End-to-End Answer Chunk Extraction and Ranking for Reading Comprehension

Yang Yu*, Wei Zhang*, Kazi Hasan, Mo Yu, Bing Xiang, Bowen Zhou
{yu, zhangwei, khasan, yum, bingxia, zhou}@us.ibm.com
IBM Watson

Abstract
This paper proposes dynamic chunk reader (DCR), an end-to-end neural reading comprehension (RC) model that is able to extract and rank a set of answer candidates from a given document to answer questions. DCR is able to predict answers of variable lengths, whereas previous neural RC models primarily focused on predicting single tokens or entities. DCR encodes a document and an input question with recurrent neural networks, and then applies a word-by-word attention mechanism to acquire question-aware representations for the document, followed by the generation of chunk representations and a ranking module to propose the top-ranked chunk as the answer. Experimental results show that DCR achieves state-of-the-art exact match and F1 scores on the SQuAD dataset (Rajpurkar et al. 2016).

Introduction
Reading comprehension-based question answering (RCQA) is the task of answering a question with a chunk of text taken from related document(s). A variety of neural models have been proposed recently either for extracting a single entity or a single token as an answer from a given text (Hermann et al. 2015; Kadlec et al. 2016; Trischler et al. 2016b; Dhingra et al. 2016; Chen, Bolton, and Manning 2016; Sordoni, Bachman, and Bengio 2016; Cui et al. 2016a; Trischler et al. 2016a). In both cases, an answer boundary is either easy to determine or already given.

Different from the above two assumptions for RCQA, in the real-world QA scenario, people may ask questions about both entities (factoid) and non-entities such as explanations and reasons (non-factoid) (see Table 1 for examples). In this regard, RCQA has the potential to complement other QA approaches that leverage structured data (e.g., knowledge bases) for both the above question types. This is because RCQA can exploit the textual evidences to ensure increased answer coverage, which is particularly helpful for non-factoid answers. However, it is also challenging for RCQA to identify answer in arbitrary position in the passage with arbitrary length, especially for non-factoid answers which might be clauses or sentences. As a result, apart from a few exceptions (Rajpurkar et al. 2016; Wang and Jiang 2016), this research direction has not been fully explored yet.

Compared to the relatively easier RC task of predicting single tokens/entities predicting answers of arbitrary lengths and positions significantly increase the search space complexity: the number of possible candidates to consider is in the order of \(O(n^2)\), where \(n\) is the number of passage words. In contrast, for previous works in which answers are single tokens/entities or from candidate lists, the complexity is in \(O(n)\) or the size of candidate lists \(l\) (usually \(l \leq 5\)), respectively. To address the above complexity, Rajpurkar et al. (2016) used a two-step chunk-and-rank approach that employs a rule-based algorithm to extract answer candidates from a passage, followed by a ranking approach with hand-crafted features to select the best answer. The rule-based chunking approach suffered from low coverage (\(\approx 70\%\) recall of answer chunks) that cannot be improved during training; and candidate ranking performance depends greatly on the quality of the hand-crafted features. More recently, Wang and Jiang (2016) proposed two end-to-end neural network models, one of which chunks a candidate answer by predicting the answer’s two boundary indices and the other classifies each passage word into answer/not-answer. Both models improved significantly over the method proposed by Rajpurkar et al. (2016).

Our proposed model, called dynamic chunk reader (DCR), not only significantly differs from both the above systems in the way that answer candidates are generated and ranked, but also shares merits with both works. First, our model uses deep networks to learn better representations for candidate answer chunks, instead of using fixed feature representations as in (Rajpurkar et al. 2016). Second, it represents answer candidates as chunks, as in (Rajpurkar et al. 2016), instead of word-level representations (Wang and Jiang 2016), to make the model aware of the subtle differences among candidates (importantly, overlapping candidates).

The contributions of this paper are three-fold. (1) We pro-

---

* Both authors contribute equally

1 State-of-the-art RC models have a decent accuracy of \(\approx 70\%\) on the widely used CNN/DailyMail dataset (Hermann et al. 2015).
The United Kingdom (UK) intends to withdraw from the European Union (EU), a process commonly known as Brexit, as a result of a June 2016 referendum in which 51.9% voted to leave the EU. The separation process is complex, causing political and economic changes for the UK and other countries. As of September 2016, neither the timetable nor the terms for withdrawal have been established: in the meantime, the UK remains a full member of the European Union. The term “Brexit” is a portmanteau of the words “British” and “exit.”

Q1. Which country withdrew from EU in 2016?
A1. United Kingdom

Q2. How did UK decide to leave the European Union?
A2. as a result of a June 2016 referendum in which 51.9% voted to leave the EU

Q3. What has not been finalized for Brexit as of September 2016?
A3. neither the timetable nor the terms for withdrawal

Problem Definition

Table 1: Example of questions (with answers) which can be potentially answered with RC on a Wikipedia passage. The first question is factoid, asking for an entity. The second and third are non-factoid.

| Question                                                                 | Answer                                                                 |
|-------------------------------------------------------------------------|------------------------------------------------------------------------|
| Q1. Which country withdrew from EU in 2016?                             | United Kingdom                                                        |
| Q2. How did UK decide to leave the European Union?                      | as a result of a June 2016 referendum in which 51.9% voted to leave the EU |
| Q3. What has not been finalized for Brexit as of September 2016?         | neither the timetable nor the terms for withdrawal                     |

Baseline: Chunk-and-Rank Pipeline with Neural RC

In this section we modified a state-of-the-art RC system for cloze-style tasks for our answer extraction purpose, to see how much gap we have for the two type of tasks, and to inspire our end-to-end system in the next section. In order to make the cloze-style RC system to make chunk-level decision, we use the RC model to generate features for chunks, which are further used in a feature-based ranker like in (Rajpurkar et al. 2016). As a result, this baseline can be viewed as a deep learning based counterpart of the system in (Rajpurkar et al. 2016). It has two main components: 1) a stand-alone answer chunker, which is trained to produce overlapping candidate chunks, and 2) a neural RC model, which is used to score each word in a given passage to be used thereafter for generating chunk scores.

Answer Chunking

To reduce the errors generated by the rule-based chunker in (Rajpurkar et al. 2016), first, we capture the part-of-speech (POS) pattern of all answer subsequences in the training dataset to form a POS pattern trie tree, and then apply the answer POS patterns to passage $P_i$ to acquire a collection of all subsequences (chunk candidates) $C_i$, whose POS patterns can be matched to the POS pattern trie. This is equivalent to putting a constraint $\text{sub}(m, n, P_i)$ to candidate answer chunk generation process that only choose the chunk with a POS pattern seen for answers in the training data. Then the sub-sequences $C_i$ are used as answer candidates for $P_i$. Note that overlapping chunks could be generated for a passage, and we rely on the ranker to choose the best candidate based on features from the cloze-style RC system. Experiments showed that for $>90\%$ of the questions on the development set, the ground truth answer is included in the candidate set constructed in such manner.

Remark: Categories of RC Tasks

Other simpler variants of the aforementioned RC task were explored in the past. For example, *quiz-style* datasets (e.g., MCTest (Richardson, Burges, and Renshaw 2013), MovieQA (Tapaswi et al. 2015)) have multiple-choice questions with answer options. *Cloze-style* datasets (Hermann et al. 2015) (Hill et al. 2015) (Ou et al. 2016), usually automatically generated, have factoid “question”s created by replacing the answer in a sentence from the text with blank. For the *answer selection* task this paper focuses on, several datasets exist, e.g. TREC-QA for factoid answer extraction from multiple given passages, bAbI (Weston, Chopra, and Bordes 2014) designed for inference purpose, and the SQuAD dataset (Rajpurkar et al. 2016) used in this paper. To the best of our knowledge, the SQuAD dataset is the only one for both factoid and non-factoid answer extraction with a question distribution more close to real-world applications.

Baseline: Chunk-and-Rank Pipeline with Neural RC

In this section we modified a state-of-the-art RC system for cloze-style tasks for our answer extraction purpose, to see how much gap we have for the two type of tasks, and to inspire our end-to-end system in the next section. In order to make the cloze-style RC system to make chunk-level decision, we use the RC model to generate features for chunks, which are further used in a feature-based ranker like in (Rajpurkar et al. 2016). As a result, this baseline can be viewed as a deep learning based counterpart of the system in (Rajpurkar et al. 2016). It has two main components: 1) a stand-alone answer chunker, which is trained to produce overlapping candidate chunks, and 2) a neural RC model, which is used to score each word in a given passage to be used thereafter for generating chunk scores.

Baseline: Chunk-and-Rank Pipeline with Neural RC
DCR works in four steps. First, the encoder layer encodes passage and question separately, by using bidirectional recurrent neural networks (RNN). Second, the attention layer calculates the relevance of each passage word to the question. Third, the chunk representation layer dynamically extracts the candidate chunks from the given passage, and create chunk representation that encodes the contextual information of each chunk. Fourth, the ranker layer scores the relevance between the representations of a chunk and the given question, and ranks all candidate chunks using a softmax layer. We describe each step in details below.

**Dynamic Chunk Reader**

The dynamic chunk reader (DCR) model is presented in Figure 1. Inspired by the baseline we built, DCR is deemed to be superior to the baseline for 3 reasons. First, each chunk has a representation constructed dynamically, instead of having a set of pre-defined feature values. Second, each passage word’s representation is enhanced by word-by-word attention that evaluates the relevance of the passage word to the question. Third, these components are all within a single, end-to-end model that can be trained in a joint manner.

**Feature Extraction and Ranking**

For chunk ranking, we (1) use neural RCQA model to annotate each $p_{ij}$ in passage $P_i$ to get score $s_{ij}$, then (2) for every chunk $c_i^{m, n}$ in passage $i$, collect scores $(s_{im}, \ldots, s_{in})$ for all the $(p_{im}, \ldots, p_{in})$ contained within $c_i^{m, n}$, and (3) extract features on the sequence of scores $(s_{im}, \ldots, s_{in})$ to characterize its scale and distribution information, which serves as the feature representation of $c_i^{m, n}$. In step (1) to acquire $s_{ij}$ we train and apply a word-level single-layer Gated Attention Reader [Dhingra et al. 2016], which has state-of-the-art performance on CNN/DailyMail cloze-style RC task. In step (3) for chunk $c_i^{m, n}$, we designed 5 features, including 4 statistics on $(s_{im}, \ldots, s_{in})$: maximum, minimum, average and sum; as well as the count of matched POS pattern within the chunk, which serves as an answer prior. We use these 5 features in a state-of-the-art ranker [Ganjisafaar, Caruana, and Lopes 2011].

---

3We tried using more than one layers in Gated Attention Reader, but no improvement was observed.
where $h_p^q$ and $h_q^q$ are hidden states from the bi-directional RNN encoders (see Figure 1). An inner product, $\alpha_{jk}$, is calculated between $h_p^q$ and every question word $h_q^k$. It indicates how well the passage word $p_j$ matches with every question word $q_k$. $\beta_j$ is a weighted pooling of $|Q|$ question hidden states, which serves as a $p_j$-aware question representation. The concatenation of $h_p^q$ and $\beta_j$ leads to a passage-question joint representation, $v_j \in \mathbb{R}^{4d}$. Next, we apply a second bi-GRU layer taking the $v_j$s as inputs, and obtain forward and backward representations $\gamma_j$ and $\delta_j \in \mathbb{R}^d$, and in turn their concatenation, $\gamma_j = [\gamma_j^T; \delta_j^T]$.

**Chunk Representation Layer** A candidate answer chunk representation is dynamically created given attention layer output. We first decide the text boundary for the candidate chunk, and then form a chunk representation using all or part of those $\gamma_j$ outputs inside the chunk. To decide a candidate chunk (boundary): we tried two ways: (1) adopt the POS trie-based approach used in our baseline, and (2) enumerate all possible chunks up to a maximum number of tokens. For (2), we create up to $N$ (max chunk length) chunks starting from any position $j$ in $P_j$. Approach (1) can generate candidates with arbitrary lengths, but fails to recall candidates whose POS pattern is unseen in training set; whereas approach (2) considers all possible candidates within a window and is more flexible, but over-generates invalid candidates.

For a candidate answer chunk $c_{i}^{m,n}$ spanning from position $m$ to $n$ inclusively, we construct chunk representation $\gamma_{m,n} \in \mathbb{R}^{2d}$ using every $\gamma_j$ within range $[m, n]$, with a function $g(\cdot)$. Formally,

$$\gamma_{m,n} = g(\gamma_m, \ldots, \gamma_n)$$

We experimented with several pooling functions (e.g., max, average) for $g(\cdot)$, and found out that, instead of pooling, the best function is to concatenate the hidden state of the first word in a chunk in forward RNN and that of the last word in backward RNN. Formally,

$$\gamma_{m,n} = [\gamma_m^T; \gamma_n^T]$$

We hypothesize that the hidden states at that two ends can better represent the chunk’s contexts, which is critical for this task, than the states within the chunk. This observation also agrees with Kobayashi et al. (2016).

**Ranker Layer** Each chunk $c_{i}^{m,n}$ is evaluated on its context similarity to the question, by taking the cosine similarity between the chunk context representation $\gamma_{m,n}$ acquired from chunk representation layer, and the question representation which is the concatenation of the last hidden state in forward RNN and the first hidden state in backward RNN. Thus, for training example $i$, we have the probability of the chunk $c_{i}^{m,n}$ as

$$\mathbb{P}(c_{i}^{m,n} | P_i, Q_i) = \text{softmax}(\gamma_{m,n}^T \cdot [h_{Q_i}^T; h_{Q_i}^T])$$

where $\gamma_{i}^{m,n}$ denotes representation of the chunk $c_{i}^{m,n}$, $h_{Q_i}^T$ is the $k$-th hidden state output from question $Q_i$’s forward and backward RNN encoder, respectively. In runtime, the chunk with the highest probability is taken as the answer. In training, the following negative log likelihood is minimized:

$$L = - \sum_{i=1}^{N} \log \mathbb{P}(A_i | P_i, Q_i)$$

Note that the $i$-th training instance is only used when $A_i$ is included in the corresponding candidate chunk set $C_i$, i.e., $\exists_{m,n} A_i = c_{i}^{m,n}$. The softmax in the final layer serves as the list-wise ranking module similar in spirit to Cao et al. (2007).

**Experiments**

**Dataset** We used the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. 2016) for the experiment. SQuAD came into our sight because it is a mix of factoid and non-factoid questions, a real-world data (crowd-sourced), and of large scale (over 100K question-answer pairs collected from 536 Wikipedia articles). Answers range from single words to long, variable-length phrase/clauses. It is a relaxation of assumptions by the cloze-style and quiz-style RC datasets in the Problem Definition section.

**Features** The input vector representation of each word $w$ to encoder RNNs has six parts including a pre-trained 300-dimensional GloVe embedding (Pennington, Socher, and Manning 2014) and five features (see Figure 1): (1) a one-hot encoding (46 dimensions) for the part-of-speech (POS) tag of $w$; (2) a one-hot encoding (14 dimensions) for named entity (NE) tag of $w$; (3) a binary value indicating whether $w$’s surface form is the same to any word in the question; (4) if the lemma form of $w$ is the same to any word in the question; and (5) if $w$ is capitalized. Feature (3) and (4) are designed to help the model align the passage text with question. Note that some types of questions (e.g., “who”, “when” questions) have answers that have a specific POS/NE tag pattern. For instance, “who” questions mostly have proper nouns/persons as answers and “when” questions may frequently have numbers/dates (e.g., a year) as answers. Thus, we believe that the model could exploit the co-relation between question types and answer POS/NE patterns easier with POS and NE tag features.

**Implementation Details** We pre-processed the SQuAD dataset using Stanford CoreNLP tool[5] (Manning et al. 2014) with its default setting to tokenize the text and obtain the POS and NE annotations. To train our model, we used stochastic gradient descent with the ADAM optimizer (Kingma and Ba 2014), with an initial learning rate of 0.001. All GRU weights were initialized from a uniform distribution between (-0.01, 0.01). The hidden state size, $d$, was set to 300 for all GRUs. The question bi-GRU shared parameters with the passage bi-GRU, while the attention-based passage bi-GRU had its own parameters. We shuffled all training examples at the beginning of each epoch and adopted a

---

[1] We tried another word-by-word attention methods as in Sanhotos et al. (2016), which has similar passage representation input to question side. However, this does not lead to improvement due to the confusion caused by long passages in RC. Consequently, we used the proposed simplified version of word-by-word attention on passage side only.
attributed to the advanced model structure and end-to-end state-of-the-art model for cloze-style RC tasks. This can be more than 12% (EM) behind even though it is based on the our DCR model (Table 3, row 2), the baseline (row 1) is feature-based ranking system. However when compared to 10% (EM) over Rajpurkar et al. (2016) (Table 2, row 1), the model contributes to the overall performance. Table 3 shows the details as well as the results of the baseline ranker. As we also studied how each component in our DCR model (1) and question-word feature (3) are the two most important features. Finally, combining the DCR model with the proposed POS-trie constraints yields a score similar to the one obtained using the DCR model with all possible n-gram chunks. The result shows that (1) our chunk representations are powerful enough to differentiate even a huge amount of chunks when no constraints are applied; and (2) the proposed POS-trie reduces the search space at the cost of a small drop in performance.

### Analysis

To better understand our system, we calculated the accuracy of the attention mechanism of the gated attention reader used in our deep learning-based baseline. We found that it is 72% accurate i.e., 72% of the times a word with the highest attention score is inside the correct answer span. This means that, if we could accurately detect the boundary around the word with the highest attention score to form the answer span, we could achieve an accuracy close to 72%. In addition, we checked the answer recall of our candidate chunking approach. When we use a window size of 10, 92% of the time, the ground truth answer will be included in the extracted Candidate chunk set. Thus the upper bound of the exact match score of our baseline system is around 66% (92% (the answer recall) × 72%). From the results, we see our DCR system’s exact match score is at 62%. This shows that DCR is proficient at differentiating answer spans dynamically.

To further analyze the system’s performance while predicting answers of different lengths, we show the exact match (EM) and F1 scores for answers with lengths up to 10 tokens in Figure 2(a). From the graph, we can see that, with the increase of answer length, both EM and F1 drops, but in different speed. The gap between F1 and exact match also widens as answer length increases. However, the model
still yields a decent accuracy when the answer is longer than a single word. Additionally, Figure 2(b) shows that the system is better at “when” and “who” questions, but performs poorly on “why” questions. The large gap between exact match and F1 on “why” questions means that perfectly identifying the span is harder than locating the core of the answer span.

Since “what”, “which”, and “how” questions contain a broad range of question types, we split them further based on the bigram a question starts with, and Figure 3 shows the breakdown for “what” questions. We can see that “what” questions asking for explanations such as “what happens” and “what happened” have lower EM and F1 scores. In contrast, “what” questions asking for year and numbers have much higher scores and, for these questions, exact match scores are close to F1 scores, which means chunking for these questions are easier for DCR.

## Related Work

Attentive Reader was the first neural model for factoid RCQA (Hermann et al. 2015). It uses Bidirectional RNN (Cho et al., 2014; Chung et al., 2014) to encode document and query respectively, and use query representation to match with every token from the document. Attention Sum Reader (Kadlec et al. 2016) simplifies the model to just predicting positions of correct answer in the document and the training speed and test accuracy are both greatly improved on the CNN/Daily Mail dataset. (Chen, Bolton, and Manning 2016) also simplified Attentive Reader and reported higher accuracy. Window-based Memory Networks (MemN2N) is introduced along with the CBT dataset (Hill et al. 2015), which does not use RNN encoders, but embeds contexts as memory and matches questions with embedded contexts. Those models’ mechanism is to learn the match between answer context with question/query representation. In contrast, memory enhanced neural networks like Neural Turing Machines (Graves, Wayne, and Danihelka 2014) and its variants (Zhang, Yu, and Zhou 2015; Gulcehre et al. 2016) were also potential candidates for the task, and Gulcehre et al. (2016) reported results on the bAbI task, which is worse than memory networks. Similarly, sequence-to-sequence models were also used (Yu et al. 2015; Hermann et al. 2015), but they did not yield better results either.

Recently, several models have been proposed to enable more complex inference for RC task. For instance, gated attention model (Dhingra et al. 2016) employs a multi-layer architecture, where each layer encodes the same document, but the attention is updated from layer to layer. EpiReader (Trischler et al. 2016b) adopted a joint training model for answer extractor and reasoner, where the extractor proposes top candidates, and the reasoner weighs each candidate by examining entailment relationship between question-answer representation and the document. An iterative alternating attention mechanism and gating strategies were proposed in (Sordoni, Bachman, and Bengio 2016) to optimize the attention through several hops. In contrast, Cui et al. (2016a, 2016b) introduced fine-grained document attention from each question word and then aggregated those attentions from each question token by summation with or without weights. This system achieved the state-of-the-art score on the CNN dataset. Those different variations all result in roughly 3-5% improvement over attention sum reader, but none of those could achieve higher than that. Other methods include using dynamic entity representation with max-pooling (Kobayashi et al. 2016) that aims to change entity representation with context, and Weissenborn’s (2016) system, which tries to separate entity from the context and then matches question to context, scoring an accuracy around 70% on the CNN dataset.

However, all of those models assume that the answers are single tokens. This limits the type of questions the models can answer. Wang and Jiang (2016) proposed a match-lstm and achieved good results on SQuAD. However, this approach predicts a chunk boundary or whether a word is part of a chunk or not. In contrast, our approach explicitly constructs the chunk representations and similar chunks are compared directly to determine correct answer boundaries.

## Conclusion

In this paper we proposed a novel neural reading comprehension model for question answering. Different from the previously proposed models for factoid RCQA, the proposed model, dynamic chunk reader, is not restricted to predicting a single named entity as an answer or selecting an answer from a small, pre-defined candidate list. Instead, it is capable of answering both factoid and non-factoid questions as it learns to select answer chunks that are suitable for an input question. DCR achieves this goal with a joint deep learning model enhanced with a novel attention mechanism and five simple yet effective features. Error analysis shows that the DCR model achieves good performance, but still needs to improve on predicting longer answers, which are usually non-factoid in nature.

## References

[Bahdanau, Cho, and Bengio 2015] Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. ICLR.

[Bengio, Louradour, Collobert, and Weston 2009] Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In Proceedings of the 26th
