Probabilistic Assumptions Matter: Improved Models for Distantly-Supervised Document-Level Question Answering

Hao Cheng*, **Ming-Wei Chang†, Kenton Lee†, Kristina Toutanova†**

*Microsoft Research
chehao@microsoft.com

†Google Research
{mingweichang, kentonl, kristout}@google.com

Abstract

We address the problem of extractive question answering using document-level distant supervision, pairing questions and relevant documents with answer strings. We compare previously used probability space and distant supervision assumptions (assumptions on the correspondence between the weak answer string labels and possible answer mention spans). We show that these assumptions interact, and that different configurations provide complementary benefits. We demonstrate that a multi-objective model can efficiently combine the advantages of multiple assumptions and outperform the best individual formulation. Our approach outperforms previous state-of-the-art models by 4.3 points in F1 on TriviaQA-Wiki and 1.7 points in Rouge-L on NarrativeQA summaries.[1]

1 Introduction

Distant supervision assumptions have enabled the creation of large-scale datasets that can be used to train fine-grained extractive short answer question answering (QA) systems. One example is TriviaQA (Joshi et al., 2017). There the authors utilized a pre-existing set of Trivia question-answer string pairs and coupled them with relevant documents, such that, with high likelihood, the documents support answering the questions (see Fig. 1 for an illustration). Another example is the NarrativeQA dataset (Kočiský et al., 2018), where crowd-sourced abstractive answer strings were used to weakly supervise answer mentions in the text of movie scripts or their summaries. In this work, we focus on the setting of document-level extractive QA, where distant supervision is specified as a set $\mathcal{A}$ of answer strings for an input question-document pair.
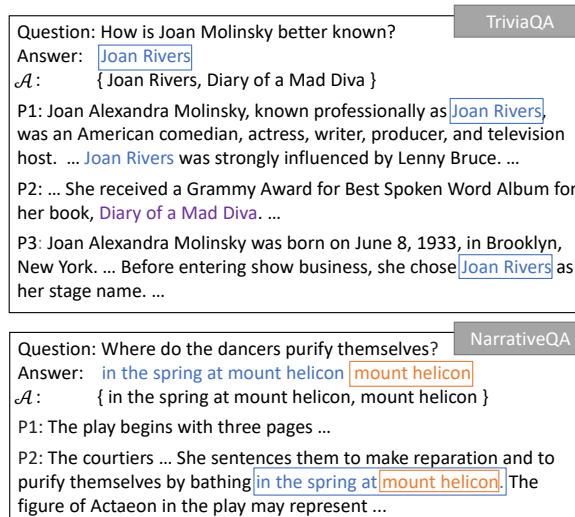

Figure 1: TriviaQA and NarrativeQA examples. In the TriviaQA example, there are three occurrences of the original answer string "Joan Rivers" (blue), and one alternate but incorrect alias "Diary of a Mad Diva" (purple). Only two "Joan Rivers" mentions (shown in blue boxes) support answering the question. In the NarrativeQA example, there are two answer stings in $\mathcal{A}$: "in the spring at mount helicon" (blue) and "mount helicon" (orange), with the latter being a substring of the former. Both mentions in P2 are correct answer spans.

Depending on the data generation process, the properties of the resulting supervision from the sets $\mathcal{A}$ may differ. For example, the provided answer sets in TriviaQA include aliases of original trivia question answers, aimed at capturing semantically equivalent answers but liable to introducing semantic drift. In Fig. 1, the possible answer string "Diary of a Mad Diva" is related to "Joan Rivers", but is not a valid answer for the given question.

On the other hand, the sets of answer strings in NarrativeQA are mostly valid since they have high overlap with human-generated answers for the given question/document pair. As shown in Fig. 1, "in the spring at mount helicon" and "mount helicon" are both valid answers with relevant mentions. In this case, the annotators chose answers

---

[1] Based on the TriviaQA-Wiki leaderboard, our approach was the SOTA when this work was submitted on Dec 04, 2019.

that appear verbatim in the text but in the more general case, noise may come from partial phrases and irrelevant mentions.

While distant supervision reduces the annotation cost, increased coverage often comes with increased noise (e.g., expanding entity answer strings with aliases improves coverage but also increases noise). Even for fixed document-level distant supervision in the form of a set of answers $\mathcal{A}$, different interpretations of the partial supervision lead to different points in the coverage/noise space and their relative performance is not well understood.

This work systematically studies methods for learning and inference with document-level distantly supervised extractive QA models. Using a BERT (Devlin et al., 2019) joint question-passage encoder, we study the compound impact of:

- **Probability space** (§2): ways to define the model's probability space based on independent paragraphs or whole documents.
- **Distant supervision assumptions** (§3): ways to translate the supervision from possible strings $\mathcal{A}$ to possible locations of answer mentions in the document.
- **Optimization and inference** (§4): ways to define corresponding training objectives (e.g. Hard EM as in Min et al. (2019) vs. Maximum Marginal Likelihood) and make answer string predictions during inference (Viterbi or marginal inference).

We show that the choice of probability space puts constraints on the distant supervision assumptions that can be captured, and that all three choices interact, leading to large differences in performance. Specifically, we provide a framework for understanding different distant supervision assumptions and the corresponding trade-off among the coverage, quality and strength of distant supervision signal. The best configuration depends on the properties of the possible annotations $\mathcal{A}$ and is thus data-dependent. Compared with recent work also using BERT representations, our study show that the model with most suitable probabilistic treatment achieves large improvements of 4.6 F1 on TriviaQA and 1.7 Rouge-L on NarrativeQA respectively. Additionally, we design an efficient multi-loss objective that can combine the benefits of different formulations, leading to significant improvements in accuracy, surpassing the best previously reported results on the two studied
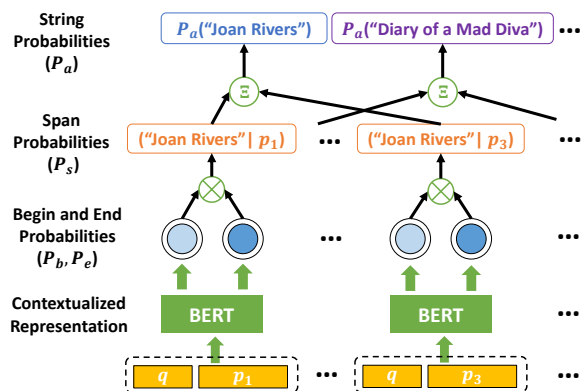

Figure 2: The document-level QA model as used for test-time inference. The lower part is a BERT-based paragraph-level answer scoring component, and the upper part illustrates the probability aggregation across answer spans sharing the same answer string. Ξ refers to either a sum or a max operator. In the given example, "John Rivers" is derived from two paragraphs.

tasks. Results are further strengthened by transfer learning from fully labeled short-answer extraction data in SQuAD 2.0 (Rajpurkar et al., 2018), leading to a final state-of-the-art performance of 76.3 F1 on TriviaQA-Wiki and 62.9 on the NarrativeQA summaries task.[2]

2 Probability Space

Here, we first formalize both paragraph-level and document-level models, which have been previously used for document-level extractive QA. Typically, paragraph-level models consider each paragraph in the document independently, whereas document models integrate some dependencies among paragraphs.

To define the model, we need to specify the probability space, consisting of a set of possible *outcomes* and a way to assign *probabilities* to individual outcomes. For extractive QA, the probability space outcomes consist of token positions of answer mention spans.

The overall model architecture is shown in Fig. 2. We use BERT (Devlin et al., 2019) to derive representations of document tokens. As is standard in state-of-the-art extractive QA models (Devlin et al., 2019; Lee et al., 2019; Min et al., 2019), the BERT model is used to encode a pair of a given question with one paragraph from a given document into neural text representations. These representations are then used to

---

[2]The code is available at https://github.com/hao-cheng/ds_doc_qa

5658

define scores/probabilities of possible answer begin and end positions, which are in turn used to define probabilities over possible answer *spans*. Then the answer string probabilities can be defined as the aggregation over all possible answer spans/mentions.

In the following, we show that paragraph-level and document-level models differ only in the space of possible outcomes and the way of computing answer span probabilities from answer position begin and end scores.

Scoring answer begin and end positions Given a question q and a document d consisting of K paragraphs $p_1, \ldots, p_K$, the BERT encoder produces contextualized representations for each question-paragraph pair (q, p_k). Specifically, for each token position i^k in p_k, the final hidden vector $\mathbf{h}^{(i,k)} \in \mathbb{R}^d$ is used as the contextualized token embedding, where d is the vector dimension.

The span-begin score is computed as $s_b(i^k) = \mathbf{w}_b^T \mathbf{h}^{(i,k)}$ using a weight vector $\mathbf{w}_b \in \mathbb{R}^d$. The span-end score $s_e(j^k)$ is defined in the same way. The probabilities for a start position i^k and an end position j^k are

$$P_b(i^k) = \frac{\exp(s_b(i^k))}{Z_b}, \qquad (1)$$

$$P_e(j^k) = \frac{\exp(s_e(j^k))}{Z_e}, \qquad (2)$$

where Z_b, Z_e are normalizing factors, depending on the probability space definition (detailed below). The probability of an answer span from i^k to j^k is defined as $P_s(i^k, j^k) = P_b(i^k) P_e(j^k)$.

The partition functions Z_b and Z_e depend on whether we use a paragraph-level or document-level probability space.

Paragraph-level model In paragraph level models, we assume that for a given question against a document d, each of its paragraphs $p_1, \ldots, p_K$ independently selects a pair of answer positions (i^k, j^k), which are the begin and end of the answer from paragraph p_k. In the case that p_k does not support answering the question q, special NULL positions are selected (following the SQuAD 2.0 BERT implementation[3]). Thus, the set of possible outcomes Ω in the paragraph-level probability space is the set of lists of begin/end position pairs, one from each paragraph: $\{[(i^1, j^1), \ldots, (i^K, j^K)]\}$, where i^k

and j^k range over positions in the respective paragraphs.

The answer positions in different paragraphs are independent, and the probability of each paragraph's answer begin and end is computed by normalizing over all possible positions in that paragraph, i.e.,

$$Z_b^k = \sum_{i \in \mathcal{I}^k \cup \{\text{NULL}\}} \exp(s_b(i)), \qquad (3)$$

$$Z_e^k = \sum_{j \in \mathcal{I}^k \cup \{\text{NULL}\}} \exp(s_e(j)), \qquad (4)$$

where $\mathcal{I}^k$ is the set of all positions in the paragraph p_k. The probability of an answer begin at i^k is $P_b(i^k) = \exp(s_b(i^k))/Z_b{}^k$ and the probability of an end at j^k is defined analogously. The probability of a possible answer position assignment for the document d is then defined as $P([(i^1, j^1), \ldots, (i^K, j^K)]) = \prod_k P_b(i^k) P_e(j^k)$.

As we can see from the above definition, due to the independence assumption, models using paragraph-level normalization do not learn to directly calibrate candidate answers from different paragraphs against each other.

Document-level model In document-level models, we assume that for a given question against document d, a single answer span is selected (as opposed to one for each paragraph in the paragraph-level models).[4] Here, the possible positions in all paragraphs are a part of a joint probability space and directly compete against each other.

In this case, Ω is the set of token spans $\{(i, j)\}$, where i and j are the begin and end positions of the selected answer. The normalizing factors are therefore aggregated over all paragraphs, i.e.,

$$Z_b^* = \sum_{k=1}^{K} \sum_{i \in \mathcal{I}^k} \exp(s_b(i)), \qquad (5)$$

$$Z_e^* = \sum_{k=1}^{K} \sum_{j \in \mathcal{I}^k} \exp(s_e(j)). \qquad (6)$$

Compared with (3) and (4), since there is always a valid answer in the document for the tasks studied here, NULL is not necessary for document-level models and thus can be excluded from the

---

[3] https://github.com/google-research/bert

[4] In this paper, we focus on datasets where the document is known to contain a valid answer. It is straightforward to remove this assumption and consider document-level NULL for future work.

	Coverage	Quality	Strength
H1	↗	↘	↗
H2	→	→	→
H3	↘	↗	↘

Table 1: Distant supervision assumptions and their corresponding tradeoffs. (↗) indicates highest value, (→) medium, and (↘) lowest value.

inner summation of (5) and (6). The probability of a possible outcome, *i.e.* an answer span, is $P(i, j) = \exp(s_b(i) + s_e(j))/(Z_b^* Z_e^*)$.

3 Distant Supervision Assumptions

There are multiple ways to interpret the distant supervision signal from $\mathcal{A}$ as possible outcomes in our paragraph-level and document-level probability spaces, leading to corresponding training loss functions. Although several different paragraph-level and document-level losses (Chen et al., 2017; Kadlec et al., 2016; Clark and Gardner, 2018; Lin et al., 2018; Min et al., 2019) have been studied in the literature, we want to point out that when interpreting the distant supervision signal, there is a tradeoff among multiple desiderata:

- **Coverage**: maximize the number of instances of relevant answer spans, which we can use to provide positive examples to our model.
- **Quality**: maximize the quality of annotations by minimizing noise from irrelevant answer strings or mentions.
- **Strength**: maximize the strength of the signal by reducing uncertainty and pointing the model more directly at correct answer mentions.

We introduce three assumptions (H1, H2, H3) for how the distant supervision signal should be interpreted, which lead to different tradeoffs among the desiderata above (see Table 1).

We begin with setting up additional useful notation. Given a document-question pair (d, q) and a set of answer strings $\mathcal{A}$, we define the set of $\mathcal{A}$-consistent token spans $\mathcal{Y}_\mathcal{A}$ in d as follows: for each paragraph p_k, span $(i^k, j^k) \in \mathcal{Y}_\mathcal{A}^k$ if and only if the string spanning these positions in the paragraph is in $\mathcal{A}$. For paragraph-level models, if for paragraph p_k the set $\mathcal{Y}_\mathcal{A}^k$ is empty, we redefine $\mathcal{Y}_\mathcal{A}^k$ to be {NULL}. Similarly, we define the set of $\mathcal{A}$-consistent begin positions $\mathcal{Y}_{b,\mathcal{A}}^k$ as the start positions of consistent spans: $\mathcal{Y}_{b,\mathcal{A}}^k = \cup_{(i,j) \in \mathcal{Y}_\mathcal{A}^k} \{i\}$. $\mathcal{Y}_{e,\mathcal{A}}^k$ for $\mathcal{A}$-consistent end positions is defined

analogously. In addition, we term an answer span (i, j) *correct* for question q, if its corresponding answer string is a correct answer to q, and the context of the specific mention of that answer string from positions i to j entails this answer. Similarly, we term an answer begin/end position *correct* if there exists a correct answer span starting/ending at that position.

H1: All $\mathcal{A}$-consistent answer spans are correct. While this assumption is evidently often incorrect (low on the quality dimension ↘), especially for TriviaQA, as seen from Fig. 1, it provides a large number of positive examples and a strong supervision signal (high on coverage ↗ and strength ↗). We include this in our study for completeness.

H1 translates differently into possible outcomes for corresponding models depending on the probability space (paragraph or document). Paragraph-level models select multiple answer spans, one for each paragraph, to form a possible outcome. Thus, multiple $\mathcal{A}$-consistent answer spans can occur in a single outcome, as long as they are in different paragraphs. For multiple $\mathcal{A}$-consistent answer spans in the same paragraph, these can be seen as mentions that can be selected with equal probability (e.g., by different annotators). Document-level models select a single answer span in the document and therefore multiple $\mathcal{A}$-consistent answer spans can be seen as occurring in separate annotation events. Table 2 shows in row one the log-probability of outcomes consistent with H1.

H2: Every positive paragraph has a correct answer in its $\mathcal{A}$-consistent set. Under this assumption, each paragraph with a non-empty set of $\mathcal{A}$-consistent spans (termed a *positive* paragraph) has a correct answer. As we can see from the TriviaQA example in Fig. 1, this assumption is correct for the first and third paragraph, but not the second one, as it only contains a mention of a noisy answer alias. This assumption has medium coverage (→), as it generates positive examples from multiple paragraphs but does not allow multiple positive mentions in the same paragraph. It also decreases noise (higher quality →) (e.g. does not claim that all the mentions of "Joan Rivers" in the first paragraph support answering the question). The strength of the supervision signal is weakened (→) relative to H1, as now the model needs to figure out which of the multiple $\mathcal{A}$-consistent mentions in each paragraph is correct.

H2 has two variations: correct *span*, assuming

	Span-Based	Position-Based
H1	$\sum_{k\in\mathcal{K}}\sum_{(i^k,j^k)\in\mathcal{Y}^k_{\mathcal{A}}}\log P_s(i^k,j^k)$	$\sum_{k\in\mathcal{K}}\sum_{i^k\in\mathcal{Y}^k_{b,\mathcal{A}}}\log P_b(i^k) + \sum_{k\in\mathcal{K}}\sum_{j^k\in\mathcal{Y}^k_{e,\mathcal{A}}}\log P_e(j^k)$
H2	$\sum_{k\in\mathcal{K}}\log\Xi_{(i^k,j^k)\in\mathcal{Y}^k_{\mathcal{A}}}P_s(i^k,j^k)$	$\sum_{k\in\mathcal{K}}\log\Xi_{i^k\in\mathcal{Y}^k_{b,\mathcal{A}}}P_b(i^k) + \sum_{k\in\mathcal{K}}\log\Xi_{j^k\in\mathcal{Y}^k_{e,\mathcal{A}}}P_e(j^k)$
H3	$\log\Xi_{k\in\mathcal{K}}\Xi_{(i^k,j^k)\in\mathcal{Y}^k_{\mathcal{A}}}P_s(i^k,j^k)$	$\log\Xi_{k\in\mathcal{K}}\Xi_{i^k\in\mathcal{Y}^k_{b,\mathcal{A}}}P_b(i^k) + \log\Xi_{k\in\mathcal{K}}\Xi_{j^k\in\mathcal{Y}^k_{e,\mathcal{A}}}P_e(j^k)$

Table 2: Objective functions for a document-question pair (d,q) under different distant supervision assumptions. Ξ refers to $\sum$ and max for MML and HardEM, respectively.

that one of the answer spans (i^k,j^k) in $\mathcal{Y}^k_{\mathcal{A}}$ is correct, and correct *position*, assuming that the paragraph has a correct answer begin position from $\mathcal{Y}^k_{b,\mathcal{A}}$ and a correct answer end position from $\mathcal{Y}^k_{e,\mathcal{A}}$, but its selected answer span may not necessarily belong to $\mathcal{Y}^k_{\mathcal{A}}$. For example, if $\mathcal{A}$ contains $\{abcd, bc\}$, then abc would have correct begin and end, but not be a correct span. It does not make sense for modeling to assume the paragraph has correct begin and end positions instead of a correct answer span (i.e., we don't really want to get inconsistent answers like abc above), but given that our probabilistic model assumes independence of begin and end answer positions, it may not be able to learn well with span-level weak supervision. Some prior work (Clark and Gardner, 2018) uses an H2 position-based distant supervision assumption with a pair-paragraph model akin to our document-level ones. Lin et al. (2018) use an H2 span-based distant supervision assumption. The impact of position vs. span-based modeling of the distant supervision is not well understood. As we will see in the experiments, for the majority of settings, position-based weak supervision is more effective than span-based for our model.

For paragraph-level and document-level models, H2 corresponds differently to possible outcomes. For paragraph models, one outcome can select answer spans in all positive paragraphs and NULL in negative ones. For document-level models, we view answers in different paragraphs as outcomes of multiple draws from the distribution. The identity of the particular correct span or begin/end position is unknown, but we can compute the probability of the event comprising the consistent outcomes. Table 2 shows the log-probability of the outcomes consistent with H2 in row two (right for span-based and left for position-based interpretation, when plugging in $\sum$ for Ξ).

H3: The document has a correct answer in its $\mathcal{A}$-consistent set $\mathcal{Y}_{\mathcal{A}}$. This assumption posits that the document has a correct answer span (or

begin/end positions), but not every positive paragraph needs to have one. It further improves supervision quality ($\nearrow$), because for example, it allows the model to filter out the noise in paragraph two in Fig. 1. Since the model is given a choice of any of the $\mathcal{A}$-consistent mentions, it has the capability to assign zero probability mass on the supervision-consistent mentions in that paragraph.

On the other hand, H3 has lower coverage ($\searrow$) than H1 and H2, because it provides a single positive example for the whole document, rather than one for each positive paragraph. It also reduces the strength of the supervision signal ($\searrow$), as the model now needs to figure out which mention to select from the larger document-level set $\mathcal{Y}_{\mathcal{A}}$.

Note that we can only use H3 coupled with a document-level model, because a paragraph-level model cannot directly tradeoff answers from different paragraphs against each other, to select a single answer span from the document. As with the other distant supervision hypotheses, span-based and position-based definitions of the possible consistent outcomes can be formulated. The log-probabilities of these events are defined in row three of Table 2, when using $\sum$ for Ξ. H3 was used by Kadlec et al. (2016) for cloze-style distantly supervised QA with recurrent neural network models.

The probability-space (paragraph vs. document-level) and the distant supervision assumption (H1, H2, and H3, each position or span-based) together define our interpretation of the distant supervision signal resulting in definitions of probability space outcomes consistent with the supervision. Next, we define corresponding optimization objectives to train a model based on this supervision and describe the inference methods to make predictions with a trained model.

4 Optimization and Inference Methods

For each distant supervision hypothesis, we maximize either the marginal log-likelihood of $\mathcal{A}$-

consistent outcomes (MML) or the log-likelihood of the most likely outcome (HardEM). The latter was found effective for weakly supervised tasks including QA and semantic parsing by Min et al. (2019).

Table 2 shows the objective functions for all distant supervision assumptions, each comprising a pairing of a distant supervision hypothesis (H1, H2, H3) and position-based vs. span-based interpretation. The probabilities are defined according to the assumed probability space (paragraph or document). In the table, $\mathcal{K}$ denotes the set of all paragraphs in the document, and $\mathcal{Y}^k$ denotes the set of weakly labeled answer spans for the paragraph p_k (which can be {NULL} for paragraph-level models). Note that span-based and position-based objective functions are equivalent for H1 because of the independence assumption, *i.e.* $P_s(i^k, j^k) = P_b(i^k)P_e(j^k)$.

Inference: Since the task is to predict an answer *string* rather than a particular mention for a given question, it is potentially beneficial to aggregate information across answer spans corresponding to the same string during inference. The score of a candidate answer string can be obtained as $P_a(x) = \Xi_{(i,j)\in\mathcal{X}} P_s(i,j)$, where $\mathcal{X}$ is the set of spans corresponding to the answer string x, and Ξ can be either $\sum$ or $\max$.[5] It is usually beneficial to match the training objective with the corresponding inference method, *i.e.* MML with marginal inference $\Xi = \sum$, and HardEM with max (Viterbi) inference $\Xi = \max$. Min et al. (2019) showed HardEM optimization was useful when using an H2 span-level distant supervision assumption coupled with max inference, but it is unclear whether this trend holds when $\sum$ inference is useful or other distant supervision assumptions perform better. We therefore study exhaustive combinations of probability space, distant supervision assumption, and training and inference methods.

5 Experiments

5.1 Data and Implementation

Two datasets are used in this paper: TriviaQA (Joshi et al., 2017) in its Wikipedia formulation, and NarrativeQA (summaries setting) (Kočiský et al., 2018). Using the same preprocessing as

Clark and Gardner (2018) for TriviaQA-Wiki[6], we only keep the top 8 ranked paragraphs up to 400 tokens for each document-question pair for both training and evaluation. Following Min et al. (2019), for NarrativeQA we define the possible answer string sets $\mathcal{A}$ using Rouge-L (Lin, 2004) similarity with crouwdsourced abstractive answer strings. We use identical data preprocessing and the evaluation script provided by the authors.

In this work, we use the BERT-base model for text encoding and train our model with the default configuration as described in (Devlin et al., 2019), fine-tuning all parameters. We fine-tune for 3 epochs on TriviaQA and 2 epochs on NarrativeQA.

5.2 Optimization and Inference for Latent Variable Models

Here we look at the cross product of optimization (HardEM vs MML) and inference (**Max** vs **Sum**) for all distant supervision assumptions that result in models with latent variables. We therefore exclude H1 and look at the other two hypotheses, H2 and H3, each coupled with a span-based (Span) or position-based (Pos) formulation and a paragraph-level (P) or a document level (D) probability space. The method used in Min et al. (2019) corresponds to span-based H2-P with HardEM training and **Max** inference. The results are shown in Fig. 3.

First, we observe that inference with **Sum** leads to significantly better results on TriviaQA under H2-P and H2-D, and slight improvement under H3-D. On NarrativeQA, inference with **Max** is better. We attribute this to the fact that correct answers often have multiple relevant mentions for TriviaQA (also see §5.6), whereas for NarrativeQA this is rarely the case. Thus, inference with **Sum** in NarrativeQA could potentially boost the probability of irrelevant frequent strings.

Consistent with (Min et al., 2019), we observe that span-based HardEM works better than span-based MML under H2-P, with a larger advantage on NarrativeQA than on TriviaQA. However, under H2-D and H3-D, span-based MML performs consistently better than span-based HardEM. For position-based objectives, MML is consistently better than HardEM (potentially because HardEM may decide to place its probability mass on begin-end position combinations that do not contain mentions of strings in $\mathcal{A}$). Finally, it can be ob-

---

[5]For inference with marginal ($\sum$) scoring, we use an approximate scheme where we only aggregate probabilities of candidates strings generated from a 20-best list of begin/end answer positions for each paragraph.

---

[6]https://github.com/allenai/document-qa

(a) TriviaQA F1

(b) NarrativeQA Rouge-L

Figure 3: Comparison of different optimization and inference choices grouped by distant supervision hypothesis based on dev set results for TriviaQA and NarrativeQA.

Objective	Infer	TriviaQA		NarrativeQA
		F1	EM	Rouge-L
Paragraph-level Models				
H1-P	Max	67.9	63.3	55.3
	Sum	70.4	66.0	53.6
H2-P	Max	71.9	67.7	59.2
	Sum	73.0	69.0	57.8
Document-level Models				
H1-D	Max	55.8	51.0	59.4
	Sum	65.2	61.2	59.1
H2-D	Max	70.3	66.2	60.1
	Sum	72.4	68.4	59.9
H3-D	Max	75.1	70.6	59.1
	Sum	75.3	70.8	59.2

Table 3: Comparison of distant supervision hypotheses using MML-Pos objectives on TriviaQA and NarrativeQA dev sets.

served that under each distant supervision hypothesis/probability space combination, the position-based MML is always the best among the four objectives. Position-based objectives may perform better due to the independence assumptions for begin/end positions of the model we use and future work may arrive at different conclusions if position dependencies are integrated. Based on this thorough exploration, we focus on experimenting with position-based objectives with MML for the rest of this paper.

5.3 Probability Space and Distant Supervision Assumptions

In this subsection, we compare probability space and distant supervision assumptions. Table 3 shows the dev set results, where the upper section compares paragraph-level models (H1-P, H2-P), and the lower section compares document-level models (H1-D, H2-D, H3-D). The performance of models with both **Max** and **Sum** inference is shown. We report F1 and Exact Match (EM) scores for TriviaQA, and Rouge-L scores for NarrativeQA.

For TriviaQA, H3-D achieves significantly bet-

ter results than other formulations. Only H3-D is capable of "cleaning" noise from positive paragraphs that don't have a correct answer (e.g. paragraph two in Fig. 1), by deciding which $\mathcal{A}$-consistent mention to trust. The paragraph-level models H1-P and H2-P outperform their corresponding document-level counterparts H1-D and H2-D. This may be due to the fact that without H3, and without predicting NULL, D models do not learn to detect irrelevant paragraphs.

Unlike for TriviaQA, H2-D models achieve the best performance for NarrativeQA. We hypothesize this is due to the fact that positive paragraphs that don't have a correct answer are very rare in NarrativeQA (as summaries are relatively short and answer strings are human-annotated for the specific documents). Therefore, H3 is not needed to clean noisy supervision, and it is not useful since it also leads to a reduction in the number of positive examples (coverage) for the model. Here, document-level models always improve over their paragraph counterparts, by learning to calibrate paragraphs directly against each other.

5.4 Multi-Objective Formulations and Clean Supervision

Here we study two methods to further improve weakly supervised QA models. First, we combine two distant supervision objectives in a multi-task manner, *i.e.* H2-P and H3-D for TriviaQA, and H2-P and H2-D for NarrativeQA, chosen based on the results in §5.3. H2 objectives have higher coverage than H3 while being more susceptible

Objective	Clean	Infer	F1	EM	Rouge-L
			TriviaQA		NarrativeQA
			Single-objective		
Par	X	Max	71.9	67.7	59.2
		Sum	73.0	69.0	57.8
	✓	Max	74.2	70.1	61.7
		Sum	**74.9**	**70.9**	**61.7**
Doc	X	Max	75.1	70.6	60.1
		Sum	75.3	70.8	59.9
	✓	Max	75.5	70.8	62.8
		Sum	**75.5**	**70.9**	**62.9**
			Multi-objective		
Par + Doc	X	Max	75.6	71.2	60.5
		Sum	75.9	71.6	60.5
	✓	Max	75.8	71.2	63.0
		Sum	**76.2**	**71.7**	**63.1**

Table 4: Dev set results comparing multi-objectives and clean supervison. ✓ indicates the QA model is pre-trained on SQUAD.

TriviaQA Wiki	Full		Verified	
	F1	EM	F1	EM
Ours (**H2-P+H3-D**)	**76.3**	**72.1**	**85.5**	**82.2**
w/o SQUAD	75.7	71.6	83.6	79.6
(Wang et al., 2018b)	71.4	66.6	78.7	74.8
(Clark and Gardner, 2018)	68.9	64.0	72.9	68.0
(Min et al., 2019)	67.1	–	–	–

NarrativeQA Summary	Rouge-L
Ours (**H2-P+H2-D**)	**62.9**
w/o SQUAD	60.5
(Nishida et al., 2019)	59.9
w/o external data	54.7
(Min et al., 2019)	58.8

Table 5: Test set results on TriviaQA Wiki and NarrativeQA Summaries. "*w/o* SQUAD" refers to our best model without pretraining on SQUAD 2.0. "*w/o* external data" refers to the model from (Nishida et al., 2019) without using MS MARCO data (Bajaj et al., 2018).

to noise. Paragraph-level models have the advantage of learning to score irrelevant paragraphs (via NULL outcomes). Note that we use the same parameters for the two objectives and the multi-objective formulation does not have more parameters and is no less efficient than the individual models. Second, we use external clean supervision from SQUAD 2.0 (Rajpurkar et al., 2018) to train the BERT-based QA model for 2 epochs. This model matches the P probability space and is able to detect both NULL and extractive answer spans. The resulting network is used to initialize the models for TriviaQA and NarrativeQA. The results are shown in Table 4.

It is not surprising that using external clean supervision improves model performance (e.g. (Min et al., 2017)). We note that, interestingly, this external supervision narrows the performance gap between paragraph-level and document-level models, and reduces the difference between the two inference methods.

Compared with their single-objective components, multi-objective formulations improve performance on both TriviaQA and NarrativeQA.

5.5 Test Set Evaluation

Table 5 reports test set results on TriviaQA and NarrativeQA for our best models, in comparison to recent state-of-art (SOTA) models. For TriviaQA, we report F1 and EM scores on the full test set and the verified subset. For NarrativeQA, Rouge-L scores are reported.

Compared to recent TriviaQA SOTA (Wang et al., 2018b), our best models achieve 4.9 F1 and 5.5 EM improvement on the full test set, and 6.8 F1 and 7.4 EM improvement on the verified subset. On the NarrativeQA test set, we improve Rouge-L by 3.0 over (Nishida et al., 2019). The large improvement, even without additional fully labeled data, demonstrates the importance of selecting an appropriate probability space and interpreting the distant-supervision in a way cognizant of the properties of the data, as well as selecting a strong optimization and inference method. With external fully labeled data to initialize the model, performance is further significantly improved.

5.6 Analysis

In this subsection, we carry out analyses to study the relative performance of paragraph-level and document-level models, depending on the size of answer string set $|\mathcal{A}|$ and the number of $\mathcal{A}$-consistent spans, which are hypothesized to correlate with label noise. We use the TriviaQA dev set and the best performing models, *i.e.* H2-P and H3-D with **Sum** inference.

We categorize examples based on the size of their answer string set, $|\mathcal{A}|$, and the size of their corresponding set of $\mathcal{A}$-consistent spans, $|\mathcal{I}|$. Specifically, we divide the data into 4 subsets and

Subset	$|\mathcal{A}|$	$|\mathcal{I}|$	Size	H2-P	H3-D	Δ
$\mathcal{Q}^{ss}$	$=1$	≤ 5	2585	66.8	67.4	0.6
$\mathcal{Q}^{ls}$	>1	≤ 5	853	68.7	70.1	1.4
$\mathcal{Q}^{sl}$	$=1$	>5	1149	82.0	84.9	2.9
$\mathcal{Q}^{ll}$	>1	>5	3034	86.3	88.4	2.1

Table 6: F1 scores on 4 subsets of TriviaQA dev, grouped by the size of their answer string sets $\mathcal{A}$ and corresponding set of possible mentions $\mathcal{I}$. Δ indicates the improvement from H2-P to H3-D.

report performance separately on each subset, as shown in Table 6. In general, we expect $\mathcal{Q}^{sl}$ and $\mathcal{Q}^{ll}$ to be noisier due to the larger $\mathcal{I}$, where $\mathcal{Q}^{sl}$ potentially includes many irrelevant mentions while $\mathcal{Q}^{ll}$ likely contains more incorrect answer strings (false aliases). We can observe that the improvement is more significant for these noisier subsets, suggesting document-level modeling is crucial for handling both types of label noise.

6 Related Work

Distant supervision has been successfully used for decades for information extraction tasks such as entity tagging and relation extraction (Craven and Kumlien, 1999; Mintz et al., 2009). Several ways have been proposed to learn with DS, e.g., multi-label multi-instance learning (Surdeanu et al., 2012), assuming at least one supporting evidence (Hoffmann et al., 2011), integration of label-specific priors (Ritter et al., 2013), and adaption to shifted label distributions (Ye et al., 2019).

Recent work has started to explore distant supervision to scale up QA systems, particularly for open-domain QA where the evidence has to be retrieved rather than given as input. Reading comprehension (RC) with evidence retrieved from information retrieval systems establishes a weakly-supervised QA setting due to the noise in the heuristics-based span labels (Chen et al., 2017; Joshi et al., 2017; Dunn et al., 2017; Dhingra et al., 2017). One line of work jointly learns RC and evidence ranking using either a pipeline system (Wang et al., 2018a; Lee et al., 2018; Kratzwald and Feuerriegel, 2018) or an end-to-end model (Lee et al., 2019).

Another line of work focuses on improving distantly-supervised RC models by developing learning methods and model architectures that can better use noisy labels. Clark and Gardner (2018) propose a paragraph-pair ranking objective, which has components of both our H2-P and H3-D

position-based formulations. They don't explore multiple inference methods or combinations of objectives and use less powerful representations. In (Lin et al., 2018), a coarse-to-fine model is proposed to handle label noise by aggregating information from relevant paragraphs and then extracting answers from selected ones. Min et al. (2019) propose a hard EM learning scheme which we included in our experimental evaluation.

Our work focuses on examining probabilistic assumptions for document-level extractive QA. We provide a unified view of multiple methods in terms of their probability space and distant supervision assumptions and evaluate the impact of their components in combination with optimization and inference methods. To the best of our knowledge, the three DS hypotheses along with position and span-based interpretations have not been formalized and experimentally compared on multiple datasets. In addition, the multi-objective formulation is new.

7 Conclusions

In this paper, we demonstrated that the choice of probability space and interpretation of the distant supervision signal for document-level QA have a large impact, and that they interact. Depending on the properties of the data, different configurations are best, and a combined multi-objective formulation can reap the benefits of its constituents.

A future direction is to extend this work to question answering tasks that require reasoning over multiple documents, e.g., open-domain QA. In addition, the findings may generalize to other tasks, e.g., corpus-level distantly-supervised relation extraction.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

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. Association for Computational Linguistics.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855. Association for Computational Linguistics.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*.

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. Association for Computational Linguistics.

Bhuwan Dhingra, Kathryn Mazaitis, and William W. Cohen. 2017. Quasar: Datasets for question answering by search and reading. *CoRR*, abs/1707.03904.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *CoRR*, abs/1704.05179.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 908–918. 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.

Bernhard Kratzwald and Stefan Feuerriegel. 2018. Adaptive document retrieval for deep question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 576–581. Association for Computational Linguistics.

Jinhyuk Lee, Seongjun Yun, Hyunjae Kim, Miyoung Ko, and Jaewoo Kang. 2018. Ranking paragraphs for improving answer recall in open-domain question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 565–569. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745.

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. Association for Computational Linguistics.

Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 510–517. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Kyosuke Nishida, Itsumi Saito, Kosuke Nishida, Kazutoshi Shinoda, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2019. Multi-style generative reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2273–2284. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Alan Ritter, Luke Zettlemoyer, Oren Etzioni, et al. 2013. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics*, 1:367–378.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 455–465, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. 2018a. R3: Reinforced reader-ranker for open-domain question answering. In *AAAI Conference on Artificial Intelligence*.

Wei Wang, Ming Yan, and Chen Wu. 2018b. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1705–1714. Association for Computational Linguistics.

Qinyuan Ye, Liyuan Liu, Maosen Zhang, and Xiang Ren. 2019. Looking beyond label noise: Shifted label distribution matters in distantly supervised relation extraction. 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 3839–3848. Association for Computational Linguistics.