Semantic Aware Answer Sentence Selection using Self-Learning based Domain Adaptation

Rajdeep Sarkar*
Data Science Institute, National University of Ireland Galway
Galway, Ireland
rajdeep.sarkar@insight-centre.org

Sourav Dutta
Huawei Research
Dublin, Ireland
sourav.dutta2@huawei.com

Haytham Assem†
Amazon Alexa AI
Cambridge, UK
hithsala@amazon.co.uk

Mihael Arcan
Data Science Institute, National University of Ireland Galway
Galway, Ireland
mihael.arcan@insight-centre.org

John McCrae
Data Science Institute, National University of Ireland Galway
Galway, Ireland
john.mccrae@insight-centre.org

ABSTRACT

Selecting an appropriate and relevant context forms an essential component for the efficacy of several information retrieval applications like *Question Answering (QA)* systems. The problem of *Answer Sentence Selection (AS2)* refers to the task of selecting sentences, from a larger text, that are relevant and contain the answer to users' queries. While there has been a lot of success in building AS2 systems trained on open-domain data (e.g., SQuAD, NQ), they do not generalize well in *closed-domain* settings, since domain adaptation can be challenging due to poor availability and annotation expense of domain-specific data. This paper proposes *SEDAN*, an effective self-learning framework to adapt AS2 models for domain-specific applications. We leverage large pre-trained language models to automatically generate *domain-specific QA pairs* for domain adaptation. We further fine-tune a pre-trained Sentence-BERT architecture to capture semantic relatedness between questions and answer sentences for AS2. Extensive experiments demonstrate the effectiveness of our proposed approach (over existing state-of-the-art AS2 baselines) on different Question Answering benchmark datasets.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Question answering**; • **Computing methodologies** → **Information extraction**.

KEYWORDS

Answer Sentence Selection, Domain Adaptation, Self Learning, Question Answering, Semantic Learning, Language Models

*Work done during internship at Huawei Research, Ireland.
†This work was conducted when the author was working at Huawei Research, Ireland.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539162

ACM Reference Format:
Rajdeep Sarkar, Sourav Dutta, Haytham Assem, Mihael Arcan, and John McCrae. 2022. Semantic Aware Answer Sentence Selection using Self-Learning based Domain Adaptation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539162

1 INTRODUCTION

Information Retrieval involves the precise extraction of succinct information chunks (generally from a larger data source) that meet the application's needs. In recent times, Question Answering (QA) approaches have become a ubiquitous part of intelligent search, chatbots, intent detection, and virtual personal assistants (e.g., Alexa, Siri, etc.) catering to a plethora of tasks. As a result, there have been increased efforts in improving the performance of such systems. Researchers have primarily focused on two related components of QA systems: *Answer Sentence Selection (AS2)* and *Machine Reading Comprehension (MRC)*. Given a user question and a relevant document, AS2 systems focus on selecting candidate sentence(s) (i.e., context), from the document, that contains the probable answer (with a high probability) to the user's question. On the other hand, MRC involves the extraction of text spans from the *selected context* that correctly answers the question – also referred to in the literature as Extractive Question Answering (EQA).

AS2 involves the selection of sentences that contain the information required to answer a given question and forms an important problem in its own right. It enables several applications like information search snippet generation, summarization and knowledge base generation, as well as in the larger context of open domain question answering [35]. The relevance of an answer sentence to a question is typically determined by measuring the semantic similarity between the question and answer. Traditional systems like BM25 were initially used to gauge information relatedness between questions and answers based on the bag-of-words (BoW) model. With the advent of neural models [6, 11, 12, 17, 18], such architectures have been shown to perform extremely well in the AS2 task. Transformer [32] based language models such as BERT [8] and RoBERTa [20] have enabled the capture of the contextual relationship between words within a sentence, to achieve state-of-the-art

Q	**Which drug should be used as an antidote in benzodiazepine overdose?**
S1	*A 54-y-old man ingested 2 g of bulk laboratory diazepam and was treated with activated charcoal, enhanced diuresis and flumazenil infusion.*
S2	The treatment resulted in awakening, but the patient had drowsiness, dysarthria, diplopia, and dizziness for 9 d. Blood levels of diazepam and its main metabolite, nordiazepam, were obtained for 1 mo.
S3	The half-lives in this benzodiazepine overdose were longer than those seen with therapeutic doses.
S4	Benzodiazepines should not be readministrated when patients awake after suicide attempts.

Table 1: Challenges in domain-specific AS2. Only sentence S1 is relevant to query Q and requires understanding of relation between *D*iazepam & *B*enzodiazepine.

results in various natural language processing tasks. Such semantic understanding of texts has also been successfully adapted to the AS2 task [11, 17, 18].

Although research on QA systems is increasingly gaining traction, large documents (like Wikipedia articles) provided to QA systems contain diverse information, which, although contextually relevant, might contain completely unrelated information to the posed question. The presence of such a large "noisy" context might adversely affect the performance of the QA pipeline. Hence, identifying potentially relevant text snippets from a large context becomes pivotal for enhanced QA performance – impressing the need for efficient and accurate AS2 methodologies. In fact, the reduction of "noisy" irrelevant context (w.r.t. to a question) has been shown to improve QA performance in production settings [11].

This challenge is more compounded in *domain-specific* settings due to the need for specialised understanding in terms of domain semantics, terminology, and relationships. For example, Table 1 shows a QA scenario from the biomedical domain. Given question **Q**, we can observe that only sentence **S1** is pertinent for answering the question. Sentences S2, S3 and S4, although relevant within the domain, are incapable of answering **Q**. Note, that an AS2 system in this scenario needs to understand the implicit relation that `Diazepam is a type of Benzodiazepine` to correctly identify the answer sentence. This depicts the need for *domain-aware AS2 models* for subsequent improvements of end-to-end QA platforms.

Domain adaptation has been relatively unexplored in the context of Answer Sentence Selection (AS2). The recent work of Garg et al. [11] suggested a two-step fine-tuning of BERT-based architecture for domain adaptation. However, such models require expert-annotated domain-specific data, wherein the creation of such datasets can be a time-consuming and expensive process. Additionally, the availability of domain experts can be challenging in several specialised areas and enterprise settings. Furthermore, to the best of our knowledge, the impact of using AS2 models on different downstream tasks such as Extractive Question Answering (EQA) has not yet been explored.

In this work, we propose the *Self-lEarning for Domain adaptation in Answer seNtence selection* (SEDAN) framework to ameliorate the above issues. We show that *SEDAN* can automatically adapt itself to a specific domain via *self-learning* without the need for any explicitly annotated training data. We utilise fine-tuned *Sentence-BERT* architecture to capture the semantic relationships between a question and answer texts – for *SEDAN* to extract appropriate answer sentences in closed-domain as well as open-domain settings. We further study the performance impact of *SEDAN* (and other baselines) on the downstream EQA task. In summary, our contributions are:

- *SEDAN*, a *self-learning* based framework for Answer Sentence Selection (AS2) using *two-stage* fine-tuning.
- Domain adaptability of *SEDAN* using synthetically generated QA pairs from domain-specific documents by leveraging large pre-trained language models.
- A Siamese network based fine-tuning of Sentence-Transformers (S-BERT) to capture semantic understanding between questions and answer sentences.
- Extensive evaluation to demonstrate strong *evidential empirical performance improvements* of *SEDAN*(over other baselines) on AS2 task, as well as in the downstream task of EQA on multiple benchmark datasets.

The framework is currently under the development process of deployment for automated chatbot solutions across various products and offerings.

2 RELATED WORK

Neural network models have been studied extensively in search, information extraction, text reranking [10], AS2 task [11, 17, 18] and for measuring textual semantic similarity [8, 20, 27]. In the context of AS2, researchers have employed Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to learn the semantic relationship between text pairs and in turn rank textual context relevant to user needs. Severyn and Moschitti [29] employed a CNN network to learn a similarity function between pairs of texts, while Tan et al. [31] proposed modelling context information with hierarchical gated recurrent neural networks.

While such models perform well, transformer [32] based AS2 models [11, 17, 18] have been shown to significantly outperform CNN and RNN based models. Garg et al. [11] proposed fine-tuning of transformer-based language models for the AS2 task. They released a large scale AS2 dataset and showcased the need for a two-step fine-tuning process for domain adaptation in AS2. Specifically, they suggest the first step of fine-tuning on a general-purpose AS2 dataset to adapt transformer-based language models for the AS2 task. The authors recommend a second stage of fine-tuning on a domain-specific dataset for effective domain adaptation. Lauriola and Moschitti [18] suggested the use of a global context in addition to the local context used by Garg et al. [11] to improve the performance of transformer-based AS2 model. In this context, hierarchical transformer architectures have also been explored [30].

The above models use the $[CLS]$ token representation for sentence or QA pair embedding to capture the semantic relatedness. However, Reimers and Gurevych [27] showed that such embeddings lead to sub-optimal embeddings and can adversely impact the performance of models on downstream tasks. Additionally, such models need expert annotated domain-specific training data for effective domain adaptation – to understand contextual relevance between questions and the answers for the particular domain. Sentence-Transformers [27] are powerful models for learning rich contextual sentence embeddings by leveraging Siamese networks. In *SEDAN* we adopt the two-stage fine-tuning process, albeit

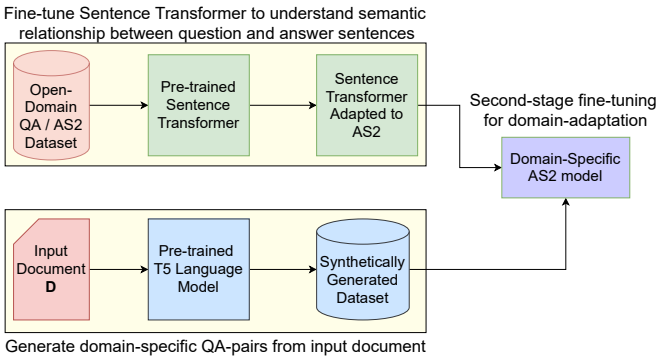

Figure 1: Modular overview of proposed framework, *SEDAN*. The first step involves fine-tuning transformer model on a generic QA dataset (e.g., SQuAD). The second fine-tuning utilises domain-specific synthetic QA pairs generated using pre-trained large language model (e.g., T5).

with the crucial difference of utilising *self-learning on Sentence-Transformers*.

The use of synthetic dataset generation for training has been posited in MRC. Alberti et al. [1] follow a two-step process for synthetic QA dataset generation, where they train a BERT [8] model to generate answers from the context passage. They train another BERT model that generates the question using the answer generated from the previous step and the passage context. Chan and Fan [5] improved upon the question generation by leveraging sequential neural architectures built on top of BERT. While BERT models have been largely used to generate synthetic data, Puri et al. [24] used a GPT-2 [25] model to generate QA pairs and used BERT model to filter QA pairs using round-trip consistency. Additionally, Lu et al. [21] utilised the generation of a synthetic dataset for neural passage retrieval. To the best of our knowledge, training AS2 models using synthetically generated datasets has not yet been fully explored. Deng et al. [7] augmented a golden labelled dataset with an additional synthetic dataset for the ranking if question and answer sentences pairs. As outlined in [3], *SEDAN* uses a T5 model to generate QA pairs and uses this synthetic dataset to adapt an AS2 model to a closed-domain. In contrast to previous AS2 work, our work uses fine-tuned Sentence Transformers to learn rich contextual information between the question text and the answer sentences. Similar to Reimers and Gurevych [27], we use a Siamese network to fine-tune a BERT model to generate sentence embeddings. Moreover, this is the first work on using a synthetic dataset for effective domain adaption of AS2 models via *self-learning*. Additionally, to analyse the effectiveness of *SEDAN* framework, we compare the performance of *SEDAN* framework on the EQA task.

3 *SEDAN* FRAMEWORK

We now formally introduce the problem statement and describe the proposed *SEDAN* framework. We discuss how domain-specific QA pairs are automatically generated for domain adaptation and the subsequent fine-tuning of Sentence-BERT for suiting the AS2 task.

Task Definition: Given a question, AS2 involves selecting the answer sentence(s) from a large context passage. Answer sentences are those sentences in the context passage that potentially contain the answer to the question posed. More formally, given a context passage C_i consisting of sentences $S_i = \{S_{i_1}, S_{i_2}, \cdots S_{i_n}\}$ and a question Q pertaining to the context passage C_i, the aim is to learn a ranking function $f : Q \times S_i \rightarrow \mathbb{R}$ that assigns a score to each of the Question-Answer sentence pairs, where a higher score indicates a higher probability of the sentence containing the answer of Q.

3.1 Synthetic Domain Dataset Generation

The synthetic dataset generation module is an essential component of *SEDAN* for *self-learning*. The synthetic data generation in *SEDAN* is similar to that detailed in [3] (can also be replaced with a different data generation scheme). We use a pre-trained T5 model to generate QA pairs from a domain-specific document D. Since transformer models can process only 512 tokens at a time, we split the document into chunks of K_{GQA} sentences, to prevent the T5 model from ignoring document text, while preserving domain-specific contextual information. Once the data is chunked into multiple short contexts, we follow a two-step process to generate QA pairs.

Answer Generation. Given the document, D, chunked into contexts, $C_1, C_2, \cdots, C_n$, we use each C_i to generate QA pairs. Similar to Puri et al. [24], given a context passage C_i, we split it into sentences $S_{i_1}, S_{i_2}, \cdots, S_{i_m}$. We use the publicly available *T5-small model*[1], pre-fine-tuned on SQuAD1.1, to generate possible answer spans given a sentence, to obtain answers a_j from sentences S_{i_j}.

Question Generation. The question generation of *SEDAN* uses another publicly available pre-trained T5-small model[2], which is trained for question generation on SQuAD1.1 dataset, using the method described in Chan and Fan [5]. Given the generated answer a_j as above, sentence S_{i_j} along with the corresponding context passage C_i, the model generates a potentially relevant question q_j based on the probability $p(q_j \mid a_j, S_{i_j}, C_i)$ using a beam search mechanism.

In our framework, to generate the synthetic dataset, we first concatenate all the context passages from the test dataset to create the domain-specific document D. From the chunked contexts, $C_1, C_2, \cdots, C_n$ (from D), we obtain the synthetically generated QA pairs forming the training data for domain adaptation of *SEDAN*. Observe, that this process of self-learning in *SEDAN* is unsupervised.

3.2 Fine-tuning Transformers for AS2

Answer Sentence Selection (AS2) involves both semantic and syntactic information to establish what information the question seeks, as well as whether a candidate sentence fulfils the requirement. Transformers [32] are a powerful architecture to capture syntactic and semantic relationships between words in natural languages. Pre-trained transformers such as BERT [8], RoBERTa[20] or DistilBERT [28] have shown to be effective in learning language models, and have been used for AS2. Prior work [11, 18] concatenated the question and the answer sentence tokens to learn their joint representation, which is then fed to a multi-layer perceptron for answer-sentence classification. Here, the common practice is to use the [*CLS*] token embedding to derive sentence representations –

[1] https://huggingface.co/valhalla/t5-small-qa-qg-hl
[2] https://huggingface.co/valhalla/t5-small-qg-hl

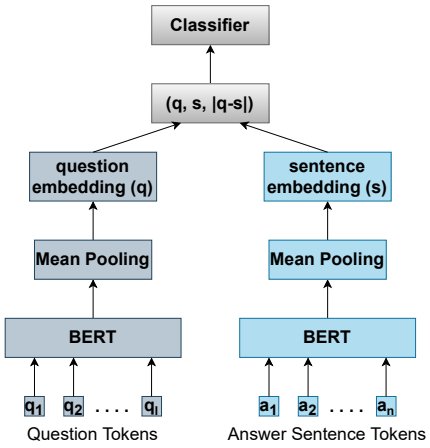

Figure 2: Siamese Network architecture for Transformer fine-tuning in *SEDAN*.

shown to lead to sub-optimal sentence representation [27]. Hence, following the architecture of Reimers and Gurevych [27], we use a Siamese network to fine-tune pre-trained sentence representations to understand the contextual similarity between questions and relevant answer sentences. To this end, *SEDAN* utilises the SBERT-BERT-BASE-NLI[3] model as the base for the *two-stage* fine-tuning.

3.2.1 Question-Sentence Semantic Understanding. The initial fine-tuning of the transformer model adapts the architecture towards generic AS2 systems and trains it to understand semantic similarity-based relevance between questions and corresponding answer sentences. In this respect, we use the SQuAD2.0 dataset, wherein given a question (Q), the related context (with sentences S_i), and the corresponding answer a, we mark the sentences that contain the answer text as *positive samples* while the other sentences are considered as *negative points*. Thus, a question-sentence pair (Q, S_i) is marked as 1 if $a \in S_i$, or else considered as 0.

The above setup is used to tune the transformer architecture based on learning via a Siamese network. The BERT model takes as input the text tokens and outputs their contextual embeddings, which are then mean-pooled to obtain the textual representation. We compute the embeddings for both the questions and sentences (within the context). The question encoding $\mathbf{q}$ and sentence embedding $\mathbf{s}$ (for the question-sentence pairs obtained above) are then concatenated along with their difference $|\mathbf{q} - \mathbf{s}|$. This question-sentence representation along with its label (whether the answer is present in the sentence) is then provided as input to a shallow classifier network to learn to identify potential answer sentences for a question. This fine-tuning network is shown in Figure 2. Mathematically it is represented as:

$$\mathbf{q} = \text{BERT}(q) \text{ and } \mathbf{s} = \text{BERT}(s),$$
$$r = concat(\mathbf{q}, \mathbf{s}, |\mathbf{q} - \mathbf{s}|) \text{ and}$$
$$p(s) = softmax(r\mathbf{W})$$

[3]https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

Dataset	Generated Train Data			Test Dataset		
	#Q	#C	#QA Pairs	#Q	#C	#QA Pairs
BioASQ	3,914	422	39,750	141	451	4,564
TextbookQA	2,787	361	12,544	448	117	22,141
DROP	766	81	8,135	406	85	4,682
HotpotQA	40,841	4,979	901,416	5,901	5,860	94,646

Table 2: Characteristics of the *synthetically generated* training datasets. BioASQ and TextbookQA are *closed-domain* data, while DROP and HotpotQA are *open-domain*.

where $\mathbf{W}$ is a learnable weight matrix. Observe, our training of the Siamese network is similar to the fine-tuning setting of S-BERT [27] on the SNLI corpus [4], and captures semantic similarity between the questions and answer sentences.

3.2.2 Self-Learning for Domain Understanding. The final fine-tuning of *SEDAN* involves training it to understand domain-specific contexts and relationships, as observed in Table 1. Here, we train *SEDAN* using the same Siamese network setting as discussed above, albeit with the synthetically generated domain-specific dataset using a pre-trained T5 model (as mentioned in Section 3.1).

Similar to the first-stage fine-tuning, for the generated QA pairs from the document, context sentences that contain the generated answers are considered as *positive samples*. The transformer model of *SEDAN* is now fine-tuned on these generated domain question-sentence pairs – enabling *domain adaptation via self-learning*.

Thus, the overall training of our proposed *SEDAN* framework, as shown in Figure 1, involves (a) Siamese-network-based fine-tuning transformer architecture to generic QA scenarios, (b) synthetic generation of domain-specific training data, and (c) domain adaptation via self-learning – providing an efficient methodology to identify potential answer sentences for domain-specific user questions.

3.2.3 Answer Sentence Selection (AS2): On arrival of a user question, the associated context is first split into sentences. Question-Sentence pairs are created and their concatenated representations are obtained from the fine-tuned transformer of the *SEDAN* framework. The representations are then fed to the trained Siamese network for classification, to identify potential sentences that are relevant to the posed user question. The sentence(s) may be returned as an answer snippet or provided to downstream EQA for exact answer extraction.

4 EXPERIMENTAL SETUP

This section describes our empirical setup for evaluating *SEDAN* against existing baseline methods. We introduce the datasets, the baselines compared with, and the implementation details.

4.1 Datasets

As discussed previously, *SEDAN* uses a two-step fine-tuning process for the domain-aware AS2 task. In the first step, we adapt the pre-trained Sentence-BERT model to the QA task using SQuAD2.0 dataset [26], containing 129,353 unique questions and a total of 334,364 QA sentence pairs. *SEDAN* utilises the synthetically generated dataset from the T5 module as detailed in Sec. 3.1 for adapting

the model to a specific domain. The datasets are obtained from huggingface.co/datasets/mrqa, and their details are shown in Tab. 2.

- **BioASQ** [22]: BioASQ involves information extraction on biomedical semantic indexing and QA pairs annotated by domain experts. We take 70% of the dataset as training and 30% as the test set. The test split contains 141 unique questions, 451 unique contexts and 4,564 QA sentence pairs. The synthetic dataset contains 3,914 and 422 unique questions and contexts respectively, with 39,750 QA sentence pairs.
- **TextbookQA** [15]: This dataset contains QA pairs from middle school science curricula. We split the dataset into 70% for training and 30% for testing. The test dataset contains 448 and 117 unique questions and contexts respectively, with 4,682 QA sentence pairs. The synthetic dataset contains 2,787 and 361 unique questions and contexts, with a total of 12,544 QA sentence pairs.
- **DROP** [9]: DROP dataset involves quantitative reasoning questions over Wikipedia paragraphs. The original dataset contains numeric answers to be inferred from the context, however, we only consider those questions where the answer is explicitly mentioned in the passage text. The test dataset contains 406 questions, 85 passages, and 4,682 QA sentence pairs. The generated synthetic dataset contains 766 unique questions over 81 passages 8,135 QA sentence pairs.
- **HotpotQA** [33]: HotpotQA is a multihop QA dataset to answer questions over Wikipedia. The test contains 5,901 unique questions, 5,860 unique contexts and 94,646 QA sentence pairs. The synthetic data contains 40,841 unique questions, 4,979 contexts with 901,416 QA sentence pairs.

Note, BioASQ and TextbookQA are *closed-domain*, while DROP and HotpotQA are *open-domain* – showcasing the performance of *SEDAN* in AS2 for diverse settings. For the above datasets, we evaluate if the sentence(s) retrieved by an AS2 model contains the ground truth answer text span within them.

Note that the original questions in the test set are not visible to *SEDAN* during the domain adaptation process, and the synthetic QA pairs (used during self-learning based training) are generated from the test documents. It should be noted that we do not use the training dataset during the training procedure.

4.2 Baseline Approaches

We compare our proposed approach against several state-of-the-art AS2 models as well as passage ranking models as baselines.

- **Dual-CTX** [18] is a RoBERTa [20] based state-of-the-art AS2 model, utilising global article context along with local passage context. It was also shown that using the TANDA approach, based on an initial fine-tuning on the ASNQ dataset, improves the performance of AS2 on other data (referred to here as Dual-CTX-TA).
- **Reranker** [10] is a lightweight and efficient approach based on language model based re-ranking for information retrieval (IR), QA and other natural language processing tasks. The model uses localised contrastive estimation loss to fine-tune a transformer model for text reranking.
- **Propagate-Selector** [34] utilises graph structures for detecting candidate answer sentences. It constructs a graph using sentences in the passage as nodes, with edges between sentences and questions, for utilising Graph Neural Network (GNN) for AS2.
- **TF-Ranking** [23] is a library for Learning-to-Rank (LTR) framework for IR tasks using the TensorFlow platform.
- **BM25** [2] is a traditional IR model based on BoW model and uses term frequency and inverse document frequency to rank answer sentences given a question.

4.3 Evaluation Measures

We evaluate the different AS2 models based on the following:

- *Mean Reciprocal Rank (MRR)* – It is defined as the average harmonic mean (over all questions) based on the rank of the first correct answer sentence reported, given a question. Mathematically, $MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_{s_i}}$, where $rank_{s_i}$ denotes the rank of the first correct answer sentence returned for question Q_i, and $|Q|$ denotes the total no. of questions.
- *Mean Average Precision (MAP)* – It computes the mean (over all questions) of the average precision in identifying the answer sentences of a question. Thus, $MAP = \frac{1}{|Q|}\sum_{i=1}^{|Q|} AP_i$, where $AP_i = \frac{1}{K_i}\sum_{j=1}^{K_i} P@j$ computes the accuracy of the answer (for query Q_i) being present within the top-j sentences returned (i.e., $P@j$), and K_i denotes the total number of answer sentences.
- *Recall@k (R@k)* – It measures the fraction of the total number of correct answer sentences (within the context) retrieved in the top-k sentences by the methodologies. We report for $k = \{1, 3, 5\}$. Mathematically, we have,
  $R@k = \frac{\text{\# correct answer sentences retrieved}}{\min(\text{\# total answer sentences, k})}$

For the Extractive Question Answering (EQA) use-case application task, we measure the performance of the models on Exact Match (EM) and *F1* metrics. EM reports the ratio of the extracted answer spans that are equivalent to the ground truth answer, while *F1* uses precision-recall of the token (word) overlap between the extracted answer and the gold answers of the questions.

For a fair comparison, we train Reranker [10] with the BERT-base model. As the source code of Dual-CTX is not available online, we evaluate its performance on its best effort re-implementation with the RoBERTa-base model[4]. Observe, that the focus of this work is to alleviate the need for annotated domain-specific training data, and hence we do not consider any document or paragraph level information that might not be available. Hence, such information is not provided to the Propagate-Selector, and individual passages are used to construct the global context for Dual-CTX.

We also evaluate the impact of the baseline AS2 techniques on the downstream EQA task. For this, we consider the state-of-the-art *SpanBERT* model [14], a BERT model fine-tuned using span masking for predicting precise answer spans robustly. We use their publicly available model[5] fine-tuned on the SQuAD2.0 dataset.

[4]Our re-implementation produced similar results on SQuAD as reported by the authors.
[5]https://huggingface.co/mrm8488/spanbert-base-finetuned-squadv2

Model	BioASQ					TextbookQA				
	MRR	MAP	R@1	R@3	R@5	MRR	MAP	R@1	R@3	R@5
Dual-CTX	68.78	62.06	55.53	59.55	70.58	18.91	14.17	8.93	8.77	11.52
Dual-CTX-TA	71.68	64.83	59.95	63.79	72.03	21.16	16.17	10.26	11.75	15.90
Propagate-Selector	45.26	47.48	26.99	48.16	66.54	11.48	12.06	4.02	8.22	12.74
Reranker	72.51	65.49	59.29	64.63	74.42	55.13	42.21	43.75	40.77	**48.49**
TF-Ranking	54.92	50.92	36.06	49.19	64.59	16.17	13.08	6.03	7.51	11.43
BM25	71.30	64.59	56.19	64.89	75.56	52.23	32.69	39.06	42.63	45.79
SEDAN	**78.07**	**70.04**	**65.26***†	**72.16***†	**80.26***†	**57.26**	**43.57**	**44.86**†	**45.02**†	48.04†

Model	DROP					HotpotQA				
	MRR	MAP	R@1	R@3	R@5	MRR	MAP	R@1	R@3	R@5
Dual-CTX	39.92	38.02	13.15	49.35	60.89	61.37	57.10	47.43	59.67	69.32
Dual-CTX-TA	59.43	55.52	41.90	54.89	79.40	**67.33**	**64.44**	**53.00**	**69.23**	**79.17**
Propagate-Selector	40.40	41.23	17.41	41.90	75.15	42.11	44.33	24.06	46.85	66.68
Reranker	56.53	54.53	36.43	62.21	76.57	61.44	59.54	43.87	64.76	77.75
TF-Ranking	42.41	40.68	21.86	40.69	67.56	33.29	31.90	16.11	27.70	42.84
BM25	65.24	63.25	49.79	66.83	79.25	47.49	44.86	27.89	47.22	64.52
SEDAN	**73.79**	**70.71**	**59.71***†	**76.51***†	**86.41***†	64.19	61.82	48.11†	66.01†	78.64†

* refers to statistically significant result for SEDAN compared to Dual-CTX-TA, Reranker and BM25 with $p < 0.1$,
while † refers to statistically significant result for SEDAN compared to Propagate-Selector and TF-Ranking with $p < 0.01$.

Table 3: Performance of the algorithms trained using synthetically generated dataset on different test dataset for the AS2 task.

4.4 Implementation Details

We use spaCy [13] to split the provided contexts into sentences during the synthetic dataset construction via T5 model (Section 3.1). We set $K_{GQA} = 10$ for BioASQ, DROP and HotpotQA, while for TextbookQA it is set to 41 – based on the average length of the test documents in each dataset.

For the AS2 task, we train the transformer models in *SEDAN*, Reranker and Dual-CTX with a batch size of 256 on 2 Tesla V100 GPUs, while for the TF-Ranking model and Propagate-Selector we use batch sizes of 256 and 16 respectively, on NVIDIA 1080Ti GeForce GPUs. We optimise the models with the Adam optimiser [16] with a learning rate of $1e-6$ and a linear warm-up over 10% of the training data. We set the weight decay parameter to 0.01 for regularisation. We train all models for 50 epochs on the BioASQ and DROP dataset, and for 10 epochs on HotpotQA and TextbookQA dataset.

For evaluating on the EQA task, we prune the context passage to contain the top $K_{pruning}$ candidate answer sentences (based on AS2 classification scores) and the candidate sentences are concatenated and presented to the SpanBERT EQA model as context for the question posed. We set $K_{pruning}$ to min (50% of the number of document sentences, K_{min}). We set K_{min} for BioASQ, TextbookQA and DROP datasets to 4, 6 and 6 respectively.

It should be noted that no annotated training dataset was used in the *SEDAN* framework. We use the test documents (unannotated text alone) to generate the synthetic dataset from the passages for self-learning based domain adaptation. Since, such chatbot applications are predominantly online offerings, the document from which a user might ask questions (i.e., the test document) has been used to ensure deployment compatibility.

Model	BioASQ		TextbookQA		DROP	
	EM	F1	EM	F1	EM	F1
Full Context	39.60	49.62	27.90	35.87	23.07	28.49
Dual-CTX-TA	36.94	46.98	15.18	20.64	23.68	**29.82**
Reranker	39.38	49.39	25.89	33.85	21.65	27.72
TF-Ranking	30.75	38.53	16.74	21.78	18.62	23.73
BM25	36.94	47.09	27.23	34.99	20.04	27.05
SEDAN	**40.04**	**50.94**	**28.79**	**36.33**	**23.88**	29.01

Table 4: Performance of AS2 models for EQA task.

5 EMPIRICAL RESULTS

This section reports the performance results obtained by *SEDAN* along with the other methods for AS2 and EQA tasks on different *closed-* and *open-domain* datasets. We also analyse the semantic relationship between question and answer sentence embeddings learnt by *SEDAN*, by ablation study. For completeness, we explore the performance of the methodologies in settings when domain-specific training data might be available.

5.1 Performance of *SEDAN* on AS2 Task

Table 3 reports the performance of *SEDAN* along with other competing approaches trained using the synthetically generated dataset. For *closed-domain* or domain-specific datasets like BioASQ and TextbookQA, we observe that *SEDAN* outperforms all other baselines. *SEDAN* achieves a performance improvement of around **5%**

Model	BioASQ					DROP				
	MRR	MAP	R@1	R@3	R@5	MRR	MAP	R@1	R@3	R@5
Domain Fine-tuning + Siamese	71.03	64.17	55.97	65.04	78.27	61.20	58.23	39.47	71.28	82.62
Domain Fine-tuning + Cosine	71.68	65.26	55.97	67.51	79.17	64.70	62.04	45.95	71.22	82.60
Full Fine-tuning + Cosine	76.31	67.83	63.71	68.65	77.94	69.20	66.81	53.64	73.00	84.02
SEDAN (Full architecture)	**78.07**	**70.04**	**65.26**	**72.16**	**80.26**	**73.79**	**70.71**	**59.71**	**76.51**	**86.41**

Table 5: Ablation study on fine-tuning and classification modules in *SEDAN*.

Model	**BioASQ**					**TextbookQA**					**DROP**				
	MRR	MAP	R@1	R@3	R@5	MRR	MAP	R@1	R@3	R@5	MRR	MAP	R@1	R@3	R@5
Dual-CTX-TA	**91.10**	**88.91**	**85.39**	**91.85**	**95.28**	21.19	16.27	10.26	12.57	16.25	65.80	62.54	49.59	67.84	80.45
Reranker	84.42	78.07	74.55	79.68	88.27	54.22	41.70	41.96	42.18	48.00	73.59	70.73	59.31	76.31	84.00
BM25	71.30	64.59	56.19	64.89	75.56	52.23	39.69	39.06	42.63	45.79	65.24	63.25	49.79	66.83	79.25
SEDAN	89.21	85.00	83.18	86.20	92.12	**57.73**	**44.97**	**45.75**	**44.86**	**50.64**	**75.79**	**73.17**	**61.94**	**79.14**	**89.03**

Table 6: Performance of baseline methods in presence of gold training data.

on MRR, MAP and R@1 measures over the state-of-the-art Dual-CTX-TA and Reranker approaches, while Propagate-Selector and TF-Ranking report the lowest performances.

On the open-domain DROP dataset, we also achieve the best performance with a significant improvement over the baselines. This can be attributed to a better understanding of the Question text-Answer sentence semantic relationship in *SEDAN* across different regions of the document. While on HotpotQA, which is a multi-hop QA dataset, Dual-CTX-TA performs the best as it captures the global context using bag-of-words, in turn capturing the inter-sentence information, with *SEDAN* being the second best.

We conduct the one-sided Binomial test to study the statistical significance of the reported results on R@k metrics. As the majority of the questions across the datasets contain only one correct answer sentence, we treat the R@k values as Bernoulli variables. We consider the null hypothesis, H_0 : *The performance of method X is comparable to that of SEDAN*; where X∈{Dual-CTX-TA, Reranker, BM25} – the top-3 best-performing approaches. In Table 3, we find that the performance of *SEDAN* is significantly better than the baselines, thereby rejecting the above null hypothesis with a p-value $p < 0.1$. That is, the performance of *SEDAN* is observed to be *significantly better* than the existing methodologies in both *open- and closed-domain* datasets. Observe, that since the sample test size is quite low for the datasets (see Table 2, we consider a confidence threshold of 90% (instead of the standard 95%) for significance testing, as noted in Leamer [19] for data with fewer samples.

It should be noted that all models are fine-tuned on the domain-specific data synthetically generated as in *SEDAN*. We later show in Section 5.3, that without such fine-tuning, the embeddings are not rich enough to capture domain-specific semantic information. In the remaining experiments, we ignore the two lowest-performing approaches, Dual-CTX and Propagate-Selector.

Overall, we find that *SEDAN* captures both closed-domain and open-domain contextual information can identify answer sentences effectively – and provides an effective self-learning based AS2 framework in diverse settings.

5.2 Performance on Downstream EQA Task

Although AS2 has been positioned to benefit downstream QA (by providing an enriched context for exact answer text extraction), its actual impact on EQA has not been studied in the literature. Here, we showcase how providing the top-k candidate answer sentences (from AS2) might impact the performance of EQA. We analyze the impact of AS2 on the performance of downstream EQA task as a use-case application, and hence compare the impact of different AS2 approaches (as opposed to other EQA specific trained models).

Table 4 reports the performance of the different models on the EQA task for BioASQ, TextbookQA and DROP datasets. It can be observed that *SEDAN* outperforms other existing baselines in both EM and F1 scores across the closed- and open-domain datasets. Interestingly, we observe that all the other baselines perform worse than the *Full Context* (for domain-based QA), where the entire original passage is provided to the EQA module without any AS2 processing. Thus, although existing AS2 models are accurate in retrieving potential answer sentences, they may not provide a good context for enhancing the EQA system. However, the context obtained from *SEDAN* provides better contextual and semantic relevant information for a question, thereby improving the performance of EQA – by learning-rich semantic information between questions and sentences. To the best of our knowledge, *SEDAN* is the first AS2 framework to not only depict accuracy improvements for answer sentence selection, but also positively impact downstream Extractive Question Answering (EQA) task.

5.3 Ablation Study of *SEDAN* Modules

One component in the *SEDAN* framework involves *two-stage fine-tuning* and *Siamese network* based representation learning. Here we perform a small-scale ablation study to understand the impact of the individual components on the BioASQ and DROP dataset.

We initially consider a generic Sentence-Transformer architecture with no domain fine-tuning and extract the sentences in the context that demonstrate the highest cosine similarity with the question. As per expectation, this variation does not perform well on

	Text
Q	What disease is characterised by an abnormal production of blood cells?
Pred.	It is characterised by an abnormal production of blood cells, usually white blood cells.
Gold	Leukemia is a cancer of the blood or bone marrow.
Q	Which domain of TIA-1 is necessary for stress granule assembly?
Pred.	The PRD of TIA-1 exhibits many characteristics of prions: concentration-dependent aggregation that is inhibited by the molecular chaperone heat shock protein (HSP)70; resistance to protease digestion; sequestration of HSP27, HSP40, and HSP70; and induction of HSP70, a feedback regulator of PRD disaggregation.
Gold	The RNA recognition motifs of TIA-1 are linked to a glutamine-rich prion-related domain (PRD).

Table 7: Sentence Retrieval Analysis of *SEDAN*

the domain-specific datasets, since it has no knowledge of semantic understanding within the domain. We next remove the first-stage fine-tuning in *SEDAN* (i.e., do not pre-align it for QA), and evaluate using both the Siamese network classification (as discussed earlier) and using the question-sentence cosine similarity score. We observe that simple single-stage domain adaptation (based on the synthetically generated QA pairs) improves AS2 performance on the closed-domain dataset. Finally, we perform the full two-step fine-tuning, but use the cosine similarity to identify candidate answer sentences (instead of the Siamese network classification). Similar to the TANDA approach [18], interestingly, the complete fine-tuning leads to a significant boost in performance over the other variations, and together with the Siamese network sentence extraction classifier (i.e., the proposed *SEDAN* framework) provides the best results across all the ablation study scenarios, as shown in Table 5.

5.4 Performance with Gold Training Data

For completeness, we also study the behaviour of the approaches in scenarios where annotated training data might be available. From Table 6, we observe that *SEDAN* outperforms the baselines on TextbookQA and DROP, while reporting the second-best (comparable) on BioASQ. In fact, we see a slight improvement in the performance of *SEDAN*, in the presence of annotated data. On the other hand, the remaining baselines showcase a marked performance enhancement (compared to the unsupervised setting). This depicts that existing methodologies are primarily geared toward expensive training procedures for acceptable performance. Thus, we see that *SEDAN* is geared towards domain-specific self-learning and is robust to diverse scenarios – effective for supervised and unsupervised AS2.

5.5 Error Analysis

This section highlights a couple of interesting scenarios underlying the working of *SEDAN*. Tab. 7 shows two examples where *SEDAN* predicts the correct answer sentence (i.e., containing the answer span), but is different from the gold answer sentence.

In the first example, we can observe that although different from the annotated example, the model selects the sentence that has a very high contextual overlap as compared to the annotated sample. Here, the pronoun "It" in the predicted answer sentence from *SEDAN*, refers to the correct answer (when read with the full document context). It is worth noticing that a mere co-reference resolution stage can address the issue to enable the downstream EQA task in choosing the correct answer span.

In the second example, we find that both the predicted sentence and the annotated ground-truth sentence, although different, contain the correct answer text – *PRD*. It can be observed that *SEDAN* relates the word *disaggregation* in the predicted sentence to the word *assembly* in the question text to come up with the answer sentence. The ground truth sentence contains the full form of the acronym *PRD* and hence might be a bit more helpful in answering the question, in terms of user experience.

5.6 Obstacles to Deployment

Model Size. The proposed *SEDAN* framework utilizes a pre-trained sentence-transformer model and two T5 models. Hence, the proper training and deployment of our system depend on the presence of sufficient computational resources to this end.

Domain Data Update. Although *SEDAN* is self-supervised, it needs to be trained individually for each application domain of usage. A single model spanning multiple domains might suffer from degraded performance due to "information collusion" across the domains. The presence of such multiple large models might impede certain deployment opportunities. Further, the models need to be re-trained with new emerging data within the domains.

Dependence of Synthetic Data Quality. The performance of *SEDAN* depends on self-supervised domain adaptation based on the synthetically generated training question-answer pairs. In the case of niche domains (e.g., molecular geometry, etc.), the use of such pre-trained language models (which are not pre-trained on such domain data) might not provide quality synthetically generated training data. This might lead to the sub-par performance of our proposed system and might hinder the adoption of such applications.

6 CONCLUSION

This paper proposes *SEDAN*, a framework for *self-learning* based *domain adaptation* for efficient Answer Sentence Selection (AS2). We showcase how a two-stage fine-tuning process using synthetically generated domain data, along with a Siamese network-based classification, enables a better understanding of the question text-answer sentence semantic relationships. We depict results on closed- and open-domain datasets and exhibit improved accuracy in answer sentence identification, as well as for downstream tasks such as EQA. Based on the evidential performance improvements, *SEDAN* is

currently under the development phase of deployment for an automated chatbot solution. In future work, we plan to explore multilinguality and inclusion of word-level as well as sentence-level information in *SEDAN* for the AS2 task.

7 ACKNOWLEDGEMENTS

This work is supported by a grant from The Government of Ireland Postgraduate Fellowship, Irish Research Council under project ID GOIPG/2019/3480. The work is also co-supported by Science Foundation Ireland under grant number SFI/12/RC/2289 2 (Insight).

REFERENCES

[1] Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA Corpora Generation with Roundtrip Consistency. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*. 6168–6173.

[2] Giambattista Amati. 2009. *BM25*. Springer US, Boston, MA, 257–260.

[3] Haytham Assem, Rajdeep Sarkar, and Sourav Dutta. 2021. Qasar: Self-Supervised Learning Framework for Extractive Question Answering. In *2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA*. IEEE, 1797–1808.

[4] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP*. The Association for Computational Linguistics, 632–642.

[5] Ying-Hong Chan and Yao-Chung Fan. 2019. A Recurrent BERT-based Model for Question Generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP*. Association for Computational Linguistics, 154–162.

[6] Yang Deng, Yuexiang Xie, Yaliang Li, Min Yang, Wai Lam, and Ying Shen. 2021. Contextualized Knowledge-Aware Attentive Neural Network: Enhancing Answer Selection with Knowledge. *ACM Trans. Inf. Syst.* 40, 1, Article 2 (Sept. 2021), 33 pages. https://@doi.org/10.1145/3457533

[7] Yang Deng, Wenxuan Zhang, and Wai Lam. 2021. Learning to Rank Question Answer Pairs with Bilateral Contrastive Data Augmentation. In *Proceedings of the Seventh Workshop on Noisy User-generated Text, W-NUT*. Association for Computational Linguistics, 175–181.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT*. Association for Computational Linguistics, 4171–4186.

[9] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. Association for Computational Linguistics, 2368–2378.

[10] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Rethink Training of BERT Rerankers in Multi-stage Retrieval Pipeline. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR (Lecture Notes in Computer Science, Vol. 12657)*. Springer, 280–286.

[11] Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI*. AAAI Press, 7780–7788.

[12] Rujun Han, Luca Soldaini, and Alessandro Moschitti. 2021. 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: Main Volume, EACL*. Association for Computational Linguistics, 3005–3010.

[13] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. https://@doi.org/10.5281/zenodo.1212303

[14] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Trans. Assoc. Comput. Linguistics* 8 (2020), 64–77.

[15] Aniruddha Kembhavi, Min Joon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. IEEE Computer Society, 5376–5384.

[16] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR*.

[17] Md. Tahmid Rahman Laskar, Jimmy Xiangji Huang, and Enamul Hoque. 2020. Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC*. European Language Resources Association, 5505–5514.

[18] Ivano Lauriola and Alessandro Moschitti. 2021. Answer Sentence Selection Using Local and Global Context in Transformer Models. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR (Lecture Notes in Computer Science, Vol. 12656)*. Springer, 298–312.

[19] Edward E Leamer. 1978. *Specification searches: Ad hoc inference with nonexperimental data*. Vol. 53. John Wiley & Sons Incorporated.

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[21] Jing Lu, Gustavo Hernández Ábrego, Ji Ma, Jianmo Ni, and Yinfei Yang. 2021. Multi-stage Training with Improved Negative Contrast for Neural Passage Retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*. Association for Computational Linguistics, 6091–6103.

[22] Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2019. Results of the Seventh Edition of the BioASQ Challenge. In *Machine Learning and Knowledge Discovery in Databases - International Workshops of ECML PKDD (Communications in Computer and Information Science, Vol. 1168)*, Peggy Cellier and Kurt Driessens (Eds.). Springer, 553–568.

[23] Rama Kumar Pasumarthi, Sebastian Bruch, Xuanhui Wang, Cheng Li, Michael Bendersky, Marc Najork, Jan Pfeifer, Nadav Golbandi, Rohan Anil, and Stephan Wolf. 2019. TF-Ranking: Scalable TensorFlow Library for Learning-to-Rank. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD*. ACM, 2970–2978.

[24] Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training Question Answering Models From Synthetic Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*. Association for Computational Linguistics, 5811–5826.

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[26] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP*. The Association for Computational Linguistics, 2383–2392.

[27] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*. Association for Computational Linguistics, 3980–3990.

[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108 (2019).

[29] Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 373–382.

[30] Luca Soldaini and Alessandro Moschitti. 2020. The Cascade Transformer: an Application for Efficient Answer Sentence Selection. In *ACL*. 5697–5708.

[31] Chuanqi Tan, Furu Wei, Qingyu Zhou, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. 2018. Context-Aware Answer Sentence Selection With Hierarchical Gated Recurrent Neural Networks. *IEEE ACM Trans. Audio Speech Lang. Process.* 26, 3 (2018), 540–549.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*. 5998–6008.

[33] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP*. Association for Computational Linguistics, 2369–2380.

[34] Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Propagate-Selector: Detecting Supporting Sentences for Question Answering via Graph Neural Networks. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC*. European Language Resources Association, 5400–5407.

[35] Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep Learning for Answer Sentence Selection. In *NIPS Deep Learning Workshop*.