Unsupervised Candidate Answer Extraction through Differentiable Masker-Reconstructor Model

Zhuoer Wang, Yicheng Wang, Ziwei Zhu, James Caverlee

Overview

We target the task of identifying valuable candidate answers that can be used as the input of Question Generation(QG) systems for generating targeted questions given the context passage. Specifically:

- **We propose a novel Differentiable MaskerReconstructor model, in light of recent progress on self-consistency learning and masked language models, for unsupervised candidate answer extraction.**
- **We release two newly created datasets with exhaustively-annotated candidate answers.**
- **We benchmark a comprehensive list of supervised and unsupervised candidate answer extraction methods and show the strong performance of our DMR model.**

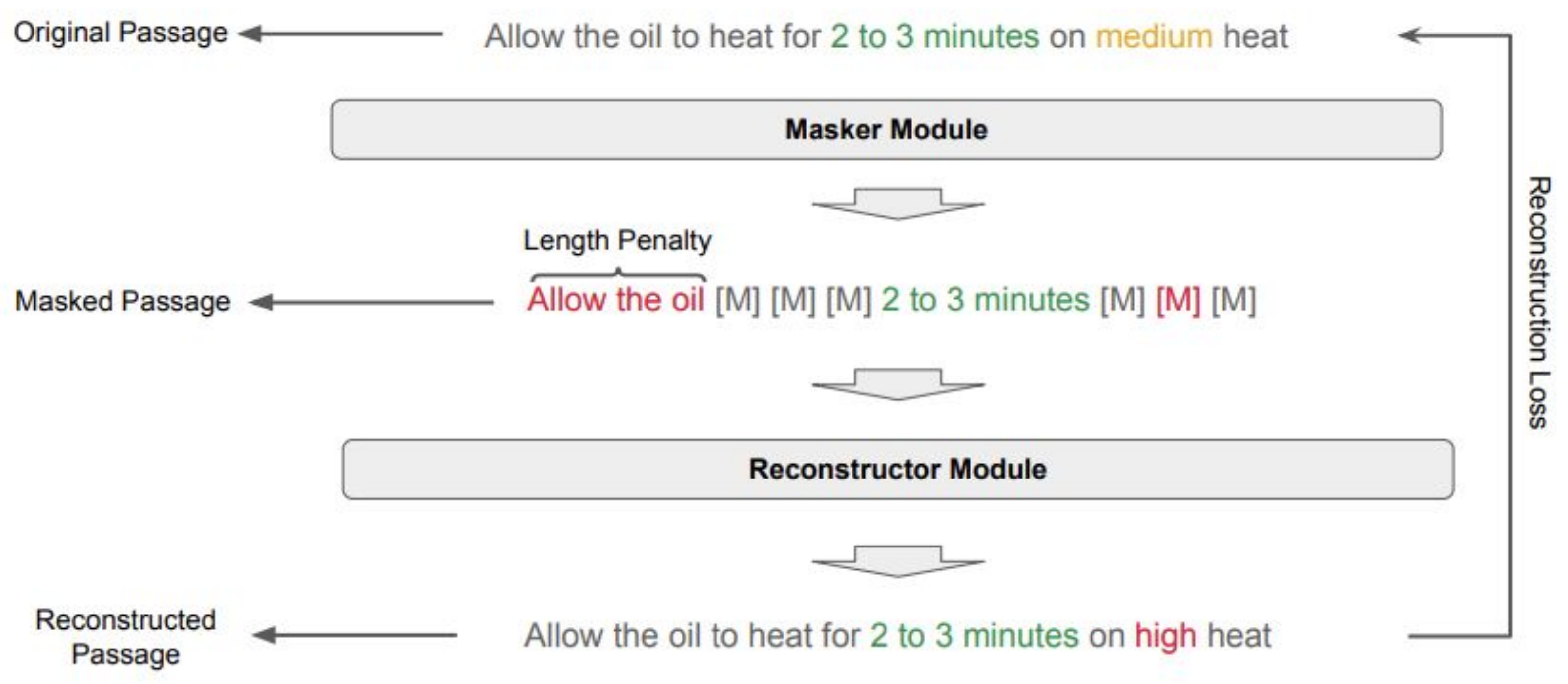
Motivation

Article Title	Content of the step *Heat the Olive Oil*
How to Make Caldo Tlalpeno	Add 2 tablespoons (30 ml) of olive oil to a large pot. Allow the oil to heat for 2 to 3 minutes on medium-high heat, or until it begins to shimmer. You can substitute vegetable or canola oil for the olive oil if you prefer.
How to Make Slow Cooker Spaghetti Sauce	Place alarge skillet on the stove, and add 2 tablespoons (30 ml) of olive oil to it. Turn the burner to medium, and allow the oil to heat for 3 to 5 minutes or until it starts to shimmer.
How to Make Shrimp Bisque	Add 3 tablespoons (45 ml) of olive oil to a large pot or Dutch oven, and place it on the stove. Turn the heat to medium, and allow the oil to heat for 5 minutes, or until it starts to shimmer. If you prefer, you can substitute butter for the olive oil.
How to Make an Omelette in a Jar	Place a large skillet on the stove, and add 1 tablespoon (15 ml) of olive oil. Turn the heat to medium-high, and allow the oil to heat until it starts to shimmer, which should take approximately 5 minutes. You can substitute butter for the olive oil if you prefer.
How to Make a White Pizza	Add 2 tablespoons (30 ml) of olive oil to a medium, heavy-bottomed saucepan. Place the pan on the stove, and heat it over medium heat until it begins to shimmer, which should take approximately 5 minutes. You can substitute vegetable oil for the olive oil.

- **Backbone Tokens** are those black tokens of the content shown in Table 1, are structural and common across various passages within the same domain, and such tokens are easily recoverable when masked.

- **Information Tokens**, in contrast to Backbone Tokens, are difficult to recover when masked, and such tokens are crucial information of a specific context passage, making them excellent candidate answers. Examples of Information Tokens are colored tokens in Table 1 that express the AMOUNT, CONTAINER, TIME, HEAT-LEVEL, SUBSTITUTION of the specific context. Besides the cooking domain, Wikipedia articles regarding a person may have similar structure that has date of birth, place of birth, occupation, education, etc. as information tokens. News articles may express who, what, where, when of events as information tokens with a shared explicit or implicit template.

The DMR Model


Original Passage
Allow the oil to heat for 2 to 3 minutes on medium heat
Masker Module
Length Penalty
Masked Passage
Allow the oil [M] [M] [M] 2 to 3 minutes [M] [M] [M]
Reconstructor Module
Reconstructed Passage
Allow the oil to heat for 2 to 3 minutes on high heat
Reconstruction Loss

The Masker-Reconstructor Model ([M] represents the special token [MASK]). Reconstruction loss guides the learning of the Reconstructor module as well as penalizes the Masker module for masking out hard-to-recover tokens (illustrated by [M] and high in red). Length penalty enforces the learning of the Masker module so that those easy-to-recover tokens (illustrated by Allow the oil in red) are more likely to be masked out as masking them would yield the gain of both length and reconstruction loss.

Datasets and Baselines

Dataset	SQuAD	WH-C
Source	Wikipedia	WikiHow
Domain	Open-domain	Cooking
Context Passage Amount	20,947	16,642
Average Passage Length	135 Tokens	66 Tokens
Answer/Context Ratio (Original)	13.35%	N/A
Answer/Context Ratio (Exhaustively-annotated)	35.11%	50.51%

- **FT-LLM:** fine-tune RoBERTa with the training set of SQuAD
- **SCOPE:** the SOTA supervised method that learns from partially annotated data with the Positive-Unlabeled Learning method
- **Noun Phrases / Named Entities:** extracted by spaCy
- **Extended NE:** extends Named Entities by finding a longer constituent that contains the NE in the constituency parse tree
- **DiverseQA:** combines Extended NE and constituents tagged as NP, ADJP, VP, and S in the constituency parse tree
- **ChatGPT:** prompt gpt-3.5-turbo with context passage and instructions

Experiment Results

- **Experiment Results on Original and Exhaustively-annotated SQuAD**

Dataset	Original (Partially-annotated)			Exhaustively-annotated		
Metric	Precision	Recall	F1	Precision	Recall	F1
Supervised Methods						
FT-LLM	**47.45**$_{(0.20)}$	33.22$_{(0.96)}$	39.07$_{(0.70)}$	**60.31**$_{(3.77)}$	19.90$_{(1.80)}$	29.84$_{(1.59)}$
SCOPE	32.28$_{(1.65)}$	**69.38**$_{(3.59)}$	**44.09**$_{(0.84)}$	45.81$_{(0.27)}$	**59.00**$_{(1.32)}$	**51.57**$_{(0.49)}$
Unsupervised Methods						
Noun Phrases	18.91	75.45	30.24	37.90	69.42	49.04
Named Entities	**25.37**	38.00	30.42	**43.54**	26.02	32.60
Extended NE	20.07	44.04	27.57	40.19	36.08	38.02
DiverseQA	16.87	**94.12**	28.62	35.69	**89.26**	50.99
|- ADJP	17.85	5.64	8.57	36.74	4.91	8.66
|- VP	16.23	47.24	24.16	30.93	38.84	34.43
|- S	15.86	28.22	20.31	29.01	24.00	26.27
ChatGPT	23.75	53.79	**32.95**	41.00	39.80	40.39
DMR (Ours)	18.14$_{(0.03)}$	80.63$_{(0.34)}$	29.61$_{(0.05)}$	37.94$_{(0.32)}$	77.98$_{(1.17)}$	**51.04**$_{(0.07)}$

- **Experiment Results on Exhaustively-annotated WH-C**

Metric	Precision	Recall	F1
Supervised Methods			
FT-LLM	**83.64**$_{(4.29)}$	20.65$_{(2.46)}$	33.03$_{(2.91)}$
SCOPE	67.23$_{(1.88)}$	**78.59**$_{(1.89)}$	**72.47**$_{(1.90)}$
Unsupervised Methods			
Noun Phrases	60.69	71.17	65.51
Named Entities	**88.18**	13.83	23.91
Extended NE	71.57	15.74	25.80
DiverseQA	52.28	**87.65**	65.49
|- ADJP	69.61	7.53	13.58
|- VP	46.42	53.58	49.74
|- S	47.76	33.28	39.23
ChatGPT	75.91	52.78	62.27
DMR (Ours)	58.20$_{(1.04)}$	78.04$_{(2.73)}$	**66.64**$_{(0.48)}$

- SCOPE consistently achieved the best performance, as measured by F1, with the help of a mass amount of annotated answers from the original SQuAD dataset.
- When such annotated data is not available, our DMR model achieved the best performance among the unsupervised methods.
- Compared to unsupervised methods with second highest F1, our DMR model has better balance of precision and recall. The performance gain is +0.05 and +1.13 on exhaustively-annotated SQuAD and WH-C respectively. We attribute the difference to the fact that WH-C is more focused on a specialized cooking domain while SQuAD consists of open-domain data. Therefore, WH-C should have a more prominent underlying structure that can be captured by the DMR model through self-consistency learning to discriminate backbone tokens and information tokens.