Paragraph Acquisition and Selection for List Question Using Amazon’s Mechanical Turk

Fang Xu and Dietrich Klakow

Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
{fang.xu, dietrich.klakow}@lsv.uni-saarland.de

Abstract
Creating more fine-grained annotated data than previously relevant document sets is important for evaluating individual components in automatic question answering systems. In this paper, we describe using the Amazon’s Mechanical Turk (AMT) to judge whether paragraphs in relevant documents answer corresponding list questions in TREC QA track 2004. Based on AMT results, we build a collection of 1300 gold-standard supporting paragraphs for list questions. Our online experiments suggested that recruiting more people per task assures better annotation quality. In order to learning true labels from AMT annotations, we investigated the influence of annotation accuracy and number of labels per HIT on the performance of those approaches. Experimental studies show that the Naive Bayesian model and EM-based GLAD model can generate results highly agreeing with gold-standard annotations, and dominate significantly over the majority voting method for true label learning. We also suggested setting higher HIT approval rate to assure better online annotation quality, which leads to better performance of learning methods.

1. Introduction

The question answering (QA) is an important common task for the information retrieval (IR), information extraction (IE) and natural language processing (NLP) communities. TREC QA evaluation cover a broad range of techniques in those communities. Most QA systems’ architecture in TREC includes IR techniques to locate supporting paragraphs from relevant documents, and IE techniques involving with syntactic or semantic processing to target exact answers from paragraphs. Comparing with documents, paragraphs reduce search granularity and provide compact context for pinpointing answers, and therefore serve as an important intermediary between whole documents and exact answers. TREC QA only provided answer patterns and related documents for yearly question sets, which are useful for evaluating the overall performance of QA systems, but for the evaluation of individual component in QA systems more compact and precise paragraphs are required. Currently there is no such dataset of question-supporting texts for TREC list question task. The purpose of our work is to contribute to the development of QA systems by providing a new corpus, which include questions, answers and paragraphs which support their containing answer to the question. The application of IR, IE and NLP techniques in QA will all benefit from the fine-grained annotated dataset.

Table 1: Example of list questions

| Question                                                                 | Answers                                           |
|-------------------------------------------------------------------------|---------------------------------------------------|
| What countries have IFC financed projects in?                            | Albania, Argentina, Bosnia, etc. 42 answers       |
| Where did Johnny Applised plant trees?                                  | Ohio, Indiana, Pennsylvania. 3 answers            |

Table 1: Example of list questions

TREC QA proposed two fact-based short-answered tasks – factoid question and list question tasks. Factoid task require one answer, while list task require to provide a list of distinct instances. As it shown in Table 1, the main challenge for the list task is to determine the number of instances to return. Kaisser et al. (2008) collected the corpus of supporting sentences for factoid questions via Amazon Mechanical Turk (AMT). In contrast to expensive and time-consuming relevance judgement by very few assessors (Voorhees and Harman, 2005), AMT offers a web-based solution to quickly and cheaply annotate supporting compact excerpt in the question-relevant documents. Our work is not only to construct the corpus of supporting paragraphs for list task, but also to investigate and compare various methods to select true annotations and improve the quality of data from AMT results.

We conduct the data collection in following steps: data generation, online annotation and automatic selection of true annotations. In the following sections, we first introduce the usage of AMT and the control of data quality by build-in functionalities from AMT. We then describe three methods to learn the true annotations from AMT data. Finally we summarize the related work and proposed future works. The ListQA corpus can be downloaded from www.lsv.uni-saarland.de.

2. Experiment Design

2.1. Mechanical Turk

AMT is a web-based marketplace where requesters design and publish their work as micro HITs (Human Intelligence Tasks) to be done by multiple workers concurrently. With a large group of people working on HITs, requesters can get results very fast with very low cost. Major categories of

1TREC 2004 answer sets are in http://trec.nist.gov/data/qa/t2004_qadata.html.
2www.mturk.com
HITs include “catalogue and data management, search optimization, database creation and content management”\(^4\). AMT provides the web interface, command line tools and developer API, so that requesters can choose their favourite way of creating, publishing and managing their HITs. One HIT consists of one or more assignments. The requester can set desired number of assignments and download results of all HITs in different formats. AMT-registered online workers can preview and work on HITs, and then get paid by requesters.

### 2.2. Data and Experiment Setup

Table 2 shows an example of answer patterns (regular expressions) and linked document IDs for questions provided by TREC. To construct supporting paragraph candidates, first we use built-in paragraph boundary tags to split each document into successive paragraphs. Then those passages matching given patterns are selected as paragraph candidates. Given that the result of pattern matching is very noisy and coarse, we created HITs to recruit people to make binary decisions on whether each paragraph supports its containing answer(s) to the corresponding question or not. We generate 2856 question-paragraph pairs for TREC 2004. To reduce the number of HITs and control the budget, every HIT contains 2 question-paragraph pairs and costs $0.02.

| Question | What countries have IFC financed projects in? |
|----------|--------------------------------------------|
| Pattern  | Sri Lankan  || Sri Lanka  |
| Doc. IDs | XIE19981108.0129, XIE19990506.0269, XIE19981101.0083 |

Table 2: Example of answer pattern and linked documents

### 2.3. Data Quality Control

Requesters can use HIT approval rate\(^5\) to control the quality of work. At the initial run, we set the rate more than 95 (frequent threshold) and recruited 3 workers per HIT. We found the result (Dataset A) very noisy, with the annotation accuracy\(^6\) of 49.37%. To minimize the negative effects from the diverse workers’ expertise and spam workers, we therefore increase the approval rate to more than 98 and recruited 5 workers per HIT, and the final AMT results are exported as Dataset B.

Based on AMT results, we manually created the gold-standard annotations to evaluate the quality of work. Among the gold-standard 2856 paragraphs, 1300 paragraphs completely support their containing answer(s) to the given question, while rest 1556 paragraphs are irrelevant or partially relevant to the questions.

Table 3 shows the annotation accuracy of Dataset B, compared with A, increases by 24.40%, from 49.37% to 73.77%.

| Dataset | A       | B       |
|---------|---------|---------|
| Approval Rate | 95      | 98      |
| Workers per HIT | 3       | 5       |
| Duration (hrs)   | 9.55    | 47.63   |
| Annotation Accuracy | 49.37%  | 73.77%  |

Table 3: Comparison of Datasets exported from AMT

| # of Agreed Workers | # of HITs   |
|---------------------|-------------|
|                     | Dataset A   |
| Two                 | 1748 (61.20%) |
| Three               | 1108 (38.80%) |
|                     | Dataset B   |
| Three               | 1068 (37.39%) |
| Four                | 1030 (36.06%) |
| Five                | 758 (26.54%)  |

Table 4: Inter Annotation Agreement

![Accuracy vs. Number of Completed Paragraphs](Dataset A)

Figure 1: Individual workers' accuracy vs. # Paragraphs they completed.

73.77%. Table 4 demonstrates the inter-annotator agreements, i.e. how often a certain number (Two to Five) of workers make the same judgement about one HIT. Figure 1 shows the relation between individual worker’s accuracy with number of their completed paragraphs. Compare dataset A and B, even though we increase the approval...
rate, there still exist some spam workers as the points in
the right part of both figures with accuracy of around 50%
for binary judgment. The manually checking of their judg-
ments indicates that they produce a large number of random
labels. The argumentation of HIT approval rates doesn’t ef-
effectively filter more spammer workers as we expected. In
practice, we rejected bad workers whose annotation rates
are below a threshold. On the other hand, as the number of
annotators per HIT increase, from 3 workers for date-
set A to 5 for B the density of workers on the right up of
figures increase, i.e. the rate of workers with higher anno-
tation accuracy w.r.t. true labels of paragraphs.

The argumentation of HIT approval rates doesn’t ef-
effectively filter more spammer workers as we expected. In
practice, we rejected bad workers whose annotation rates
are below a threshold. On the other hand, as the number of
annotators per HIT increase, from 3 workers for date-
set A to 5 for B the density of workers on the right up of
figures increase, i.e. the rate of workers with higher anno-
tation accuracy increase. This is a joint effect of increasing
HIT approval rate and recruiting more people per HIT. To
reduce the proportion of spammer workers, we emphasized
employing more workers per HIT along with setting higher
HIT approval rate.

3. True Annotation Learning

Regarding the variety of individual worker’s reliability and
each HIT’s complexity, AMT worker’s labels are not per-
fect (see Table 4 and Figure 1). How to optimally combine
labels from multiple labelers and learn the true label is of
great significance to automatic data annotation. Hereby to
learn true annotations from AMT results we compare three
approaches: supervised Naive Bayesian model (Snow et al.,
2008), unsupervised GLAD model (Whitehill et al., 2009),
and the Majority Voting (MV) as the baseline.

3.1. Naive-Bayesian-Type Model

With the principle that the majority rules, the majority vot-
ing method assumes all works exhibit identical expertise
and therefore have equal vote. However, in online annota-
tion scenario, if the majority are noisy or adversarial work-
ers who give the same incorrect label for a specific para-

graph, the majority voting would favour the major incor-
rect label and ignore true labels in the minority. Snow et al.
(2008) introduced a multinomial Naive-Bayes-Type (NBT
) model to estimate the worker’s expertise and weight each
worker’s label with their performance likelihood.

Each paragraph $i$ has a true label $x_i \in \{0, 1\}$.
Let $L_i = \{l_{ij} : j = 1, \ldots, J\}$ be the set of labels given by workers.
The conditional probability of a paragraph’s true label $x_i$
given its labels $L_i$ is calculated to determine the true label.
Using Bayes rules,

$$P(x_i | L_i) = \left( \prod_{j} P(l_{ij}|x_i) \right) \frac{P(x_i)}{P(L_i)}$$

where each worker’s label are assumed as conditionally in-
dependent of others’ given the true label $x_i$.

During the training stage, the estimation of each worker’s
performance likelihood $P(l_{ij}|x)$ is derived from incorporat-
ing his annotation accuracy w.r.t. true labels of paragraphs
he completed, e.g., $P(l_{ij} = w|x = t)(w, t \in \{0, 1\})$ mea-
sures the ratio of the worker $j$’s labels are class $w$ given

$$P(L_i = w|X = t) = \frac{\sum_{k \in \Phi_j} \delta(l_{jk} = w \lor x_k = t) + 1}{\sum_{k \in \Phi_j} \delta(x_k = t) + |\Phi||S|}$$

where, $\Phi_j$ is the set of paragraphs worker $j$ completed. $\Phi$
is the complete set of all paragraphs. $|S|$ is the number of
assignments per HIT.

Given all workers’ response likelihood for paragraph $i$, the
true label $x_i$ is judged using the posterior log odds:

$$Q(R) = \log \frac{P(x_i = 1 | L_i)}{P(x_i = 0 | L_i)} = \sum_j \log \frac{P(l_{ij} | x_i = 1)}{P(l_{ij} | x_i = 0)} + \log \frac{P(x_i = 1)}{P(x_i = 0)}$$

If the log odds $Q(R)$ is positive, the label of a paragraph is
class 1.

3.2. GLAD Model

The Generative model of Labels, Abilities and Difficulties
(GLAD) (Whitehill et al., 2009) simultaneously learns the
true label, item difficulty and the labeler expertise in an un-
supervised manner. Following their method, we model the difficulty of para-
graph $i$ using the parameter $1/\beta_i \in [0, \infty)$ where $\beta_i > 0$.
Here $1/\beta_i = \infty$ means the paragraph is very hard to judge.
$1/\beta_i = 0$ means the paragraph is so easy that most workers
will always judge correctly.

The worker $j$’s ability is modeled by the parameter $\alpha_j \in
(-\infty, +\infty)$. Here an $\alpha_j = +\infty$ means the worker always
makes correct labels, while $\alpha_j = -\infty$ means the worker
always judges incorrectly. Then for worker $j$ to paragraph
$i$, the posterior probability is defined as,

$$P(l_{ij} = x_i|\alpha_j, \beta_i) = \frac{1}{1 + e^{-\alpha_j \beta_i}}$$

The Expectation-Maximization (EM) algorithm is used to
obtain maximum likelihood estimates of true labels $X$ and
parameters $\alpha, \beta$ given the observed data. Each iteration
of the EM algorithm consists of an Expectation(E)-step and a
Maximization(M)-step.

1. **E step:** The posterior probabilities of all $x_i \in \{0, 1\}$
given the $\alpha, \beta$ from last M step and the worker labels:

$$P(x_i | l, \alpha, \beta) \propto P(x_i) \prod_j P(l_{ij}|x_i, \alpha_j, \beta_i)$$

2. **M step:** To maximize the standard auxiliary function
$Q$, which is defined as the expectation of the joint log-
likelihood of the observed and hidden variables ($l, X$) given
the parameters ($\alpha, \beta$) estimated during the last
E-step:

$$Q(\alpha, \beta) = \sum_j E[\ln P(x_i)] +$$

$$+ \sum_{ij} E[\ln P(l_{ij}|x_i, \alpha_j, \beta_i)]$$

$\delta(x)$ is 1 if its logical argument $x$ is true and 0 otherwise.
Gradient ascent Algorithm is employed to find values of \( \alpha, \beta \) that locally maximized \( Q \).

For dataset A, as a large proportion of labels are judged incorrectly, \( \alpha \) need be made very low for \( \alpha > 0 \). We used Gaussian priors \((\mu = 0.0001, \sigma = 0.0001)\) as priors for \( \alpha \) and the \( X \) are initialized with 0.0001. For dataset B, optimal priors \( \alpha \) values are Gaussian priors \((\mu = 0.9, \sigma = 0.9)\) the \( X \) are initialized with 0.5. We re-parameterized \( \beta = e^{\beta'} \) and imposed a Gaussian prior \((\mu = 0.0001, \sigma = 0.0001)\) on \( \beta' \) for dataset A and \((\mu = 0.9, \sigma = 0.9)\) for dataset B. The label of a paragraph is class 1 when \( P(x_i = 1|l, \alpha, \beta) > 0.5 \).

### 3.3. Results

To compare the effectiveness of learning methods, the gold-standard annotations are used as ground truth judgements. We measured the effectiveness in term of proportion of correctly inferred labels. Table 5 showed the accuracy of each approach against two different levels of annotation accuracies. The NBT model is trained and tested via 20-fold cross validation on the whole dataset. The application of both methods brings a significant accuracy growth over the baseline in learning the true annotations. Contrary to results presented in (Whitehill et al., 2009), the NBT model makes fewer errors than the GLAD model. The probable reason is as following: Snow et al.'s method makes use of pre-labeled ground truth labels; Although Whitehill et al. (2009) claimed the GLAD’s advantage of modeling task difficulty might be very important, experimental results with different values of \( \beta \) rarely changed in our case, therefore GLAD’s performance is somehow weaken by unsuccessfully modelling the paragraph difficulty.

| \( \alpha, \beta \) | A | B | C\(_{3W} \) | B\(_{3W} \) |
|-----------------|---|---|--------|--------|
| MV              | 49.61% | 82.98% | 67.09% | 79.06% |
| GLAD            | 54.52% | 89.81% | 67.51% | 85.04% |
| NBT             | 61.75% | 91.36% | 81.79% | 87.47% |

Table 5: Accuracies of the approaches on dataset A and B with different annotation accuracies (AA)

In order to explore the influence of setting HIT approval rate on performance of learning methods, we perform a simple simulation: for each paragraph in dataset B, 3 labels are randomly chosen from 5 labels and totally collected as Dataset B\(_{3W} \), on which we test those three approaches. The simulation are repeated 100 times to smooth out variability between trials and the average accuracy is shown in Table 5. Comparison between dataset A and B\(_{3W} \) indicates that improving HIT approval rate can result in better AMT online annotation accuracy and therefore lead to significant performance improvements. From dataset B\(_{3W} \) to B, we can see that recruiting more labelers per HIT can also obviously boost performance.

We merge dataset A and B into dataset C ( 8 labelers per HIT and annotation accuracy 63.58%), on which we further investigate the effect of varying the number of labelers per HIT. Figure 2 demonstrates the analytical relationship between the accuracy of estimated labels and the number of labelers, for different approaches. As expected the performance of NBT and majority voting model improves with larger numbers of labelers, while GLAD model shows unstable performance and doesn’t show advantage over the majority voting method. Dataset A, C\(_{3W} \) and B\(_{3W} \) all employ 3 lablers per HIT. Comparisons of their results in Tabel 5 indicate that as the annotation accuracy increase steady in those three datasets, the performance of all methods increases. The GLAD model works noticeably better on dataset with better quality ( e.g. dataset B and B\(_{3W} \) ). When the annotation accuracy are low ( 49.37% of dataset A), all methods tend to show low accuracy due to the influence of large amount of noisy and adversarial labels.

![Figure 2](image-url)

After all, our results highly suggest setting higher HIT approval rate (normally 98%) for the practice with AMT assures higher online annotation accuracy, therefore those three approaches can recover the true labels more accurately. Additionally, Naive-Bayesian-type method mainly rely on prior of workers’ performance likelihood on the training data. If a number of new workers appear only in the testing data, their response likelihood can not be estimated during the training stage, while GLAD model don’t suffer from this new worker problem. When ground truth labels are not available and AMT annotations show reasonable accuracy, GLAD still can have beneficial practical applications in unsupervised learning of true labels.

### 4. Related Work

Mechanical Turk’s advantage of low cost, speedy workflow and huge workforce has attracted increasing interests in IR and NLP communities. The upcoming workshops in
NAACL\textsuperscript{9} and Coling\textsuperscript{10} aim at promoting wide and creative use of AMT in various domains. Several works have explored the effectiveness of using AMT for resources annotation and performance evaluation.

Snow et al.\textsuperscript{(2008)} thoroughly reviewed works of annotation collection via AMT, empirically examined 5 natural language processing tasks and proposed a technique for improving annotation quality. Callison-Burch \textsuperscript{(2009)} showed that AMT can be for complex tasks such as creating multiple reference translations and reading comprehension tests. Both papers demonstrated that although AMT labelers are often individually less reliable and stable, non-expert labelers in aggregation can produce judgements highly agree with gold-standard experts.

For the search evaluation, Kaisser et al. \textsuperscript{(2008)} launched a survey via AMT on customizing the summary length of search results. Studies on AMT results suggest that search results best presented different lengths of summary snippets for different types of queries. Alonso et al. \textsuperscript{(2009)} evaluated the quality of search results produced by their time-based clustering algorithm combined with temporal snippets.

Previous efforts at QA corpus construction focus on annotating more precise annotated data. Kaiser et al. \textsuperscript{(2008)} constructed a corpus of question-sentence pairs for the TREC factoid question and employed experts to further clean the corpus and tagged how sufficiently a sentence supports its question. To create a Why QA corpus, Morzin et al. \textsuperscript{(2008)} first asked workers to write a why question based on part of a Wikipedia article, then presented HITs to select answer sentences from the original articles, and in the final task workers paraphrased each question to provide variation of questions. In this paper, we collected corpus for the TREC list questions and further explored three methods to automatically boost the quality of corpus.

5. Conclusion and Future Work

We constructed a new corpus of supporting paragraphs collections for list question in TREC QA 2004. We also investigated how to control annotation quality through the functionality provided by AMT and suggested that recruiting more people per task along with setting higher HIT approval rate assures better annotation quality. We compared three approaches of selecting accurate annotations in AMT results, and investigated the influence of mislabeling data and number of labelers per HIT on their performance. Experiments show that, with careful design of tasks and appropriate approaches to select true labels, high-quality labels can be automatically learned from AMT non-expert annotations. We also suggested that better online annotation quality leads to better performance of learning methods.

Furthermore, we will continue collecting supporting paragraphs for TREC 2005-2007 list questions. With a large collection of data, obvious areas for future work are paragraph retrieval and answer extraction for list questions. The question-paragraph corpus and AMT results is available via \url{www.lsv.uni-saarland.de}.

Acknowledgements

Thanks to Matt Lease, Omar Alonso, and Grzegorz Chrupała for enlightening discussions and comments. Fang Xu was funded by the German research council DFG through the Partnership for Research and Education “PIRE” on the International Research Training Group “IRTG” grant.

6. References

Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. 2009. Clustering and Exploring Search Results Using Timeline Constructions. In Proceedings of CIKM, Hong Kong, China.

Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazons Mechanical Turk. In Proceedings of EMNLP, Singapore.

Michael Kaisser and John B. Lowe. 2008. Creating a Research Collection of Question Answer Sentence Pairs with Amazons Mechanical Turk. In Proceedings of International Conference on Language Resources and Evaluation, LREC, Marrakech, Morocco.

Michael Kaisser, Marti A. Hearst, and John B. Lowe. 2008. Improving search results quality by customizing summary lengths. In Proceedings of ACL-08: HLT, pages 701–709, Columbus, Ohio, June. Association for Computational Linguistics.

Joanna Mrozinski, Edward Whittaker, and Sadaoki Furui. 2008. Collecting a Why-Question Corpus for Development and Evaluation of an Automatic QA-system. In Proceedings of ACL-08: HLT, pages 443–451, Columbus, Ohio, USA.

Rion Snow, Brendan O Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of ACL-08:HIT, Columbus, Ohio, USA.

E. Voorhees and D. Harman. 2005. Trec experiment and evaluation in information retrieval. MIT Press, MA, USA.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose Vote Should Count more: Optimal Integration of Labels from Labelers of Unknown Expertise. In Proceedings of Advances in Neural Information Processing Systems, Vancouver, Canada.

\textsuperscript{9}http://sites.google.com/site/amtworkshop2010

\textsuperscript{10}http://www.ukp.tu-darmstadt.de/scientific-community/coling-2010-workshop
A third investment involves the Korea Trade Enhancement Facility (KTEF), a US $100 million trade enhancement facility established by IFC with Sumitomo Bank Ltd to expand trade finance to South Korea.

IFC’s investment will finance the first stage of development of the Bolivar Block in Colombia’s Middle Magdalena Valley. This phase will include drilling nine wells and constructing facilities and transmission pipelines to produce up to 30,000 barrels of oil per day which will be exported via Covenas on the country’s Caribbean coast.

More than 66 million Dollars have been committed by IFC, the private sector lending arm of the World Bank, to projects in Kenya since 1970, the East African weekly reported today.

Addressing a meeting at the Lahore Chamber of Commerce and Industry, he said that the IFC would continue its financial assistance in Pakistan’s investment activities by further expanding its operation.

The IFC is a member of the World Bank Group, and the largest multilateral source of equity and loan financing for private sector projects in developing countries. Up to date, the IFC has invested over 11 million dollars for six projects in Mozambique.

WASHINGTON, October 23 (Xinhua) -- The International Finance Corporation (IFC) today announced the approval of 37.5 million U.S. dollars in loan and equity to finance a shipping company in the Philippines.

WASHINGTON, May 22 (Xinhua) -- The International Finance Corporation (IFC) has agreed to provide up to 112.35 million U.S. dollars for an expansion of a ceramic roof tile manufacture project in Indonesia.

China is IFC’s fastest growing client. IFC had provided 1.2 billion U.S. dollars in financing for 37 projects in China by the end of June, with the total project cost standing at 2.95 billion U.S. dollars.

WASHINGTON, August 13 (Xinhua) -- The International Finance Corporation (IFC) announced Friday that it will lend 35 million U.S. dollars to Uzel Makina Sanayi A.S. of Turkey to help the tractor maker modernize.

All made the remarks when referring to Malaysia’s re-inclusion to the International Finance Corporation’s (IFC) indices.

WASHINGTON, June 18 (Xinhua) -- The International Finance Corporation (IFC) announced on Friday that it will invest 13.2 million U.S. dollars in La Universal, S.A., one of Ecuador’s leading confectionery and food companies.

Table 6: Examples of the question ID, paragraph ID and answer string following with the supporting paragraph for the question “What countries has the IFC financed projects in?”