Hierarchical Attention Flow for Multiple-choice Reading Comprehension

Haichao Zhu[1], *Furu Wei*[2], *Bing Qin*[1], *Ting Liu*[1]

[1]Harbin Institute of Technology

[2]Microsoft Research

Multiple-choice Reading Comprehension

Passage:
... In 1993, New York State ordered stores to charge money on beverage containers. Within a year, consumers had returned millions of aluminum cans and glass and plastic bottles. Plenty of companies were eager to accept the aluminum and glass as raw material for new products, but because few could figure out what to do with the plastic, much of it wound end up ...

Multiple-choice Reading Comprehension

Passage:
... In 1993, New York State ordered stores to charge money on beverage containers. Within a year, consumers had returned millions of aluminum cans and glass and plastic bottles. Plenty of companies were eager to accept the aluminum and glass as raw material for new products, but because few could figure out what to do with the plastic, much of it wound end up ...

Question:
What regulation was issued by New York State concerning beverage containers?

Multiple-choice Reading Comprehension

Passage:
... In 1993, New York State ordered stores to charge money on beverage containers. Within a year, consumers had returned millions of aluminum cans and glass and plastic bottles. Plenty of companies were eager to accept the aluminum and glass as raw material for new products, but because few could figure out what to do with the plastic, much of it wound end up ...

Question:
What regulation was issued by New York State concerning beverage containers?

Options:

A. A fee should be charged on used containers for recycling.
B. Throwaways should be collected by the state for recycling.
C. Consumers had to pay for beverage containers and could get their money back on returning them.
D. Beverage companies should be responsible for collecting and reusing discarded plastic soda bottles.

Models Adapted for Multiple-choice RC

- Models Proposed on Cloze-style RC
  - Stanford Attentive Reader [Chen et al. 2016]
  - Gated-Attention Reader [Dhingra et al. 2017]

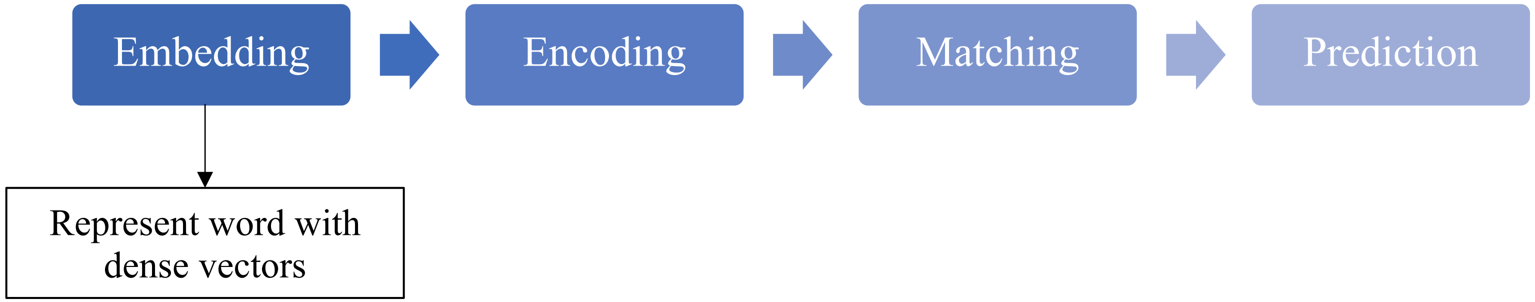
- Neural Pipeline

Embedding → Encoding → Matching → Prediction

Models Adapted for Multiple-choice RC

- Models Proposed on Cloze-style RC
  - Stanford Attentive Reader [Chen et al. 2016]
  - Gated-Attention Reader [Dhingra et al. 2017]

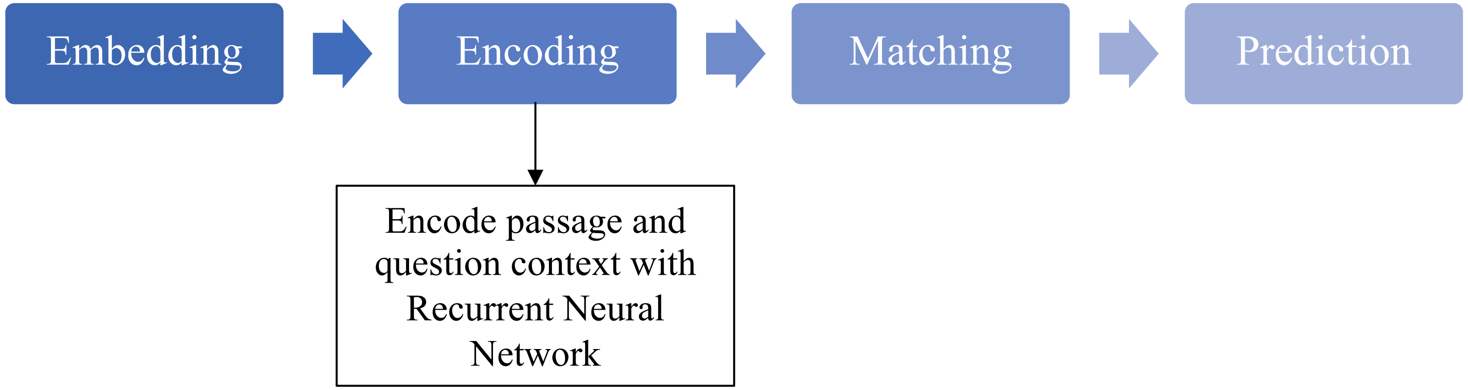
- Neural Pipeline



Models Adapted for Multiple-choice RC

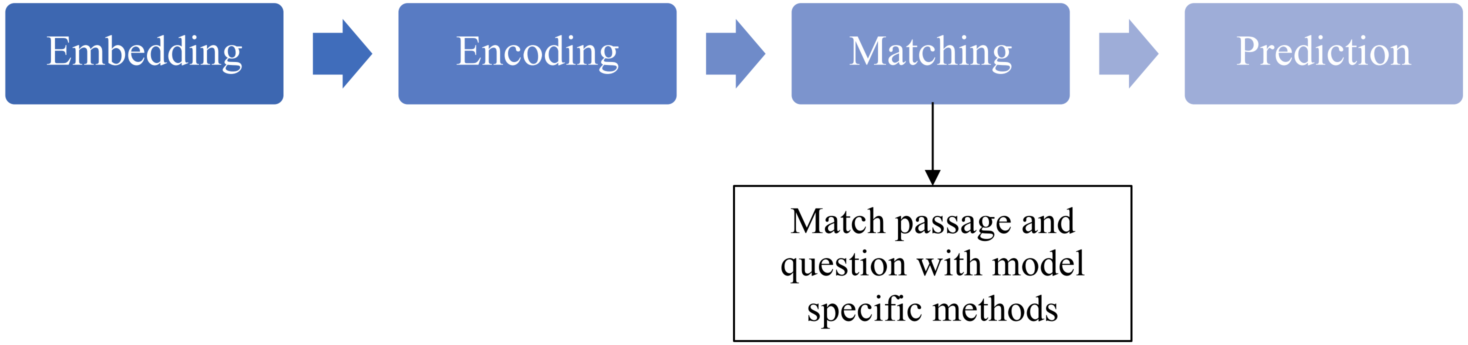
- Models Proposed on Cloze-style RC
  - Stanford Attentive Reader [Chen et al. 2016]
  - Gated-Attention Reader [Dhingra et al. 2017]

- Neural Pipeline



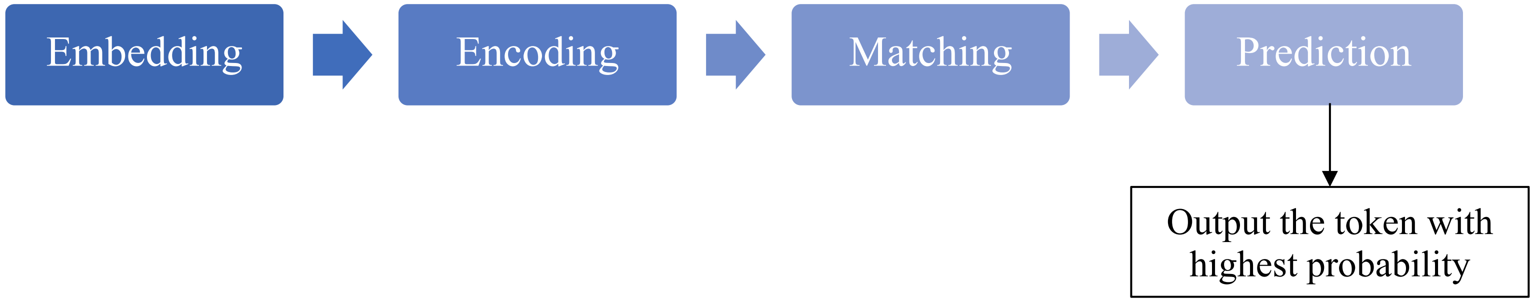
Models Adapted for Multiple-choice RC

- Models Proposed on Cloze-style RC
  - Stanford Attentive Reader [Chen et al. 2016]
  - Gated-Attention Reader [Dhingra et al. 2017]
- Neural Pipeline

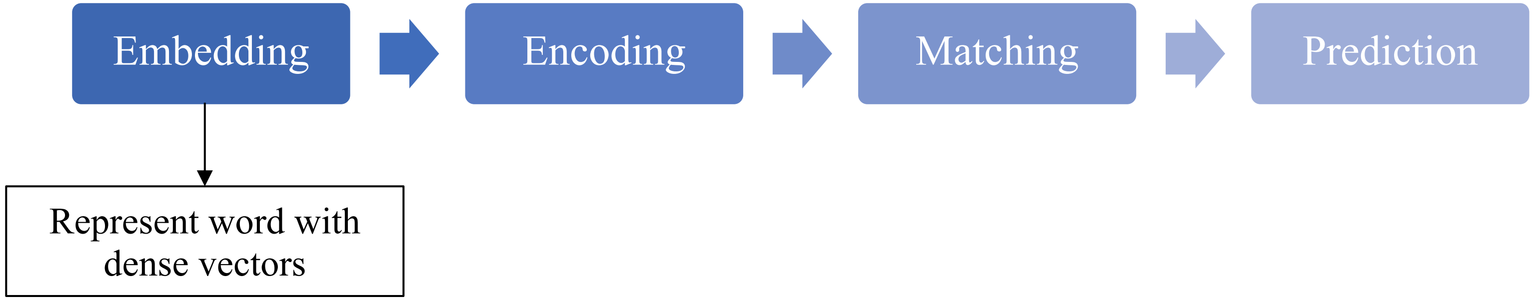


Models Adapted for Multiple-choice RC

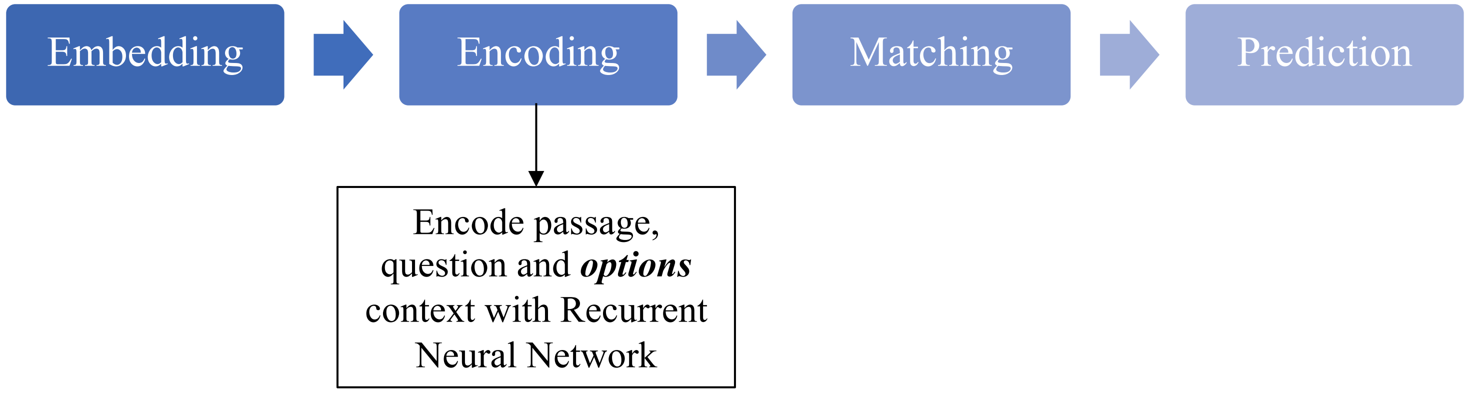
- Models Proposed on Cloze-style RC
  - Stanford Attentive Reader [Chen et al. 2016]
  - Gated-Attention Reader [Dhingra et al. 2017]
- Neural Pipeline



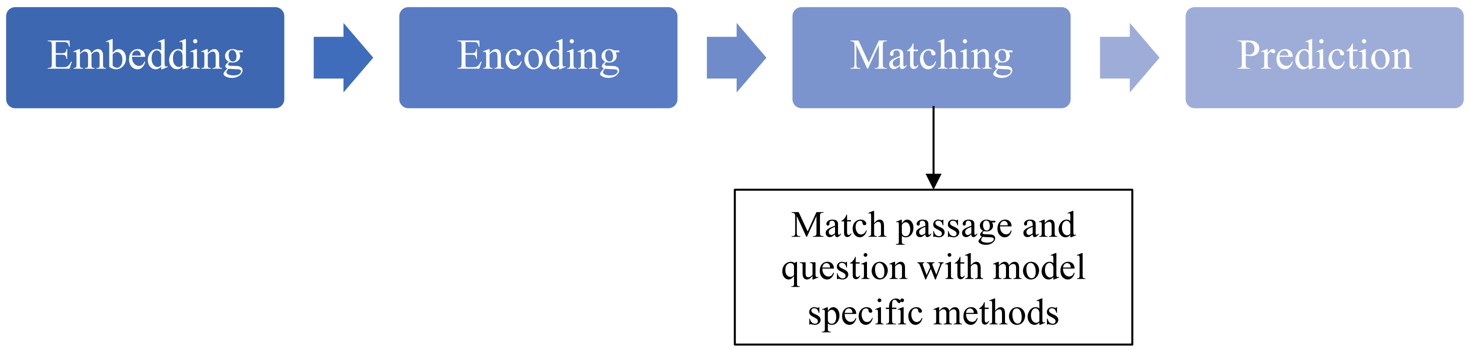
Models Adapted for Multiple-choice RC



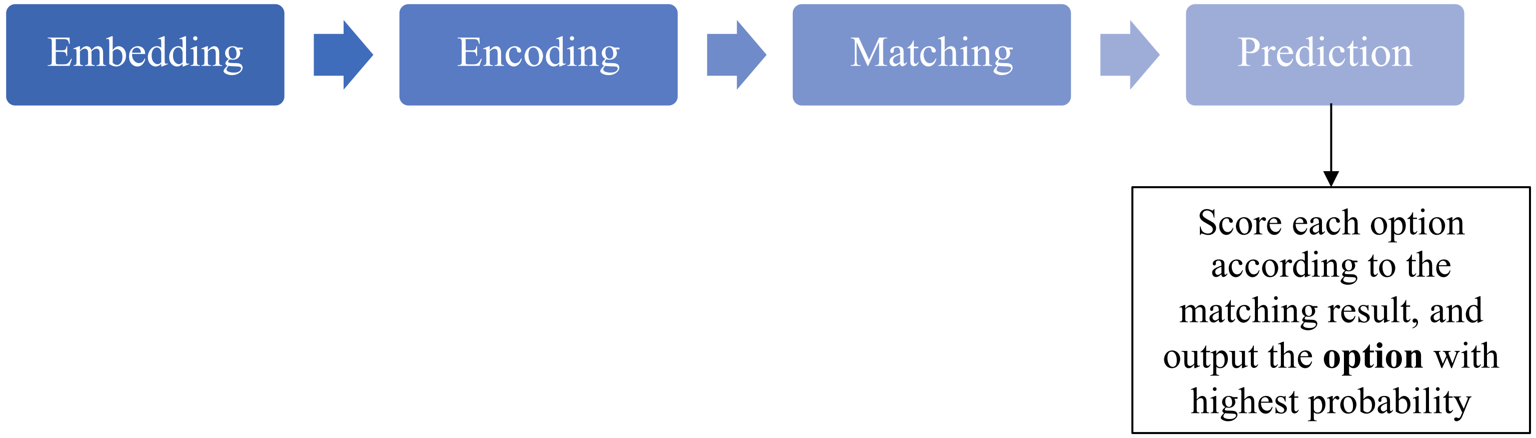
Models Adapted for Multiple-choice RC



Models Adapted for Multiple-choice RC

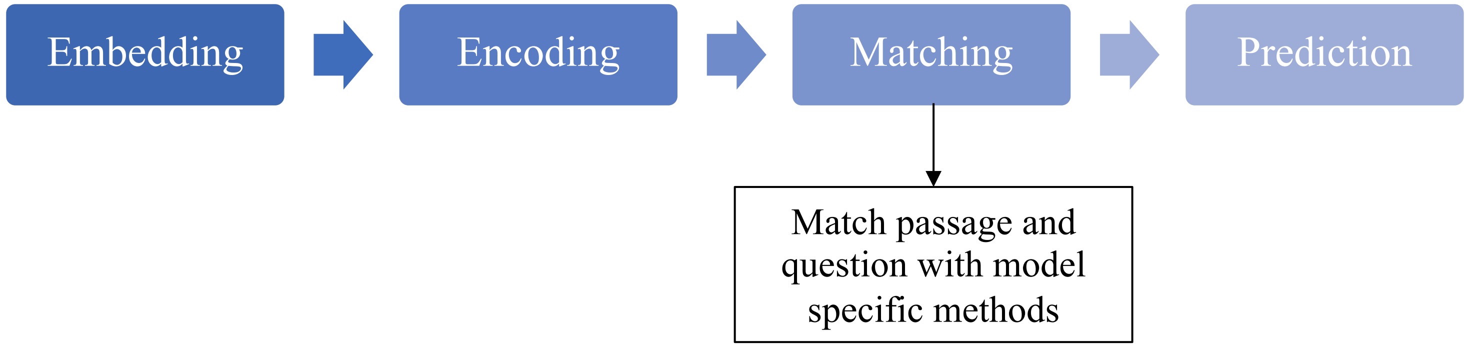


Models Adapted for Multiple-choice RC



Models Adapted for Multiple-choice RC

- Adapted for Multiple-choice RC [Lai et al. 2017]
  - Incorporating candidate options into pipeline



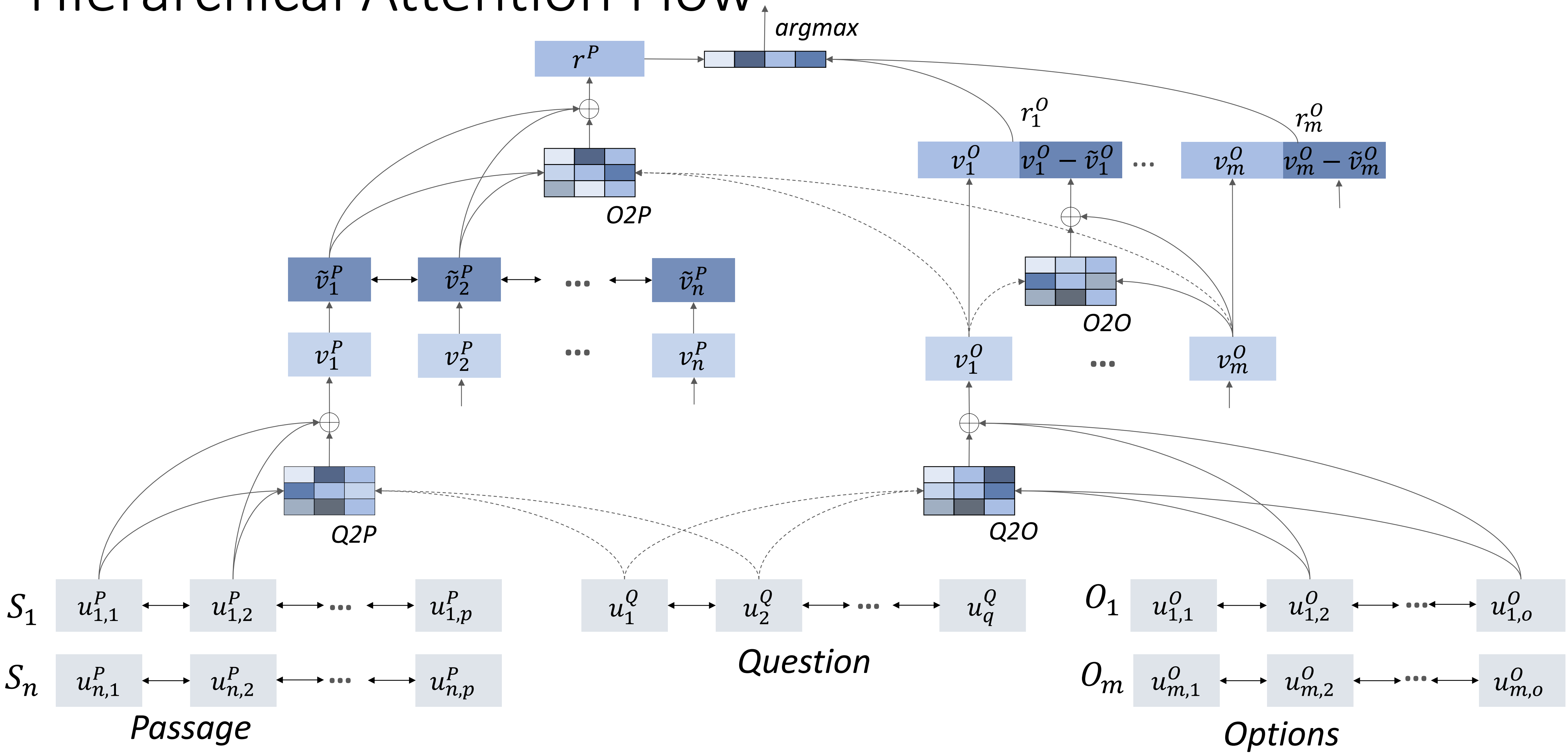
  - **Candidate options are not fully utilized!**

Motivation

- We want to incorporate candidate options into **evidence gathering** (matching procedure).
  - Questions like:
    - *It can be concluded from the passage that _ .*
- Human beings **compare options** with each other before choosing the candidate answer.

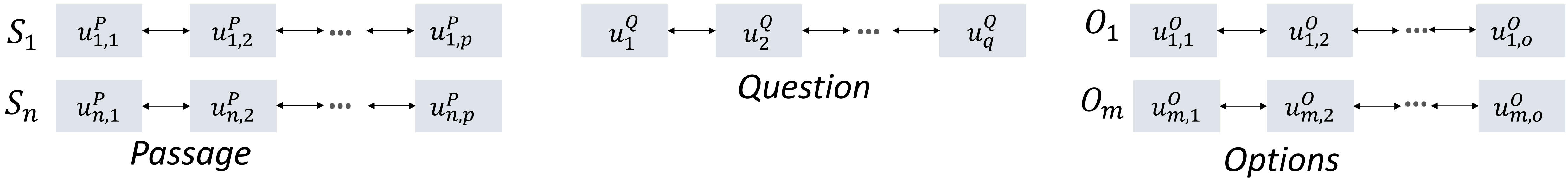
- Passages are document-level texts with **hierarchical structure**, which should be considered.
  - word → phrase → sentence → paragraph → document

Hierarchical Attention Flow



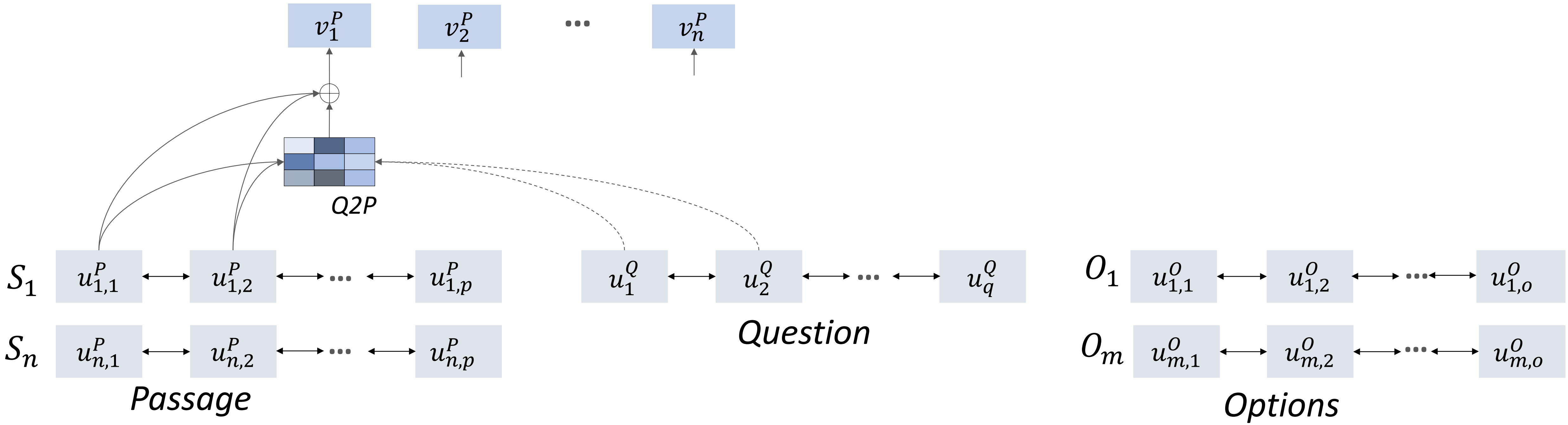
Hierarchical Attention Flow

We first encode **word context** within passage sentences, question and options separately with bidirectional RNN.



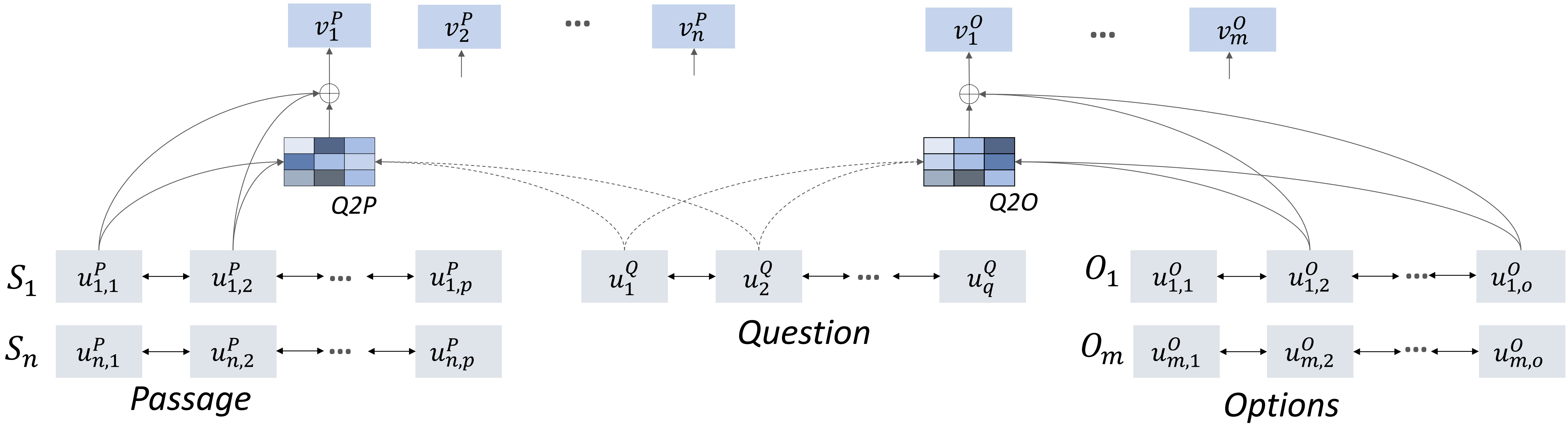
Hierarchical Attention Flow

Through **question to passage word-level** attention, we summarize a passage sentence into vector representation $\boldsymbol{v_i^P}$



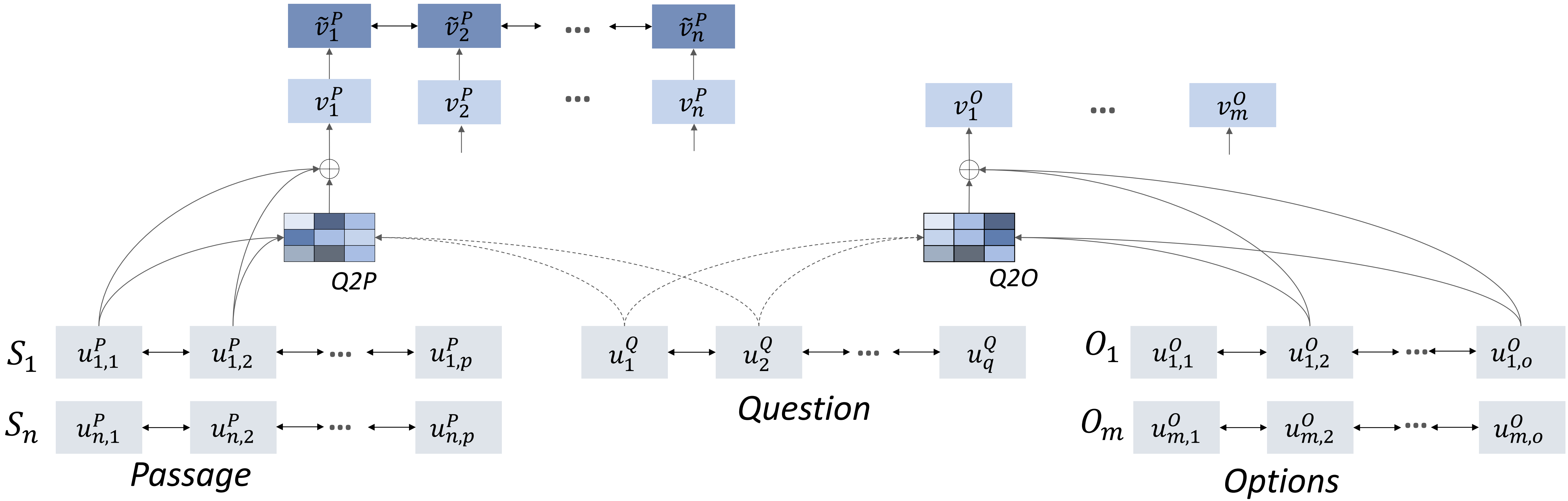
Hierarchical Attention Flow

Similarly, we get the vector representation $\boldsymbol{v_i^O}$ of the i-th candidate option via **question to option word-level** attention



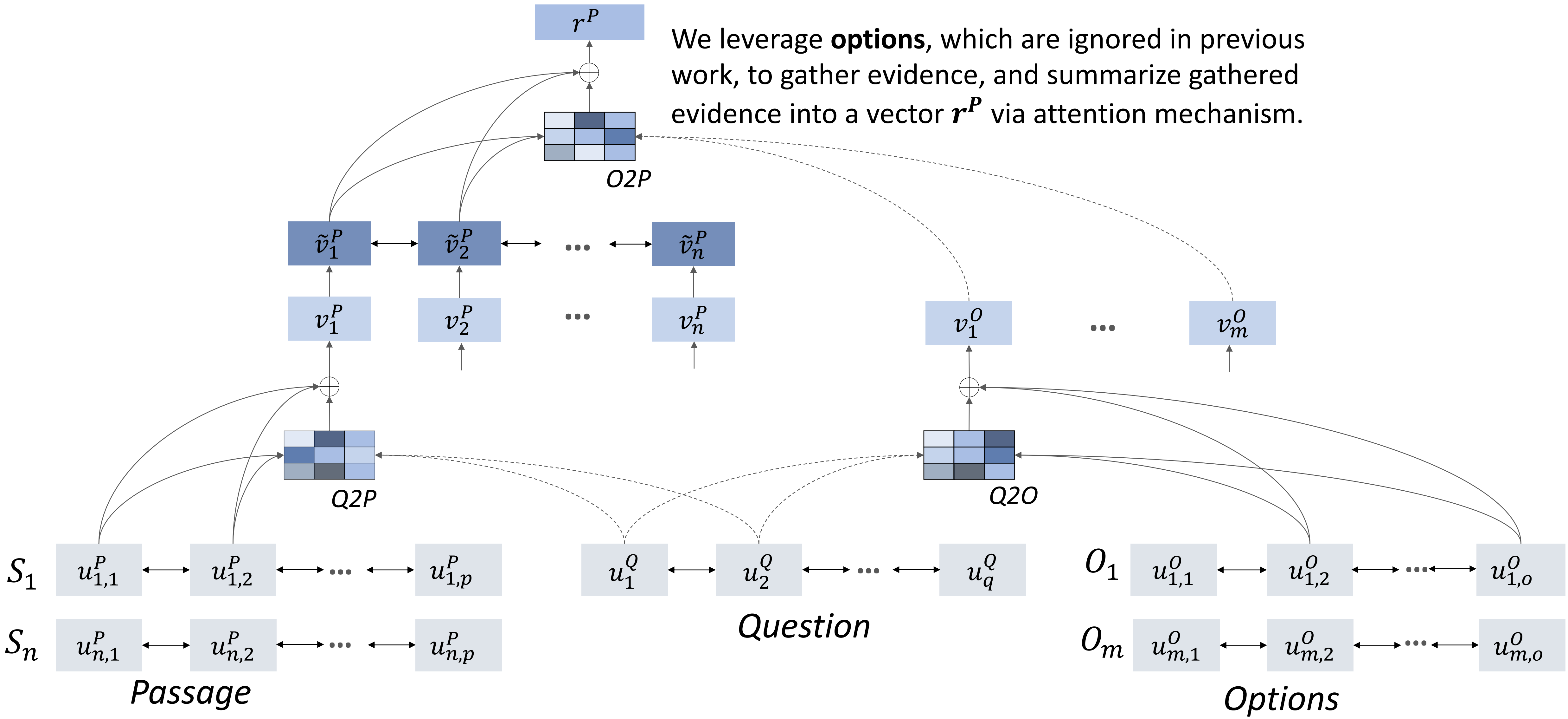
Hierarchical Attention Flow

Another bi-RNN is used to encode **sentence context** of passage sentences.



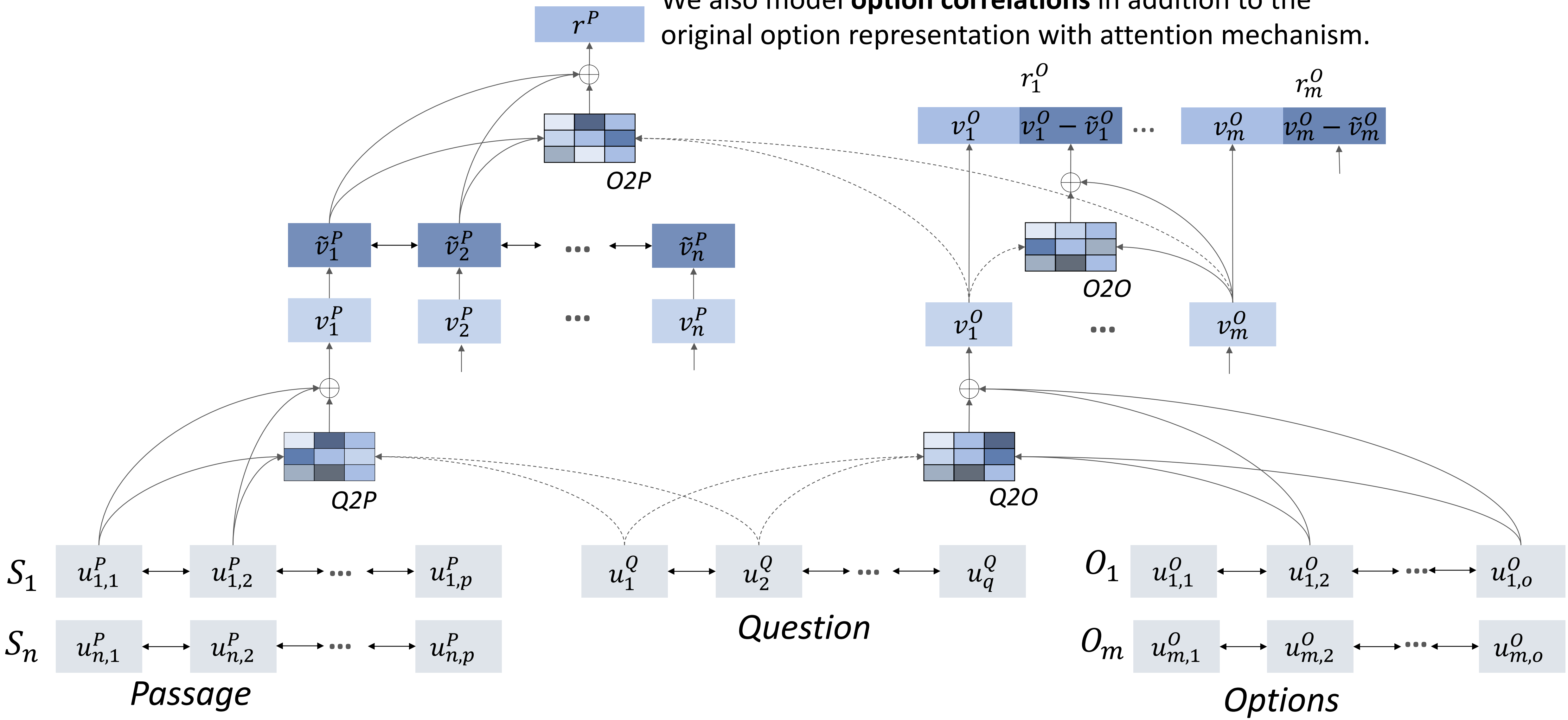
Hierarchical Attention Flow

We leverage **options**, which are ignored in previous work, to gather evidence, and summarize gathered evidence into a vector $\boldsymbol{r^P}$ via attention mechanism.



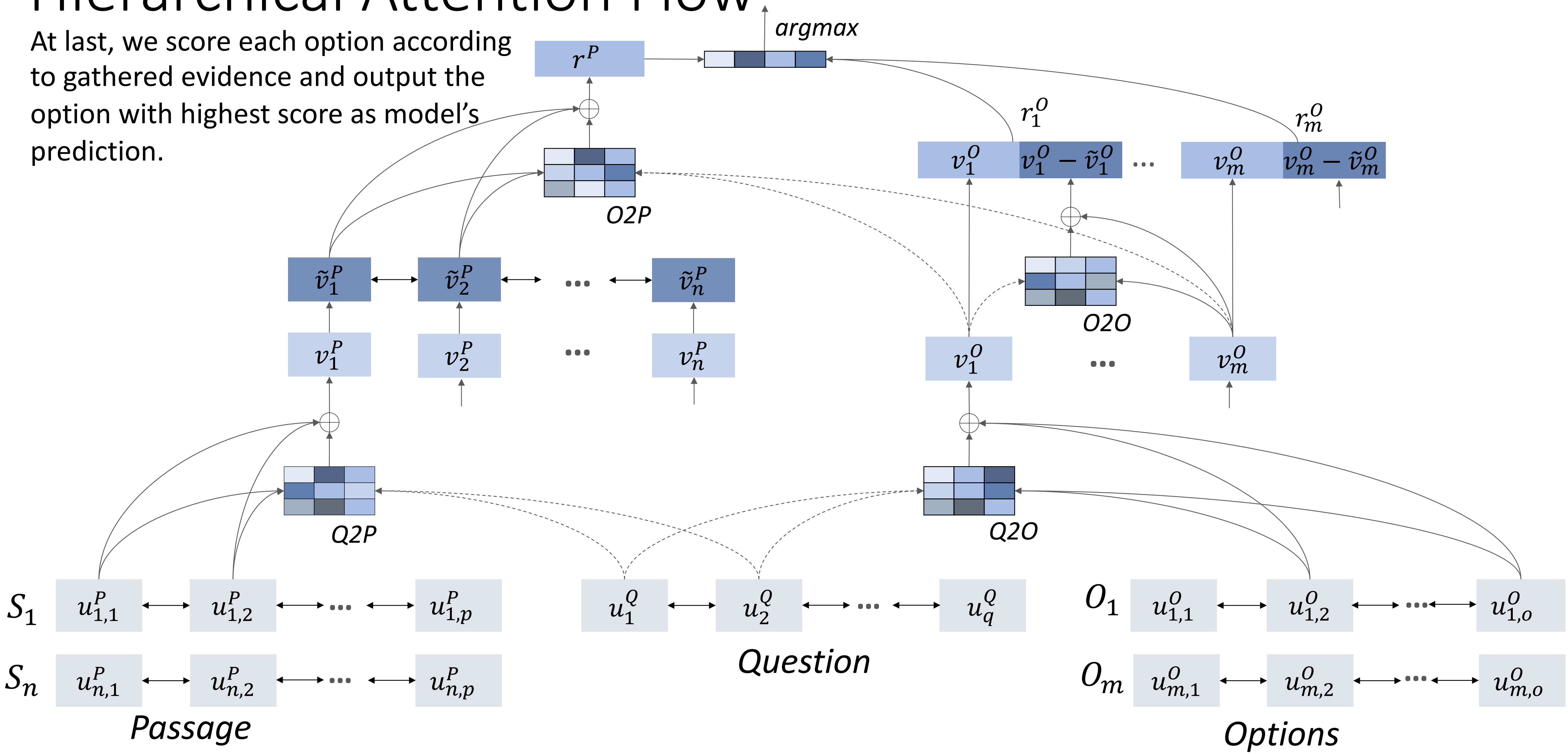
Hierarchical Attention Flow

We also model **option correlations** in addition to the original option representation with attention mechanism.



Hierarchical Attention Flow

At last, we score each option according to gathered evidence and output the option with highest score as model's prediction.



Multiple-choice RC Dataset

- Large-scale ReAding Comprehension Dataset From Examinations (**RACE**)
  - Lai et al. 2017 *EMNLP*
  - **4** candidate options for each question
  - RACE-**M** & RACE-**H** subsets

	RACE-M	RACE-H	RACE
#w/p	249.9	374.9	342.9
#s/p	17.2	19.2	18.7
#w/q	10.1	11.4	11.0
#w/o	4.9	6.8	6.3

#w/p and #s/p represent the average number of words and sentences in the passage.
#w/q and #w/o are the average length of the question and option.
Training, development and test sets share the similar statistics.

Experiments

	RACE-M	RACE-H	RACE
Random	24.6	25.0	24.9
Sliding Window	37.3	30.4	32.2

Table: Accuracy on test set of RACE-M, RACE-H and RACE.

Experiments

	RACE-M	RACE-H	RACE
Random	24.6	25.0	24.9
Sliding Window	37.3	30.4	32.2
GA Reader (100D)	43.7	44.2	44.1
Stanford AR (100D)	44.2	43.0	43.3

Table: Accuracy on test set of RACE-M, RACE-H and RACE.

Experiments

	RACE-M	RACE-H	RACE
Random	24.6	25.0	24.9
Sliding Window	37.3	30.4	32.2
GA Reader (100D)	43.7	44.2	44.1
Stanford AR (100D)	44.2	43.0	43.3
Ours (100D)	**46.2**	44.1	44.7

Table: Accuracy on test set of RACE-M, RACE-H and RACE.

Experiments

	RACE-M	RACE-H	RACE
Random	24.6	25.0	24.9
Sliding Window	37.3	30.4	32.2
GA Reader (100D)	43.7	44.2	44.1
GA Reader (300D)	42.4	44.5	43.9
Stanford AR (100D)	44.2	43.0	43.3
Stanford AR (300D)	44.9	43.7	44.1
Ours (100D)	**46.2**	44.1	44.7

Table: Accuracy on test set of RACE-M, RACE-H and RACE.

Experiments

	RACE-M	RACE-H	RACE
Random	24.6	25.0	24.9
Sliding Window	37.3	30.4	32.2
GA Reader (100D)	43.7	44.2	44.1
GA Reader (300D)	42.4	44.5	43.9
Stanford AR (100D)	44.2	43.0	43.3
Stanford AR (300D)	44.9	43.7	44.1
Ours (100D)	**46.2**	44.1	44.7
Ours (300D)	45.0	**46.4**	**46.0**

Table: Accuracy on test set of RACE-M, RACE-H and RACE.

Summary

- We exploit candidate options to boost evidence gathering and enhance option representation with correlations.

- We utilize hierarchical structure to model interactions among the passage, question and candidate options at word level and sentence level.

- The proposed model achieves significant improvement over two neural baseline models.

Thanks!

Q&A