Answer Sentence Selection using Local and Global Context in Transformer Models

Ivano Lauriola and Alessandro Moschitti

Amazon Alexa AI,
Manhattan Beach, California, USA
{lauivano,amosch}@amazon.com

Abstract. An essential task for the design of Question Answering systems is the selection of the sentence containing (or constituting) the answer from documents relevant to the asked question. Previous neural models have experimented with using additional text together with the target sentence to learn a selection function but these methods were not powerful enough to effectively encode contextual information. In this paper, we analyze the role of contextual information for the sentence selection task in Transformer based architectures, leveraging two types of context, local and global. The former describes the paragraph containing the sentence, aiming at solving implicit references, whereas the latter describes the entire document containing the candidate sentence, providing content-based information. The results on three different benchmarks show that the combination of the local and global context in a Transformer model significantly improves the accuracy in Answer Sentence Selection.

Keywords: Question Answering · Answer Sentence Selection · pre-trained Transformer · Deep learning.

1 Introduction

Recent research in Question Answering (QA) mainly addresses two tasks: (i) Answer Sentence Selection (AS2), which, given a question and a set of answer sentence candidates (e.g., retrieved by a search engine), consists in selecting the sentence that correctly answers the question with the highest probability; and (ii) Machine Reading (MR) comprehension [2], which, given a question and a reference text, involves finding an exact text span answering it. AS2 research originated from the TREC competitions [24], which target large databases of unstructured text. It has the advantage of high efficiency, which enables its use in real-world applications, e.g., see the study in [5].

Neural models have significantly contributed to both directions with new techniques [25, 14, 12]. In particular, recent approaches to neural language models, e.g., ELMO [13], GPT [16], BERT [4], RoBERTa [9], XLNet [3] have led to major advancements in several NLP subfields. These methods capture dependencies between words and their compounds by pre-training neural networks on

Table 1. An example of correct answer sentence requiring larger context to be selected.

Question	**When was Lady Gaga born?**
Prev.	Lady Gaga is an American singer, songwriter, and actress.
Target	She was born in 1986.
Next	Both of her parents have Italian ancestry, and...

Table 2. Each of the three sentences can be a correct answer. Only the global document information, e.g., the title and the link between document concepts, allows us to select the correct sentence.

Question	**Which role did Bradley Cooper play with Lady Gaga?**
doc. title	Avengers: endgame - Movie plot
sentence	Rocket Raccoon was voiced by Bradley Cooper.
doc. title	A star is born - Movie plot
sentence	Jackson "Jack" Maine (Bradley Cooper), a famous country rock singer...
doc. title	American sniper - Movie plot
sentence	Chris Kyle, the leading actor, was played by Bradley Cooper.

large amounts of data. Interestingly, the resulting models can be easily applied to different tasks by fine-tuning them on the target training data. The impact of such methods on AS2, also thanks to transfer learning, is impressive. For example, [5] exceeded the state of the art by 50% (relative error reduction) on WikiQA [28] and TREC-QA [24] datasets. Although this result seems hard to improve, we note that most previous work does not exploit contextual information in addition to the candidate sentence with a few exceptions, e.g., [22]. This aspect produces a suboptimal solution as there can be many cases that contain ambiguities, and they cannot be solved without other references or context. Formally, the term *context* refers to additional linguistic information coming from the source of a candidate answer sentence, which can be, for instance, the document containing the sentence, the paragraph, the domain, and so on.

For example, Table 1 shows a simple question asking for the birthdate of *Lady Gaga*. The answer is the middle sentence contained in a paragraph of three sentences. Clearly, an AS2 classifier cannot select the middle sentence with high reliability since the sentence does not reveal that *she* refers to *Lady Gaga*. On the other hand, AS2 is effective as it targets just one sentence at a time: selecting an entire paragraph to be sent to the users, often provides them with too much irrelevant information[1]. A further example is described in Table 2, where the question asks for the role of *Bradley Cooper* in a specific movie. In the same example, we retrieved three sentences belonging to three different documents containing movie plots. Each of the three sentences may reasonably be a correct answer. Also, the title of the movie is not enough to select the right answer and it can be too far from the "local" context window showed in the previous example.

[1] Of course, a solution based on a summarization approach would be optimal but poses complicated challenges, which have prevented to obtain better solutions than AS2 (to our knowledge).

However, *"A star is born - Movie plot"* is the only document that contains references to *Lady Gaga*. This related information allows us to recognize the correct answer. The two examples describe two different problems in common QA scenarios. In the first case, the sentence requires a local context to solve the pronoun *she*. Conversely, the candidate requires global information from the whole document to recognize the correct movie in the second example.

It should be noted that (i) previous neural network work, e.g., by [22], used context for AS2 in a hierarchical gated recurrent network but their accuracy is 10-12 points below the state of the art by [5] (as measured on the same exact dataset). Thus, it is not clear if their context is really useful for improving AS2 models. (ii) MR models clearly use a larger context but (a) they are not efficient enough to analyze hundreds of documents for each question, and (b) they target the selection of any subset of the document. This prevents them to be fast and accurate for AS2.

In this paper, we propose to model local and global contexts for AS2 by using multiple sentences and Bag-of-Word (BOW) features in Transformer networks [23]. More specifically, we consider candidates as a triplet (s_{i-1}, s_i, s_{i+1}), where s_i is the target answer sentence and s_{i-1} and s_{i+1} are the preceding and the next sentence of s_i, respectively. We integrate this triplet in Transformer architectures by using one single RoBERTa [10] model encoding the three sentences in three embeddings. Then, we add document-level BOW representation in the classification layer. We tested our models on three different datasets, Google NQ and SQuAD adapted for the AS2 task, as well as the well-known WikiQA, comparing with the very recent state of the art in AS2 [5]. The results clearly show that local and global contexts can improve AS2 models.

2 Related Work

We consider retrieval-based QA systems, which are mainly constituted by (i) a search engine, retrieving top-k documents related to the questions; and (ii) an Answer Sentence Selection (AS2) model, which reranks passages/sentences extracted from the documents. The task of reranking answer sentence candidates provided by a retrieval engine can be modeled with a classifier scoring the candidates.

Recent work has proposed neural networks that apply a series of non-linear transformations to the input question and answer text, represented as compositions of word or character embeddings; and (ii) then measure the similarity between the obtained representations. Question-to-question and answer-to-answer patterns are typically important to derive if an answer is correct for a question. For example, the CNN by [18] has two separate embedding layers for the question and answer, and a relational embedding, which aims at connecting them. More recent work uses attention mechanism, e.g., Compare-Aggregate [30], inter-weighted alignment networks [20], and pre-trained Transformer models [5].

In particular, the latter has shown to largely outperform any previous approach in AS2: a simply binary classifier is built by adding a linear layer on top

of the Transformer architecture, and is fine-tuned with positive and negative answers. The training of such model can be carried out by using a cross-entropy binary loss function. Additionally, the approach was highly boosted using out of domain data, as a first fine-tuning step, followed by a second fine-tuning on the target data. This procedure was referred to as the TANDA model, i.e., transfer the pre-trained models on the task, then adapt it to the target domain.

However, the proposed Transformer methods only focus on the similarity between the question and the candidate sentence pairs, without taking any additional information into account. Contextual information was already introduced in neural networks for solving AS2, e.g., [22], by combining question/answer pairs with context information, selected by applying a similarity between question and document sentences.

MR research has produced state-of-the-art models, e.g., [15, 26, 1]. By definition, MR is supposed to exploit a larger context than standard AS2 models, as their input is an entire abstract. Transformer models limit the input to 512 tokens, which prevent to encode large documents, e.g., webpages. Thus, we cannot consider them as global models. In contrast, they surely fit the definition of local context. However, as pointed out in the introduction, they are not enough efficient to analyze hundreds of documents for each question, which is a requirement of real-world applications [11, 21]. Also, they optimize the selection text sub-sequences, which, is a stronger requirement that does not lead to a better sentence selection model. Indeed, in our experiments, we show that state-of-the-art MR systems used for selecting answers are outperformed by AS2 models.

Differently from previous solutions, our model is built with state-of-the-art Transformer models for AS2. Moreover, our approach is more modular and can be easily extended with additional context definitions. We also improve the results from [22] by a huge margin (+12% on WikiQA and +5% on SQuAD).

3 Transformer Models for Answer Sentence Selecting

AS2 is the task of identifying sentences that contain the answer to a given question. The task can be modeled with a scoring function that outputs a probability of correctness for each question/sentence pair, (q, s_i). Such function can be implemented with a Transformer model as shown in [5]. In the remainder of this section, we formalize the task and describe a state-of-the-art model based on the Transformer.

3.1 AS2 definition

Let $\mathcal{Q}$ and $\mathcal{S}$ be the sets of questions and sentences, respectively, the AS2 task can be defined as a ranking function $r : \mathcal{Q} \times \mathcal{S} \to \mathbb{R}$, which assigns a score to each possible question/answer pair, where the higher the score is, the higher the probability of selecting a correct answer candidate is. In other words, we want to learn r, such that for each $q \in \mathcal{Q}$, we select

$$a = \arg \max_{s_i \in \mathcal{S}(q)} r(q, s_i)$$

as the final answer, where $S(q) \subseteq S$ is the set of answer sentence candidates for the input question q. For example, $\mathcal{S}(q)$ can be built by retrieving sentences from text repositories such as the web [5, 29]. In this work, we define and develop the ranking function r with Transformer models.

3.2 Selecting Sentences with a Transformer model

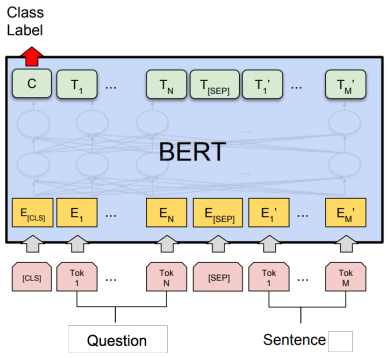

Fig. 1. The q/a pair is codifies as a whole sequence with special delimiters.

The Transformer is a popular neural network designed to learn language models, e.g., dependencies between words, in a context. Transformer models have recently been shown to produce a remarkable impact on AS2, when used as ranker [19, 5, 7]. Besides architectural definitions, an important advantage of Transformer models is their ability to be pre-trained on large-scale corpora, using masked language and next sentence prediction tasks [4].

More specifically, Fig. 1 shows the approach of using a pre-trained Transformer for AS2. The question and answer candidate pairs are codified as a joint sequence of tokens with specialized delimiters and separators, i.e., *[CLS]* $q^1 \dots q^n$ *[SEP]* $s^1 \dots s^m$ *[EOS]*, where x^j defines the j-th token of the sequence x. *[CLS]*, *[SEP]*, and *[EOS]* are special tokens used to mark the beginning of the sequence, the separation between question and candidate answer tokens, and the end of the text, respectively. Several Transformer blocks are applied, and then the representation associated with *[CLS]* is used in a linear fully-connected layer to compute the final score associated with the question/answer pair. The same concepts can be applied to RoBERTa or other pre-trained Transformer models.

4 Contextual Transformer for AS2

To our knowledge, no Transformer model for AS2 uses context, except for the information on the sentences. This is critical as a sentence may contain references to other parts of the text and to external entities (see the example in Table 1). We enhance the standard Transformer model for AS2 with two types of context: local and global. The former aims at resolving the coreferences between the constituents in the candidate answer sentence and its neighborhood sentences (typically part of the paragraph containing the answer). In contrast, the global context introduces information concerning the topics and concepts of the entire document containing the answer sentence.

4.1 Local context

Given the target answer sentence candidate, s_i, we extend the standard AS2 model using the preceding, s_{i-1}, and the following, s_{i+i}, sentences. The (local) contextual ranker r_{L} takes four elements as input and provides the following answer:

$$a_{\text{L}} = \arg \max_{s_i \in \mathcal{S}(q)} r_{\text{L}}(q, s_{i-1}, s_i, s_{i+1}),$$

where $\mathcal{S}(q)$ is the set of relevant sentences for the question q and r_{L} is our ranking function. To implement r_{L} in the RoBERTa model, the input sequence becomes *[CLS]* q *[SEP]* s_{i-1} *[SEP]* s_i *[SEP]* s_{i+1} *[EOS]*. Additionally, RoBERTa encodes each input word by using three pieces of information: the token, the sentence, and the positional embeddings.

The first is a standard word-embedding. The positional embedding describes a token as a function of its position in the sequence. Finally, the sentence embedding defines a token as a function of the sentence that contains it. The sentence embedding helps the model to distinguish between different input sentences: it can be seen as a particular word embedding of size four, one entry for each element of the input tuple, $(q, s_{i-1}, s_i, s_{i+1})$. This embedding plays a crucial role in our model to learn that the instance label is exclusively associated with the middle sentence. According to the canonical procedure, the three embeddings are then summed to produce the final representation of the sentences to be fed as input to the Transformer. This process is described in Fig. 2 (see dashed squares). When the preceding sentence s_{i-1} is not available, we consider an empty sequence in our input encoding, that is, *[CLS]* q *[SEP] [SEP]* s_i *[SEP]* s_{i+1} *[EOS]*. In this case, the model is still able to recognize the different parts of the input thanks to the sentence embedding and the two consecutive separators. The same strategy holds when the following sentence s_{i+1} is missing. Note that the local context is not limited in co-ref resolution as it also encodes semantic information from the whole sentences.

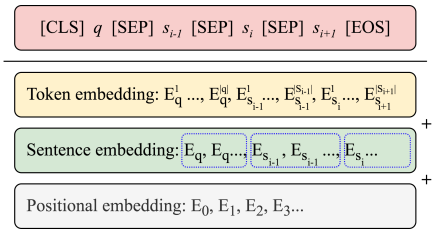

Fig. 2. BERT/RoBERTa input sequences.

4.2 Global context

The local context models the information related to the paragraph containing the candidate answer, and it is helpful to solve coreference problems. However, the local context is small and does not include other important information. Global document-based features can provide additional information to the local context, e.g., the main document topic, which can be used to select the correct sentence. Our global context describes the document content rather than the structure of

the paragraph containing the answer. The global ranker r_{G} is defined as

$$a_{\mathrm{G}} = \arg \max_{s_i \in \mathcal{S}(q)} r_{\mathrm{G}}(q, s_i, d(s_i)),$$

where $d(s_i)$ is the document containing s_i. There are several ways to take global information into account. We concatenate a *bag-of-words* (BOW) based feature vector to the CLS representation developed in the last Transformer layer. Specifically, given a candidate answer s_i, we firstly compute the representation associated with the CLS token at the last layer, v_{CLS}. Then, we extract BOW features from $d(s_i)$. The BOW vector v_{BOW} contains the frequency of each input word from the document. It should be noted that the BOW vector contains 50265 components as we considered the same vocabulary used by RoBERTa model. Hence, the direct concatenation of v_{CLS} and v_{BOW} is not adequate: it may suffer from scaling issues as v_{CLS} consists of only 768 components. To solve this problem, we apply a random projection over v_{BOW}, that is, $\widetilde{v}_{BOW} = v_{BOW}^{\top}\mathbf{W}$. $\mathbf{W} \in \mathbb{R}^{50265 \times 768}$ is the random projection matrix. Finally, we normalize the two vectors and concatenate them. The classification is then performed using the RoBERTa's classification head.

4.3 Combined context

Local and global contexts contain different information, thus their combination can provide a better model. The complete architecture using global and local contexts, here named Dual-CTX, is depicted in Fig. 3. A RoBERTa model receives the question and the candidate sentence with local context encoded as described in Section 4.1. The output of the Transformer is then combined with the global representation by using the strategy introduced in Section 4.2. The architecture is modular and extensible, local and global feature extraction modules can be easily exchanged. This flexibility can lead to the definition of several models. However, our main objective is to show the benefits of global and local contexts in the AS2 task. The exhaustive evaluation of all different context combinations and strategies is beyond our scope.

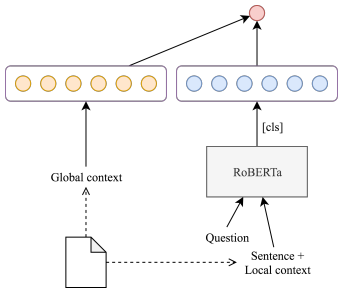

Fig. 3. Model combining global and local contexts.

5 Empirical assessment

We carried out comparative experiments to evaluate the local and global contexts and their combination.

5.1 Corpora

We used two AS2 corpora, ASNQ, and WikiQA, to empirically assess the proposed contextual architecture. Additionally, to better collocate our research in a broader QA work, we tested our model on SQuAD, which is a standard MR dataset, adapted for AS2.

ASNQ, Answer Sentence Natural Question [5] is a large-scale open-domain corpus for AS2. The corpus is built by transforming the recently proposed Natural Question (NQ) dataset [8] corpus from MR into AS2. In short, the corpus consists of 57,242 distinct natural questions for training and 2,672 for development. For each question, candidate answers have been extracted from a single Wikipedia page. The original NQ defines a long answer (typically a paragraph) and a short answer inside the associated page, whereas the ASNQ splits the document into sentences, whose binary label is 1 if the sentence contains the short answer, 0 otherwise. The corpus contains 21,307,630 question/answer pairs, with an average of 356 answer candidates per question.

Table 3. Questions (Q) and question/answer (QA) pairs available for training.

Corpus	Q	QA pairs
ASNQ	59914	21,307,630
WikiQA	2,117	20,374
SQuAD	87,355	448,108

WikiQA [28] is an open-domain corpus containing queries sampled from Bing logs. Based on the user clicks, the questions have been associated with a Wikipedia page (only the summaries were used). We used the clean setting for which only questions having at least one good and one wrong answer are considered. The resulting corpus consists of 2,118 training, 126 development, and 243 test questions, with about 10 candidate answers per question on average. We merged the dev. and test sets as they are too small to derive reliable results from each of them individually. Overall, we have 2,117 questions and 20,374 question/answer pairs.

SQuAD 1.1, Stanford Question Answering Dataset [17], is a large-scale corpus consisting of questions crowdsourced on a set of 20,000 Wikipedia articles. The dataset was designed for MR. We transformed it into a corpus for AS2 task, by applying the same procedure described by [5]. In short, we split each input paragraph into sentences and labeled those containing the annotated answers as correct candidates, and all the others as negative candidates. After this preprocessing, our corpus contains 87,355 questions and 448,108 question/answer pairs. Please note that the results presented in this paper are not directly comparable to the SQuAD leaderboard[2].

The main characteristics of the datasets are briefly described in Table 3.

5.2 Models

We implemented our methods with RoBERTa pre-trained models, using the shared checkpoint [27]. We fine-tuned the checkpoint on our data by using (i)

[2] https://rajpurkar.github.io/SQuAD-explorer/

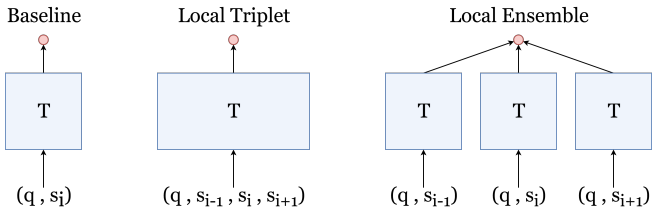

Fig. 4. Our three approaches to encode local context: a simple Transformer with question/answer pairs (left), the contextual multi-sentence architecture (center), and the ensemble of Transformers (right).

the Adam optimizer set with the warmup linear scheduler and a learning rate peak of $1e-6$; (ii) the binary cross-entropy loss; (iii) a batch size of 64 examples on a single GPU to train on WikiQA and SQuAD; and (iii) a batch size of 512 examples distributed on 8 GPUs to train on the ASNQ corpus (which is much larger). We used the official dev. set to derive the results, thus we set the hyperparameters, i.e., learning rate, scheduler, and batch size, on a small portion of the training set (as our dev. set). We train and test our models on SQuAD and WikiQA four times and take the average results to account for their variability.

Finally, we also used the models generated with TANDA (transfer and adapt) approach [5] for WikiQA. The authors apply a first fine-tuning on ASNQ and then a second fine-tuning on the target data. TANDA is the current state of the art, 7-10 points better than any other approach on WikiQA. Our models based on local context are depicted in Fig. 4, and described below:

- Transformer: the Transformer model for AS2 introduced in Sec. 3.2. It receives the question/answer pair as input without any context.
- Local Triplet (LOC_T): the proposed Transformer-based method described in Sec. 4.1, which relies on three different sentences, i.e., the previous, the target, and the next;
- Local Ensemble (LOC_E): an ensemble of three Transformer models encoding the three pairs, q/s_{i-1}, q/s_i, and q/s_{i+1} and a final linear layer fed with the concatenation of the *[CLS]* embeddings of the three models. The latter do not share their weights except those from *[CLS]*. The ensemble is the most expensive approach.

The baseline models for encoding global context are:

- Global BOW (GLOB_B): the global context described in Sec. 4.2 consisting of a simple Transformer model with a (compressed) BOW feature set on the top;
- Global Embedding (GLOB_E): a document embedding constituted by the average of the embeddings derived from all document sentences. We extract the sentence embedding using RoBERTa fine-tuned on ASNQ. We concatenate the average with the *[CLS]* representation output by the AS2 Transformer model.

We set the max sequence length of the input text to 128 tokens for LOC_T, GLOB_B/E, and each branch of LOC_E, whereas the contextual architecture LOC_T uses sequences up to 256 tokens, which cover a larger input.

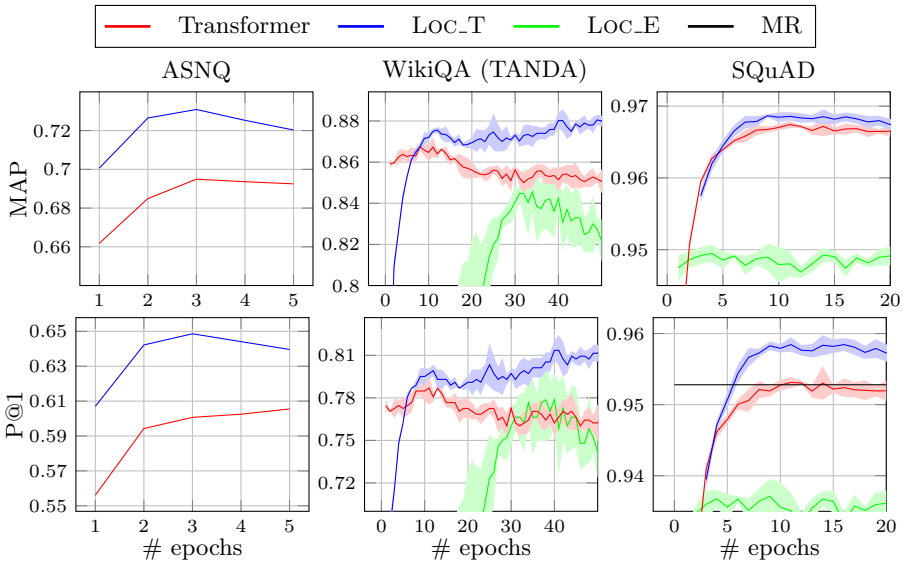

Fig. 5. Local context results (including standard deviation) computed on the dev. sets.

5.3 Results

We tested different context models on three different datasets using the state-of-the-art model in AS2 as our baseline, i.e., the transformer model made available in [5]. The latter improves all previous AS2 models 7-10 points, on WikiQA and TREC-QA datasets.

Local context Fig. 5 shows the Mean Average Precision (MAP) and the Precision at 1 (P@1) for each epoch for the Transformer, Loc_t, and Loc_e models. The plots show two main results: first, the superior accuracy of Loc_t is evident on all corpora, demonstrating that the local context has a positive impact on the AS2 model accuracy. Additionally, the performance of Loc_e method shows that the mere use of more information is not sufficient: its arrangement into the model is fundamental. Indeed, the simple aggregation of the three summarized context vectors seems not able to capture sentence dependencies: disarranged information produces noise, with a consequent drop in performance.

Next, we used an MR Transformer [27] to implement a sentence selector model. Our MR approach achieves 0.881 of F1 score on MR task (showing competitive results on the SQuAD leaderboard with respect to single models). Then, we simply select the sentence from which the MR extracts the answer span to solve the AS2 task on SQuAD: the model achieves a P@1 of 0.952. Fig. 5 shows that such model (straight line) is comparable to our baseline (single Transformer models), whereas Loc_t achieves better performance, 0.96. Although this is a loose comparison, it suggests that our approach may be applied to develop new

Table 4. Input examples from WikiQA and SQuAD.

q	**What happened to "The Glades" tv series?**
s_{i-1}	The Glades was renewed by A&E for a third season on October 18, 2011, which aired from June 3 to August 12, 2012.
s_i	The show has been renewed for a fourth season.
q	**What field of computer science is primarily concerned with determining the likelihood of whether or not a problem can ultimately be solved using algorithms?**
s_i	Closely related fields in theoretical computer science are analysis of algorithms and computability theory.
s_{i+1}	A key distinction between analysis of algorithms and computational complexity theory is that the former. . . , whereas the latter asks a more general question about all possible algorithms that could be used to solve the same problem.

MR methods. Table 4 illustrates interesting examples of answers correctly selected by LOC_T but misclassified by the baseline (which does not exploit any context). For example, the baseline could not reliably link *the show* to *The Glades*: this prevented to select the correct s_i as the top answer. In contrast, LOC_T contains such name in s_{i-1}.

Global context Fig. 6 shows the MAP and the P@1 achieved by the simple Transformer and the two global models, i.e., GLOB_B and GLOB_E. We also report the results of the combined model, which includes local and global contexts. Finally, we evaluated the models when applied to WikiQA without the TANDA approach, showing their behavior in a scenario, where there is no large and general data for the first fine-tuning step of TANDA.

The figure shows that both global methods, i.e., BOW and document embedding, improve the standard model both on WikiQA and SQuAD. We did not apply GLOB_B and GLOB_E to ASNQ as the training has a very large computational cost. This means that we cannot apply TANDA to WikiQA with such context. In any case, the global context produces an increase of accuracy on WikiQA and SQuAD (w/o TANDA). Concerning the combined model, DUAL-CTX improves the overall performance on WikiQA (w/o TANDA) and SQuAD. It does not improve the MAP of LOC_T on WikiQA when TANDA is used, but P@1 receives a significant boost. This result provides evidence that global and local features describe different (and potentially orthogonal) information.

It should be noted that we used BOW in the DUAL-CTX rather than the document embedding for computational reasons. The BOW representation can be efficiently computed, and it does not require dedicated hardware. Conversely, the document embedding requires the application of a RoBERTa model to each sentence composing the document. Moreover, the BOW representation can be highly improved, for instance, by learning the projection matrix. This is an interesting research line we would like to explore in the future.

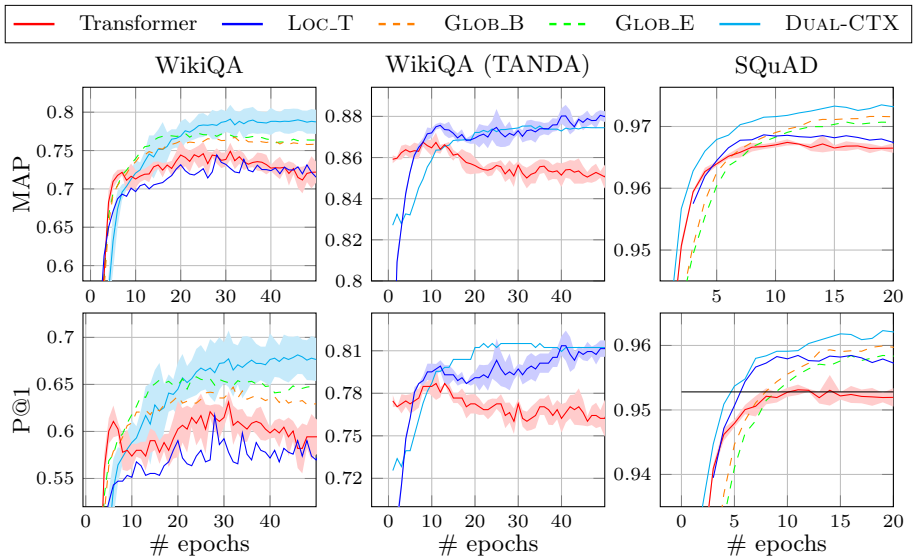

Fig. 6. Global context - empirical results computed on the development sets. The standard deviation is not always exposed to improve the readability.

6 Conclusion

AS2 is an important IR task, which provides an effective and efficient solution for the design of automated QA systems. Previous state-of-the-art models for AS2 only considered the question and the answer sentence candidate, without taking the context into account, and, to our knowledge, previous work did not use a context beyond the target sentence with Transformer models.

In this paper, we define two types of context, local and global. The former tries to solve implicit references in a candidate sentence, and it consists of the previous and successive sentences of a candidate answer. Conversely, the global context injects document related information, such as the main content and topics. We proposed Transformer-based architectures that leverage the different contexts for AS2. Our empirical assessment shows that our proposed approach remarkably improves over the TANDA model, which is the state of the art for AS2, on three different AS2 datasets, i.e., ASNQ, WikiQA, and SQuAD 1.1 adapted for AS2. It should be stressed that we used the model made available by the TANDA's authors, thus our results are perfectly comparable with their model. We also release our contextualized checkpoints and the SQuAD adaption for AS2[3]. In addition to some follow up in [6], interesting future extensions of our work regard the extraction of features from the entire rank of documents retrieved for a question. Clearly, learning to rank features can also improve the selection of answer sentences.

[3] https://github.com/alexa/wqa-contextual-qa

References

1. Alberti, C., Lee, K., Collins, M.: A bert baseline for the natural questions. arXiv preprint arXiv:1901.08634 (2019)
2. Chen, D., Fisch, A., Weston, J., Bordes, A.: Reading Wikipedia to answer open-domain questions. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1870–1879. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-1171
3. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., Salakhutdinov, R.: Transformer-XL: Attentive language models beyond a fixed-length context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2978–2988. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1285
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
5. Garg, S., Vu, T., Moschitti, A.: Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 7780–7788 (2020)
6. Han, R., Soldaini, L., Moschitti, A.: Modeling context in answer sentence selection systems on a latency budget. In: Proceedings of The 16th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Online (2021)
7. Kumar, S., Mehta, K., Rasiwasia, N., et al.: Improving answer selection and answer triggering using hard negatives. In: EMNLP-IJCNLP (2019)
8. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al.: Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics **7**, 453–466 (2019)
9. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
10. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L.S., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. ArXiv **abs/1907.11692** (2019)
11. Matsubara, Y., Vu, T., Moschitti, A.: Reranking for efficient transformer-based answer selection. In: Huang, J., Chang, Y., Cheng, X., Kamps, J., Murdock, V., Wen, J., Liu, Y. (eds.) Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020. pp. 1577–1580. ACM (2020). https://doi.org/10.1145/3397271.3401266, https://doi.org/10.1145/3397271.3401266
12. Nogueira, R., Cho, K.: Passage re-ranking with BERT. CoRR **abs/1901.04085** (2019), http://arxiv.org/abs/1901.04085
13. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. CoRR **abs/1802.05365** (2018), http://arxiv.org/abs/1802.05365

14. Qiao, Y., Xiong, C., Liu, Z., Liu, Z.: Understanding the behaviors of BERT in ranking. CoRR **abs/1904.07531** (2019), http://arxiv.org/abs/1904.07531
15. Qu, C., Yang, L., Qiu, M., Croft, W.B., Zhang, Y., Iyyer, M.: Bert with history answer embedding for conversational question answering. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1133–1136 (2019)
16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2018), https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
17. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822 (2018)
18. Severyn, A., Moschitti, A.: Learning to rank short text pairs with convolutional deep neural networks. In: SIGIR. ACM (2015)
19. Shao, T., Guo, Y., Chen, H., Hao, Z.: Transformer-based neural network for answer selection in question answering (2019)
20. Shen, G., Yang, Y., Deng, Z.H.: Inter-weighted alignment network for sentence pair modeling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1179–1189. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1122, https://www.aclweb.org/anthology/D17-1122
21. Soldaini, L., Moschitti, A.: The cascade transformer: an application for efficient answer sentence selection. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5697–5708. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.504, https://www.aclweb.org/anthology/2020.acl-main.504
22. Tan, C., Wei, F., Zhou, Q., Yang, N., Du, B., Lv, W., Zhou, M.: Context-aware answer sentence selection with hierarchical gated recurrent neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2017)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
24. Wang, M., Smith, N.A., Mitamura, T.: What is the Jeopardy model? a quasi-synchronous grammar for QA. In: EMNLP-CoNLL. pp. 22–32. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), https://www.aclweb.org/anthology/D07-1003
25. Wang, S., Jiang, J.: A compare-aggregate model for matching text sequences. CoRR **abs/1611.01747** (2016), http://arxiv.org/abs/1611.01747
26. Wang, Z., Ng, P., Ma, X., Nallapati, R., Xiang, B.: Multi-passage bert: A globally normalized bert model for open-domain question answering. arXiv preprint arXiv:1908.08167 (2019)
27. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
28. Yang, Y., Yih, W.t., Meek, C.: Wikiqa: A challenge dataset for open-domain question answering. In: EMNLP. pp. 2013–2018 (2015)
29. Yang, Y., Yih, W.t., Meek, C.: WikiQA: A challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2013–2018. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). https://doi.org/10.18653/v1/D15-1237, https://www.aclweb.org/anthology/D15-1237

30. Yoon, S., Dernoncourt, F., Kim, D.S., Bui, T., Jung, K.: A compare-aggregate model with latent clustering for answer selection. CoRR **abs/1905.12897** (2019), http://arxiv.org/abs/1905.12897