Rethinking the objectives of extractive question answering

Martin Fajcik, Josef Jon, Santosh Kesiraju, Pavel Smrz
Brno University of Technology
{ifajcik,ijon,kesiraju,smrz}@fit.vutbr.cz

Abstract

This paper describes two generally applicable approaches towards the significant improvement of the performance of state-of-the-art extractive question answering (EQA) systems.
Firstly, contrary to a common belief, it demonstrates that using the objective with independence assumption for span probability $P(a_s, a_e) = P(a_s)P(a_e)$ of span starting at position a_s and ending at position a_e may have adverse effects. Therefore we propose a new compound objective that models joint probability $P(a_s, a_e)$ directly, while still keeping the objective with independence assumption as an auxiliary objective.
Our second approach shows the beneficial effect of distantly semi-supervised shared-normalization objective known from (Clark and Gardner, 2017). We show that normalizing over a set of documents similar to the golden passage, and marginalizing over all ground-truth answer string positions leads to improvement of results from smaller statistical models.
Our findings are supported via experiments with three QA models (BidAF, BERT, ALBERT) over six datasets. The proposed approaches do not use any additional data.
Our code, analysis, pretrained models and individual results will be available online.

1 Introduction

The common goal of extractive question answering (EQA) is to find the span boundaries – the start and the end of the span from text evidence, which answers a given question. Therefore, a natural choice of the objective to this problem is to model the probabilities of the span boundaries. In the last years, there was a lot of effort put into building better neural models underlying the desired probability distributions. However, there has been a little progress seen towards the change of the objective itself. For instance, the "default" choice of objective for modelling the probability over spans in SQuAD (Rajpurkar et al., 2016) – maximization of independent span boundary probabilities $P(a_s)P(a_e)$ for answer at position $\langle a_s, a_e \rangle$ – has stayed the same over the course of years in many influential works (Xiong et al., 2017; Seo et al., 2016; Chen et al., 2017; Yu et al., 2018; Devlin et al., 2019; Cheng et al., 2020) since the earliest work on this dataset – the submission of Wang and Jiang (2016). Based on the myths of worse performance of different objectives, these works adopt the deeply rooted assumption of independence. This assumption may lead to obviously wrong predictions, as shown in Figure 1.

Question: What was the name of atom bomb dropped by USA on Hiroshima?
Passage: ...The Allies issued orders for atomic bombs to be used on four Japanese cities were issued on July 25. on August 6, one of its b - 29s dropped a little boy uranium gun-type bomb on Hiroshima. three days later, on August 9, a fat man plutonium implosion-type bomb was dropped by another b - 29 on Nagasaki...
Ground truth: little boy

P	Predictions from BERT-base
33.3	little boy uranium gun-type bomb on Hiroshima. three days later, on August 9, a fat man
32.15	little boy
23.51	fat man
3.60	a fat man
2.08	a little boy uranium gun - type bomb on hiroshima. three days later, on august 9, a fat man
1.03	a little boy

Figure 1: An example of an error which comes with an independence assumption. The model assigns high probability mass to boundaries around "little boy", and "fat man" answers. However, during decoding, the start of one and the end of another answer is picked up.

In addition, this assumption leads to degenerate distribution $P(a_s, a_e)$, as high probability mass is assigned to many trivially wrong[1] answers.

[1]We define 'trivially wrong' as not resembling any string

Only recently, large language representation models (LRMs) like XLNet (Yang et al., 2019), ALBERT (Lan et al., 2019) or ELECTRA (Clark et al., 2020) started modelling the span probability via conditional probability factorization $P(a_e|a_s)P(a_s)$. However, is it unknown whether this objective improves any performance at all, as none of the works reported results on its effect, not even described its existence (except ELECTRA paper). Additionally, this objective requires beam search which slows down inference in test time. In this work, we try to break the myths about the objectives that have been widely used previously. We also introduce a new compound objective, that deals with modelling joint probability $P(a_s, a_e)$ directly while keeping the traditional independent objective as an auxiliary objective. Conducted experiments demonstrate that our objective is superior to previously used objectives across the various choices of models or datasets.

Secondly, orthogonal to the first approach, this work explores the possibility of distant semi-supervised training of EQA systems. Our approach is based on the multi-passage shared-normalization training objective (SNO) of Clark and Gardner (2017). The authors demonstrate that training EQA model with the SNO damages its performance in a closed-domain setting[2]. Their results support the belief that closed-domain EQA system cannot benefit from distant semi-supervision with SNO. However, we disprove this belief, by showing that fine-tuning EQA models by including topically similar passages from the closed-domain dataset and using distantly supervised ground truth annotation can be beneficial towards improving EQA performance in closed-domain setting by improving its performance and also its robustness for smaller models.

In summary, our work contributions are:

- introduction of the novel *compound* objective and its comparison with the traditional objectives based on assumption of independence or conditional factorization,
- we show that multi-passage shared normalization objective improves the closed-domain EQA robustness and performance for smaller models,
- a thorough evaluation on the wide spectrum of models and datasets comparing different objectives supported by statistical tests.

form human would answer, e.g., the first or the second last answer of Figure 1.

[2]See Table 4 in Clark and Gardner (2017).

2 Compound Objective

This section describes the common approach to the EQA, with its independent modelling of the answer span start and end positions. We propose an extension towards relieving this assumption by defining a compound objective.

The EQA can be defined as follows: Given a question q and a passage or a set of passages D, find an answer string a from D such that a answers the question q. This can be expressed by modelling a categorical probability mass function (PMF) that has its maximum in the answer endpoint indices $a = \langle a_s, a_e \rangle$ from the passage D as $P(a_s, a_e|q, D)$ for each question-passage-answer triplet (q, D, a) from dataset $\mathcal{D}$. During inference, the most probable answer span $\langle a_s, a_e \rangle$ is predicted. Although there are works that were able to explicitly model the joint probability (Lee et al., 2016), modelling it directly results in a number of categories which is quadratic to the passage's length. Optimizing such models with the amount of data available today may lead toward poor parameter estimates. Therefore, state-of-the-art approaches modelling the answer span PMF introduce the independence assumption $P(a_s, a_e|q, D) = P(a_s|q, D)P(a_e|q, D)$. The factorized PMFs are usually computed by the shared model, as introduced in Wang and Jiang (2016).

Our work proposes an extension to the described approach, replacing the standard *independent* objective, that maximizes the probability $P(a_s|q, D)P(a_e|q, D)$, with the multi-task *compound* objective (1) computed via a shared model[3].

$$\sum_{(q,D,a)\in\mathcal{D}} \log P_\theta(a_s, a_e)P_\theta(a_s)P_\theta(a_e) \qquad (1)$$

The auxiliary *independent* objective $P(a_s)P(a_e)$ can be seen as a form of regularization for the *joint* objective $P_\theta(a_s, a_e)$, which is used in test time.

For most of the systems modelling the *independent objective* with neural networks, the final endpoint probabilities are derived from start/end position passage representations $\boldsymbol{H}_s$,$\boldsymbol{H}_e \in \mathbb{R}^{d\times L}$ as shown for $b \in \{s, e\}$.

$$P_\theta(a_b) = \mathrm{softmax}(\boldsymbol{w}_b^\top \boldsymbol{H}_b + \boldsymbol{b}_b) \qquad (2)$$

[3]For brevity, q, D dependencies were omitted.

The passage representations $\boldsymbol{H}_s$,$\boldsymbol{H}_e$ are often pre-softmax layer representations from neural network with passage and question at the input. Symbols d and L denote the model-specific dimension and the passage length, respectively.

For the *joint* objective $P_\theta(a_s, a_e)$, an arbitrary vector-to-vector similarity function f_{sim} can be used for obtaining each span score (e. g., the dot product $\boldsymbol{H}_s^\top \boldsymbol{H}_e$)[4].

$$P_\theta(a_s, a_e) = \text{softmax}(\text{vec}(f_{sim}(\boldsymbol{H}_s, \boldsymbol{H}_e))) \tag{3}$$

3 Training with Distant Semi-Supervision

We follow the shared normalization objective (SNO) introduced in Clark and Gardner (2017) for multi-paragraph answer selection. Given a set of topically close passages D, with each passage containing its set of position indices p, ground truth span annotating function[5] $GT(p) \subseteq p$ that returns a set of answer indices for span boundary type $b \in \{start, end, joint\}$ and scores for each span boundary s, the shared normalization objective can be generally defined as:

$$P(a_b) = \frac{\sum_{p \in D} \sum_{a_b \in GT(p)} e^{s_{a_b}}}{\sum_{p \in D} \sum_{a_b \in p} e^{s_{a_b}}} \tag{4}$$

The complete objective is therefore obtained as $\mathcal{L} = -\sum \log P(a_{start})P(a_{end})P(a_{joint})$. The passage set D of context size $|D|$ for question q in this experiment is obtained from a top-K most probable passages scored via passage retrieval model $P(p|q)$ over all passages in the dataset.

4 Experimental Setup

We use Transformers (Wolf et al., 2019) for LRM implementation. Our experiments were done on a machine with four 16GB GPUs using PyTorch (Paszke et al., 2019). We used Adam optimizer with a decoupled weight decay (Loshchilov and Hutter, 2017). For experiments with LRMs, the SQuAD1.1 default hyperparameters and scheduler settings were used, as proposed by specific LRM authors.

Dataset	Train	Test
SQuAD1.1	87,599	10,570
SQuAD2.0	130,319	11,873
Adversarial SQuAD	-	3,560
Natural Questions	104,071	12,836
NewsQA	74,160	4,212
TriviaQA	61,688	7,785

Table 1: Number of examples per each dataset used in this paper.

4.1 Datasets

We evaluate our approaches on a wide spectrum of datasets. The statistics to all datasets are shown in Table 1. To focus only on the extractive part of QA and to keep the format the same, we use curated versions of the last 3 datasets as released in MrQA shared task (Fisch et al., 2019).

SQuAD v1.1 (Rajpurkar et al., 2016), is a popular dataset composed from question, paragraphs and answer span annotation collected from the subset of Wikipedia passages.

SQuAD v2.0 (Rajpurkar et al., 2018) is an extension of SQuAD v1.1 with additional 50k questions and passages, which are topically similar to the question, but do not contain an answer.

Adversarial SQuAD (Jia and Liang, 2017) tests, whether the system can answer questions about paragraphs that contain adversarially inserted sentences, which are automatically generated to distract computer systems without changing the correct answer or misleading humans. In particular, our system is evaluated in ADDSENT adversary setting, which runs the model as a black box for each question on several paragraphs containing different adversarial sentences and picks the worst answer.

Natural Questions (Kwiatkowski et al., 2019) dataset consists of real users queries obtained from Google search engine. Each example is accompanied by a relevant Wikipedia article found by the search engine, and human annotation for long/short answer. The long answer is typically the most relevant paragraph from the article, while short answer consists of one or multiple entities or short text spans. We only consider short answers in this work.

NewsQA (Trischler et al., 2017) is a crowd-

[4] Here, we slightly abuse the notation for the sake of generality. See Subsection 4.2 for specific applications.

[5] The function returns all positions of an answer string in the passage. This is the distant supervision.

sourced dataset based on CNN news articles. Answers are short text spans and the questions are designed such that they require reasoning and inference besides simple text matching.

TriviaQA (Joshi et al., 2017) consists of question-answer pairs from 14 different trivia quiz websites and independent evidence passages collected using Bing search from various sources such as news, encyclopedias, blog posts and others. Additional evidence is obtained from Wikipedia through entity linker.

4.2 Applied models

Our experiments are based on three EQA models: **BERT** (Devlin et al., 2019) and **ALBERT** (Lan et al., 2019) are LRMs based on the self-supervised pretraining objective. During fine-tuning, each model is applied on the concatenation of question and passage representations. Outputs $\boldsymbol{H} \in \mathbb{R}^{d \times L}$ corresponding to the passage inputs of length L are then reduced to boundary probabilities by two vectors $\boldsymbol{w_s}, \boldsymbol{w_e}$ as $P(a_b) = \text{softmax}(\boldsymbol{w}_b^\top \boldsymbol{H} + \boldsymbol{b}_b)$. To compute joint probability $P(a_s, a_e)$, start representations are computed using $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b} \in \mathbb{R}^d$ (broadcasted) as $\boldsymbol{H}_s = \boldsymbol{W}\boldsymbol{H} + \boldsymbol{b}$ and end representations as $\boldsymbol{H}_e = \boldsymbol{H}$. A dot product is used as the similarity measure.

$$P(a_s, a_e) = \text{softmax}(\text{vec}(\boldsymbol{H}_s^\top \boldsymbol{H}_e)) \quad (5)$$

See Appendix B for experiments with different similarity measures. For modelling conditional probability factorization objective, we closely follow the implementation from (Lan et al., 2019), and provide exact details in the Appendix C.

BiDAF (Seo et al., 2016) dominated the state-of-the-art systems in 2016 and motivated a lot of following research work (Clark and Gardner, 2017; Yu et al., 2018). Question and passage inputs are represented via the fusion of word-level embeddings from GloVe (Pennington et al., 2014) and character-level word embeddings obtained via a convolutional neural network. Independently represented questions and passages are then combined into a common representation via two directions of attention over their similarity matrix $\boldsymbol{S}$. The similarity matrix is computed via multiplicative-additive interaction (6) between each pair of question vector $\boldsymbol{q_i}$ and passage vector $\boldsymbol{p_j}$, where ; denotes concatenation and $\circ$ stands for the Hadamard product.

$$\boldsymbol{S_{ij}} = f_{sim_{ma}}(\boldsymbol{q_i}, \boldsymbol{p_j}) = \boldsymbol{w}^\top[\boldsymbol{q_i}; \boldsymbol{p_j}; \boldsymbol{q_i} \circ \boldsymbol{p_j}] \quad (6)$$

Common representations are then concatenated together with document representations and passed towards two more recurrent layers – first to obtain answer-start representations $\boldsymbol{H}_s = [\boldsymbol{G}; \boldsymbol{M}]$ and second to obtain answer-end representations[6] $\boldsymbol{H}_e = [\boldsymbol{G}; \boldsymbol{M}^2]$. The joint probability $P(a_s, a_e)$ is then computed over scores from vectorized similarity matrix of $\boldsymbol{H}_s$ and $\boldsymbol{H}_e$ using the same similarity function $f_{sim_{ma}}$ between each two column vectors.

4.3 Passage retrieval

The passage set D of context size $|D|$ for question q in an experiment with distant semi-supervision is obtained from a top-K most probable passages scored via passage retrieval model $P(p|q)$ over all passages in the dataset. For passage retrieval $P(p|q)$, we separately train passage encoder and question encoder, producing passage vector $\boldsymbol{p}$ and question vector $\boldsymbol{q}$ representations.

To represent passages, we have used Bayesian subspace multinomial model (BSMM) (Kesiraju et al., 2019). BSMM is a generative log-linear model that learns to represent passages in the form of Gaussian distributions and achieves state-of-the-art results in topic identification. In our experiments, we use 100-dimensional means of the Gaussian distributions as the document embeddings.[7]. The model is trained on dataset's training passages and used to extract representations for all passages in the training and validation data, therefore creating a matrix $\boldsymbol{D} \in \mathbb{R}^{\mathcal{L} \times d}$ of $\mathcal{L}$ passages each of dimension d.

Our question encoder is based on RoBERTa (Liu et al., 2019) LRM. Given a question $y = (y_1, y_2, ..., y_n)$, vector $\boldsymbol{q} = \boldsymbol{W}(LRM(y)[CLS])$ is extracted as a linear projection $\boldsymbol{W}$ of `CLS`-level output of RoBERTa. The encoder is trained via log-bilinear objective (7), defined on the set of dataset's questions Q, using default RoBERTa parameters and scheduler. We use dropout δ set to 0.35.

$$\sum_{q \in Q} P(p = i|q) = \sum_{q \in Q} \text{softmax}_i(\boldsymbol{D}\delta(\boldsymbol{q})) \quad (7)$$

In test time, the nearest passages are scored via cosine similarity of $\boldsymbol{p}$ and $\boldsymbol{q}$. To include passages with distant supervision more often in the context, we discard 50% of passages which do not contain

[6] For details, see formulae 2 to 4 in Seo et al. (2016).

[7] To profit from co-variance information, we have also experimented with KL-divergence based retrieval, gaining only negligible improvement.

Model	Obj	SQ1	SQ2	AdvSQ	TriviaQA	NQ	NewsQA
BERT	I	81.31/88.65	**73.89**/76.74	47.04/52.62	62.88/69.85	65.66/78.20	52.39/67.17
	J	81.33/88.13	72.66/75.04	48.10/53.54	**63.93**/69.90	**67.82**/78.72	52.71/66.43
	JC	80.96/87.86	-	45.33/50.59	62.68/69.81	65.66/78.33	52.39/67.05
	I+J	***81.83***/88.52	73.53/76.14	***48.32***/53.47	*63.68*/69.71	*67.73*/78.78	***52.90***/66.79
	I+J(DSS)	<u>82.15</u> /88.90	<u>74.77</u>/77.29	<u>49.91</u>/55.32	-	67.85/78.94	52.50/66.29
ALBERT	I	88.55/94.62	87.07/90.02	68.12/73.54	74.72/80.34	70.79/83.39	59.91/74.97
	J	88.84/94.64	86.87/89.71	68.90/74.17	75.57/80.86	73.32/83.95	60.15/74.30
	JC	88.50/94.51	-	66.75/72.03	-	72.34/83.37	58.44/72.69
	I+J	***89.02***/94.77	**87.13**/89.98	***69.57***/74.76	***75.36***/80.48	***73.36***/84.12	***60.45***/74.48
	I+J(DSS)	89.04/94.82	87.16/89.95	68.20/73.60	-	-	-

Table 2: Results of different objectives through the spectrum of datasets. Bold results mark best EM across the objectives. Cursive on I+J row marks results, which are improved significantly over the independent objective. Underscore on I+J(DSS) row marks results which improve significantly over compound results (I+J row).

exact answer overlap from the total ranking before taking top-K passages for distant semi-supervision.

4.4 Prediction filtering

In recent literature, there are two common ways of filtering the prediction probabilities:

Length filtering. Probabilities $P(a_s = i, a_e = j)$ are set to 0 iff $j - i > \zeta$, where ζ is a length threshold. Following (Devlin et al., 2019), we set $\zeta = 30$.

Top-K surface form filtering. Following Das et al. (2019), the top-K probabilities are aggregated for the top-K span predictions for all their surface forms (same strings) in the paragraph. This is done by summing all probabilities bound to same surface form into the most probable position of that surface form and setting rest to 0. We use $K = 100$.

4.5 Statistical testing

To improve the soundness of the presented results, we use several statistical tests. An exact match (EM) metric can be viewed as an average of samples from Bernoulli distribution. As stated via central limit theorem, a good assumption might be the EM comes from the normal distribution. We train 10 models for each presented LRM's result, obtaining 10 EMs for each sample. *Anderson-Darling* normality test (Stephens, 1974) is used to check this assumption – whether the sample truly comes from the normal distribution. Then we use the *one-tailed paired t-test* to check whether the case of improvement for compound objective w.r.t. independent objective is significant and analogically whether the improvement of compound objective with DSS w.r.t. compound objective is significant. The improvement is significant iff *p-value*<0.05. We use the reference implementation from (Dror et al., 2018).

5 Results and Discussion

We now show the effectiveness of proposed approaches. Our main results – the performance of *independent* (I), *joint* (J), *joint-conditional* (JC) and *compound* (I+J) objectives – are shown in Table 2. We also train models with *compound* objective inside DSS framework (see equation 4); denoted as I+J(DSS). Each of the presented results is averaged from 10 training runs[8].

Firstly, we note that the largest improvements can be seen for an exact match (EM) performance metric. In fact, in some cases *joint* or *compound* objective lead to degradation of F1, while improving EM (e.g., on SQuAD and NewsQA datasets for BERT). Upon manual analysis based on 100 differences between *independent* and *compound* models on SQuADv1.1, we found that in 14% of cases the *independent* model chooses larger span encompassing multiple potential answers, thus obtaining non-zero F1 score. However, this happens only in 2% of cases for *compound* models. Instead, in these borderline cases, we have found the *compound* and *joint* models to pick just one of these potential answers, obtaining either full match or no F1 score at all.

Next, we remark that *compound* objective outperformed others in most of our experiments. In BERT case, the *compound* objective performed

[8] We will release each result measurement along with result statistics.

Model		I	J	JC	I+J	I+J (DSS)
BiDAF	-	65.22/75.01	61.77/71.24	-	66.02/75.54	68.02/76.97
	LF	65.66/75.55	61.77/71.24	-	66.03/75.53	68.04/76.99
	SF	**65.73/75.58**	61.78/71.23	-	**66.05/75.55**	**68.06/77.00**
BERT	-	80.98/88.40	81.30/88.11	80.87/87.81	81.80/88.50	82.13/88.89
	LF	81.31/88.65	**81.33/88.13**	80.96/87.86	**81.83/88.52**	**82.15/88.90**
	SF	**81.38/88.68**	81.23/87.97	-	81.65/88.36	82.04/88.80
ALBERT	-	88.39/94.51	88.82/**94.64**	88.43/94.47	89.01/**94.77**	89.02/**94.82**
	LF	**88.55/94.63**	**88.84/94.64**	88.50/94.51	**89.02/94.77**	**89.04/94.82**
	SF	88.53/94.00	88.28/94.10	-	88.68/94.49	88.69/94.50

Table 3: Achieved results with different filtering approaches.

significantly better than *independent* objective on 5 out of 6 datasets. In ALBERT case, the *compound* objective performed significantly better than *independent* objective 5 from 6 times and it was on par in the last case. Comparing *compound* to *joint* objective in BERT case, the two behave almost equally, with compound objective significantly outperforming joint objective on the two SQuAD datasets and no significant differences for the other 4 datasets. However, we found the convergence of joint objective to be very sensitive to hyperparameter choice (BERT experiments of our preliminary results failed to converge), while we have never observed this behaviour with the compound objective. However, in ALBERT case, the *compound* objective significantly improves results over *joint* objective in all but one case and is on par in this last case.

Considering the results of DSS objective framework, we found it to improve the results of smaller models in all cases for BERT and BidAF (see Table 3). Interestingly, we found it not beneficial at all in ALBERT's case, indicating that the ALBERT's pre-training captures all relevant knowledge to be obtained from distant semi-supervision.

Upon closer inspection of results, we found possible reasons for result degradation of the *compound* model on SQuAD2, and also its large improvements gained on NQ dataset. For SQuAD2, the accuracies of no-answer detection for *indepedent*/*joint*/*compound* objectives in case of BERT models are 79.89/78.12/79.32. We found the same trend for ALBERT. We hypothesize, that this inferior performance of *joint* and *compound* models may be caused by the model having to use more parameters while learning a more complex problem of K^2 classes of all possible spans. To confirm that *compound* model is better at answer extraction step, we run models trained on SQuAD2 data with an answer, while masking model's no-answer option. The results shown in Table 4 support this hypothesis. On the other side, we found the large improvements over NQ might be exaggerated by the evaluation approach of MRQA, wherein the case of multi-span answers, choosing one of the spans from multi-span answer counts as correct. Upon closer result inspection, we found that the independent model here was prone to select the start of one start of the span from multi-span answer and end from another answer from multi-span.

Besides we found that models trained via *compound* objective does not benefit from previously used heuristics such as length filtering (LF) or surface form filtering (SF) significantly, as shown in Table 3. Therefore, we find it unnecessary to use these heuristics anymore. Note that results from SF already include LF. In this experiment, we also include our preliminary results with BiDAF, which show significant improvement of its performance on SQuADv1.1 dataset from both presented approaches.

	Objective	EM	F1
BERT	I	80.70	88.71
	J	81.38	81.51
	I+J	81.51	88.69
	DSS	81.85	88.99
ALBERT	I	87.40	94.10
	J	87.74	94.31
	I+J	87.90	94.38
	DSS	87.97	94.31

Table 4: Performance of SQuADv2.0 models on answerable examples of SQuADv2.0.

6 Analysis

Apart from example in Figure 1, we provide more examples of different predictions between models trained with *independent* and *compound objective* in Table 5. In general, by doing manual analysis of errors, we noticed three types of errors being fixed by the *compound* objective model in BERT:

1. The model assigns high probability to answer surrounded by the paired tags (e. g. quotes). It chooses the answer without respecting the symmetry between paired tags (third row of Table 5).

2. Uncertainty of the model causes it to assign high probabilities to two spans containing the same answer string. During decoding the start/end boundaries of two same answers are picked up (first row of Table 5). This can be alleviated with the surface form filtering.

3. Uncertainty of the model causes it to assign high probabilities of two different answer boundaries. During decoding the start/end boundaries of two different answers are picked up (fourth row in Table 5).

Interestingly, we found it hard to directly identify any of the mentioned error types in ALBERT's predictions[9].

To show, that spans retrieved with *compound* objective approach differ from the ones retrieved when using *independent* objective, we took the top 20 most probable spans from each model and averaged their length. This was done for each example in the development data. The histogram of these averages is shown in Figure 2. For a fair comparison, these predictions were filtered via length filtering.

7 Related work

Learning objective. One of the earliest works in EQA from Wang and Jiang (2016) experimented with generative models based on index sequence generation via pointer networks (Vinyals et al., 2015) and now traditional boundary models that focus on the prediction of start/end of an answer span. Their work shown substantial improvement of boundary models over the index sequence generative models.

[9]The full difference of BERT's and ALBERT's predictions can be found at https://tinyurl.com/y5f5uevc.

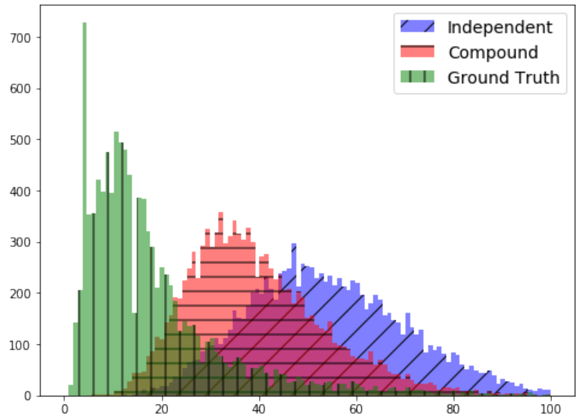

Figure 2: Histograms of average character length of top-20 predicted answers from BERT trained with different objectives compared with character length of ground-truth answers.

Authors of recent LRMs like XLNet (Yang et al., 2019), ALBERT (Lan et al., 2019) or ELECTRA (Clark et al., 2020) started modelling conditional probability factorization $P(a_e|a_s)P(a_s)$ for answer extraction in some cases[10]. Although the objective is not described in mentioned papers (except for ELECTRA), we follow the recipe for modelling the conditional probability from their implementation in this work. We believe this is the first official comparison of this objective w.r.t. others.

The most similar to our work is RASOR system (Lee et al., 2016). In their work, authors compared various objectives – binary answer classification of every input token, BIO sequence classification with CRF layer on top of their model, and most importantly joint objective, which turns out to work the best. However, in our experiments, training with the joint objective alone does not perform that well, while also being less stable to the choice of hyperparameters. Also, we use dot product as a similarity function while (Lee et al., 2016) used a feed-forward neural network.

Semi-supervision and distant supervision. Dhingra et al. (2018) exploit the structure of dataset passages to create cloze-style questions. They pretrain a powerful neural network on the cloze-style pseudo-questions and fine-tune the model on the labelled examples.

Chen et al. (2017) shows an improvement in open-domain QA when using distantly supervised ground truths with the same surface form as the

[10]For instance, ALBERT uses conditional objective for SQuADv2.0, but not for SQuADv1.1.

Question	Passage	Independent	Compound	Ground Truth
What company won a free advertisement due to the QuickBooks contest?	QuickBooks sponsored a "Small Business Big Game" contest, in which Death Wish Coffee had a 30-second commercial aired free of charge courtesy of QuickBooks. Death Wish Coffee beat out nine other contenders from across the United States for the free advertisement.	Death Wish Coffee had a 30-second commercial aired free of charge courtesy of QuickBooks. Death Wish Coffee	Death Wish Coffee	Death Wish Coffee
In what city's Marriott did the Panthers stay?	The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.	San Jose State practice facility and stayed at the San Jose	San Jose	San Jose
What was the first point of the Reformation?	Luther's rediscovery of "Christ and His salvation" was the first of two points that became the foundation for the Reformation. His railing against the sale of indulgences was based on it.	Christ and His salvation"	Christ and His salvation	Christ and His salvation
How many species of bird and mammals are there in the Amazon region?	The region is home to about 2.5 million insect species, tens of thousands of plants, and some 2,000 birds and mammals. To date, at least 40,000 plant species, 2,200 fishes, 1,294 birds, 427 mammals, 428 amphibians, and 378 reptiles have been scientifically classified in the region. One in five of all the bird species in the-world live in the rainforests of the Amazon, and one in five of the fishspecies live in Amazonian rivers and streams. Scientists have describedbetween 96,660 and 128,843 invertebrate species in Brazil alone.	2,000 birds and mammals. To date, at least 40,000 plant species, 2,200 fishes, 1,294 birds, 427	427	2,000
What was found to be at fault for the fire in the cabin on Apollo 1 regarding the CM design?	NASA immediately convened an accident review board, overseen by both houses of Congress. While the determination of responsibility for the accident was complex, the review board concluded that "deficiencies existed in Command Module design, workmanship and quality control." At the insistence of NASA Administrator Webb, North American removed Harrison Storms as Command Module program manager. Webb also reassigned Apollo Spacecraft Program Office (ASPO) Manager Joseph Francis Shea, replacing him with George Low.	deficiencies existed in Command Module design, workmanship and quality control."	Harrison Storms	deficiencies

Table 5: Examples of predictions from SQuADv1.1 using BERT trained with independent and compound objective.

original ground truth.

Clark and Gardner (2017) present a shared normalization objective (SNO) to alleviate the problems with multi-paragraph answer selection. Using a model trained via SNO, and TF-IDF based retrieval system retrieving passages from whole Wikipedia, they show a significant improvement in open-domain QA. However, when evaluating the model trained with their objective in the closed-domain setting over paragraphs from the same article of Wikipedia as the golden paragraph, they found a slight decrease in performance. Contrarily, we found that their objective can significantly improve closed-domain performance with different retrieval approach.

Min et al. (2019) show, that when using SNO, in most cases it is beneficial to switch objective to HardEM – maximizing only the probability of the most probable ground truth answer span, instead of summing over all answer occurrences – after some period of training with standard SNO.

Concurrently with our work, Cheng et al. (2020) explored different objective assumptions for BERT on multi-paragraph datasets TriviaQA and NarrativeQA. Experimenting with different DSS assumptions, they found our variant of DSS, denoted as position-based `H3-D` in their work, works as the best on TriviaQA. Note that in *compound* case, we simply add sum over joint positions into this assumption. While the objective does not perform as good on NarrativeQA, the cause of this is according to their hypothesis that NarrativeQA usually contains the answer for every paragraph and therefore TriviaQA results are closer to our application. Note that we retrieve our own set via retrieval model that considers BSMM paragraph representations over golden paragraphs, not via TF-IDF over document's texts split into paragraphs. Finally, our TriviaQA is slightly different as we use its MRQA version.

8 Conclusion

The paper looks back at the last years of progress in QA and finds two understudied phenomena which relate to common beliefs that do not necessarily hold true. Two new approaches in the paper show the new directions that can be applied in order to improve statistical EQA systems, without using any additional data. Firstly, we experimentally demonstrated that introducing the *compound* objective as a new objective for EQA improves the results of state-of-the-art systems. Secondly, we shown that the distant semi-supervised objective, known from open-domain EQA, is beneficial towards the performance of closed-domain EQA systems too, without the need for additional data. Using the proposed approaches, we were able to reach significant improvements through the wide spectrum of datasets, including +2.87 EM on Adversarial SQuAD and +2.07 EM on NaturalQuestions. In addition, we also identified the reason for performance decrease with *compound* objective on SQuADv2.0 – no-answer classifier trained within the same model performs worse – and we leave the solution for this deficiency for future work.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Hao Cheng, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2020. Probabilistic assumptions matter: Improved models for distantly-supervised document-level question answering. *arXiv preprint arXiv:2005.01898*.

Christopher Clark and Matt Gardner. 2017. Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. *arXiv preprint arXiv:1905.05733*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Bhuwan Dhingra, Danish Pruthi, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. *arXiv preprint arXiv:1804.00720*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. *arXiv preprint arXiv:1910.09753*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Mandar Joshi, Eunsol Choi, Daniel S 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.

Santosh Kesiraju, Oldřich Plchot, Lukaš Burget, and Suryakanth V Gangashetty. 2019. Learning document embeddings along with their uncertainties. *arXiv preprint arXiv:1908.07599*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Kenton Lee, Shimi Salant, Tom Kwiatkowski, Ankur Parikh, Dipanjan Das, and Jonathan Berant. 2016. Learning recurrent span representations for extractive question answering. *arXiv preprint arXiv:1611.01436*.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. A discrete hard em approach for weakly supervised question answering. 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)*, pages 2844–2857.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

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*, pages 2383–2392.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Michael A Stephens. 1974. Edf statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347):730–737.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

A Examples of answer span distribution

This section provides a deeper insight towards most probable elements of answer span PMF.

Question: What was the first point of the Reformation?
Passage: Luther's rediscovery of "Christ and His salvation" was the first of two points that became the foundation for the Reformation. His railing against the sale of indulgences was based on it.
Ground Truth: Christ and His salvation

Confidence	Predictions from BERT-base
59.7	Christ and His salvation"
35.4	Christ and His salvation
2.3	Christ
1.3	"Christ and His salvation"
0.8	"Christ and His salvation
0.1	Christ and His salvation" was
0.1	"Christ

Figure 3: Example of answer span distribution from model trained via *independent* objective.

Confidence	Predictions from BERT-base-*compound*
71.8	Christ and His salvation
10.9	"Christ and His salvation"
4.7	Christ and His salvation"
4.6	Luther's rediscovery of "Christ and His salvation
3.1	"Christ and His salvation
1.2	Luther's rediscovery of "Christ and His salvation"
0.8	Luther's rediscovery

Figure 4: Example of answer span distribution from model trained via *compound* objective.

Question: How many species of bird and mammals are there in the Amazon region?
Passage: The region is home to about 2.5 million insect species, tens of thousands of plants, and some 2,000 birds and mammals. To date, at least 40,000 plant species, 2,200 fishes, 1,294 birds, 427 mammals, 428 amphibians, and 378 reptiles have been scientifically classified in the region. One in five of all the bird species in the world live in the rainforests of the Amazon, and one in five of the fish species live in Amazonian rivers and streams. Scientists have described between 96,660 and 128,843 invertebrate species in Brazil alone.
Ground Truth: 2,000

Confidence	Predictions from BERT-base
37.0	2,000 birds and mammals. To date, at least 40,000 plant species, 2,200 fishes, 1,294 birds, 427
34.6	427
27.7	2,000
0.2	1,294 birds, 427
0.2	427 mammals
0.1	2,000 birds
0.1	2,000 birds and mammals

Figure 5: Example of answer span distribution from model trained via *independent* objective.

Confidence	Predictions from BERT-base-*compound*
71.7	427
21.5	2,000
5.1	2,000 birds and mammals. To date, at least 40,000 plant species, 2,200 fishes, 1,294 birds, 427
0.8	some 2,000
0.2	427 mammals
0.1	1,294 birds, 427
0.1	2,000 birds and mammals. To date, at least 40,000 plant species, 2,200 fishes, 1,294

Figure 6: Example of answer span distribution from model trained via *compound* objective.

B Exploration of similarity functions

We have experimented with 4 types of similarity functions in our experiments. For each start representation $\boldsymbol{h_s} \in \mathbb{R}^d$ and end representation $\boldsymbol{h_e} \in \mathbb{R}^d$, both column vectors from the matrix of boundary vectors $\boldsymbol{H}_s$, $\boldsymbol{H}_e \in \mathbb{R}^{d \times L}$ respectively. Note that d here is model specific dimension and L is passage length. The similarity functions above these representations are defined as:

- A dot product:

$$f_{dot}(\boldsymbol{h_s}, \boldsymbol{h_e}) = \boldsymbol{h_s}^\top \boldsymbol{h_e} \tag{8}$$

- A weighted dot product:

$$f_{wdot}(\boldsymbol{h_s}, \boldsymbol{h_e}) = \boldsymbol{w}^\top [\boldsymbol{h_s} \circ \boldsymbol{h_e}] \tag{9}$$

- An additive similarity:

$$f_{add}(\boldsymbol{h_s}, \boldsymbol{h_e}) = \boldsymbol{w}^\top [\boldsymbol{h_s}; \boldsymbol{h_e}] \tag{10}$$

- An additive-similarity combined with weighted product:

$$f_{add-wdot}(\boldsymbol{h_s}, \boldsymbol{h_e}) = \boldsymbol{w}^\top [\boldsymbol{h_s}; \boldsymbol{h_e}; \boldsymbol{h_s} \circ \boldsymbol{h_e}] \tag{11}$$

- A multi-layer perceptron (MLP) as proposed by Lee et al. (2019):

$$f_{MLP}(\boldsymbol{h_s}, \boldsymbol{h_e}) = \boldsymbol{w}^\top \sigma(\boldsymbol{W}[\boldsymbol{h_s}; \boldsymbol{h_e}] + \boldsymbol{b}) + \boldsymbol{b}_2 \tag{12}$$

  where $\sigma(x) = ln(relu(x))$ and ln denotes layer normalization (Ba et al., 2016).

Our results are presented in Table 6.

	BidAF-IJ		BERT-IJ	
f_{dot}	64.4	74.35	**81.83**	**88.52**
f_{add}	**66.28**	75.25	81.52	88.47
f_{wdot}	65.73	75.03	81.35	88.29
$f_{add-wdot}$	66.02	**75.54**	81.45	88.44
f_{MLP}	-	-	81.61	88.44

Table 6: A comparison of similarity functions in the models trained via *compound* objective.

C Conditional objective

Some of the recent LRMs also predict start and end indices of a span jointly. For comparison with our joint objective, we reimplemented the conditional objective used in ALBERT (Lan et al., 2019).

First, the probabilities $\boldsymbol{P}(a_s)$ for the start position are computed in the same manner as for the independent objective – by applying a linear transformation layer on top of representations $\boldsymbol{H} \in \mathbb{R}^{d \times L}$ from the last layer of the LRM, where d is the model dimension and L denotes the input sequence length.

$$\boldsymbol{P}(a_s) = \boldsymbol{w}_s^\top \boldsymbol{H} + \boldsymbol{b_s} \tag{13}$$

During the validation, top k ($k = 10$ in our experiments) start positions are selected from these probabilities, while in the training phase, we apply teacher forcing by only selecting the correct start position. Representation of i-th start position $\boldsymbol{h}_i$ from the last layer of the LRM corresponding to the selected position is then concatenated with representations corresponding to all the other positions $k = 0..L$ into matrix $\boldsymbol{C}$.

$$\boldsymbol{C} = \begin{bmatrix} — & [\boldsymbol{h_0}; \boldsymbol{h_i}] & — \\ — & [\boldsymbol{h_1}; \boldsymbol{h_i}] & — \\ & \vdots & \\ — & [\boldsymbol{h_n}; \boldsymbol{h_i}] & — \end{bmatrix} \tag{14}$$

Subsequently, a layer with $\tanh$ activation is applied on this matrix $\boldsymbol{C}$, followed by a linear transformation to obtain the end probabilities:

$$\boldsymbol{P}(a_e | a_s = i) = \boldsymbol{w}_c^\top \tanh(\boldsymbol{W}\boldsymbol{C} + \boldsymbol{b'}) + \boldsymbol{b} \tag{15}$$

For each start position we again select top k end positions, to obtain k^2-best list of answer spans. In contrast to the official ALBERT implementation, we omitted a layer normalization after *tanh* layer.