MICRON: Multigranular Interaction for Contextualizing RepresentatiON in Non-factoid Question Answering

Hojae Han*, **Seungtaek Choi***, **Haeju Park**, **Seung-won Hwang**
Department of Computer Science, Yonsei University, Seoul, Korea
{stovecat, hist0613, phj0225, seungwonh}@yonsei.ac.kr

Abstract

This paper studies the problem of non-factoid question answering, where the answer may span over multiple sentences. Existing solutions can be categorized into representation- and interaction-focused approaches. We combine their complementary strength, by a hybrid approach allowing multi-granular interactions, but represented at word level, enabling an easy integration with strong word-level signals. Specifically, we propose **MICRON: Multigranular Interaction for Contextualizing RepresentatiON**, a novel approach which derives contextualized *uni*-gram representation from *n*-grams. Our contributions are as follows: First, we enable multi-granular matches between question and answer *n*-grams. Second, by contextualizing word representation with surrounding n-grams, MICRON can naturally utilize word-based signals for query term weighting, known to be effective in information retrieval. We validate MICRON in two public non-factoid question answering datasets: WikiPassageQA and InsuranceQA, showing our model achieves the state of the art among baselines with reported performances on both datasets.

1 Introduction

Non-factoid questions, unlike factoid questions answered by short facts like a word or a phrase, may get answered by a long answer spanning across multiple sentences. Following the definition in (Guo et al., 2019), neural approaches for this task can be roughly categorized into **representation-** and **interaction-focused** approaches.

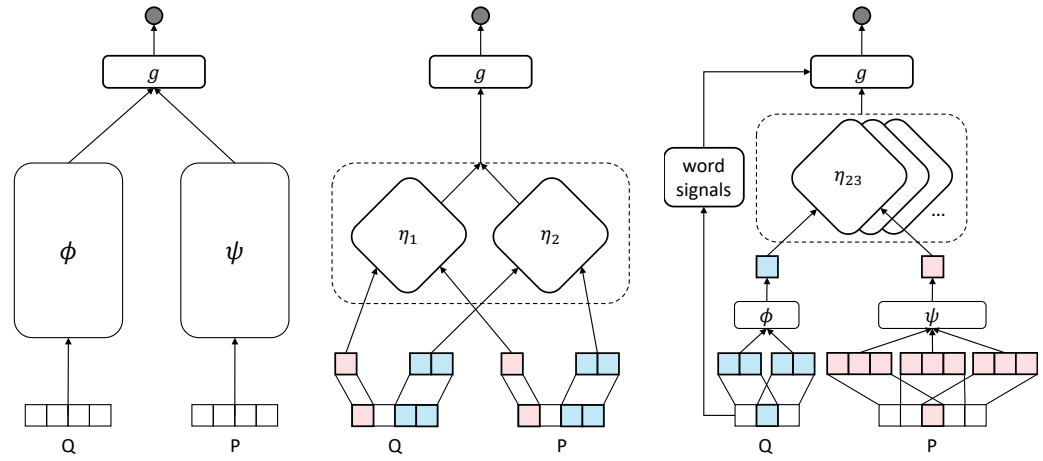
First, **representation-focused** approaches (Rücklé and Gurevych, 2017; Shao et al., 2019) encode query and answer into vectors of the same size, and match the two by computing vector similarity. Models in this category have advantages of efficiency, as representations can be pre-computed and indexed for efficient retrieval. However, structural information, such as some question word matching another in answer, is missing in this representation. In addition, structural information can be diluted when squeezing a long text into a single vector. This weakness is often complemented by auxiliary information such as attention (Tan et al., 2016; Santos et al., 2016; Wang and Jiang, 2016). Figure 1a illustrates a representative architecture in this category (Shao et al., 2019).

Second, **interaction-focused** approaches aim to preserve structural information above. A naive structural information is a matrix storing pairwise word interaction, or 1:1. However, due to a typical length difference between a question and a long answer in our problem setting, most answer words are left unmatched, except a few *uni*-gram in the answer. Later work relaxes 1:1 constraint, to 1:N and M:N, by allowing a match to *n-gram* (1:N) or a match between query and answer *bi*-grams (2:2). A state-of-the-art in this category (Rücklé et al., 2019), shown in Figure 1b, uses *bi*-gram Convolutional Neural Network (CNN) to represent query/answer *bi*-grams and their interactions. Similar architecture was generalized for N:N or N:M matches (Song et al., 2019; Chen et al., 2018), which may introduce a new challenge of multi-granular interaction we discuss later.

Our work is of combining the strength of the two, as shown in Figure 1c. We illustrate our technical contributions using the following running example:

Example 1 *Consider a running example of matching a question, "Who is in charge of this education process", with a matching passage on "the institution of higher learning". Interaction between a query bi-gram "education process" and the 5-gram "the institution of higher learning" is a key*

*The authors contribute equally to this paper.



(a) Representation-focused (b) Interaction-focused (c) Ours:

Figure 1: Comparative illustration of three approaches.

indicator explaining this match. In addition, external word-level importance signals, such as Inverse Document Frequency (IDF), are observed to be simple yet most powerful (Guo et al., 2016), in matching a short query (or, question) with a long document, as in information retrieval or non-factoid question answering scenarios. For our example question with eight words, the IDF weight is highest for education, appearing rarely in other questions, while that is lower for common words.

Below are our two key contributions, inspired by the above running example.

1) Multigranular interaction: Figure 1c shows a dotted area, where interaction between *m*- and *n*-grams are represented. This enables matching between different sized n-grams: η_{25} enables the interaction between the *bi*-gram *"education process"* and the 5-gram *"the institution of higher learning"* in our running example. However, existing multigranular interaction (Chen et al., 2018) cannot combine word-level signal, such as a high IDF score of word "*education*".

2) N-gram contextualized word representation: Our next step is to combine this matching signal into a contextualized word representation For example, we represent word *higher* as an aggregation of its participating consecutive 5-grams, where *"... the institution of higher"* and *"the institution of higher ..."* disambiguate that the term should not be matched a question on *"high school"*. Similarly, question word *education* is represented by surrounding 2-grams: *"education process"* and *"of education"*. Contextualizing into word-level representation makes it natural to combine with word-level IDF scores in the model, and also enables indexing (Hwang and Chang, 2005). This shares the spirit of contextualized embedding, such as BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018), but specialized for **short-distance** phrase context **localized** within question and passage.

We summarize the main contributions of this paper as follows. First, we utilize multigranular interaction to extract important information from the question/passage matching by proposing **MICRON: Multigranular Interaction for Contextualizing RepresentatiON**. Second, we leverage strong word-level signals, which we will discuss later.

We evaluate our method in two public non-factoid QA datasets: WikiPassageQA (Cohen et al., 2018) and InsuranceQA (Feng et al., 2015). The results show that our model achieves the state of the art among baselines with reported performances on both datasets. Our source code is freely-available at `https://github.com/stovecat/MICRON` for further study.

2 Our approach

In this section, we introduce our method in detail. MICRON mainly consists of three modules: encoding module, matching module, and scoring module. We use a Siamese architecture for encoding module, which is a common setting in our target problem (Rücklé and Gurevych, 2017; Shao et al., 2019; Rücklé et al., 2019).

Encoding Module For a word vector sequence $W \in \mathbb{R}^{|W| \times d}$ with dimensionality d, we encode

it by n-gram CNN as the following:

$$\Gamma_n(W) = \texttt{n-gramCNN}(W) \quad (1)$$

where n is the window size of n-gram CNN. Each $\Gamma_n(W) \in \mathbb{R}^{|W| \times d}$ represents n-gram semantics.

As a distinction from other interaction-focused approaches, we introduce an additional **Contextualization Layer** Φ, which returns a word representation, contextualized by surrounding n-gram phrases of the word belongs to. In our work, we define Φ as the arithmetic mean of n-gram representations, formalized as follows[1]:

$$[\Phi_n(W)]_k = \frac{\sum_{i=1}^{n} [\Gamma_n(W)]_{k-i+1}}{n} \quad (2)$$

where $[\Gamma_n(W)]_k$ is k-th row vector of $\Gamma_n(W)$, and each row of $\Phi_n(W) \in \mathbb{R}^{|W| \times d}$ is the contextualized n-gram representation, corresponding to each word.

Matching Module Query and candidate answer $Q \in \mathbb{R}^{|Q| \times d}$ and $P \in \mathbb{R}^{|P| \times d}$ can be encoded into $\Phi_n(Q)$ and $\Psi_m(P)$. We build an interaction matrix η_{nm} by computing dot product between $\Phi_n(Q)$ and $\Psi_m(P)$:

$$\eta_{nm} = \Phi_n(Q)\Psi_m(P)^T \quad (3)$$

Output matrix $\eta_{nm} \in \mathbb{R}^{|Q| \times |P|}$ contains the relevance scores of all pairs between n-grams in query and m-grams in answer.

From η_{nm}, we conduct a row-wise max-pooling to obtain A_{nm}, relaxing the length constraint in interactions (Rücklé et al., 2019).

$$[A_{nm}]_i = \max_j([\eta_{nm}]_{ij}) \quad (4)$$

Scoring Module We then aggregate the best matching scores A_{nm} across all combinations of question n-grams and answer m-grams from $F = \{1, 2, 3, 5\}$ following the convention of (Shao et al., 2019), yielding the cumulative score for each question word $\gamma \in \mathbb{R}^{|Q|}$:

$$\gamma = \sum_{n \in F} \sum_{m \in F} A_{nm} \quad (5)$$

Finally, we obtain the relevance score Ω from γ vector. Note that we could adopt any effective word-based signals $\tau \in \mathbb{R}^{|Q|}$, known a priori. By applying dot product between γ and τ, we can contrast matching scores by word importance. Specifically,

$$\Omega = \begin{cases} \gamma \cdot \tau, & \text{if } \tau \text{ exists} \\ \sum_{i=1}^{|\gamma|} \gamma_i, & \text{otherwise} \end{cases} \quad (6)$$

A widely adopted example of τ is IDF, computed either globally (treating all passages in the dataset as a corpus) or locally (treating only candidate passages of given question as a corpus) (Blair-Goldensohn et al., 2003). Note that effective word-level signals may depend on the characteristic of dataset. We will further show empirically which measure is more effective for each dataset and explain why in later section.

Loss function Our model is trained by the loss function studied in (Cohen and Croft, 2016):

$$L = \sum_{q \in Q} (1 - (\mu_{q_r} - \max q_{nr}))BCE_q \quad (7)$$

where BCE_q is the standard binary cross entropy for the question, μ_{q_r} is the mean score of all relevant answers and $\max q_{nr}$ is the max score of all irrelevant answers for q.

3 Experiments

3.1 Dataset

We evaluate MICRON on two non-factoid question answering datasets: 1) **WikiPassageQA** (Cohen et al., 2018) is a recent Wikipedia based collection. There are high contextual similarity between answers and non-answers since all candidate answers are from the same document. 2) **InsuranceQA** (Feng et al., 2015) is another well-known large-scale non-factoid QA dataset from insurance domain constructed by putting the ground truth answers into the pool and randomly sampling negative answers[2]. Table 2 shows the statistics of two datasets.

3.2 Baselines

We divide the models into following four categories: 1) IR scores, 2) Representation-focused, 3) Interaction-focused, and 4) MICRON in Table 1. As state of the art in one dataset is not likely to be that in another, we focused on baselines either open source or reported results on both datasets.

We implement two interaction-focused and one representation-focused baselines: **N-gram CNN**

[1]We omit Ψ for simplicity. Φ and Ψ are the same in our architecture.

[2]Among two test sets available, V1 and V2, we used the latter.

Model	InsuranceQA	WikiPassageQA							
	Accuracy	MAP	MRR	P@5	P@10	nDCG	R@5	R@10	R@20
IR Based									
BM25[1]	24.9	53.73	62.58	19.47	11.51	66.59	63.34	73.11	83.09
Representation-focused									
CNN[1]	24.4	27.33	31.48	-	-	-	-	-	-
BiLSTM[1]	32.4	46.16	52.89	-	-	-	-	-	-
Att-BiLSTM (Tan et al., 2016)[1]	37.9	47.04	54.36	-	-	-	-	-	-
AP-BiLSTM (Santos et al., 2016)[1]	31.9	46.98	55.20	-	-	-	-	-	-
LW-BiLSTM (Rücklé and Gurevych, 2017)[1]	36.9	47.56	54.33	-	-	-	-	-	-
Interaction-focused									
CA-Wang (Wang and Jiang, 2016)[1]	37.0	48.71	56.11	-	-	-	-	-	-
COALA (Rücklé et al., 2019)[1]	38.0	60.58	69.40	-	-	-	-	-	-
COALA p-means (Rücklé et al., 2019)[1]	39.9	59.29	68.48	-	-	-	-	-	-
COALA syntax-aware (Rücklé et al., 2019)[1]	39.5	60.48	68.75	-	-	-	-	-	-
N-gram CNN	45.8	56.81	64.53	21.63	13.20	69.29	69.52	81.85	91.35
Unigram CNN	36.5	55.45	64.93	20.53	12.64	68.57	65.54	78.49	89.13
Unigram CNN w/ global IDF	35.4	56.68	68.52	20.91	12.72	69.86	66.94	79.27	89.68
Unigram CNN w/ local IDF	29.7	58.66	68.96	22.50	13.17	71.23	71.43	81.91	91.97
Ours									
MICRON	**49.8**	59.38	67.17	22.21	13.13	71.31	71.09	81.73	91.46
MICRON w/ global IDF	46.7	59.44	67.87	22.07	13.13	71.40	70.82	81.27	90.75
MICRON w/ local IDF	48.0	**63.00**	**71.03**	**23.17**	**13.82**	**74.14**	**73.77**	**85.30**	**93.07**

Table 1: Results of the different models on the InsuranceQA and WikiPassageQA. [1] are reported in (Rücklé et al., 2019) . COALA syntax-aware (Rücklé and Gurevych, 2017) is a variant of COALA using dependency parse trees (Schuster and Manning, 2016).

Dataset	# of Questions			Answer
	Train	Valid	Test	Length
WikipassageQA	3332	417	416	153
InsuranceQA	10391	1592	1625	112

Table 2: Statistics of datasets

builds N:N matching matrices respectively. The size of N is the same with our method for fair comparison. **Unigram CNN** uses 1:1 word matching, and is able to utilize word-based signals as query term weighting value.

3.3 Implementation Details

For word embeddings, we use 300d pre-trained Glove (Pennington et al., 2014). The sequence length of the passage are all different for each dataset: 400 tokens for WikiPassageQA, 200 tokens for InsuranceQA. The dropout is applied after every layers with a keep rate of 0.7. All weights except embedding matrices are constrained by L2 regularization with constant values of 10^{-7} and 10^{-5} respectively for WikiPassageQA and InsuranceQA. We use Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 10^{-6} and 10^{-4} for each dataset. The learning parameters were chosen by the best performance on the dev set.

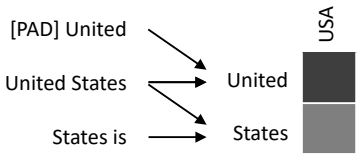
3.4 Results

Table 1 shows the results on WikiPassageQA and InsuranceQA datasets. We observe that our proposed approach, named MICRON, significantly outperforms both representation-focused and interaction-focused baselines in various evaluation metrics, achieving the best performance in both datasets.

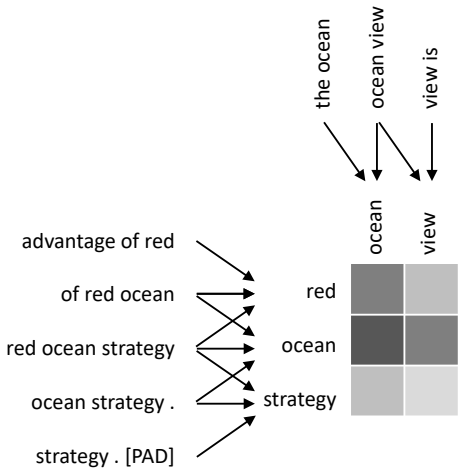
Our finding could be summarized as below: First, we manifest the effectiveness of multigranular interaction. Compared to N-gram CNN, MICRON allows matching between different n-grams (e.g., 2:3, 3:5) and achieves the improvement on both datasets by 4.0% point accuracy, 2.57% point MAP respectively.

Second, we relax length constraint in n-gram from COALA and achieve relative gain in InsuranceQA dataset. However, this gain is marginal when the phrase is short as in WikiPassageQA dataset, considering the better performance of COALA over N-gram CNN.

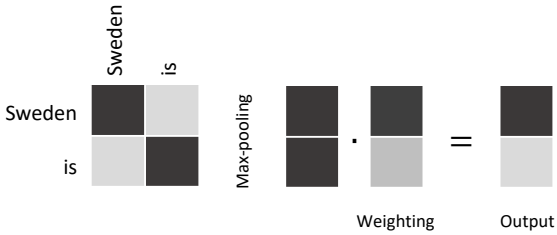
Third, word-based signals may help considerably in WikiPassageQA, where both global and local IDF scores of words are vary significantly (high variance). This variance is especially high for local IDF, which serves as a strong signal as consistently observed in (Blair-Goldensohn et al., 2003). In contrast, in InsuranceQA, the variance



(a) An example of multigranular interaction over different n-grams (2:1).



(b) An example of contextualized representation for the word *ocean*, lowering the matching score for negative candidate.



(c) Effect of word-based signals of enhancing query term weighting.

Figure 2: Qualitative examples of MICRON in the WikiPassageQA dataset.

of word signals are low. Consequently, the use of IDF cannot contribute to performance, or even contributes negatively.

3.5 Qualitative Examples

We illustrate several qualitative examples of MICRON in Figure 2. In Figure 2a, multigranular interaction (2:1) between the *bi*-gram *"United States"* and the *uni*-gram *"USA"* allows the matching. Figure 2b shows the case of where the contextualized representation enables to lower the matching score between *"red ocean"* and *"ocean view"*. From Figure 2c, we can see the word based signals can control the impact of each contextualized word scores: amplifying the matching of *"Sweden"*-*"Sweden"* and reducing the *"is"*-*"is"* matching.

4 Conclusion

In this paper, we study non-factoid question answering. Specifically, our approach is inspired by the complementary strength of representation- and interaction-focused approaches. We combine the strength of the two, by allowing multigranular interactions, but represented per-word basis, contextualized by participating n-grams. For this purpose, we propose MICRON, allowing to match flexible n-grams and to combine with word-based query term weighting, achieving the state of the art among baselines with reported performances on both datasets[3].

Acknowledgments

This work is supported by Microsoft Research Asia and IITP funded by MSIT (2017-0-0177; XAI). Hwang is a corresponding author.

References

Sasha Blair-Goldensohn, Kathleen McKeown, and Andrew Hazen Schlaikjer. 2003. A hybrid approach for answering definitional questions.

Haolan Chen, Fred X Han, Di Niu, Dong Liu, Kunfeng Lai, Chenglin Wu, and Yu Xu. 2018. Mix: Multi-channel information crossing for text matching. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 110–119. ACM.

Daniel Cohen and W Bruce Croft. 2016. End to end long short term memory networks for non-factoid question answering. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pages 143–146. ACM.

Daniel Cohen, Liu Yang, and W Bruce Croft. 2018. Wikipassageqa: A benchmark collection for research on non-factoid answer passage retrieval. *arXiv preprint arXiv:1805.03797*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[3]After the submission of this paper, BERT-based approach (Xu et al., 2019) outperformed MICRON in WikiPassageQA. However, MICRON is more energy-efficient (1M parameters) than BERT (110M parameters) and still outperforms BERT in InsuranceQA

Minwei Feng, Bing Xiang, Michael R Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 813–820. IEEE.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 55–64. ACM.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. 2019. A deep look into neural ranking models for information retrieval. *arXiv preprint arXiv:1903.06902*.

Seung-won Hwang and Kevin Chang. 2005. Optimizing access cost for top-k queries over web sources: A unified cost-based approach. In *ICDE*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Andreas Rücklé and Iryna Gurevych. 2017. Representation learning for answer selection with lstm-based importance weighting. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Andreas Rücklé, Nafise Sadat Moosavi, and Iryna Gurevych. 2019. Coala: A neural coverage-based approach for long answer selection with small data. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*.

Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*.

Sebastian Schuster and Christopher D Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *LREC*, pages 23–28. Portorož, Slovenia.

Taihua Shao, Xiaoyan Kui, Pengfei Zhang, and Honghui Chen. 2019. Collaborative learning for answer selection in question answering. *IEEE Access*, 7:7337–7347.

Yang Song, Qinmin Vivian Hu, and Liang He. 2019. P-cnn: Enhancing text matching with positional convolutional neural network. *Knowledge-Based Systems*.

Ming Tan, Cicero Dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 464–473.

Shuohang Wang and Jing Jiang. 2016. A compare-aggregate model for matching text sequences. *arXiv preprint arXiv:1611.01747*.

Peng Xu, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Passage ranking with weak supervsion. *arXiv preprint arXiv:1905.05910*.