Passage Similarity and Diversification in Non-factoid Question Answering

Lakshmi Vikraman, Ali Montazeralghaem, Helia Hashemi, W. Bruce Croft, James Allan
University of Massachusetts Amherst, Amherst, MA, USA
{lvnair,montazer,hhashemi,croft,allan}@cs.umass.edu

ABSTRACT

The rise in popularity of mobile and voice search has led to a shift in focus from document retrieval to short answer passage retrieval for non-factoid questions. Some of the questions have multiple answers, and the aim is to retrieve a set of relevant answer passages, which covers all these alternatives. Compared to documents, answers are more specific and typically form more defined types or groups. Grouping answer passages based on strong similarity measures may provide a means of identifying these types. Typically, kNN clustering in combination with term-based representations have been used in Information Retrieval (IR) scenarios. An alternate method is to use pre-trained distributional representations such as GloVe and BERT, which capture additional semantic relationships. The recent success of trained neural models for various tasks provides the motivation for generating more task-specific representations. However, due to the absence of large datasets for incorporating passage level similarity information, a more feasible alternative is to use weak supervision based training. This information can then be used to generate a final ranked list of diversified answers using standard diversification algorithms.

In this paper, we introduce a new dataset NFPassageQA_Sim, with human annotated similarity labels for pairs of answer passages corresponding to each question. These similarity labels are then processed to generate another dataset NFPassageQA_Div, which consists of answer types for these questions. Using the similarity labels, we demonstrate the effectiveness of using weak supervision signals derived from GloVe, fine-tuned and trained using a BERT model for the task of answer passage clustering. Finally, we introduce a model which incorporates these clusters into a MMR (Maximal Marginal Relevance) model, which significantly beats other diversification baselines using both diversity and relevance metrics.

CCS CONCEPTS

• **Computing methodologies** → **Unsupervised learning**; • **Information systems** → **Question answering**; *Clustering and classification*; **Information retrieval diversity**; *Similarity measures*; *Novelty in information retrieval*; *Document structure*;

KEYWORDS

Question Answering; Diversification; Clustering

ACM Reference Format:
Lakshmi Vikraman, Ali Montazeralghaem, Helia Hashemi, W. Bruce Croft, James Allan. 2021. Passage Similarity and Diversification in Non-factoid Question Answering. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3471158.3472249

1 INTRODUCTION

Answer retrieval is an emerging topic in information retrieval, where the goal is to display answer text relevant to a question. This is especially applicable to scenarios such as mobile search, where display space is limited, and short answers provide a good alternative to a list of documents. In this paper we focus on non-factoid questions, which have multiple descriptive answers associated with them.

We present the first effort in using passage-level information to cluster answers for the purpose of generating a diverse set of answer passages to non-factoid questions. To this end, we show that distributed representations that address the semantic gap between passages and a weak supervision strategy that models a good similarity function, can be used to generate effective passage clusters. These clusters can then be used as inputs to a classic diversification model to generate a diverse ranked answer list. To ensure proper evaluation, we also introduce two new datasets NFPassageQA_Sim, which capture inter-passage similarities and NFPassageQA_Div, which contain various answer types (or subtopics) associated with a question.

The passage clustering task is significant because it captures the notion of grouping answers belonging to an answer type. A classic IR task related to this is Cluster-based Retrieval, where clusters of documents are retrieved in response to a query [16–19, 27, 36]. However, one of the main limitations of these models is that they group documents using term based Language Modeling (LM) strategies, which use term distribution information for estimating similarities, and do not capture additional semantic information. Consequently, the document clusters tend to be quite diffuse and difficult to define. The recent success of many neural models in IR shows the efficacy of using semantic models for various tasks, but not many datasets exist for the passage clustering task. The only existing dataset with information corresponding to various answer types is the YahooL29 dataset [22]. However, this dataset suffers from some limitations: it includes questions from only a single domain in the Yahoo QA forum and it was created using prepositional phrase clusters extracted from the answers without comparing full answer passages.

ICTIR '21, July 11, 2021, Virtual Event, Canada
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8611-1/21/07…$15.00
https://doi.org/10.1145/3471158.3472249

To create a good evaluation dataset, we perform human annotation to collect a new dataset, NFPassageQA_Sim with questions sampled from the test set of the ANTIQUE dataset [13]. In contrast to the YahooL29 dataset, the questions in the NFPassageQA_Sim dataset cover multiple domains with a larger and more complete set of candidate answers. This also differs from data created for tasks such as textual entailment [2, 37] where similarity is not based on questions or queries. Aside from the passage similarity dataset, we also generate a new dataset, NFPassageQA_Div, from the passage similarity labels and group the answers into various answer types. This contrasts with standard IR TREC diversity tasks [4, 32] where the queries are very short with subtopics covering various facets of the query.

Using these datasets, we investigate the effectiveness of using a weak supervision method for the answer passage clustering task. Various distributed pre-trained representations such as GloVe [25] and BERT [12] incorporate contexual information, but may not include task-specific information, which is generally captured by training models on large training sets. Due to a limitation in data size, we exploit a weak supervision strategy to train models. Weak supervision provides an alternative to human labelled data by leveraging other easily available sources. There is existing work in IR where weak supervision methods using BM25 ranking as weak signals have been shown to be effective for document ranking [11]. More recently, Xu et al. [39] demonstrated the application of weak supervision to passage retrieval tasks where a relatively small training set was used to fine-tune BERT for this task. We use a similar weak supervision strategy for answer passage clustering using three sources of weak labels: the Language Model (LM), GloVe [25], and pre-trained BERT [12]. We train a BERT model using different objective functions and learn a good similarity function, which is helpful for grouping similar answers together.

Once we create effective clusters, we employ a variation of a classic diversification approach to display the various answers effectively. There are two standard diversification models in IR: Implicit and Explicit. The implicit type assumes that each document represents its own topic and diversifies based on document similarity. Maximal Marginal Relevance (MMR) [3] is an example of this type. The other models query topics explicitly and diversifies the result set based on them. We introduce a modified version of the MMR model, `MMR Cluster`, which incorporates the automatically generated clusters into a MMR framework and creates a diversified ranked list.

We evaluate the answer passage clusters produced by a BERT model trained with weak signals generated using GloVe representations and show that the similarity clusters from these models significantly perform better than the baselines. We apply these clusters in `MMR Cluster` and show that the final diversified output significantly outperforms standard diversification baselines using various diversity and relevance metrics when evaluated with the new NFPassageQA_Div dataset. The best performing weak supervision based cluster outperforms the best baseline by 8.3% based on Precision-IA metric, 7% based on the S-Recall metric, and 5% based on the α-NDCG metric which shows the efficacy of this method.

2 RELATED WORK

Non-factoid Question Answering: The answer passage retrieval task is becoming increasingly more important in IR. One of the initial work in this area was the feature-based learning to rank model introduced by Yang et al. [40]. Subsequently, various deep learning models [34] have proved to be more effective for this task. To this end, Cohen and Croft [5, 6] demonstrated the effectiveness of LSTM models as well as the efficacy of hybrid models, which combine character and term level information to capture semantic relationships. More recently, there has been progress on finding effective ways to sample negative data to improve the performance of such models [7]. Besides these, BERT models have been demonstrated to perform very well for passage retrieval [21, 23], as well as open domain question answering tasks. [1, 35, 41]

Diversification Models: Search Result Diversification models in IR can be broadly classified into two types : Implicit and Explicit models. Implicit diversification assumes that the each document is its own topic. One of the most popular implicit approaches is MMR [3]. Various supervised implicit techniques have been proposed recently such as Zhu et al. [46] where the model is trained to optimize both novelty and relevance. Xia et al. [38] trained a neural tensor model which learns document representations automatically without using any handcrafted features. Explicit approaches model query subtopics explicitly and generate a ranked list, which is optimized based on topic coverage. xQUAD and PM-2 are examples of this approach where xQUAD [29] uses query reformulations as topics and PM-2 [9] is a proportionality based diversification model. Term Level Diversification Model proposoed by Dang et al. [8] uses xQUAD and PM-2 in conjunction with topic terms. Hu et al. [14] proposed a hierarchical variant of xQUAD and PM-2 which models topics as a hierarchy instead of a list. More recently, Sarwar et al. [30] proposed a linear programming formulation for topic proportionality used in combination with PM-2 diversification algorithm.

Clustering models in IR: The cluster hypothesis [15] states that relevant documents which satisfy an information need, tend to cluster together. This hypothesis triggered research in the area of cluster based models, where a cluster of documents is retrieved in response to a query. Various cluster-based models have been proposed, such as those leveraging topic models [36] or language models [18] for clustering. Yi and Allan [42] demonstrated that using nearest neighbors for smoothing works nearly as well as topic models for this task. Liu et al. [19] showed the effectiveness of using geometric mean representations of documents. Recently, Kurland and Krikon [16] and Raiber and Kurland [27] demonstrated using Language Modeling and MRF techniques to identify good clusters. Sheetrit et al. showed that the cluster hypothesis applies to passages as well as documents [31]. In this paper we used the same settings used in the traditional cluster based models. One major difference was that we used semantic representations in addition to LM based models.

Text similarity models: Various similarity models have been studied in NLP. However, their focus is on sentence pair tasks such as textual entailment [2, 37] or paraphrasing [24], which relates to studying similarity between short phrases. Fine-tuned BERT models have been shown to perform very well for many of these tasks

[12]. More recent work demonstrates the advantages of applying supplementary training [26] and multi task learning [20] on BERT for the similarity task. Most of the models are trained on large scale datasets and their effectiveness has been demonstrated using fully supervised models. Besides that, almost all the models studied in NLP are based on sentence level or phrase level information, while we study the similarity at passage level, which is a more complex task.

Weak Supervision models: Weak Supervision methods have been found to be effective for various Information Retrieval tasks such as document ranking [11] and QPP (Query Performance Prediction) [44]. The document ranking model used BM25 as the weak supervision signal, while the QPP model trained a neural model to predict the weight of the multiple signals contributing to the end task. Xu et al. [39] applied weak supervision using BERT and combined multiple signals using majority voting technique and also by learning a simple generative model to predict the labels [28]. The theoretical basis behind the effectiveness of weak supervision models was shown by Zamani and Croft [43]. Recently, Dehghani et al. [10] proposed a network which uses a model trained on a small set of true labels to control the gradient updates on another network trained for a particular task using weak labels. In this paper, we explore how weak supervision can be used to improve the performance of the passage clustering task.

3 TASK DEFINITION

The end-to-end task can be broken down into two parts: the answer passage clustering task and diversification. Figure 1 illustrates the high-level architecture flow, where the answer passage similarity clusters are used as input to a diversified model to generate a re-ranked list of answers. The answer passage clustering task setting is similar to the ones used in cluster based IR models [16–19, 27, 36]. Given a question q and a passage collection C, a standard retrieval model can be used to retrieve a list of n passages $\mathcal{P}$. A cluster is generated for each of these passages $P_i \in \mathcal{P}$, with respect to every other passage $P_j : P_i \neq P_j$ and select the most similar passages. An effective clustering method must be able to perform the following:

- Cluster relevant answer passages together
- Cluster relevant answer passages of the same type together

To generate the clusters, we determine the nearest neighbors corresponding to each passage and rank them based on the similarity score.

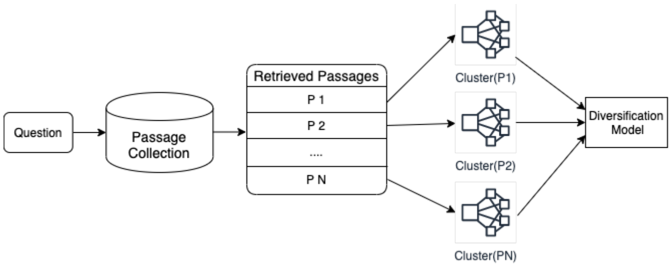

Figure 1: Passage Similarity and Diversity pipeline

This is kNN clustering similar to settings in [16–19, 27, 36], which has been shown to outperform other approaches such as k-means and our preliminary experiments confirm this. The generated passage clusters are then input into a diversified model where an initial ranked list R generated for a question q is diversified based on the clusters.

4 DATA

In this section, we first describe the annotation task for collecting the passage similarity dataset, NFPassageQA_Sim[1]. Next, we define how these similarity annotations were processed, to create a corresponding clustering dataset NFPassageQA_Div[1], with various answer types (or subtopics) for each question.

4.1 NFPassageQA_Sim dataset

Annotation Task Definition: The definition of similarity is vague and can have multiple interpretations. For example, the two passages "`Diabetes can be managed by good diet and exercise`" and "`Diabetes is a hereditary disease which affects about 5% of the population`" refer to "`diabetes`" and can be interpreted as similar. However, they would be considered dissimilar given a question "`How to treat diabetes?`", since the second passage doesn't answer the question. To eliminate such ambiguities and to ground the annotation process, we only include relevant passages for annotation. A formal definition of the task is given below.

Given a question q and a corresponding set of relevant answer passages $\mathcal{P}$, the data annotation task involves assigning a similarity label to passage pairs (P_i, P_j), where $P_i \in \mathcal{P}$, $P_j \in \mathcal{P} : P_i \neq P_j$. Here we assume the relationship to be symmetric i.e., $sim(P_i, P_j) = sim(P_j, P_i)$.

Table 1: NFPassageQA_Sim dataset statistics

Num Questions	128
Total Num of Triples	18,216
Num Label 4	244 (1.34%)
Num Label 3	4,462 (24.49%)
Num Label 2	12,650 (69.44%)
Num Label 1	860 (4.72%)
AvgLen Questions	9.4
AvgLen Passages	61

Data Annotation: As the input to the annotation process, we used the questions from the test collection of the publicly available ANTIQUE [13] dataset. A subset of 128 questions, which contains at least 10 relevant answers from the ANTIQUE dataset with labels {3,4} was filtered from the test collection. For each question q with n relevant answers, we create a set with m items $\mathcal{I} = \{(P_i^1, P_j^1), .., (P_i^m, P_j^m)\}$, where $m = \frac{n(n-1)}{2}$, consisting of all possible relevant answer pairs.

[1] https://ciir.cs.umass.edu/downloads/NFPassageQA

We employed workers from the Amazon Mechanical Turk (MTurk)[2] platform to perform the annotation. The workers were required to have a HIT (Human Intelligence Task) approval rate of 98% or higher, a minimum of 10000 approved HITs and be located in US, Canada, Australia or Great Britain. They were paid \$0.13 per HIT. Each input triple consisting of a question and a corresponding answer pair were assigned to three different workers. Detailed labeling instructions with examples were also provided to aid them with the task. After reading through instructions, they assign a similarity label 0~4 to the triple as illustrated in Table 2. The data collection was performed in 7 batches. The label with a majority agreement among the workers was chosen as the ground truth. For cases with no majority agreement or with a majority label of 0, another round of annotation was performed to break the tie. We perform a set of filtering steps to remove instances that do not have sufficient agreement among the workers. First, we discard the instances where the ground truth could not be determined even after the second round of annotation. This also includes cases with a majority label of 0. Next, for those instances where we obtain a majority label, we remove cases where there is no majority agreement in terms of overall similarity. For example, an instance with votes [1,2,3,3] has a majority label agreement for label 3 but does not have agreement based on overall similarity (which is 2:2, since labels [1,2] indicate dissimilarity and [3,3] similarity). The filtering brings down the overall number of instances from 18629 to 18314. The final statistics of the dataset are shown in Table 1.

To ensure annotation quality, we added test triples with highly objective labels into each batch. This helps us identify workers who randomly click on labels without reading the instructions. We also conducted manual checks on 10% of the data to determine the quality. After identifying and rejecting around 8% of spurious data in the initial batches, we established a fully closed qualification restricted to around 70 workers in the subsequent batches. A Label 0 ("Not Sure") was also added to discourage workers from assigning a random label when unsure about the answer, especially due to a lack of domain knowledge.

Discussion: The descriptions of the labels with examples is illustrated in Table 2. Labels 3, 4 indicate high similarity, indicating answers belonging to the same type, while label 2 indicates the passages belong to different answer types. Label 1 was added to capture any non-relevant passages, incorrectly labeled as relevant. Label 0 was added to reduce annotation noise and was removed from the final set of judgements.

Table 1 reports the final data statistics. Around 70% of the annotations correspond to Label 2, while Labels 3 and 4 cover around 26%. The high percentage of Label 2 is not surprising, considering the nature of the data used for annotation. The questions were extracted from a CQA discussion forum, where users tend to give alternate answers to the questions. Predictably, Label 4 occurs very infrequently in the dataset. We also performed manual checks to confirm that the answers were not being mislabeled as non-relevant (Label 1), since it has a relatively high coverage (5%).

4.2 NFPassageQA_Div dataset

Dataset Creation Definition: The passage similarity annotations provide us with similarity values pertaining to all pairs of relevant passages corresponding to each question. This information can then be used to generate answer types (or subtopics) for these questions. A formal definition is given below:

Given a question q, a corresponding set of relevant answer passages $\mathcal{P}$ and a set of similarity annotations between passage pairs $sim(P_i, P_j)$, where $P_i \in \mathcal{P}$ and $P_j \in \mathcal{P}$, the dataset creation task involves automatically identifying the various answer types (or subtopics) $\mathcal{T}$ and assigning passages in $\mathcal{P}$ to them.

Dataset Construction: Since the similarity annotations contain a relatively high ratio (5%) of non-relevant pairs, the first step is the identification of non-relevant passages, which can then be removed to reduce noise. All passages appearing in more than 40% of passage pairs, corresponding to each question and annotated with Label 1 is marked as non-relevant. 596 passage pairs with these newly identified non-relevant passages are then removed. This process was also manually cross-checked to ensure that no relevant passages were discarded. Due to the change in number of relevant passages per question, we retain only questions with at least 10 relevant passages remaining after the previous step, which reduces the number of questions to 93. Based on the similarity values (Labels 3 and 4), we next construct all possible passage combinations and identify the longest non-overlapping passage clusters for each question. These are considered to be the answer types (or subtopics). Each of the remaining passages is then added to the answer type if at least one of the passages in the cluster is similar to this passage.

Relevance judgements are then assigned to each passage with respect to each answer type. The original passages within each non-overlapping cluster (representing a unique answer type) and other passages subsequently added based on partial similarity to original cluster elements are assigned a relevance value of 1 indicating `relevance`. All the other passages are considered `non-relevant` with value 0.

Discussion: An example of a question and the two answer types generated from it is given in Table 3. The passages have been shortened due to the space restrictions. Due to the nature of the dataset, a passage can belong to multiple answer types. For instance, the last passage in both answer types in the Table 3 corresponds to both `Traps` and `Cats`. The dataset consists of 93 questions and Table 4 gives the distribution of answer types corresponding to the questions. For example, 32 questions have 2 answer types as indicated in the table.

5 WEAK SUPERVISION

In this section we describe how using weak labels generated from different sources can be used to train deep learning models. Such models have been demonstrated to work well for core IR tasks such as document ranking [11]. However, they require millions of weak labels to learn the ranking function. Xu et al. [39] showed that large pre-trained models such as BERT [12], fine-tuned with small number of weak labels aggregated from different sources can be used for passage ranking. The fine-tuning process then forces the model to learn task specific relationships from these weakly labeled data.

[2] https://www.mturk.com/

Table 2: Label Descriptions

Label Type	Label	Description	Example
Similar	4	Both passages answer the question Both passages contain the same information, however they maybe worded differently	Question: What do you mean by weed? Passage 1: Weed could mean the bad thing that grow in the garden or back and front yard or it could mean the drug Passage 2: It could mean weeds outside on the lawn or the drug
	3	Both passages answer the question. The passages belong to the same answer type. They may also contain information associated with a different type or other non-relevant information	Question: What do you mean by weed? Passage 1: Weed could mean the bad thing that grow in the garden or back and front yard or it could mean the drug Passage 2: Marijuana and lots of it
Dissimilar	2	Both passages answer the question. The passages belong to different answer types	Question: How can i get a cork out of,not into a wine bottle without a corkscrew? Passage 1: Use a screwdriver to put a wood screw into it, then pull the wood screw out with a pair of pliers, better yet get a $1 corkscrew Passage 2: If you have a syringe you can push it through the cork to the inside of the bottle press the air into the bottle and the pressure inside will force the cork out
	1	At least one of the passages does not answer the question	Question: How to cook Angus Burger? Passage 1: I usually cook burgers until they quit bleeding on both sides, then maybe just a little longer the cooking time will vary depending on the thickness of the burger. Passage 2: I'm not sure what the difference is other than the difference between a houstine and a angus but just because a bull is castrated doesnt make him an ox it just makes him a steer.
Not Sure	0	Not sure about the answer	N/A

Table 3: Example Answer Types

Question	How do I get rid of mice humanely?
Use traps *(Answer Type 1)*	Home Depot sells live traps.... Put down a humane trap.... Get a cat or use mouse traps...
Use natural predators such as cats. *(Answer Type 2)*	Just get or borrow a friend's cat... invite my cat over,she is a great mouser! Get a cat or use mouse traps...

Table 4: NFPassageQA_Div Answer Type Distribution

#Types	1	2	3	4	5
#Questions	39 (41.9%)	32 (34.4%)	15 (16.1%)	4 (0.04%)	3 (0.03%)

We adopt this weak supervision strategy and generate pseudo-labels to learn an improved similarity function and semantic vector representation for passage clustering, using the BERT [12] model. The following sub-sections describe the pseudo-labeling sources as well as the process of generating them from these sources, along with the model architecture used to train these models.

Pseudo-Labelers: Formally, the pseudo-labeling process can be defined as follows. Start with a question q and create a ranked list of answer passages $\mathcal{P}$. Then, for a passage $P_i \in \mathcal{P}$ with $rank_{QL}(q, P_i) \leq 10$, the weak labeling process generates a list of n passages, ranked based on the similarity score, $\mathcal{R} = (P_j^1, P_j^2, ...P_j^n)$ with $P_j^k \in \mathcal{P}$ and $\forall P_j^k, P_j^k \neq P_i$. For our experiments, the list $\mathcal{P}$ is created using the Query Likelihood (QL) model and the top $n = 200$ passages are clustered, both reflecting the typical setting for cluster based models [27, 31] and so the convention we use for our experiments. We identify three weak labeling functions based on the different text representations used to encode the passages and the questions as given below.

Language Model with Dirichlet smoothing (LM): Language Model is selected as the term frequency based representation. An alternative representation is tf-idf, however we choose LM since it has been demonstrated to perform better in cluster based settings [17]. Given a passage P, the language model can be estimated using maximum likelihood estimation: $Pr_P^{MLE} = \frac{c(w,P)}{|P|}$ where $c(w, P)$ is the count of word w in passage P. Dirichlet smoothing [45] can be applied to interpolate this estimate with collection (C), $Pr_P^{Dir(\mu)} = \frac{c(w,P)+\mu Pr(w|C)}{|P|+\mu}$. These values are calculated at passage (P) level. For this model, the similarity score, $sim(P_i, P_j)$ between two passages P_i and P_j is calculated by using cross-entropy ($\mathcal{H}$) [27, 31] similarity between the Maximum Likelihood (${Pr_P}_i^{MLE}$) and Dirichlet ($Pr_{P_j}^{Dir(\mu)}$) estimates : $sim(P_i, P_j) = \exp(-\mathcal{H}(Pr_{P_i}^{MLE}, Pr_{P_j}^{Dir(\mu)}))$.

Global Vectors for Word Representation (GloVe): Distributional models such as GloVe [25] incorporate global term co-occurrence counts along with the local contextual information, to create representations which capture semantic relationships along with term statistics. We first generate passage representations by combining the vectors using idf-weighting of terms present in the query as well as the passage. Query vectors are also used, since this adds contextual information necessary for clustering and was found to perform better. We use Euclidean distance as the scoring function.

Bidirectional Transformers for Language Understanding (P-BERT): Instead of focusing on term statistics or local word contexts, longer sequence context information can be used to model representations at sentence/passage level. BERT [12] representations capture sentence/passage level information by conditioning on both left and right contexts across all the layers of a deep neural model. For each token, BERT generates an embedding using position, segment and token embeddings. BERT pre-training is performed using two unsupervised tasks: Masked Language Modeling and Next Sentence Prediction. The vector representations corresponding to a passage/sentence can be generated by giving two inputs : query q and passage P. Similar to generating GloVe representations, we use query terms in addition to passage information to provide more context. We use the [CLS] token embedding as the representation corresponding to the input sequences. The scoring function is the Euclidean distance.

Model Architecture The BERT [12] model is used as the framework for the weak supervision experiments. For an input question q and passage pair (P_i,P_j), the model must learn a similarity function and output a score for the triple. The training for this task is performed by feeding the inputs to BERT and fine-tuning the model based on two different loss functions.

Point-wise model: The point-wise loss function is the default cross entropy loss used for the sentence pair classification experiments in BERT [12]. The two inputs to the model are $(q + P_i)$ and $(q + P_j)$ where "+" indicates that question terms have been concatenated with the corresponding passage terms. For a triple (q, P_i, P_j) with $\hat{s}(q, P_i, P_j; \theta)$ as the scoring function learned by the model under the parameters θ and $s(q, P_i, P_j)$, the ground-truth generated by the weak labeler, the training loss can be defined as follows :

$$\mathcal{L}(q, P_i, P_j; \theta) = s(q, P_i, P_j) \log \hat{s}(q, P_i, P_j; \theta) \tag{1}$$

Pair-wise model: The pair-wise model is similar to the Rank model [11] and the passage ranking model described by Xu et al. [39]. The loss function employed is the pair-wise hinge loss function. For an input pair, $[(q, P_i, P_j), (q, P_i, P_j^{'})]$ with point-wise scoring functions $\hat{s}(q, P_i, P_j; \theta)$ and $\hat{s}(q, P_i, P_j^{'}; \theta)$, and the weak labels $s(q, P_i, P_j)$, and $s(q, P_i, P_j^{'})$, the model is trained to minimize the hinge loss as follows:

$$\begin{aligned}\mathcal{L}(q, P_i, P_j, P_j^{'}; \theta) = \max\{0, \epsilon - sign(s(q, P_i, P_j) - s(q, P_i, P_j^{'})) \\ (\hat{s}(q, P_i, P_j; \theta) - \hat{s}(q, P_i, P_j^{'}; \theta))\}\end{aligned} \tag{2}$$

The point-wise model[3] used in this case is different from the default BERT model. The inputs to the model are same as the default version, $(q{+}P_i)$ and $(q{+}P_j)$ where "+" indicates that question terms have been concatenated with the corresponding passage terms. The BERT scoring model generates hidden states for the [CLS] token for the input and the final hidden layer is fed into a dense layer. We consider two variants of the model, one with a linear output activation (Pair-wise Linear) and the other with tanh activation (Pair-wise tanh). During test time, the corresponding point-wise scores are used to generate the similarity scores.

6 ANSWER PASSAGE DIVERSIFICATION

Answer passage clustering models help in grouping similar passages together. However, the final aim is to be able to display different answer types to the users. Diversification models can capture this information, since they combine relevance and diversity during re-ranking. In this paper, we use an extension of the Implicit Diversification model MMR (Maximal Marginal Relevance) [3] to diversify the answers. Given a question q, an initial ranked list R of answer passages generated using a standard retrieval model $R = \{p_1, p_2, ...p_n\}$, S representing the ranked list of diversified answers, $sim(p_i, p_j)$ the similarity score between passages p_i and p_j, $Clus_m(p)$ the m most similar passages to passage p, p^* the answer passage selected at each step of ranking, the standard MMR model is defined as follows:

$$p^* = \underset{p_i \in (R-S)}{\text{argmax}}\,(1-\delta) rel(p_i, q) + \delta \max_{p_j \in S} sim(p_i, p_j) \tag{3}$$

We modify this for cases where the passage p_j in S is ranked within the top 10 of the ranked list R (i.e highly relevant to the query). Here, the diversity component $\max_{p_j \in S} sim(p_i, p_j)$ is replaced by $\max_{p_j \in S} \max_{p_k \in Clus_m(p_j)} sim(p_i, p_k)$.

For these cases, instead of finding the maximum similarity between an element in R and passage p_j in S (as in (3)), we consider the maximum similarity with top m most similar cluster elements with respect to p_j. We only expand passages within S which are highly relevant to the query to limit the noise which could be introduced by the non-relevant passages. We also experimented with other settings such as using cluster elements in R and found this setting to be the best. We will call this `MMR Cluster` to distinguish it from the other variants.

7 EXPERIMENTAL SETUP

Data Overview: The evaluation of the answer passage similarity experiments is conducted using the newly collected NFPassageQA_Sim dataset with 128 questions. For the weak supervision experiments, we used the ANTIQUE dataset collection. 200 questions were randomly sampled from the training set to create a validation set and the remaining 2226 questions were used for training. In order to evaluate the output generated by diversified model, we used the newly generated dataset NFPassageQA_Div with 93 questions.

Weak Supervision Training and Test Setup: For training, the pseudo-labeling process described in Section 5.2 is used to generate a ranked list of passages for each (q, P_i) pair. Instead of adding the set of all P_j to training data, we add only the top 10 passages. To

Table 5: Weak Supervision Experimental settings

#Train questions	# Train point-wise instances	# Train pair-wise instances	#Test questions
2226	222600	400000	128

[3] References to point-wise model throughout the rest of the paper indicate the default BERT model

create weak labels for point-wise models, the top 5 passages from the ranked list are labeled as positive (1) and the next 5 passages are labeled as negative (0). For the pair-wise models, we need a pair of passages from the ranked list to create the training instances. These pairs are generated using a sliding window method. For each passage in the ranked list, the next 5 passages below it in the ranked list are considered to have lower scores and added as training data. From these generated instances, we randomly sample a subset for our experiments. At test time, we follow the same initial process described in Section 5.2. The point-wise scores are generated for each test triple (q,P_i,P_j) and the clustering is performed based on these scores. We perform clustering over all passages P_j – i.e., a set of 200 passages – for each $|P_i|$ with $rank(P_i) \leq 10$ (same as the settings in Section 5.2). The baseline methods also follow the same convention for correct comparison. The experimental settings are summarized in Table 5.

Diversification: The initial retrieval run is obtained using the Query Likelihood model. The diversity re-ranking is performed over the top 100 retrieved answers. We consider a number of standard baselines and compare against them.

- **Query Likelihood (QL):** This is the initial retrieval run generated using default Dirichlet prior smoothing (μ=2500).
- **MMR :** This is the classic MMR implementation [3], which uses a greedy implicit diversification algorithm to generate a diversified output. We use two different versions of this as baseline : MMR Sparse and MMR P-BERT. MMR Sparse is the classic IR approach using a sparse vector representation for terms, where the different dimensions contain term frequency information. MMR P-BERT uses the BERT representation ([CLS]) for the passages.
- **Term Level Diversification:** This was introduced by Dang et al. [8] and is an explicit diversification model, where a set of topic terms are first generated by an algorithm called DSPApprox which is then used in combination with xQUAD and PM-2 algorithms to generate a final diversified list. We only use xQUAD as the baseline since xQUAD consistently outperformed the PM-2 model for this dataset. This is also consistent with the findings in [33].

Implementation Details: The Language Model experiment settings are similar to cluster based retrieval [31]. 3-fold cross-validation was performed to set μ parameter for test questions. The default parameter value of $\mu = 10$ was set for the train questions during the LM pseudo-labeling process. $300d$ pre-trained GloVe [25] vectors[4] are used for the GloVe experiments. For the BERT [12] pre-trained experiments, Layer1 hidden vectors of the [CLS] output generated using BERT-Base (Uncased) pre-trained model[5] are used. For the weak supervision experiments, the models trained on BERT are implemented using TensorFlow[6] and fine-tuned after initializing with the BERT-Base (Uncased) pre-trained model. The maximum sequence length is set to 128, with each input truncated to length 64. The batch size is set to 20. The initial learning rate was selected from $[1e^{-5}, 2e^{-5}, 3e^{-5}]$ by tuning on the validation set. The dropout parameter is set to 0.1. The experiments were conducted on a single GeForce GTX 1080 GPU. The train time for the point-wise model was around 1 hour and pair-wise models took around 2 hours for training. The inference time was around 50 minutes for both model types. The k-nearest neighbor (kNN) clustering was conducted using the kDTree algorithm in the sklearn toolkit. The parameter k is set to 200. The parameters for the diversity baselines are set by cross-validation. δ value for the `MMR Cluster` approach is set to 0.5 and parameter m is set to 40 for WS GloVe model and 60 for P-BERT to reflect the best performance in each case. In order to maintain consistency with our approach, topic terms for the term-level diversification baseline are generated from top 200 retrieved set, same as the setting for answer passage clustering.

Evaluation: To evaluate the passage similarity models effectively, ranking metrics (instead of default clustering metrics) are used to determine if the model retrieves relevant and similar passages at higher ranks. Precision@k and Recall@k, with k=10,20 are employed for assessing this. For each test question, the metric value is calculated for each relevant passage and then averaged over all of them. The values returned for all the questions are then averaged to generate the final evaluation score. To evaluate diversification models, standard diversity metrics such as Precision-IA@k, S-Recall@k and α-NDCG@k are used. We also measure relevance values using Precision@k, Recall@k and NDCG@k metrics. For the metrics to evaluate diversification models, we set k=10. Statistical significance is measured using the paired two-tailed t-test with p-value < 0.05.

8 RESULTS AND ANALYSIS

Table 6 reports the results for the passage clustering task using weakly supervised labels. The column"Sim Clusters" refers to clustering relevant passages of the same type and "Rel Clusters" refers to clustering relevant passages together. Pseudo-labels derived using GloVe perform the best, with the pointwise and pairwise models significantly improving over the corresponding baseline. The best performing GloVe model also significantly outperforms both LM and BERT baselines, which shows that the weak labels from GloVe combines well with BERT when fine-tuned.

Table 7 shows the results on various diversity models on the NFPassageQA_Div dataset. The Term Level Diversification model using xQUAD is a competitive baseline outperforming both the QL baseline and MMR with standard BERT representation. In general, using clusters from unsupervised BERT representations and weak supervision with MMR performs well across all diversity and relevance metrics. The cluster generated using GloVe weak labels (WS GloVe) performs the best, significantly outperforming all the baselines.

Impact of using clustering for diversity : We first investigate how clustering helps in retrieving higher quality answers for the question answering system. To this end, we first analyze the Win/Tie/Loss statistics for the top performing baseline and clustering model with respect to S-Recall as given in Table 8. Since the dataset consists of questions with a single and multiple answer types , we measure this for two cases - all questions and questions with multiple answer types (>1). As S-Recall is a metric which measures how well a system discovers new subtopics, this would be relevant only for the second type. As seen in the table, the clustering technique retrieves significantly more answer types than the baseline. We also investigate how clustering contributes to these improvements and found two main reasons for this : (a) In 6 out of

[4] https://nlp.stanford.edu/projects/glove/ [5] https://github.com/google-research/bert
[6] https://www.tensorflow.org/

Table 6: Results on NFPassageQA_Sim dataset for clustering relevant passages (Rel Clusters) and clustering relevant passages of the same type (Sim Clusters) for the three main pseudo-labelers. † indicates significance with respect to corresponding baselines. Statistical significance is measured using the paired two-tailed t-test with p-value<0.05. P-BERT refers to pre-trained BERT. The scores for the best performing trained model with respect to each baseline has been marked in bold.

Pseudo-labeler	Model	Rel Clusters				Sim Clusters			
		P@10	P@20	R@10	R@20	P@10	P@20	R@10	R@20
LM	Baseline	0.1057	0.0917	0.0960	0.1780	0.0523	0.0418	0.0942	0.1537
	Point-wise	0.0706	0.0624	0.0770	0.1478	0.0380	0.0287	0.0660	0.1041
	Pair-wise Linear	0.0803	0.0739	0.0813	0.1575	**0.0427**	**0.0346**	**0.0779**	0.1225
	Pair-wise tanh	**0.0871**	**0.0794**	**0.0937**	**0.1811**	0.0420	0.0335	0.0738	**0.1321**
GloVe	Baseline	0.1478	0.1114	0.1558	0.2468	0.0663	0.0460	0.1384	0.1813
	Point-wise	**0.2087**†	**0.1663**†	**0.2403**†	**0.4070**†	**0.0964**†	**0.0686**†	**0.1887**†	**0.2736**†
	Pair-wise Linear	0.2002†	0.1632†	0.2192†	0.3722†	0.0924†	0.0674†	0.1799†	0.2617†
	Pair-wise tanh	0.1982†	0.1631†	0.2204†	0.3895†	0.0935†	0.0667†	0.1762†	0.2506†
P-BERT	Baseline	0.2279	0.1745	0.2720	0.4580	0.0838	0.0626	0.1662	0.2483
	Point-wise	0.1874	0.1457	**0.1983**	0.3277	0.0815	0.0549	0.1585	0.2118
	Pair-wise Linear	0.1928	0.1479	0.1952	**0.3284**	**0.0855**	**0.0594**	**0.1661**	**0.2399**
	Pair-wise tanh	**0.1932**	**0.1498**	0.1868	0.3224	0.0795	0.0578	0.1595	0.2215

Table 7: Results on NFPassageQA_Div dataset for different diversification methods. Q, S, T, B indicates significance with respect to the baselines QL, MMR Sparse, TLD and MMR P-BERT respectively. Here TLD refers to Term level Diversification [8]. P-BERT refer to the unsupervised models while WS GloVe refer to the weak supervision model trained with GloVe signals. Statistical significance is measured using the paired two-tailed t-test with p-value<0.05. The scores for the best performing model has been marked in bold.

Type	Model	Diversity			Relevance		
		Prec-IA	S-Recall	α-NDCG	Prec	Recall	NDCG
Baselines	QL	0.2104	0.6905	0.4671	0.3182	0.1043	0.3435
	MMR Sparse	0.0711	0.4743	0.2996	0.1290	0.0404	0.1807
	TLD xQUAD	0.2166	0.7057	0.4705	0.3301	0.1079	0.3503
	MMR P-BERT	0.1732	0.6118	0.4083	0.2763	0.0891	0.2996
Cluster	`MMR Cluster` P-BERT	0.2290^{Q,S,T,B}	0.7317^{Q,S,B}	0.4867^{Q,S,T,B}	0.3451^{Q,S,T,B}	0.1130^{Q,S,B}	0.3646^{Q,S,T,B}
	`MMR Cluster` WS GloVe	**0.2344**^{Q,S,T,B}	**0.7530**^{Q,S,T,B}	**0.4939**^{Q,S,T,B}	**0.3569**^{Q,S,T,B}	**0.1162**^{Q,S,T,B}	**0.3723**^{Q,S,T,B}

Table 8: Win/Tie/Loss statistics for models compared with the QL baseline with respect to various metrics.

Metric	Models	W/T/L All Questions	W/T/L Multi-Answer Questions
S-Recall	TLD xQUAD	3/87/3	2/49/3
	MMR Cluster WS GloVe	**11/80/2**	**9/43/2**
Prec-IA	TLD xQUAD	11/76/6	6/44/4
	MMR Cluster WS GloVe	**30/59/4**	**17/34/3**
α-NDCG	TLD xQUAD	25/45/23	16/24/14
	MMR Cluster WS GloVe	**44/37/12**	**28/18/8**

9 cases, it was found that the improvement in answer type S-Recall was caused by the presence of the currently selected passage (using MMR) within the cluster of an already retrieved relevant passage in set S. For example, for the query "`How do you prevent chicken from drying out when you cook it?`" with 5 answer types, an already relevant selected passage in set S, containing 2 answer types "`sprinkle water, wrap in foil`" has another answer with answer type "`coat chicken and fry in oil`" in its cluster and the similarity score for this would be higher than other passages and is hence retrieved. This demonstrates how "Rel Clusters" or clustering relevant passages correlates with improvement in this metric. (b) In the remaining 3 out of 9 cases, we found that non-relevant passage was responsible for retrieving relevant passage due to its presence in its cluster. These passages though non-relevant, had some contextual similarity to the expected answers. For example, for the query "`How do I get rid of mice humanely?`", a non-relevant passage "`Ask them to leave politely`", had a relevant answer with answer type "`Use live traps`" within its cluster, which was subsequently retrieved by the algorithm.

We also studied the behavior with respect to the Precision-IA metric, which measures the average number of relevant passages retrieved for each answer type. This metric would be pertinent for both cases where questions have a single answer type and for those with multiple answer types. We found the behaviour similar to that of S-Recall. Out of the 30 questions which improved compared to the QL baseline as given in Table 8, the gains for 22 of these can be attributed to their presence in the clusters of relevant passages present in set S. This demonstrated how "Sim Clusters" or clustering relevant passages of the same type correlated with improvements in the Prec-IA metric. The improvements in the remaining 7 questions, are due to the presence of relevant passages in clusters of non-relevant passages same as in the S-Recall metric.

The combination of retrieving more answer types and relevant passages for each answer type contributes to the improvement in α-NDCG. Significant gains in α-NDCG metric also demonstrates that more relevant passages are ranked higher. This also directly correlates with the improvement in various relevance metrics such as NDCG, Precision and Recall as shown in Table 7.

Impact of size of the cluster: In order to get the best performance from the diversity model, we need to pick the right number of top similar elements (m) or cluster size. To that end, we study how α-NDCG value of the models changes with increase in m. As shown in Figure 2, we plot two different clusters generated using P-BERT representations and weakly labeled GloVe with BERT against different values of $m = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. For both cases, we observe that the efficacy of the model decreases after a threshold. This was also observed with S-Recall and Precision-IA metrics. This behavior can be attributed to the additional noise introduced by less similar passages, which is added as we increase the cluster size.

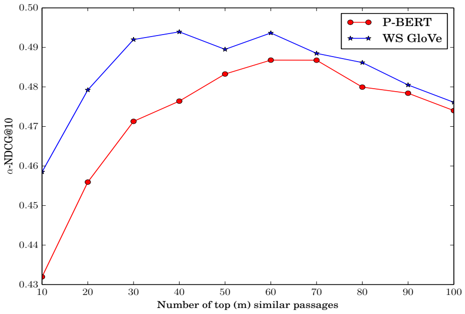

Figure 2: α-NDCG across Cluster size

Performance Comparison with Term Level Diversification baseline: We qualitatively compared the answers retrieved by the Term Level Diversification baseline, which is a high-performing model for document diversity and the cluster based MMR models and observed that the terms used for diversification in the term based model is insufficient to differentiate between non-relevant and relevant answers. For example, for the query “`How to get rid of warts?`, some of the top terms used by the term based model are “`remove, try, tape, work, freeze`”. While some of these terms do refer to methods for wart removal such as “using tape” or “freeze”, this also retrieved other non-relevant passages with the same terms. This issue is mitigated to a large extent in cluster based models due to the contextual information captured by them.

Performance comparison between different clustering models: In order to compare the performance of various clustering models in combination with MMR, we performed the experiments with clusters of size $m = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ and reported the maximum value in Table 9. The weakly supervised model performed the best amongst all the models with the unsupervised GloVe model performing the worst. This shows the efficacy of using weak supervised clusters as opposed to clusters generated using unsupervised representations.

Table 9: α-NDCG metric comparison for MMR models using different clustering models

Clustering Models	α-NDCG
MMR Cluster Glove	0.4667
MMR Cluster P-BERT	0.4867
MMR Cluster WS Glove	**0.4939**

9 CONCLUSION AND FUTURE WORK

Passage clustering is an essential component of a question answering system aimed at finding multiple answers to questions. In this paper, we show empirically how passage clustering models can be used in combination with diversification models to retrieve different answer types. We describe the creation of a passage similarity based dataset: NFPassageQA_Sim and the automatic generation of the diversity based dataset: NFPassageQA_Div. We also propose a weak supervision method to tackle the task of answer passage clustering. Since weak supervision models are expected to capture additional information compared to unsupervised representations, we use various pseudo-labels generated using the unsupervised representations described earlier and fine-tune a BERT model for this task. We found that a BERT pointwise model trained using GloVe pseudo-labels to be the most effective for this task. Since the end task is displaying the various answer types, we employ a diversification approach to create a re-ranked list. We propose a modified MMR approach, which uses the similarity between cluster elements while greedily selecting the next passage. We demonstrate that expanding the answer set using these clusters results in significant improvements across various diversity and relevance metrics in comparison with standard diversification baselines. The results suggest that this greedy approach finds more passages which are similar to existing relevant passages and also results in an overall increase in the number of answer types.

Several different avenues are open to extension from this work. This paper gives initial experiments which show how similarity based clusters can be incorporated into a diversified model. A more natural extension of this work would be creating a single model which has an objective function optimizing similarity and diversity jointly. This could also be included as part of a cluster based model to display answers under a different setting. This type of model could also be used in combination with a summarization model to display different answer types.

ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Chris Alberti, Kenton Lee, and Michael Collins. 2019. A bert baseline for the natural questions. *arXiv preprint arXiv:1901.08634* (2019).

[2] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

[3] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*. ACM, 335–336.

[4] Charles L Clarke, Nick Craswell, and Ian Soboroff. 2009. *Overview of the trec 2009 web track.* Technical Report. WATERLOO UNIV (ONTARIO).

[5] Daniel Cohen and W Bruce Croft. 2016. End to end long short term memory networks for non-factoid question answering. In *ICTIR*. ACM, 143–146.

[6] Daniel Cohen and W Bruce Croft. 2018. A Hybrid Embedding Approach to Noisy Answer Passage Retrieval. In *ECIR*. Springer, 127–140.

[7] Daniel Cohen, Scott M Jordan, and W Bruce Croft. 2019. Learning a Better Negative Sampling Policy with Deep Neural Networks for Search. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval.* 19–26.

[8] Van Dang and Bruce W Croft. 2013. Term level search result diversification. In *SIGIR*. ACM, 603–612.

[9] Van Dang and W Bruce Croft. 2012. Diversity by proportionality: an election-based approach to search result diversification. In *SIGIR*. ACM, 65–74.

[10] Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe, and Jaap Kamps. 2017. Learning to learn from weak supervision by full supervision. *arXiv preprint arXiv:1711.11383* (2017).

[11] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 65–74.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. 2019. ANTIQUE: A non-factoid question answering benchmark. *arXiv preprint arXiv:1905.08957* (2019).

[14] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. 2015. Search result diversification based on hierarchical intents. In *CIKM*. ACM, 63–72.

[15] Nick Jardine and Cornelis Joost van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information storage and retrieval* 7, 5 (1971), 217–240.

[16] Oren Kurland and Eyal Krikon. 2011. The opposite of smoothing: a language model approach to ranking query-specific document clusters. *JAIR* 41 (2011), 367–395.

[17] Oren Kurland and Lillian Lee. 2004. Corpus structure, language models, and ad hoc information retrieval. In *SIGIR*. ACM, 194–201.

[18] Xiaoyong Liu and W Bruce Croft. 2004. Cluster-based retrieval using language models. In *SIGIR*. ACM, 186–193.

[19] Xiaoyong Liu and W. Bruce Croft. 2007. Evaluating Text Representations for Retrieval of the Best Group of Documents. In *ECIR*. 454–462.

[20] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. *arXiv preprint arXiv:1901.11504* (2019).

[21] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[22] Adi Omari, David Carmel, Oleg Rokhlenko, and Idan Szpektor. 2016. Novelty based ranking of human answers for community questions. In *SIGIR*. ACM, 215–224.

[23] Harshith Padigela, Hamed Zamani, and W Bruce Croft. 2019. Investigating the successes and failures of BERT for passage re-ranking. *arXiv preprint arXiv:1905.01758* (2019).

[24] Ellie Pavlick and Chris Callison-Burch. 2016. Simple PPDB: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* 143–148.

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[26] Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088* (2018).

[27] Fiana Raiber and Oren Kurland. 2013. Ranking document clusters using markov random fields. In *SIGIR*. ACM, 333–342.

[28] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and R Christopher. 2016. Data Programming: Creating Large Training Sets. *Quickly. arXiv [stat. ML]* (2016).

[29] Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *WWW*. ACM, 881–890.

[30] Sheikh Muhammad Sarwar, Raghavendra Addanki, Ali Montazeralghaem, Soumyabrata Pal, and James Allan. 2020. Search Result Diversification with Guarantee of Topic Proportionality. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval.* 53–60.

[31] Eilon Sheetrit, Anna Shtok, Oren Kurland, and Igal Shprincis. 2018. Testing the Cluster Hypothesis with Focused and Graded Relevance Judgments. In *SIGIR*. ACM, 1173–1176.

[32] Ian M Soboroff, Nick Craswell, Charles L Clarke, and Gordon Cormack. 2011. *Overview of the trec 2011 web track.* Technical Report.

[33] Lakshmi Vikraman, W Bruce Croft, and Brendan O'Connor. 2018. Exploring Diversification In Non-factoid Question Answering. In *ICTIR*. ACM, 223–226.

[34] Di Wang and Eric Nyberg. 2015. A long short-term memory model for answer sentence selection in question answering. In *ACL*, Vol. 2. 707–712.

[35] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167* (2019).

[36] Xing Wei and W Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 178–185.

[37] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* Association for Computational Linguistics, 1112–1122. http://aclweb.org/anthology/N18-1101

[38] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2015. Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. In *SIGIR*. ACM, 113–122.

[39] Peng Xu, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Passage Ranking with Weak Supervsion. *arXiv preprint arXiv:1905.05910* (2019).

[40] Liu Yang, Qingyao Ai, Damiano Spina, Ruey-Cheng Chen, Liang Pang, W Bruce Croft, Jiafeng Guo, and Falk Scholer. 2016. Beyond factoid QA: effective methods for non-factoid answer sentence retrieval. In *ECIR*. Springer, 115–128.

[41] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718* (2019).

[42] Xing Yi and James Allan. 2009. A comparative study of utilizing topic models for information retrieval. In *ECIR*. Springer, 29–41.

[43] Hamed Zamani and W. Bruce Croft. 2018. On the Theory of Weak Supervision for Information Retrieval. In *Proceedings of the Fourth International Conference on the Theory of Information Retrieval (ICTIR '18).* 147 – 154.

[44] Hamed Zamani, W Bruce Croft, and J Shane Culpepper. 2018. Neural query performance prediction using weak supervision from multiple signals. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.* 105–114.

[45] Chengxiang Zhai and John Lafferty. 2017. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum*, Vol. 51. ACM, 268–276.

[46] Yadong Zhu, Yanyan Lan, Jiafeng Guo, Xueqi Cheng, and Shuzi Niu. 2014. Learning for search result diversification. In *SIGIR*. ACM, 293–302.