Inject Rubrics into Short Answer Grading System

Tianqi Wang[1,3] Naoya Inoue[1,3] Hiroki Ouchi[3] Tomoya Mizumoto[2,3] Kentaro Inui[1,3]
[1]Tohoku University [2]Future Corporation [3]RIKEN Center for Advanced Intelligence Project

Task Introduction

Short Answer Grading (SAG) is a task of assessing the **correctness of short answers** to questions automatically.

- Answers are **scored by graders with rubrics**
- Time consuming especially when limited graders are available

Key Idea

- Consider student answers as **combination of multiple key concepts.**
- Answers are scored based on key concept identification

Contribution

- The first study that explores how to incorporate rubric information into neural SAG
  - A general framework to extend existing neural SAG models with a component for exploiting rubric information

An example of SAG

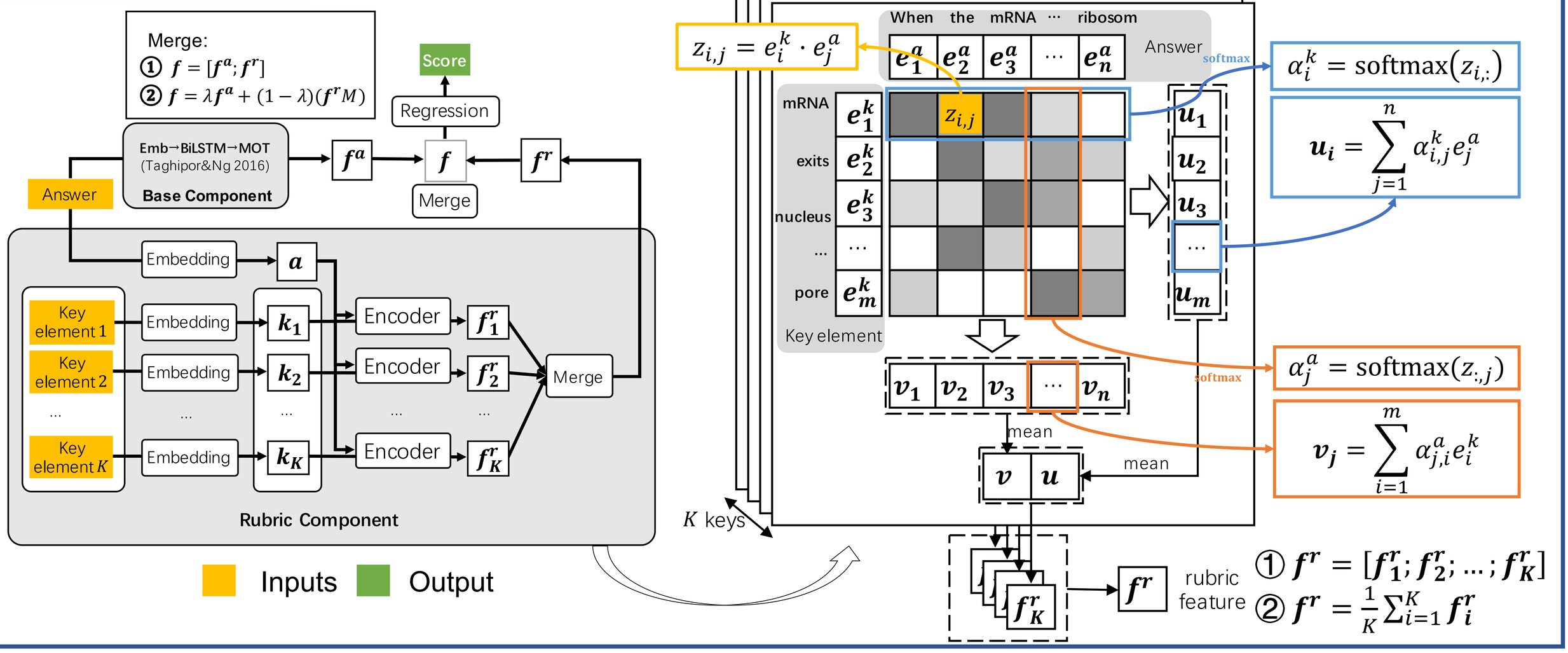
Prompt
Starting with mRNA leaving the nucleus, list and describe four major steps involved in protein synthesis.

Rubric
3 points: 4 key elements
2 points: 3 key elements
1 point: 1 or 2 key elements
0 points: Other

Key elements
1. **mRNA exits nucleus via nuclear pore**.
2. **mRNA travels through the cytoplasm to the ribosome** or enters the rough endoplasmic reticulum.
3. mRNA bases are read in triplets called codons (by rRNA).
4. …

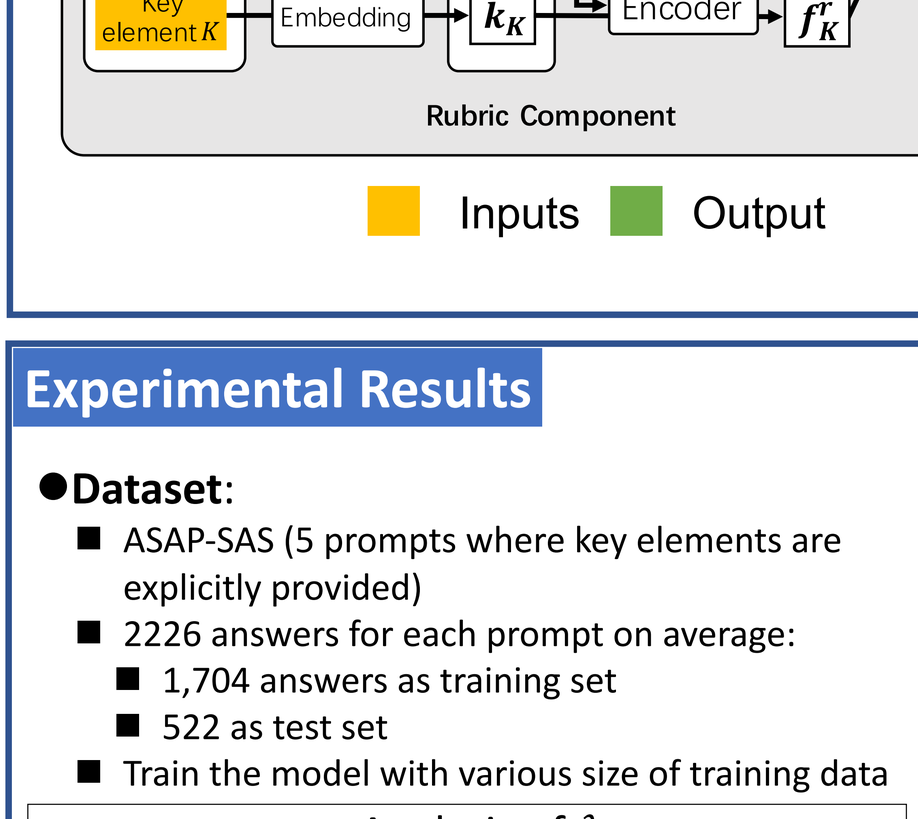
Answer (1 point)
When the mRNA leaves the nucleus, it **travels through the cell. It moves to a ribosome**. The ribosome makes tRNA. Then, protein is synthesized.

Proposed Model

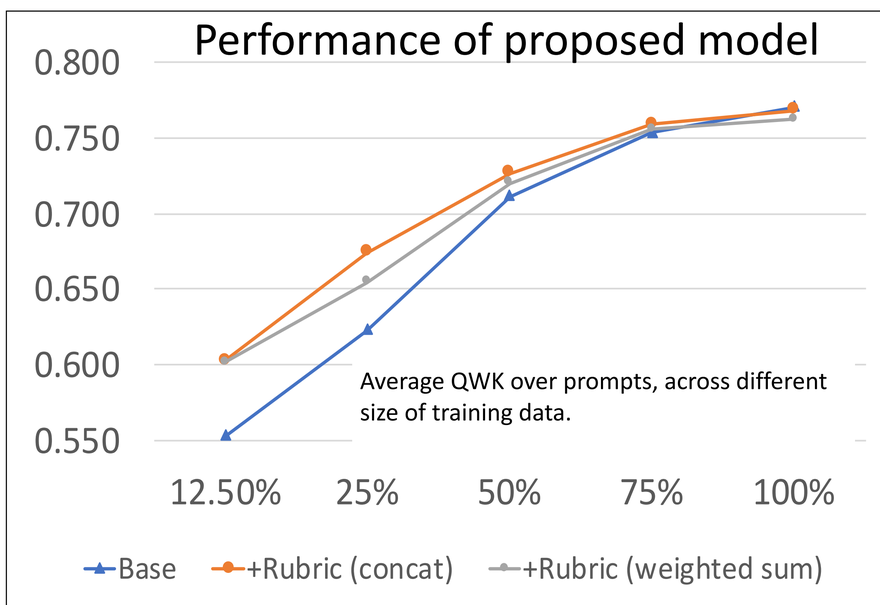


Experimental Results

- **Dataset**:
  - ASAP-SAS (5 prompts where key elements are explicitly provided)
  - 2226 answers for each prompt on average:
    - 1,704 answers as training set
    - 522 as test set
  - Train the model with various size of training data



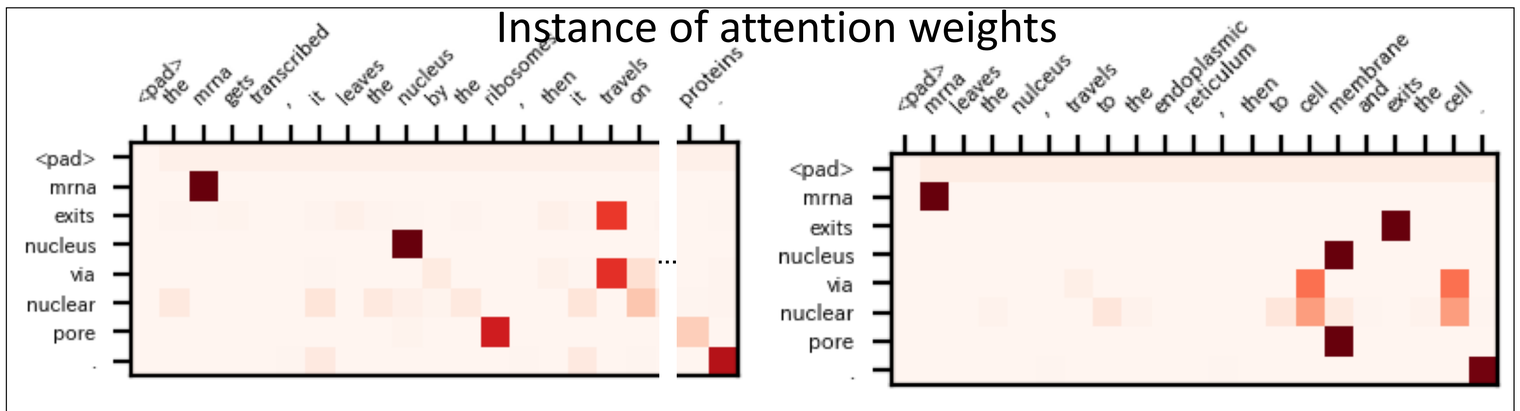
- Value of λ leaned from different size of training data
- **Higher** λ means **less** contribution from rubrics
- Rubric component contributes more when less training data is available



Average QWK over prompts, across different size of training data.

Prompt		1	2	5	6	10	mean
12.5%	B	.588	.331	.617	.611	**.618**	.553
	+R	**.599***	**.424***	**.689***	**.679***	.617	**.602**
		+.012	+.093	+.073	+.068	-.001	+.049
25%	B	.656	.473	.641	.627	**.719**	.623
	+R	**.661***	**.529***	**.687***	**.697***	.698	**.654**
		+.005	+.056	+.046	+.070	-.020	+.031
50%	B	.748	.637	.748	.718	**.705**	.711
	+R	**.747***	**.643**	**.784***	**.723***	.702*	**.720**
		+.000	+.006	+.036	+.006	-.004	+.009
75%	B	.776	.700	**.798**	.748	.744	.753
	+R	**.783***	**.704**	.787*	**.750***	**.784***	**.762**
		+.007	+.004	-.010	+.002	+.040	+.009
100%	B	**.792**	**.713**	**.804**	.788	**.753**	**.770**
	+R	.789	.695*	.786*	**.790***	.748	.762
		-.003	-.018	-.018	+.002	-.005	-.008

- Left: Performance was improved especially under low resource settings
- Right: The benefit obtained from rubric component varies with prompts.



- **Left: The model successfully found words and phrases most related to the key element, helping the model improve the performance.**
- **Right: The model incorrectly aligned words in the answer and key element.**