Abstract

Answer sentence selection (AS2) modeling requires annotated data, i.e., hand-labeled question-answer pairs. We present a strategy to collect weakly supervised answers for a question based on its reference to improve AS2 modeling. Specifically, we introduce Reference-based Weak Supervision (RWS), a fully automatic large-scale data pipeline that harvests high-quality weakly-supervised answers from abundant Web data requiring only a question-reference pair as input. We study the efficacy and robustness of RWS in the setting of TANDA, a recent state-of-the-art fine-tuning approach specialized for AS2. Our experiments indicate that the produced data consistently bolsters TANDA. We achieve the state of the art in terms of P@1, 90.1%, and MAP, 92.9%, on WikiQA.

1 Introduction

NLP advances have been brought to customers worldwide in many online services (Hassan et al., 2018; Kim et al., 2020). Such results combine efforts in both (i) advancing the state of the art in modeling and (ii) collecting more and better data that can maximizes model potentials. The latter is the focus of this paper.

Creating datasets for QA requires expensive hand-labeling work. We explore the possibility to reduce this cost by automatically leveraging abundant text data from the Web to collect weakly-supervised data, i.e., question-answer pairs. Specifically, we propose the Reference-based Weak Supervision (RWS), a fully automatic data pipeline to harvest high-quality answers from the Web.

RWS operates in two stages: (i) collecting answer candidates from Web documents and (ii) labeling them, i.e., assigning the correct or incorrect labels. More specifically, we build a large index of more than 100MM Web documents from Common Crawl’s data. Given a question-reference pair, the question is used as query to retrieve a set of relevant documents from the index. Then, we extract sentences from those documents to build a large pool of answer candidates, which are finally scored by an automatic evaluator based on the provided reference. We use AVA for our purposes, a recently introduced automatic evaluator for AS2 (Vu and Moschitti, 2021).

The experimental results suggest that the weakly supervised data produced by RWS adds new supervision capacity to the original dataset, enabling trained models to advance the state of the art. Specifically, we first verify that models trained only on RWS can approach the performance of models trained with the original clean data, just dropping 1-4%. We show that RWS complements the original data by measuring its improvement of AS2 models on WikiQA and TREC-QA datasets.

In a nutshell, our contributions include: (i) a large-scale data pipeline that generates labeled question-answer pairs using publicly available Web data, i.e., Common Crawl; and (ii) a large automatically labelled dataset derived from the data and labels of ASNQ (Garg et al., 2020) with RWS.

2 Background

In this section we provide the background of our work. We first describe AS2 task formally, and then introduce TANDA, the current state-of-the-art model for AS2 (Garg et al., 2020). Finally, we present AVA employed in our pipeline.

2.1 Answer Sentence Selection (AS2)

The task of reranking answer candidates can be modeled with a classifier scoring the candidates. Let \( q \) be a question, \( T_q = \{t_1, \ldots, t_n\} \) be a set of answer candidates for \( q \), we define \( R \) a ranking function that orders the candidates in \( T_q \) according to a score, \( p(q, t_i) \), indicating the probability of \( t_i \)
Where is the world second largest aquarium?

Located in the Southeast Asian city-state of Singapore, Marine Life Park contains twelve million gallons of water, making it the second-largest aquarium in the world. The Marine Life Park, situated in southern Singapore, was the largest oceanarium in the world from 2012 to 2014, until it was surpassed by Chimelong Ocean Kingdom.

Table 1: A sample input for the automatic evaluator, which compares the semantic similarity between a reference $r$ and an answer candidate $t$, biased by $q$.

| $q$: Where is the world second largest aquarium? | $r$: Located in the Southeast Asian city-state of Singapore, Marine Life Park contains twelve million gallons of water, making it the second-largest aquarium in the world. | $t$: The Marine Life Park, situated in southern Singapore, was the largest oceanarium in the world from 2012 to 2014, until it was surpassed by Chimelong Ocean Kingdom. |
| --- | --- | --- |

Table 1 shows an example input for $A$, which measures the correctness of an answer $t$ with respect to a question $q$, using a reference answer $r$. Formally, it is modeled as a function: $\mathcal{A}(q, r, t_i) \rightarrow \{0, 1\}$, where the output is a binary correct/incorrect label.

AVA is designed to classify an answer as correct or incorrect like an AS2 model but exploits the semantic similarity between $t$ and $r$, biased by $q$. We studied multiple configurations to optimize AVA for our task of generating weakly supervised data. In our experiments, we use the best reported setting, which relies on a Transformer-based approach with Peer-Attention to model the interaction among $q$, $t$, and $r$ (see Vu and Moschitti, 2021) for a detailed technical description. We built AVA using a dataset of 245 questions, each having roughly 100 annotated answers. The number of correct and incorrect answers are 5.3K and 20.7K, respectively. This generates approximately 500K point-wise training examples for AVA using the method described in (Vu and Moschitti, 2021). We verified that our training set is disjoint with respect to all datasets studied in this paper to generate weakly supervised data.
4 Experiments

We study the efficacy of RWS by testing its impact on TANDA models for AS2. We first describe our experimental setup, datasets, and then apply RWS to AS2-NQ. We report the results of TANDA when RWS’s data is used during the transfer stage.

4.1 Setup

Large Web Index Having the ability to query from a large index of Web documents is required in our data pipeline. In particular, we need to retrieve a large number of documents, given a question, and we process hundreds of thousands of questions. As public search engines do not allow for such large-scale experimentation, we created our search engine constituted by a large index of more than 100MM English documents, collected from 19 Common Crawl’s crawls from 2013 to 2020.

Parameter Settings We employ two standard pre-trained models in our experiments: RoBERTa (Liu et al., 2019) and ELECTRA (Clark et al., 2020). We verify our findings on both Base and Large configurations. We use HuggingFace’s Transformer library (Wolf et al., 2019) and set the learning-rates to $1\times10^{-6}$ and $1\times10^{-5}$ for the transfer and adapt stages of TANDA, respectively, across all experiments. The other hyper-parameters are set to default values.

4.2 Datasets

We evaluated the impact of RWS on AS2 using the two most popular public datasets: WikiQA and TREC-QA. In addition, we also created AS2-NQ by extending ASNQ. The dataset has more than 84K questions, i.e., 27K more questions than ASNQ, each having typically one reference answer.

| Dataset   | #Q     | #A     | #A+ | #A− |
|-----------|--------|--------|-----|-----|
| ASNQ      | 57,242 | 20,745 | 60,285 | 0.914 | 0.922 |
| AS2-NQ    | 84,121 | 27,208 | 86,756 | 0.923 | 0.975 |
| RWS       | 84,089 | 2,103  | 69,945 | 0.923 | 0.975 |

Table 2: WikiQA dataset statistics: reporting the total number of questions (#Q) and answers (#A) for each split: Train, Dev, and Test.

We verified the quality of the new dataset by comparing TANDA models trained with ASNQ and AS2-NQ. In particular, Table 4 reports the results of the models when transferred on ASNQ or AS2-NQ, measured on WikiQA and TREC-QA. The results suggest that the end-to-end performance gain given by AS2-NQ is negligible, although 47% more data is added. This indicates that the curve amount of training data/accuracy has reached a plateau. However, in Section 4.3, we show that higher performance can still be achieved with our weakly supervised data from RWS.

RWS We apply RWS to AS2-NQ following these steps: First, we collect question-reference pairs

| TANDA   | Transfer on | WikiQA | TREC-QA |
|---------|-------------|--------|---------|
|         |             | MAP   | MRR    | MAP | MRR |
| RoBERTa-Base | ASNQ (2020) | 0.889 | 0.901 | 0.914 | 0.922 |
|          | AS2-NQ     | 0.898 | 0.910 | 0.923 | 0.928 |
| % diff   |             | +0.09 | +0.09 | +0.08 | +0.05 |
| RoBERTa-Large | ASNQ (2020) | 0.920 | 0.933 | 0.943 | 0.974 |
|          | AS2-NQ     | 0.923 | 0.935 | 0.936 | 0.975 |
| % diff   |             | +0.33 | +0.33 | -0.07 | +0.15 |

Table 3: Total number of questions (#Q), answers (#A), correct and incorrect (#A+ and #A−) of ASNQ, AS2-NQ, and the weakly-supervised dataset generated from AS2-NQ via our RWS pipeline.

Table 4: TANDA’s performance on two datasets ASNQ and AS2-NQ using RoBERTa Base and Large. % diff. reports the percentage differences.

We verified the quality of the new dataset by comparing TANDA models trained with ASNQ and AS2-NQ. In particular, Table 4 reports the results of the models when transferred on ASNQ or AS2-NQ, measured on WikiQA and TREC-QA. The results suggest that the end-to-end performance gain given by AS2-NQ is negligible, although 47% more data is added. This indicates that the curve amount of training data/accuracy has reached a plateau. However, in Section 4.3, we show that higher performance can still be achieved with our weakly supervised data from RWS.

RWS We apply RWS to AS2-NQ following these steps: First, we collect question-reference pairs
from AS2-NQ by using only pairs with correct answers. We set $K_1$ and $K_2$ at 1,000 and 25, i.e., for each question, we run a query and select 1,000 relevant documents from our Elasticsearch index. This typically generates a set of 10,000 candidates. Then, we select the 25 most probable candidates using an off-the-shelf AS2 reranker tuned on ASNQ by Garg et al. (2020). While a large number of questions are shared between ASNQ and AS2-NQ, the candidates from our index are disjoint. We apply AVA to label each triple, $(q, r, t_i)$, thus generating labelled pairs, $(q, t_i)$. A pair is labeled as correct if its A VA score, produced by $A(q, r, t_i)$, is at least 0.9, otherwise it is labeled as incorrect.

### 4.3 Integrating RWS into TANDA

We study the contribution of RWS in fine-tuning models for AS2. Specifically, we compare the following transfer configurations for TANDA. First, we report the baselines using (i) vanilla BERT Base and Large models without transferring data; and (ii) TANDA-RoBERTa transferred with ASNQ. We then replace ASNQ (iii) by AS2-NQ and (iv) by RWS at transfer stage, measuring the results of each transfer. Finally, we use both datasets, AS2-NQ and RWS, at transfer stage in the following orders: AS2-NQ→RWS and RWS→AS2-NQ. We use precision at 1 (P@1), mean average precision (MAP), and mean reciprocal rank (MRR) as evaluation metrics.

#### General results

Table 5 shows that RWS used alone does not improve the baselines trained on ASNQ or AS2-NQ. This is intuitive as the quality of weakly supervised data is supposed to be lower than supervised data. However, when RWS is used as the first level of fine-tuning (i.e., TANDA approach), for any dataset and any model (see model AS2-NQ→*), we observed a significant improvement. In particular, when RWS→AS2-NQ is used with RoBERTa-Large, the model establishes the new state of the art in AS2.

#### WikiQA

RWS achieves additional performance gains when combining it with AS2-NQ during the transfer steps. In particular, we note 1%–4% performance gains over the TANDA transferred on AS2-NQ. On WikiQA, it seems better using RWS before AS2-NQ, i.e., RWS→AS2-NQ.

#### TREC-QA

Using RWS during the transfer step improves the performance on TREC-QA. While the measures are better over the baselines, i.e., using ASNQ or AS2-NQ alone, we observe a different trend during the transfer stage. Specifically, it seems more beneficial to transfer RWS later, i.e., AS2-NQ→RWS. We conjecture that this is due to the differences between WikiQA and TREC-QA. That is, the former is very similar to AS2-NQ and AS2-NQ is used at transfer stage, measuring the results of each transfer. Finally, we use both datasets, AS2-NQ and RWS, at transfer stage in the following orders: AS2-NQ→RWS and RWS→AS2-NQ. We use precision at 1 (P@1), mean average precision (MAP), and mean reciprocal rank (MRR) as evaluation metrics.

#### General results

Table 5 shows that RWS used alone does not improve the baselines trained on ASNQ or AS2-NQ. This is intuitive as the quality of weakly supervised data is supposed to be lower than supervised data. However, when RWS is used as the first level of fine-tuning (i.e., TANDA approach), for any dataset and any model (see model AS2-NQ→*), we observed a significant improvement. In particular, when RWS→AS2-NQ is used with RoBERTa-Large, the model establishes the new state of the art in AS2.

#### WikiQA

RWS achieves additional performance gains when combining it with AS2-NQ during the transfer steps. In particular, we note 1%–4% performance gains over the TANDA transferred on AS2-NQ. On WikiQA, it seems better using RWS before AS2-NQ, i.e., RWS→AS2-NQ.

#### TREC-QA

Using RWS during the transfer step improves the performance on TREC-QA. While the measures are better over the baselines, i.e., using ASNQ or AS2-NQ alone, we observe a different trend during the transfer stage. Specifically, it seems more beneficial to transfer RWS later, i.e., AS2-NQ→RWS. We conjecture that this is due to the differences between WikiQA and TREC-QA. That is, the former is very similar to AS2-NQ and AS2-NQ is used at transfer stage, measuring the results of each transfer. Finally, we use both datasets, AS2-NQ and RWS, at transfer stage in the following orders: AS2-NQ→RWS and RWS→AS2-NQ. We use precision at 1 (P@1), mean average precision (MAP), and mean reciprocal rank (MRR) as evaluation metrics.

#### General results

Table 5 shows that RWS used alone does not improve the baselines trained on ASNQ or AS2-NQ. This is intuitive as the quality of weakly supervised data is supposed to be lower than supervised data. However, when RWS is used as the first level of fine-tuning (i.e., TANDA approach), for any dataset and any model (see model AS2-NQ→*), we observed a significant improvement. In particular, when RWS→AS2-NQ is used with RoBERTa-Large, the model establishes the new state of the art in AS2.

#### WikiQA

RWS achieves additional performance gains when combining it with AS2-NQ during the transfer steps. In particular, we note 1%–4% performance gains over the TANDA transferred on AS2-NQ. On WikiQA, it seems better using RWS before AS2-NQ, i.e., RWS→AS2-NQ.

#### TREC-QA

Using RWS during the transfer step improves the performance on TREC-QA. While the measures are better over the baselines, i.e., using ASNQ or AS2-NQ alone, we observe a different trend during the transfer stage. Specifically, it seems more beneficial to transfer RWS later, i.e., AS2-NQ→RWS. We conjecture that this is due to the differences between WikiQA and TREC-QA. That is, the former is very similar to AS2-NQ and AS2-NQ is used at transfer stage, measuring the results of each transfer. Finally, we use both datasets, AS2-NQ and RWS, at transfer stage in the following orders: AS2-NQ→RWS and RWS→AS2-NQ. We use precision at 1 (P@1), mean average precision (MAP), and mean reciprocal rank (MRR) as evaluation metrics.
References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In International Conference on Learning Representations.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. TANDA: Transfer and adapt pre-trained transformer models for answer sentence selection.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360. Online. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menenez, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Tingsong Jiang, Jing Liu, Chin-Yew Lin, and Zhifang Sui. 2018. Revisiting distant supervision for relation extraction. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611. Vancouver, Canada. Association for Computational Linguistics.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tür. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *TACL*.

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.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2137–2147, Melbourne, Australia. Association for Computational Linguistics.

Gehui Shen, Yunlun Yang, and Zhi-Hong Deng. 2017. Inter-weighted alignment network for sentence pair modeling. In *EMNLP’17*, pages 1179–1189, Copenhagen, Denmark.

Thuy Vu and Alessandro Moschitti. 2021. Ava: an automatic evaluation approach to question answering systems. In *NAACL*.

Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. 2007. What is the Jeopardy model? a quasi-synchronous grammar for QA. In *EMNLP-CoNLL’07*, pages 22–32, Prague, Czech Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In Proceedings of the 2015 conference on empirical methods in natural language processing, pages 2013–2018.

Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2019. A compare-aggregate model with latent clustering for answer selection. *CoRR*, abs/1905.12897.