Simple Analogical Reasoning using Word Occurrences for Question Answering Passage Retrieval

Hapnes Toba
Information Retrieval Laboratory
Universitas Indonesia
Depok 16424, Indonesia
hapnes.toba@ui.ac.id

Ruli Manurung
Information Retrieval Laboratory
Universitas Indonesia
Depok 16424, Indonesia
maruli@cs.ui.ac.id

Mirna Adriani
Information Retrieval Laboratory
Universitas Indonesia
Depok 16424, Indonesia
mirna@cs.ui.ac.id

ABSTRACT
Recent approaches to question answering (QA) attempt to model the relation between existing question-answer pairs, and apply this model to the construction of answers to novel questions. In this paper we study whether a question and its answer can be related using simple word occurrence features, and whether this relational model can improve the passage retrieval task of a QA pipeline. We argue that in this context, words appearing in answers in analogous question-answer pairs may represent the information needs of a query, thus may also appear in other passages which have some analogical or related features. We attempt to leverage this through query expansion strategies. In our experimental setting, analogy is measured as the similarity between word occurrences of related question and answer pairs, and is modeled using the Bayesian Analogical Reasoning (BAR) framework. The resulting model is used to rank retrieved candidate answer passages. Experiments using the ResPubliQA 2009 and 2010 collections show that the analogy-based query expansion does not perform better than the baseline method, but may suggest better performance using more sophisticated linguistic features.

Keywords
Bayesian Analogical Reasoning, ResPubliQA, Mean Reciprocal Rank, Passage Retrieval.

1. INTRODUCTION
Question Answering (QA) is a specific form of information retrieval (IR) that seeks to produce an exact answer given a natural language question [1]. Most recent QA architectures are highly dependent on third-party search engines [2], such as Indri and Lucene, or web-based search engines, e.g. Google and Yahoo. Given a query that is typically a reformulation of the natural language question, these search engines work by filtering the *n* most relevant textual passages, which could range from paragraphs to entire documents, from which the final answer is constructed. Unlike the conventional search task, which depends on a number of fixed search terms, i.e. keywords, in a QA task the question first needs to be analyzed in order to produce a final answer that reflects some specific information need. Consider the question: *"On what day did the Chernobyl nuclear accident happen?"*. A conventional search engine (IR) might, for instance, be able to retrieve the following passage: *"Whereas, following the accident at the Chernobyl nuclear power-station on 26 April 1986, considerable quantities of radioactive materials were released into the atmosphere, contaminating food stuffs and feeding stuffs in several European countries to levels significant from the health point of view;"*. However, this question has a specific information need about the time of occurrence (according to the terms *"what day"* and *"happen"* in the question) for an event about the *"Chernobyl nuclear accident"*. Humans who are proficient in the required language can easily understand this kind of interpretation to produce a final answer (*26 April 1986*) from the retrieved passage. Precisely identifying the answer in this manner remains a challenge for a QA system.

Obviously, the difficult task of constructing a final answer will be made easier if the final answer is already included in a limited set of passage retrieval results. In this context, the performance of an underlying search engine is important [2] to retrieve relevant passages. Recent work in IR strategies that are specific to the QA task are mostly focused on linguistic and semantic constraints [3], relevance feedback [4], semantic role indexing [5] or by topic indexing [6]. Despite these recent approaches, performing QA passage retrieval in a more conventional IR way, i.e. using the so-called "bag-of-words" features consisting of appropriate question terms, could be preferable if important search terms are already stated in the question.

Recently, a new approach has been developed that focuses on the relational data between existing question-answer pairs [7]. Typical QA systems consider questions and answers as independent elements, where the task is constructing the appropriate answer for a given question. From this perspective, there is no gain to be obtained from an existing collection of questions and answers. However, by assuming that answers are related to their questions through certain types of implicit links, it is theoretically possible to learn these links from existing data, e.g. a gold standard corpus of question-answer pairs, and to apply the learned model for relating unseen questions to their appropriate answers. Wang et. al. [7] showed that in a community QA situation textual mismatch between a question and its passage candidates can be learnt by performing analogical reasoning that relates a question to its answer using textual, statistical and social elements features. Inspired by this work, in this paper we study whether a question and its answer can be related using simple word occurrence features, and whether this relational model can be applied to improve passage retrieval of "analogous" question-answer pairs. We argue that in a QA passage retrieval context, words appearing in answers in analogous question-answer pairs can actually contribute some positive influence to represent the information needs that also appear in other passages which have some analogical or related features. This fact may even extend to closed class function words, i.e. stopwords, which are typically removed in conventional IR systems. In our experimental setting, an analogy is a measure of similarity between word occurrences of related question and answer pairs. We use the question-answer pairs from the ResPubliQA[1] 2009 paragraph selection gold standard as our training set, and the ResPubliQA 2010 collection as our testing data. Our previous work [8] showed that paragraph

[1] http://celct.isti.cnr.it/ResPubliQA/

selection is a challenging task and one of the methods to improve paragraph retrieval is by using word occurrences as contextual information.

2. BAYESIAN ANALOGICAL REASONING

Wang et. al. [7] employed the Bayesian Analogical Reasoning (BAR) framework that was originally introduced by Silva et. al. [9]. The basic idea of BAR is to learn a prior from related objects (question and answer pairs in our case), and update it during the retrieval process of a query to obtain a marginal probability that relates the query with the objects that have been learnt.

Assume there is a space of unseen functions $Q \times A \rightarrow \{0,1\}$. If two objects, *Q* and *A* are members of a set *S*, which are related by an unknown function *f(Q,A) = 1*, what needs to be quantified is how similar the function *f(Q,A)* is to another unseen function *g(. , .)*, that classifies all pairs of $(Q^i, A^j) \in S$ as being linked where $g(Q^i,A^j) = 1$. The functions *f(., .)* and *g(. , .)* are unseen, and thus we need a prior that will be used to integrate over the function space.

For each pair $(Q^i \in Q, A^j \in A)$, there exists a retrieval result of $X^{ij} = [\Phi_1(Q^i, A^j) \ldots \Phi_k(A^i, A^j)]$ defined by the mapping $\Phi : Q \times A \rightarrow \Re^K$.

This feature space mapping computes a *K*-dimensional vector of attributes of the question answer pairs, that is hoped to have a relevant link prediction between the objects in the pairs.

If there is an unseen label L^{ij}, with $L^{ij} \in \{0,1\}$ as a predicted indicator of the existence of a relation between Q^i and A^j in a learnt set, then we will have a parameter vector $\Theta = [\Theta_1 \ldots \Theta_K]$, which could be learnt by performing the logistic regression model:

$$P(L^{ij} = 1 \mid X^{ij}, \Theta) = logistic(\Theta^T X^{ij}) \quad (1)$$

where *logistic(x)* is defined as $(1 + e^{-x})^{-1}$.

In the retrieval process, a query is compared by the functions for links predictions by marginalizing over the parameters of the functions. If we have L^S as the vector of link predictions for S, then each $L \in S$ has the value $L = 1$, indicating that every pair of objects in *S* is linked. The final score of a retrieval process indicating the order of predicted links between the query and the related objects that have been learnt is computed as follows:

$$score(A^i, B^j) = \log P(L^{ij} = 1 \mid X^{ij}, S, L^S = 1) - \log P(L^{ij} = 1 \mid X^{ij}) \quad (2)$$

Silva et. al. uses the variational Bayesian logistic regression [11] to compute this scoring function. See [10] for more fundamental proofs and information retrieval scenarios.

3. EXPERIMENTAL SETTING

3.1 Methodology

We used the JRC-ACQUIS[2] and EUROPARL[3] document collections that were suggested by the ResPubliQA organizer. We first created an index that was based on paragraph segmentations using the Indri indexing tools. In total this produces about 1.5 million indexed passages. Indri is a search engine that is specially designed for passage retrieval [12], thus we deemed it appropriate to the retrieval task in this study.

During the training set preparation, by using the ResPubliQA 2009 gold standard, we built a binary word occurrence indexing, which indicates whether a word exists in a question answer pair or not. We have in total 5,671 word features that are present in 500 question answer pairs. Further preparation that has been made for this training set is to utilize the Singular Value Decomposition (SVD) matrix operation that is very effective in applications such as Latent Semantic Analysis (LSA). We decompose our data into SVD 25-dim, in order to reduce the word features dimensionality. This SVD decomposition is the main feature for the whole experiments. The complete methodology of the whole experiment can be seen in Figure 1.

After the data preparation, the training stage is performed by using equation (1) in Section 2 above. After the training stage, we performed a retrieval process by using the ResPubliQA 2010 questions as the query set. In total we have 200 questions in the test set. Each question will first be passed into the Indri search engine as a bag-of-words (BOW) query, and we consider the top-1 passage retrieval as the relevant candidate. Thus, there are 200 question-first retrieval pairs for all the questions set. For each extracted features from the question-first retrieval pairs, i.e. the word occurrence, the BAR algorithm will compute the rank pairs of analogical set, based on the learnt priors by using equation (2). These ranked pairs show the measure of "relatedness" of a question in another set, i.e. the training set. By using this methodology, question type analysis, which is usually performed in a QA pipeline to obtain and represent some information needs

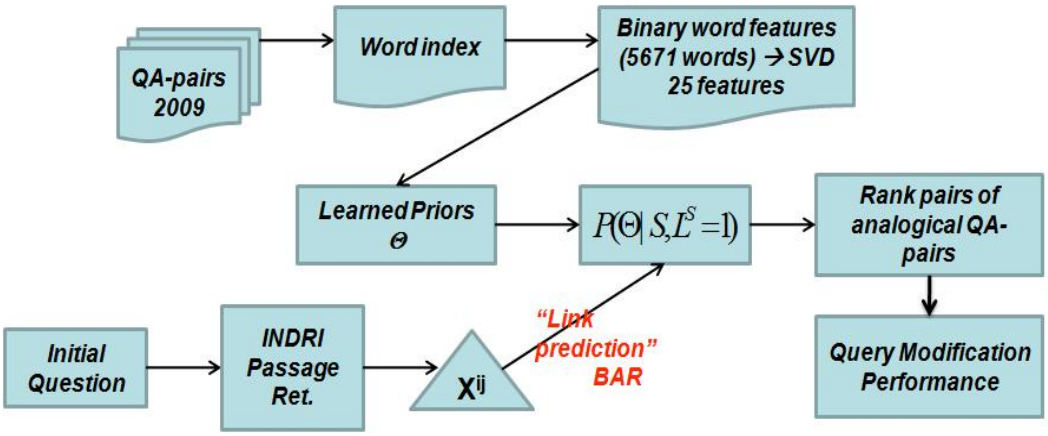

Figure 1 Experiment Methodology

[2], is replaced by the analogical reasoning algorithm. An example of the overall process is shown in Table 1.

To evaluate the influence of analogical pairs in the QA passage retrieval, we modified the initial query set, i.e. the BOW query, in a number of scenarios as explained in the next sub-section.

3.2 Query Modification Scenario

As mentioned earlier in the paper, we argue that in a QA passage retrieval context, words appearing in answers in analogous question-answer pairs may represent the information needs of a query. This fact may even extend to stopwords. For example, if a question is related to locational attributes, e.g. *"Where is…?"* or *"What is the address…?"*, it seems plausible that answers would contain location prepositions such as *at*, *on*, *in*, *near*, etc. This is the type of knowledge that would typically be manually encoded in the form of question types and their expected answer patterns. However, in our approach, we hope to learn this implicit knowledge automatically through analogical reasoning over an existing collection of question-answer pairs. In [7], this is termed as utilizing existing knowledge to "bridge the lexical and semantic gap", i.e. when the relevant keywords in the correct answer are not found in the query, and thus the search engine fails to retrieve the correct passage. Thus, in our experiment we formulate five variants of query expansion modifications, as follows:

Table 1 Example of process in experiment

Initial question	*"On what day did the Chernobyl nuclear accident happen?"*
Top-1 Indri Passage Retrieval	*"As someone who has reservations regarding the long-term safety of nuclear power stations I am not entirely happy with this report. I accept that those who work in the industry are dedicated professionals. I accept accidents are few and far between. Yet the problem is the consequences when there is an accident are potentially so extreme. We cannot forget the nuclear accident in the Urals revealed by Roy Medvedev, the Three Mile Island accident that involved the events of the film that went on release only days before the accident, nor the Tokaimura accident in Japan, nor finally Chernobyl, whose consequences live with us all today but most tragically have been visited on the people and children living in the immediate vicinity or born to parents who were there."*
Extracted word occurrence features	*1 2 4 6 8 12 28 32 36 48 58 61 68 73 97 108 133 168 185 195 200 220 233 267 274 367 406 437 441 498 561 577 598 599 646 648 665 892 895 959 978 1149 1174 1192 1246 1307 1361 1371 1412 1432 1703 1897 1953 2010 2232 2340 2348 2510 2516 2638 3723 4193 4240 4336 4394 4419 4468 4533 4687 4816 4820 4821 4822 5205 5320 5358 5429*
Best analogous question-answer pair from learning set	*Which rivers were considered as a buffer zone?* *The water catchment areas of the rivers Gandarillas, Escudo, Miera y Campiazo are considered as a buffer zone.*

Table 2 Example of query modifications

Initial / baseline	*on what day did the Chernobyl nuclear accident happen*
QE1	*chernobyl a which accident as nuclear day happen were*
QE2	*zone buffer nuclear day rivers chernobyl accident considered happen*
QE3	*the a which zone did buffer on nuclear day rivers chernobyl what accident as considered happen were*
QE4	*day chernobyl nuclear accident happen*
QE5	*day chernobyl nuclear accident happen the are of as a*

1. Remove stopwords from the initial question, and add the stopwords from the best analogous question, coded as QE1.
2. Unify all words from the initial question and the best analogous question, and then remove the stopwords, coded as QE2.
3. Unify all words from the initial question and the best analogous question, coded as QE3.
4. Remove stopwords from the initial question, coded as QE4.
5. Remove stopwords from the initial question, and add the stopwords from the best analogous answer, coded as QE5.

By performing these query modifications, we try to evaluate how the information needs are maintained by using the word occurrences in the best analogous question-answer pair. We consider that all words in a question will be important to form some information needs and not only on specific terms during passage retrieval, hence the modifications to retain and/or remove stopwords.

Referring to the example from the previous sub-section, if we perform the five query modifications above to the original query, we will obtain the queries as shown in Table 2. Note that the sequence of words in a query is not considered under the bag-of-words (BOW) retrieval model.

Table 3 MRR Performance

ResPubliQA Question Type	Baseline	QE1	QE2	QE3	QE4	QE5
Definition	**0.31**	0.24	0.10	0.10	0.27	0.26
Factoid	**0.55**	0.48	0.40	0.34	0.50	0.46
Reason / Purpose	**0.64**	0.56	0.45	0.45	0.63	0.61
Procedure	**0.38**	**0.38**	0.25	0.27	0.37	0.37
Opinion	0.52	0.51	0.51	0.38	**0.60**	0.54
Other	**0.59**	0.52	0.44	0.44	0.56	0.49
Overall	**0.50**	0.45	0.36	0.32	0.49	0.46

3.3 Performance Evaluation

To measure the passage retrieval performance, we use the mean reciprocal rank (MRR), which is defined as:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank\_i} \quad (3)$$

where:

N = number of questions;

$rank\_i$ = the rank of the relevant answer of question i in the top-n retrieval results. If no relevant answer is retrieved in the top-n retrieval, the reciprocal rank value will be set as 0.

Since a question can only have one relevant answer, a reciprocal rank can also be considered as the precision of a retrieval task. We use the top-*20* passages retrieved for performance evaluation. The bag-of-words queries are used as the baseline performance and will be compared with each of the five query modification scenarios.

4. EXPERIMENTAL RESULTS

4.1 MRR Evaluation

The MRR performance for each query modification, as well as the baseline query, can be seen in Table 3. From this table we can see that the MRR of the baseline query retrieval has the best performance, with an overall score of 0.50, and outperforms the other query modification scenarios in all but one question type. The single exception occurs for the 'Opinion' question type, where it can be seen that QE4 has the best performance. QE4 corresponds to the common practice of extracting relevant 'keywords' from the question by removing stopwords.

4.2 Positive Influence

Table 4 gives the number of positive influences that has been made for each query modification against the baseline. A positive influence means that a query modification improves the rank of the correct passage in the retrieval results, for example:

Initial question	*what approach does the montreal protocol take towards the production of bromochloromethane*	Relevant passage at Rank 2
QE1	*the protocol montreal system which on shall production be approach bromochloromethane*	Relevant passage at Rank 1

It can be seen in Table 4 that QE4 gives the best result with 29 rank improvements out of 200 questions, with QE5 next best with 21 improvements.

4.3 Discussions

Considering these results, QE2 and QE3 certainly yield significantly worse results. It seems that using words from the question part of the analogous question-answer pair does not help bridge the lexical gap.

In theory, QE5 should be able to bridge this gap. The content words from the original question should capture the subject material that is being queried, e.g. the Chernobyl accident, and the words from the answer of the most analogous QA pair should provide the information needs. The crucial factor, however, is which of these words best convey these needs. It seems clear that content words from an analogous answer are irrelevant. Looking at the example in Table 1, it is clear that the phrases *"water*

Table 4 Number of Positive Influence

ResPubliQA Question Type	QE1	QE2	QE3	QE4	QE5
Definition	1	1	0	3	**4**
Factoid	**5**	4	2	4	3
Reason / Purpose	4	2	1	**6**	3
Procedure	5	4	2	**7**	**7**
Opinion	2	5	0	**7**	4
Other	1	0	0	**2**	0
Overall	18	16	5	**29**	21

catchment areas" and *"buffer zones"* serve no purpose for querying the Chernobyl accident. However, by taking the simplistic approach of only retaining stopwords, in Table 2 we can see that the resulting words are *the, are, of, as*, and *a*, which do not seem to capture any specific information need. Thus, more sophisticated linguistic features should be employed. Another aspect to be tried is not to utilize the features of the single most analogous QA pair, but to instead aggregate patterns from the *n*-most analogous QA pairs, as recurring patterns would indicate an indicative feature of the information need.

Some other minor observations that can be made by analyzing the results in Tables 3 and 4 are as follows:

1. Passage retrieval by using the whole question words (baseline) and by removing the stopwords (QE4) have a comparable rank arrangement.
2. It seems that if we replace the stopwords in the baseline query with other stopwords from the best analogous question (QE1), the passage retrieval still have comparable results. In this sense, the information needs of a question are still maintained by replacing them with the stopwords from the best analogous question.
3. The performance of retrieval results will be far decreased if many unrelated words, which are not considered as information needs, are included in the query, as suggested by the results of QE2 and QE3.
4. The 'Definition' question type is the most difficult query type to handle, with the smallest number of word occurrences' overlapping in a question answer pair. In other words, definitional questions have the most textual mismatch between the question and the answer. For example:

Question	What is an SME?
Relevant passage	*whereas the EU SMEs, defined as enterprises with fewer than 250 employees and a turnover not exceeding EUR 50 million, account for 23 million enterprises (99% of the total) and 75 million jobs (70%) in the European Union,*

5. CONCLUSIONS

Our conclusions during this study can be summarized as follows:

1. By replacing some stopwords with other words from the best analogous pair, the information needs of a question can still be maintained.
2. The rank improvements during passage retrieval are mostly influenced by the stopwords that included in the question answer pairs.
3. By using analogical reasoning the role of question type analysis can be reduced. The predicted relation between a question and its answer has promising features that also relates how a question should be answered.
4. There are a number of other features that should be investigated as future works to develop a more rigid model in passage retrieval for answering a question, such as: word statistics, word position or textual entailment features for answer validation process.

6. REFERENCES

[1] Andrenucci, A., Sneiders, E. 2005. Automated Question Answering: Review of the Main Approaches. In *Proceedings of the Third International Conference on Information Technology and Applications* (ICITA '05), 514-519, July 4-7, 2005, Sydney, Australia. IEEE Computer Society.

[2] Bilotti, M.W. 2009. *Linguistic and Semantic Passage Retrieval Strategies for Question Answering*. Dissertation Thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

[3] Bilotti, M. W., Elsas, J., Carbonell, J., Nyberg, E. 2010. Rank Learning for Factoid Question Answering with Linguistic and Semantic Constraints. In *Proceedings of 19th ACM Conference on Information and Knowledge Management* (CIKM '10), 459-468, October 26-30, 2010, Toronto, Canada. ACM, New York, NY, USA.

[4] Pizzato, L.A., Molla D., Paris C. 2006. Pseudo-Relevance Feedback using Named Entities for Question Answering. In *Proceedings of the 2006 Australasian Language Technology Workshop* (ALTW '06), November 30-December 1, 2006, Sydney, Australia.

[5] Pizzato, L.A., Molla D., 2008. Indexing on Semantic Roles for Question Answering. In *Proceedings of the COLING 2008: 2nd Workshop on Information Retrieval for Question Answering*, 74-80, August 24, 2008, Manchester, UK. ACL, PA, USA.

[6] Ahn, K., Webber, B. 2008. Topic Indexing and Information Retrieval for Factoid QA. In *Proceedings of the 2nd ACL Workshop on Information Retrieval for Question Answering* (IRQA '08), 66-73, August 24, 2008, Manchester, UK. ACL, PA, USA.

[7] Wang, X-J., Tu, X., Feng, D., Zhang, L. 2009. Ranking Community Answers by Modeling Question-Answer Relationship via Analogical Reasoning. In *Proceedings of the 32nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '09), 179-186, July 19-23, 2009, Boston, MA, USA.

[8] Toba, H., Sari, S., Mirna, A., Manurung, R. 2010b. Contextual Approach for Paragraph Selection in Question Answering Task. In *Working Notes of CLEF ResPubliQA 2010*.

[9] Silva, R., Heller, K., Ghahramani, Z. 2007. Analogical Reasoning with Relational Bayesian Sets. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics* (AISTATS '07), March 21-24, 2007, San Juan, Puerto Rico.

[10] Silva, R. Heller, K. Ghahramani, Z., Airoldi, E. 2010. Ranking Relations Using Analogies in Biological and Information Networks. *The Annals of Applied Statistics*, 4(2):615-644. Institute of Mathematical Statistics.

[11] Jaakkola, T., Jordan, M. 2000. Bayesian Parameter Estimation via Variational Bounds. *Statistics and Computing*, 10:25-37.

[12] Buscaldi, D., Rosso, P. , Gómez-Soriano, J.M., Sanchis, E. 2009. Answering Questions with an *n*-gram based Passage Retrieval Engine. *Journal of Intelligent Information System*, 34(2):113-134, April 1, 2010. Springer Netherlands.