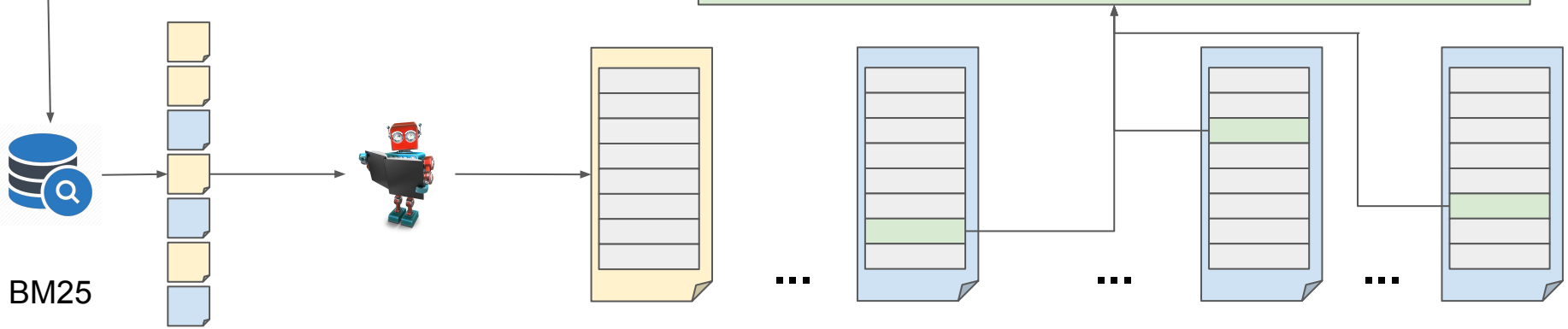
Information Retrieval



(PMID:25264242) Orteronel (TAK-700) is a substituted imidazole that was developed for the treatment of **castration-resistant prostate cancer** but was dropped in phase III clinical trials.

(PMID:25264242) Comparing the clinical efficacy of abiraterone acetate, enzalutamide, and orteronel in patients with metastatic **castration-resistant prostate cancer** by performing a network meta-analysis of eight randomized controlled trials.

(PMID:25264242) Orteronel is a nonsteroidal, selective inhibitor of 17,20-lyase that was recently in phase 3 clinical development as a treatment for **castration-resistant prostate cancer**.



Documents containing the answer

Documents not containing the answer

Sentences containing the answer

Sentences not containing the answer

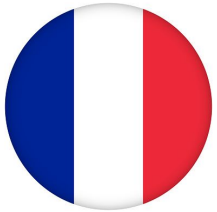
Pappas et al., "Data Augmentation for Biomedical Factoid Question Answering", BIONLP 2022

Back Translation

We do this for both questions and snippets

Orteronel was developed for treatment of which cancer?

Google Translate (ENG->FR)



Orteronel a été développé pour le traitement de quel cancer ?

Google Translate (FR->ENG)

Orteronel was developed for treating which cancer?

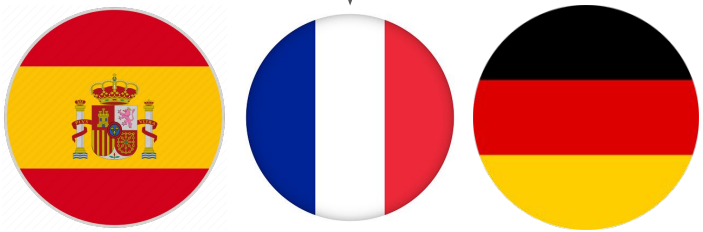
Pappas et al., "Data Augmentation for Biomedical Factoid Question Answering", BIONLP 2022

Back Translation

Orteronel was developed for treatment of which cancer?

Google Translate (ENG->FR)

We do this for both questions and snippets



Orteronel a été développé pour le traitement de quel cancer ?

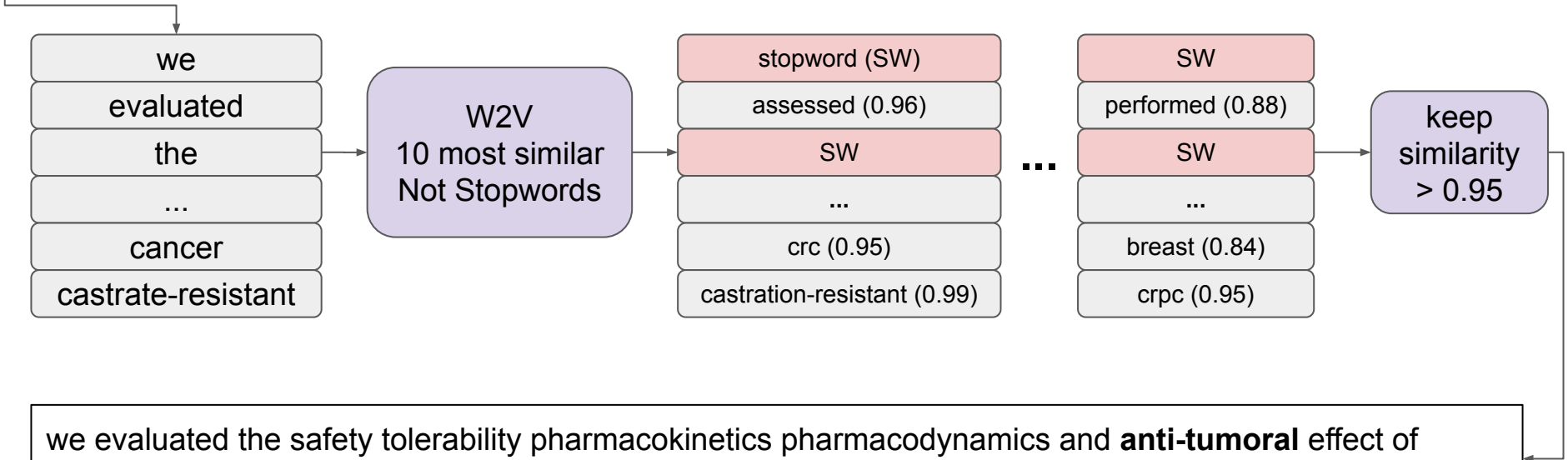
Google Translate (FR->ENG)

Orteronel was developed for treating which cancer?

Pappas et al., "Data Augmentation for Biomedical Factoid Question Answering", BIONLP 2022

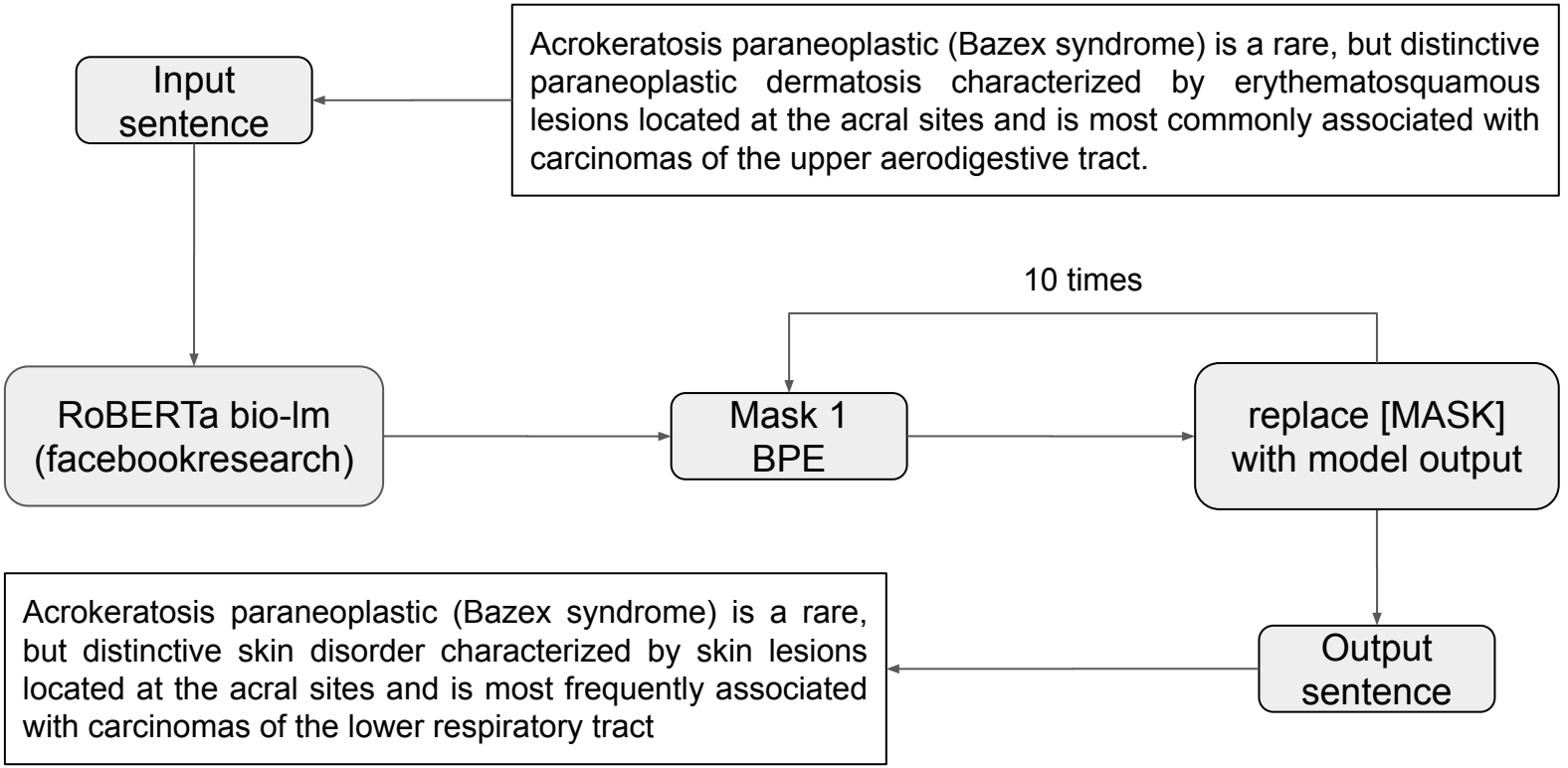
WE Embedding Substitution

we evaluated the safety tolerability pharmacokinetics pharmacodynamics and **antitumor** effect of orteronel with or without prednisolone in japanese patients with castration resistant prostate cancer **castrate-resistant**



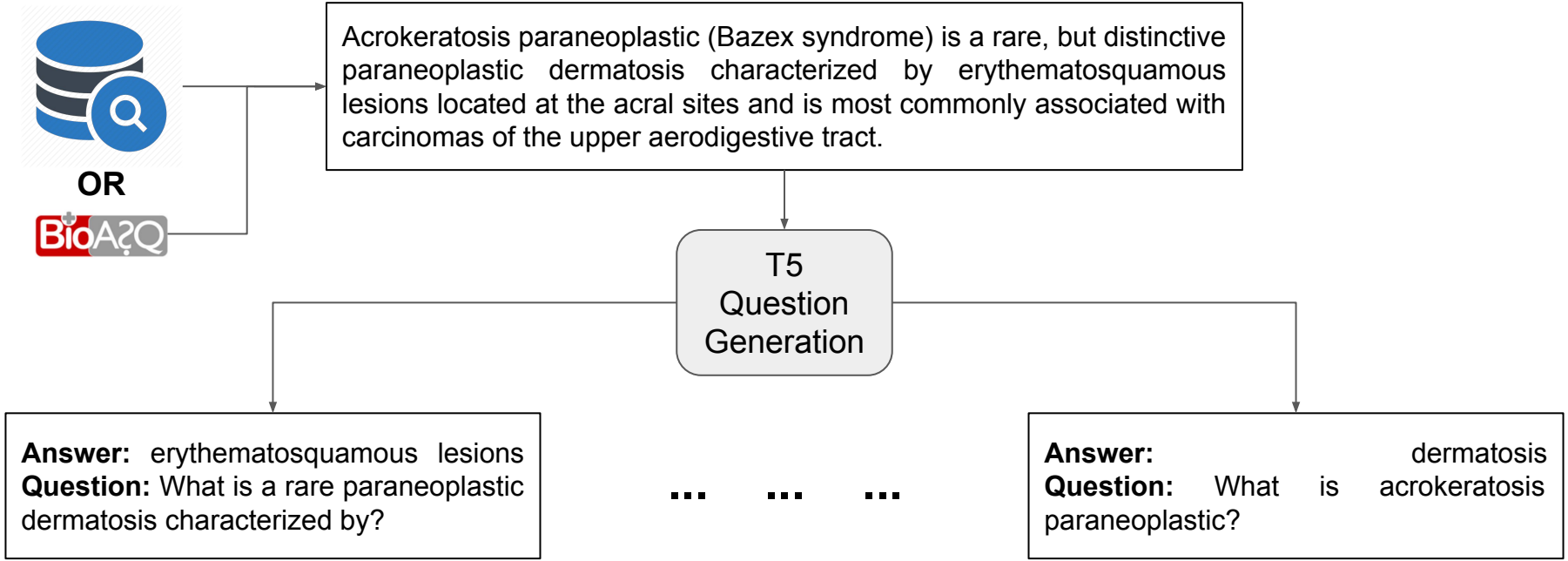
we evaluated the safety tolerability pharmacokinetics pharmacodynamics and **anti-tumoral** effect of orteronel with or without prednisolone in japanese patients with castration resistant prostate cancer **crpc**

Bert Masking



Pappas et al., "Data Augmentation for Biomedical Factoid Question Answering", BIONLP 2022

T5 Question Generation



https://github.com/biswa380/t5_multitask_api#usage

Pappas et al., "Data Augmentation for Biomedical Factoid Question Answering", BIONLP 2022

Context Increasing

Gold Data

Q: "Orteronel was developed for treatment of which cancer?"

S: “Orteronel is an investigational, partially selective inhibitor of CYP 17,20-lyase in the androgen signalling pathway, a validated therapeutic target for metastatic castration-resistant prostate cancer.”

A: “castration-resistant prostate cancer”

S1	Orteronel is an investigational, partially selective inhibitor of CYP 17,20-lyase in the androgen signalling pathway, a validated therapeutic target for metastatic castration-resistant prostate cancer.
S2	We assessed orteronel in chemotherapy-naive patients with metastatic castration-resistant prostate cancer.
S3	In this phase 3, double-blind, placebo-controlled trial, we recruited patients with progressive metastatic castration-resistant prostate cancer and no previous chemotherapy from 324 study centres in 43 countries.
...	...
Sk	On the basis of these and other data, orteronel is not undergoing further development in metastatic castration-resistant prostate cancer.

- Let gold instance be **(Q, A, S_m)**
- Create
  - **(Q, A, S_{m-2} + S_{m-1} + S_m)**
  - **(Q, A, S_{m-1} + S_m + S_{m+1})**
  - **(Q, A, S_m + S_{m+1} + S_{m+2})**

Pappas et al., “Data Augmentation for Biomedical Factoid Question Answering”, BIONLP 2022

Results

Method	+train ex.	PRAUC (dev)	PRAUC (test)
albert (squad-v2)	0	80.25	77.78
+ BIOASQ	2,848	89.57	76.78
+WORD2VEC +BIOASQ	**2,848+10,000**	**95.60 (+6.03)**	**84.99 (+8.21)**
+BIOLM +BIOASQ	2,848+50,000	94.45 (+4.88)	82.76 (+5.98)
+CONTEXT +BIOASQ	2,848+6,428 (ALL)	94.21 (+4.64)	81.63 (+4.85)
+BIOMRC +BIOASQ	2,848+10,000	93.15 (+3.58)	82.04 (+5.26)
+BTR +BIOASQ	2,848+15,593 (ALL)	92.66 (+3.09)	81.27 (+4.49)
+T5@PUBMED +BIOASQ	2,848+50.000	90.69 (+1.12)	80.26 (+3.48)
+IR +BIOASQ	2,848+289 (ALL)	89.80 (+0.23)	78.66 (+1.88)

Conclusion and Next Steps

- Conclusion
  - Seven data augmentation (DA) methods in factoid question answering
  - Data augmentation can lead to performance gains
  - WORD2VEC word substitution performed best
  - Code and Data available

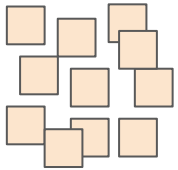
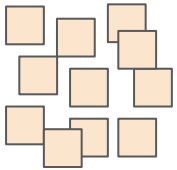
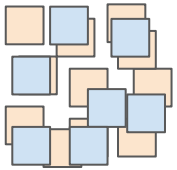
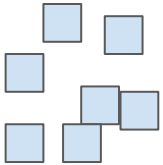
- Future Work
  - Online Data Augmentation
    - Active Learning
  - Other Domains and Datasets
  - BioASQ competition

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

@dvpappas, @AUEBNLPGroup

pappasd@aueb.gr

Pre-Train, FineTune and Combine

Name	PreTr	FineTune	Comb.
Step 1:Train			
Step 2: Train			
Step 3: Evaluate			

 Augmented data  BioASQ Train data  BioASQ Eval data

Data for Biomedical Reading Comprehension

DATASET	No of Instances	Type
BIOREAD	~ 16.4 M	Cloze-style
BMKC	~ 843 K	Cloze-style
BIOMRC	~ 812 K	Cloze-style
emrQA	~ 456 K	Ranking
PubmedQA	~ 273.5 K	yes/no/maybe
MedQA	~ 61 K	Cloze-style
MedQuAD	~ 47.5 K	Paragraphs
BioASQ FACTOID	**~ 2.8 K**	**Factoid**
COVID-QA	~ 2 K	Factoid (and snippets)

Pappas et al., "Data Augmentation for Biomedical Factoid Question Answering", BIONLP 2022

More on BIO etc.

When we use Sigmoid and BCE every score is computed independently to the others

When we use softmax on columns we can only detect one answer

HOW WE REPRESENT THE GOLD TRUTH
(The target of the model)

		IO SIGMOID BCE	IO SOFTMAX ON ROWS CROSS ENTROPY		BIO SOFTMAX ON ROWS CROSS ENTROPY			BIOES SOFTMAX ON ROWS CROSS ENTROPY					BE SOFTMAX ON COLUMNS CROSS ENTROPY		BE SOFTMAX CROSS ENTROPY			BE SIGMOID BCE	
	[CLS]	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	ποιό	1	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0
	ενέσιμο	1	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
	διάλυμα	1	1	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0	1
	αντι	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	##μετωπίζει	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	άμεσα	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	την	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	χοληστερίνη	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	[SEP]	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	το	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
This is a multi-token answer	repatha	1	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0
	solution	1	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
	injection	1	1	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0	1
	αντι	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	##μετωπίζει	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	την	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	χοληστερίνη	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	διότι	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	το	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
This is a singleton	repatha	1	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0
	περιέχει	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	εβολοκου	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	##μάμπη	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
	[SEP]																		