Macquarie University participation in the BioASQ 2021 challenges

Diego Mollá Urvashi Khanna

Department of Computing,
Macquarie University,
Sydney

LTG, 24 May 2021

Contents

Contents

BioASQ Synergy on Biomedical Semantic QA for COVID-19

(This text is based on BioASQ's page: `http://bioasq.org/`)

- In this task, biomedical experts pose unanswered questions for the developing problem of COVID-19.
- Participating systems are required to provide answers, which will in turn be assessed by the experts and fed back to the systems, together with updated questions.
- This task involves IR, QA, summarization and more on an continuously expanding dataset of documents for COVID-19.
- Through this process, this task aims to facilitate the incremental understanding of COVID-19 and contribute to the discovery of new solutions.

BioASQ Task b on Biomedical Semantic QA

(This text is based on BioASQ's page: `http://bioasq.org/`)

- This task uses benchmark datasets containing development and test questions, in English, along with gold standard (reference) answers constructed by a team of biomedical experts.
- The participants have to respond with relevant concepts, articles, snippets and RDF triples, from designated resources, as well as exact and 'ideal' answers.

BioASQ Synergy — Required information

Given a question, return the following information:

- A list of at most 10 relevant documents.
- A list of at most 10 relevant snippets.
- For each question designated as "ready to answer", return an ideal answer and exact answers.
  - Ideal answer: a paragraph-sized summary.
  - Exact answer: depends on the question type.
    - yesno: "yes" or "no".
    - factoid: a list of up to 5 entity names, ordered by decreasing confidence.
    - list: a list of up to 100 entries.
    - summary: N/A.

Incremental Approach

- The same question may be asked in multiple rounds.
- The data set (CORD-19) may change in each round (because new publications are added).
- The feedback given from past rounds can be incorporated in the new rounds.
- If a document or snippet has been assessed for the same question in a past round, it should not be submitted in the next round.
- Answers can be repeated in subsequent rounds, even if they have been assessed in past rounds. This is because the correct answer might change as new evidence is found.

BioASQ Task B Phase B — Required information

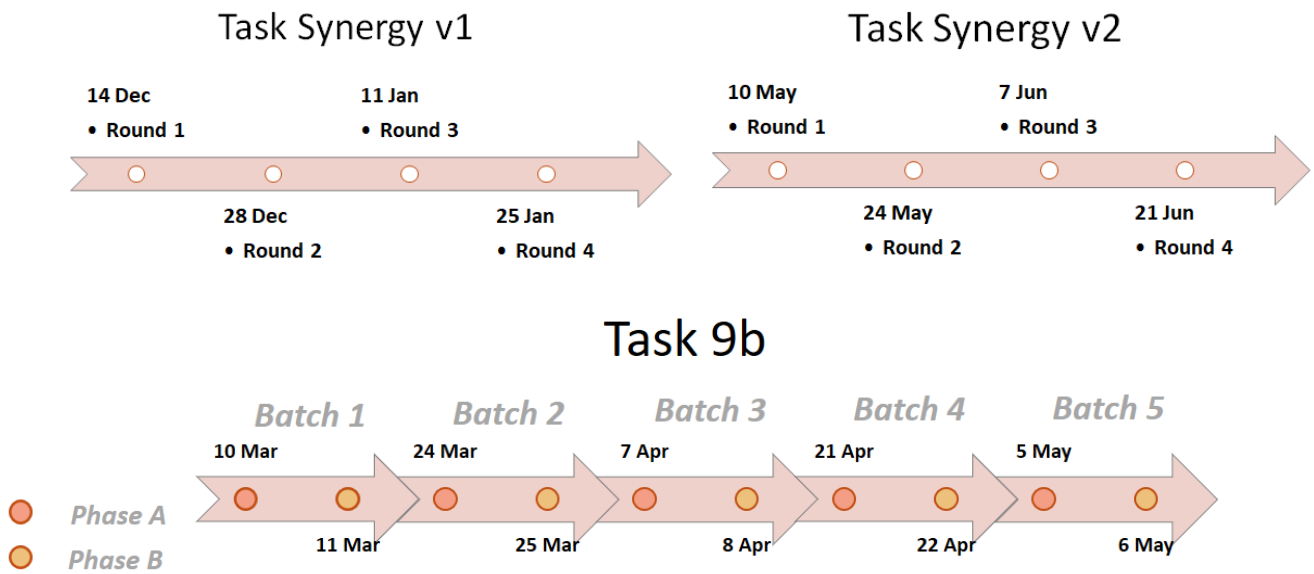
Given

- A question and question type
- A list of relevant documents
- A list of relevant snippets
- A list of concepts and RDF triples (but we don't use these)

Return

- Ideal answer: a paragraph-sized summary.
- Exact answer: depends on the question type.
  - yesno: "yes" or "no".
  - factoid: a list of up to 5 entity names, ordered by decreasing confidence.
  - list: a list of up to 100 entries.
  - summary: N/A.

Schedule


Task Synergy v1
14 Dec
• Round 1
28 Dec
• Round 2
11 Jan
• Round 3
25 Jan
• Round 4
Task Synergy v2
10 May
• Round 1
24 May
• Round 2
7 Jun
• Round 3
21 Jun
• Round 4
Task 9b
Batch 1
Batch 2
Batch 3
Batch 4
Batch 5
10 Mar
11 Mar
24 Mar
25 Mar
7 Apr
8 Apr
21 Apr
22 Apr
5 May
6 May
Phase A
Phase B

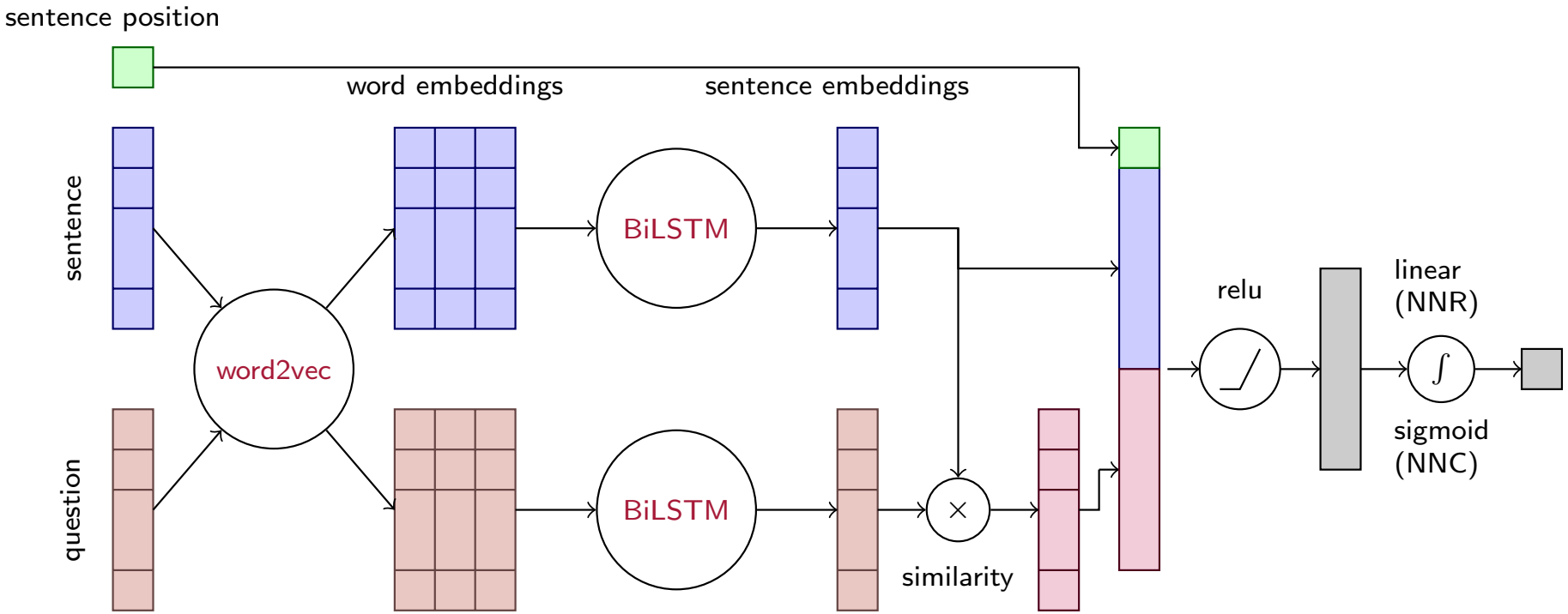
Contents

MQ's Approach to Synergy Task

- Use the Synergy API to find documents.
  - Also, a run uses sBERT tuned for retrieval.
- Use simple, unsupervised approaches to retrieve the snippets.
- Use MQ's QA entry system for BioASQ8b, trained on BioASQ8b data.
  - We did not train or fine-tune on covid-19.

MQ's participation at BioASQ8b

The system takes as input a question and a list of relevant snippets (not a list of relevant documents).



Document Retrieval

1. Get the top n documents from BioASQ's API.

   (increasing n each round because an increasing large of documents need to be removed)
   - Round 1: $n = 50$
   - Round 2: $n = 100$
   - Round 3: $n = 150$
   - Round 4: $n = 200$
2. Round 4: Runs MQ-4 and MQ-5 used sBERT for document retrieval.
   - MQ-4: Tuned with TREC data (to be confirmed).
   - MQ-5: Tuned with TREC and BioASQ data (to be confirmed).
3. Remove documents that appear in past feedback.
4. Return the top 10 documents.

Document Retrieval

1. Get the top n documents from BioASQ's API.

   (increasing n each round because an increasing large of documents need to be removed)
   - Round 1: $n = 50$
   - Round 2: $n = 100$
   - Round 3: $n = 150$
   - Round 4: $n = 200$
2. Round 4: Runs MQ-4 and MQ-5 used sBERT for document retrieval.
   - MQ-4: Tuned with TREC data (to be confirmed).
   - MQ-5: Tuned with TREC and BioASQ data (to be confirmed).
3. Remove documents that appear in past feedback.
4. Return the top 10 documents.

Snippet Retrieval

Run MQ1

1. Get the top *n* documents from BioASQ's API.
   - Round 4, runs MQ3–MQ4: used sBERT for document retrieval.
2. Remove documents annotated as negatives in past feedback.
3. Single document summarisation: extract top 3 sentences for each document. Each sentence will be a snippet.
   - Encoding: tf.idf
   - Scoring: cosine similarity with the query.
   - Sentences are returned in order of appearance.
4. Rounds 2–3, runs MQ3–MQ4: input passages obtained via sBERT tuned for passage retrieval (TBC).
5. Rounds 1–3, runs MQ2–MQ4: Re-rank the sentences using BioASQ8b's LSTM model.
6. Remove sentences (that is, snippets) listed in the feedback.
7. Return the first 10 sentences.

Snippet Retrieval

Runs MQ2-4

1. Get the top *n* documents from BioASQ's API.
   - Round 4, runs MQ3–MQ4: used sBERT for document retrieval.
2. Remove documents annotated as negatives in past feedback.
3. Single document summarisation: extract top 3 sentences for each document. Each sentence will be a snippet.
   - Encoding: tf.idf
   - Scoring: cosine similarity with the query.
   - Sentences are returned in order of appearance.
4. Rounds 2–3, runs MQ3–MQ4: input passages obtained via sBERT tuned for passage retrieval (TBC).
5. Rounds 1–3, runs MQ2–MQ4: Re-rank the sentences using BioASQ8b's LSTM model.
6. Remove sentences (that is, snippets) listed in the feedback.
7. Return the first 10 sentences.

Snippet Retrieval

1. Get the top *n* documents from BioASQ's API.
   - Round 4, runs MQ3–MQ4: used sBERT for document retrieval.
2. Remove documents annotated as negatives in past feedback.
3. Single document summarisation: extract top 3 sentences for each document. Each sentence will be a snippet.
   - Encoding: tf.idf
   - Scoring: cosine similarity with the query.
   - Sentences are returned in order of appearance.
4. Rounds 2–3, runs MQ3–MQ4: input passages obtained via sBERT tuned for passage retrieval (TBC).
5. Rounds 1–3, runs MQ2–MQ4: Re-rank the sentences using BioASQ8b's LSTM model.
6. Remove sentences (that is, snippets) listed in the feedback.
7. Return the first 10 sentences.

Question Answering

Runs MQ1 and MQ2

1. Get the top *n* documents from BioASQ's API.
2. Remove documents annotated as negatives in past feedback.
3. Single document summarisation: extract top 3 sentences for each document.
   - Encoding: tf.idf
   - Scoring: cosine similarity with the query.
   - Snippets are returned in order of appearance.
4. Re-rank the sentences using BioASQ8b's LSTM model.
   - The number of sentences depends on the question type.
   - Sentences are returned in order of appearance.

Question Answering

Rounds 1–3, runs MQ3 and MQ4

1. Obtain passage embeddings using sentence transformers fine-tuned for passage retrieval.
2. Return top 3 passages using cosine similarity.
   - A passage may contain multiple sentences.
3. Split passages into sentences.
4. Re-rank the sentences using BioASQ8b's LSTM model.
   - The number of sentences depends on the question type.
   - Sentences are returned in order of appearance.

Question Answering

Round 4, runs MQ4 and MQ5

1. Get the top *n* documents using sBERT for document retrieval.
2. Remove documents annotated as negatives in past feedback.
3. Single document summarisation: extract top 3 sentences for each document.
   - Encoding: tf.idf
   - Scoring: cosine similarity with the query.
   - Snippets are returned in order of appearance.
4. Re-rank the sentences using BioASQ8b's LSTM model.
   - The number of sentences depends on the question type.
   - Sentences are returned in order of appearance.

Preliminary Evaluation

Based on the feedback returned by BioASQ

Run	Doc. Prec.	Snip. Prec.	Ideal A. SU4
Round 1 MQ 1	0.219	0.134	–
Round 1 MQ 2	0.219	0.132	–
Round 2 MQ 1	0.149	0.061	0.241
Round 2 MQ 2	0.149	0.061	0.247
Round 2 MQ 3	0.149	0.061	0.141
Round 2 MQ 4	0.149	0.058	0.154
Round 3 MQ 1	0.087	0.056	0.213
Round 3 MQ 2	0.087	0.060	0.238
Round 3 MQ 3	0.087	0.059	0.096

- Doc. Prec. : Document precision.
- Snip. Prec. : Snippet precision.
- Ideal A. SU4: ROUGE SU4-F1 of ideal answer.
- Feedback for ideal answers in 18 questions only (round 3: 39).

Submission Results – Document F1

Run	Round 1	Round 2	Round 3	Round 4
Best	0.3457	0.3237	0.2628	0.2375
Median	0.2474	0.2387	0.1810	0.1839
Worst	0.0802	0.0560	0.0179	0.0168
MQ-1	0.2474	0.1654	0.0973	0.1053
MQ-2	0.2474	0.1654	0.0973	0.1053
MQ-3		0.1654	0.0973	0.1053
MQ-4		0.1654		0.1510
MQ-5				0.1762

Red: sBERT tuned for retrieval

Submission Results – Snippets F1

Run	Round 1	Round 2	Round 3	Round 4
Best	0.2712	0.1885	0.2026	0.1909
Median	0.2021	0.1634	0.1645	0.1461
Worst	0.0396	0.0204	0.0037	0.0078
MQ-1	0.1414	0.0704	0.0462	0.0640
MQ-2	0.1380	0.0706	0.0462	0.0657
MQ-3		0.0709	0.0473	0.0634
MQ-4		0.0695		0.0798
MQ-5				0.0912

Red: sBERT tuned for document retrieval

Dark red: sBERT tuned for passage retrieval

Submission Results – Ideal Answers F1

Run	Round 1	Round 2	Round 3	Round 4
Best		0.0749	0.1170	0.1254
Median		0.0565	0.0883	0.0857
Worst		0.0096	0.0181	0.0221
MQ-1		0.0567	0.0883	0.0971
MQ-2		0.0565	0.0926	0.0912
MQ-3		0.0436	0.0467	0.0515
MQ-4		0.0500		0.0857
MQ-5				0.0757

Red: sBERT tuned for document retrieval

Dark red: sBERT tuned for passage retrieval

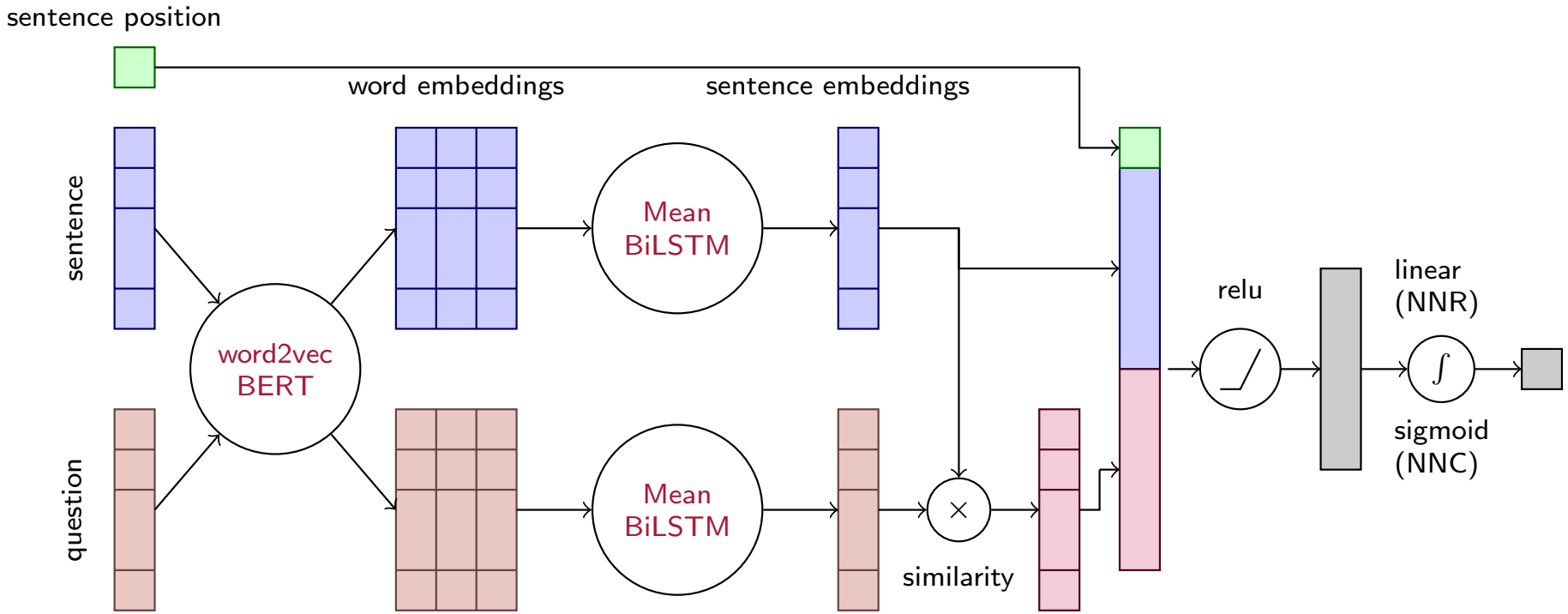
Contents

What's different between BioASQ8b and BioASQ9b runs?

- Question and sentence are combined as input to same BERT model.
- We let BERT model the interaction between question and sentence.

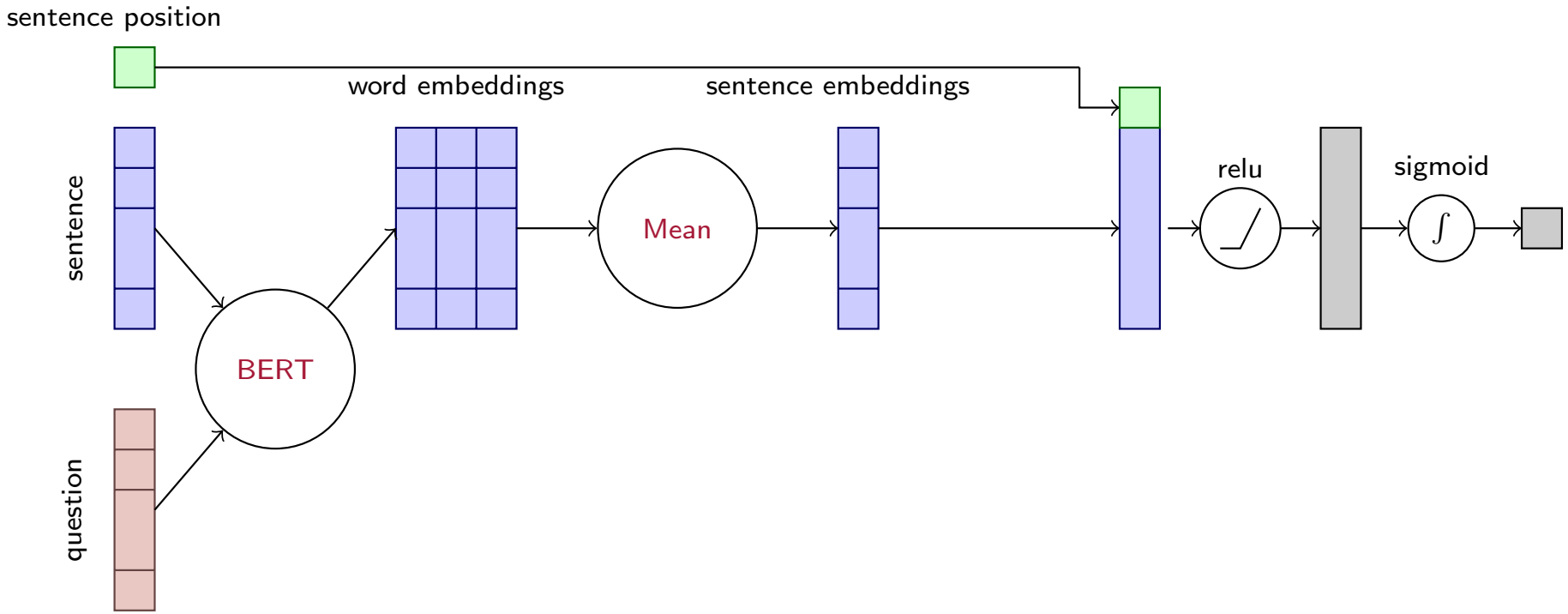
MQ's participation at BioASQ8b

The system takes as input a question and a list of relevant snippets (not a list of relevant documents).



MQ's participation at BioASQ9b

The system takes as input a question and a list of relevant snippets (not a list of relevant documents).



Adding Question and Sentence to BERT

- We use the standard approach to encode question and sentence as original BERT for QA-SQuAD (Devlin et al, 2019).
- We then mask the output embeddings of the question before computing the average of embeddings.

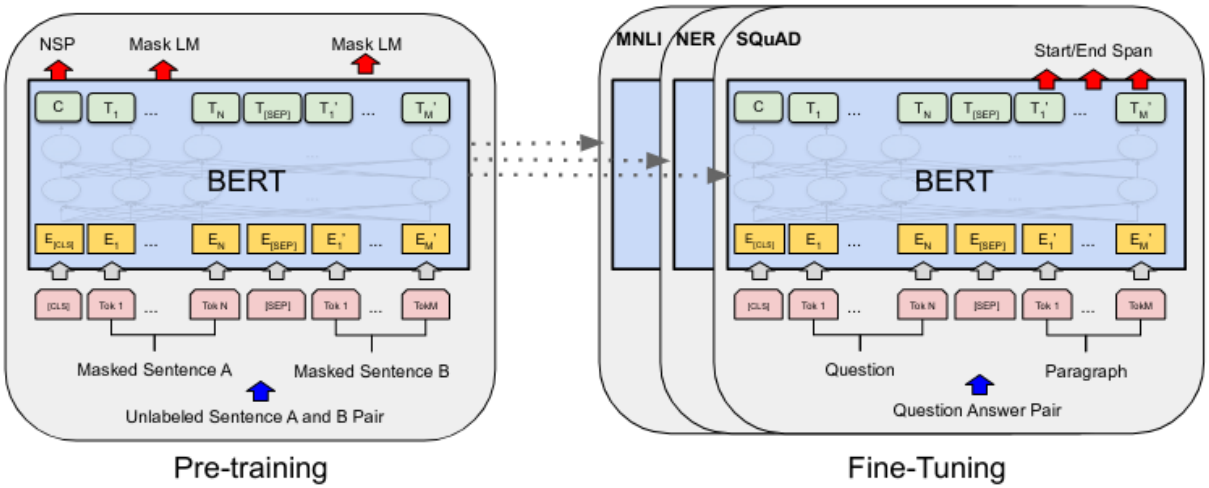


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

Results of 10-fold Cross-Validation

System	Number of Parameters Full	Trained	Epochs	Dropout	SU4-F1
BERT	109,520,791	38,551	8	0.8	0.2779
BioBERT	108,348,823	38,551	1	0.7	0.2798
DistilBERT	66,401,431	38,551	1	0.6	0.2761
ALBERT	222,800,535	204,951	5	0.5	0.2866
ALBERT-SQuAD2	222,800,535	204,951	5	0.7	0.2846
ALBERT-QA	222,800,535	204,951	5	0.4	**0.2875**

- BERT, BioBERT, DistilBERT: base model, embeddings size 768
- ALBERT variants: xxlarge model, embeddings size 4096

Submission — Preliminary Results

		ROUGE-SU4				
Run	System	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
Best		0.3410	0.3974	0.3266	0.4402	0.3893
Median		0.2536	0.1990	0.2647	0.3388	0.2666
Worst		0.1154	0.1186	0.1017	0.0886	0.1331
MQ-1	BERT	0.3032	0.3560	0.3057	0.3585	0.3511
MQ-2	BioBERT	0.3103	0.3615	0.3265	0.3612	**0.3733**
MQ-3	DistilBERT	0.3007	**0.3753**	0.3204	**0.3681**	0.3711
MQ-4	ALBERT	**0.3205**	0.3676	0.3100	0.3560	0.3570
MQ-5	ALBERT-QA		0.3610	**0.3266**	0.3559	0.3589

Conclusions

Synergy

- QA trained on BioASQ8b (relatively?) robust despite poor input snippets.
- Further work: find a better way to re-rank snippets.

BioASQ

- Straightforward BERT no worse than more specialised architectures.
- BioBERT better than BERT.
- All models have similar performance.
- Fine-tuning BERT based on exact QA answers seems to help.

Questions?

Conclusions

Synergy

- QA trained on BioASQ8b (relatively?) robust despite poor input snippets.
- Further work: find a better way to re-rank snippets.

BioASQ

- Straightforward BERT no worse than more specialised architectures.
- BioBERT better than BERT.
- All models have similar performance.
- Fine-tuning BERT based on exact QA answers seems to help.

Questions?

Conclusions

Synergy

- QA trained on BioASQ8b (relatively?) robust despite poor input snippets.
- Further work: find a better way to re-rank snippets.

BioASQ

- Straightforward BERT no worse than more specialised architectures.
- BioBERT better than BERT.
- All models have similar performance.
- Fine-tuning BERT based on exact QA answers seems to help.

Questions?