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

Hojae Han*, Seunghaek 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
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: \( \eta_{25} \) 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 the 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:

\[ \Gamma_n(W) = \text{n-gram CNN}(W) \]  

(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 \( \Phi \), which returns a word representation, contextualized by surrounding n-gram phrases of the word belongs to. In our work, we define \( \Phi \) as the arithmetic mean of n-gram representations, formalized as follows:\(^1 \)

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

(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 \( \eta_{nm} \) by computing dot product between \( \Phi_n(Q) \) and \( \Psi_m(P) \):

\[ \eta_{nm} = \Phi_n(Q)\Psi_m(P)^T \]  

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

\[ [A_{nm}]_l = \max_j ([\eta_{nm}]_{ij}) \]  

(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 \( \gamma \in \mathbb{R}^{[Q]} \):

\[ \gamma = \sum_{n \in F} \sum_{m \in F} A_{nm} \]  

(5)

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

\[ \Omega = \begin{cases} \gamma \cdot \tau, & \text{if } \tau \text{ exists} \\ \sum_{i=1}^{d} \gamma_i, & \text{otherwise} \end{cases} \]  

(6)

A widely adopted example of \( \tau \) 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):

\[ L = \sum_{q \in Q} (1 - (\mu_q - \max q_{nr}))BCE_q \]  

(7)

where \( BCE_q \) is the standard binary cross entropy for the question, \( \mu_q \) 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 \( \Psi \) for simplicity. \( \Phi \) and \( \Psi \) are the same in our architecture.

---

\(^2\)Among two test sets available, V1 and V2, we used the latter.
Table 1: Results of the different models on the InsuranceQA and WikiPassageQA.

<table>
<thead>
<tr>
<th>Model</th>
<th>InsuranceQA</th>
<th>WikiPassageQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>MAP</td>
</tr>
<tr>
<td>IR Based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BM25</td>
<td>24.9</td>
<td></td>
</tr>
<tr>
<td>CNN</td>
<td>24.4</td>
<td></td>
</tr>
<tr>
<td>BiLSTM</td>
<td>32.4</td>
<td></td>
</tr>
<tr>
<td>Att-BiLSTM (Tan et al., 2016)</td>
<td>37.9</td>
<td></td>
</tr>
<tr>
<td>AP-BiLSTM (Santos et al., 2016)</td>
<td>31.9</td>
<td></td>
</tr>
<tr>
<td>LW-BiLSTM (Rücklé and Gurevych, 2017)</td>
<td>36.9</td>
<td></td>
</tr>
<tr>
<td>Representation-focused</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNN</td>
<td>24.4</td>
<td></td>
</tr>
<tr>
<td>BiLSTM</td>
<td>32.4</td>
<td></td>
</tr>
<tr>
<td>Att-BiLSTM (Tan et al., 2016)</td>
<td>37.9</td>
<td></td>
</tr>
<tr>
<td>AP-BiLSTM (Santos et al., 2016)</td>
<td>31.9</td>
<td></td>
</tr>
<tr>
<td>LW-BiLSTM (Rücklé and Gurevych, 2017)</td>
<td>36.9</td>
<td></td>
</tr>
<tr>
<td>Interaction-focused</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA-Wang (Wang and Jiang, 2016)</td>
<td>37.0</td>
<td>48.71</td>
</tr>
<tr>
<td>COALA (Rücklé et al., 2019)</td>
<td>38.0</td>
<td>60.58</td>
</tr>
<tr>
<td>COALA p-means (Rücklé et al., 2019)</td>
<td>39.9</td>
<td>59.29</td>
</tr>
<tr>
<td>COALA syntax-aware (Rücklé et al., 2019)</td>
<td>39.5</td>
<td>60.48</td>
</tr>
<tr>
<td>N-gram CNN</td>
<td>45.8</td>
<td>56.81</td>
</tr>
<tr>
<td>Unigram CNN</td>
<td>36.5</td>
<td>55.45</td>
</tr>
<tr>
<td>Unigram CNN w/ global IDF</td>
<td>35.4</td>
<td>56.68</td>
</tr>
<tr>
<td>Unigram CNN w/ local IDF</td>
<td>29.7</td>
<td>58.66</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MICRON</td>
<td>49.8</td>
<td>59.38</td>
</tr>
<tr>
<td>MICRON w/ global IDF</td>
<td>46.7</td>
<td>59.44</td>
</tr>
<tr>
<td>MICRON w/ local IDF</td>
<td>48.0</td>
<td>63.00</td>
</tr>
</tbody>
</table>

Table 1: Results of the different models on the InsuranceQA and WikiPassageQA. The IR Based
models 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).

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Questions</th>
<th>Answer Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>WikiPassageQA</td>
<td>3332</td>
<td>417</td>
</tr>
<tr>
<td>InsuranceQA</td>
<td>10391</td>
<td>1592</td>
</tr>
</tbody>
</table>

Table 2: Statistics of datasets

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
How to survive in United States

(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.

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.


3After 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.


