Answer Presentation with Contextual Information: A Case Study using Syntactic and Semantic Models

Rivindu Perera and Parma Nand

School of Computer and Mathematical Sciences,
Auckland University of Technology
Auckland 1010, New Zealand
{rivindu.perera,parma.nand}@aut.ac.nz

Abstract. Answer presentation is a subtask in Question Answering that investigates the ways of presenting an acquired answer to the user in a format that is close to a human generated answer. In this research we explore models to retrieve additional, relevant, contextual information corresponding to a question and present an enriched answer by integrating the additional information as natural language. We investigate the role of Bag of Words (BoW) and Bag of Concepts (BoC) models to retrieve the relevant contextual information. The information source utilized to retrieve the information is a Linked Data resource, DBpedia, which encodes large amounts of knowledge corresponding to Wikipedia in a structured form as triples. The experiments utilizes the QALD question sets consisted of training and testing sets each containing 100 questions. The results from these experiments shows that pragmatic aspects, which are often neglected by BoW (syntactic models) and BoC (semantic models), form a critical part of contextual information selection.

Keywords: Contextual information, Semantic models, Syntactic models, DBpedia

1 Introduction

Question Answering (QA) systems mostly comprise of four steps; question processing, answer search, answer extraction, and answer presentation. These four steps collectively contribute for the overall performance of QA systems. Several studies have examined the first three steps, focusing on delivering correct answers as short statement facts [1–3]. This paper focuses on the last stage of QA systems aiming to present answers as if it was delivered by human being. This involves three aspects; searching for extra relevant contextual information, ranking and selecting it, and presenting it as text similar to human composed text.

Following the recent roadmap proposed by Mendes and Coheur [4], this study presents peripheral contextual information for factoid questions which require factual answers. We study two main factoid question types (single entity and

multiple entity) apply various models to retrieve the contextual information. The paper examines the performance of syntactic and semantic models to retrieve peripheral contextual information for both aforementioned question types implemented on a generic framework targeting QA systems.

The paper is structured as follows. Section 2 introduces the triple weighting methods that are considered in the research. The section explores the process of adopting various BoW and BoC models to retrieve contextual information to enrich answers. In Section 3, we present the experimental framework with results and discuss the findings from the experiment. Section 4 describes the relevant related work. Section 5 concludes the paper with an outlook on future work.

2 Content selection using weighted triples

This section presents models to rank triples focusing on open domain questions as communicative goals. Our objective is to select a set of triples from a linked data resource (i.e. DBpedia) which can be used to generate a more informative answer for a given question. We investigate the problem from two perspectives; as a Bag of Words (BoW) and as a Bag of Concepts (BoC).

2.1 Problem as a Bag of Words

Token similarity In this approach triples are ranked by calculating the cosine similarity between the question/answer and the triple. Both question/answer and triples are tokenized and the cosine similarity was computed using (1).

$$sim_{cosine}(\overrightarrow{Q}, \overrightarrow{T}) = \frac{\overrightarrow{Q} \cdot \overrightarrow{T}}{|\overrightarrow{Q}||\overrightarrow{T}|} = \frac{\Sigma_{i=1}^{n} Q_i T_i}{\sqrt{\Sigma_{i=1}^{n} Q_i^2}\sqrt{\Sigma_{i=1}^{n} T_i^2}} \tag{1}$$

Here, Q and T represent the question and the triple respectively.

Term Frequency – Inverse Document Frequency (TF-IDF) In our problem we considered the triple collection as a document collection and the query was provided as an augmented domain corpus. The TF-IDF is then able to provide a rank to each term (t) present in the triple (T) compared to the rest of the triples. The weight of a triple is the sum of weights assigned to all the terms present in the triple. The TF-IDF takes a document collection and rank each document based on the presence of query terms. The TF-IDF can be explained as follows:

$$TF - IDF(Q, T) = \sum_{i \in Q,T} tf_i.idf_i = \sum_{i \in Q,T} tf_i.log_2 \frac{N}{df_i} \tag{2}$$

Where tf represents the term frequency, N stands for number of documents in the collection and df is the number of documents with the corresponding term. Q represents the question, however in our experiment we tested the possibility of utilizing a domain corpus instead of the original question or the question with the answer.

Okapi BM25 The Okapi ranking function can be defined as follows:

$$Okapi(Q,T) = \sum_{i \in Q,T} \left[log \frac{N}{df_i} \right] . \frac{(k_1+1)\, tf_{i,T}}{k_1 \left((1-b) + b \left(\frac{L_T}{L_{ave}} \right) \right) + tf_{i,T}} . \frac{(k_3+1)\, tf_{i,Q}}{k_3 + tf_{i,Q}} \tag{3}$$

Where, L_T and L_{ave} represent the length of the triple and average of length of a triple. The Okapi also uses set of parameters where b is usually set to 0.75 and k_1 and k_3 are ranging between 1.2 and 2.0. The and k_3 can be determined through optimization or can be set to range within 1.2 and 2.0 in the absence of development data.

Residual Inverse Document Frequency (RIDF) The idea behind the RIDF is to find content words based on actual IDF and predicted IDF. The widely used methods to IDF prediction is Poisson and K mixture. Since K mixture fits with term distribution very well, we modelled that lower the residual (between actual IDF and IDF predicted by K mixture), the term tends to be a content term. Given term frequencies in triple collection, predicted IDF can be used to measure the RIDF for a triple as follows:

$$RIDF = \sum_{i \in T} \left(idf_i - \widehat{idf_i} \right) = \sum_{i \in T} \left(idf_i - log \frac{1}{1 - P\left(0; \lambda_i\right)} \right) \tag{4}$$

Where λ_i represents the average number of occurrences of term and $P\left(0; \lambda_i\right)$ represents the Poisson prediction of *df* where term will not be found in a document. Therefore, $1 - P\left(0; \lambda_i\right)$ can be interpreted as finding at least one term and can be measured using:

$$P(k; \lambda_i) = e^{-\lambda_i} \frac{\lambda_i^k}{k!} \tag{5}$$

Based on the same RIDF concept, we can moderate this to work with term distribution models that fits well with actual df such as K mixture. The definition of the K-mixture is given below.

$$P(k; \lambda_i) = (1-\alpha)\delta_{k,0} + \frac{\alpha}{\beta+1} \left(\frac{\beta}{\beta+1} \right)^k \tag{6}$$

In K-mixture based RIDF we interpreted the deviation from predicated *df* to make the term as a non-content term.

2.2 Problem as a Bag of Concepts

Latent Semantic Analysis This method analysed how triples in the collection can be concept wise ranked and retrieved related to the question and answer where triples are represented in a semantic space. To retrieve the triples based on this new representation the question and answer must also be transformed

to the latent semantic space. Our initial experiment identified that the transformation of question and answer to latent semantic space cannot perform well for contextual information selection. Due to this fact in the experiment we used the augmented domain corpus as the query.

Corpus based Log Likelihood Distance The idea behind the implementation of this method is to identify domain specific concepts compared to the general concepts and rank triples which contain such concepts. For this we employed the domain corpus (see Section 2.3) and a general reference corpus (see Section 2.4). We utilized the log likelihood distance [5] to measure the importance as mentioned below:

$$w_t = 2 \times \left(\left(f_t^{dom} \times log \left(\frac{f_t^{dom}}{f\_exp_t^{dom}} \right) \right) + \left(f_t^{ref} \times log \left(\frac{f_t^{ref}}{f\_exp_t^{ref}} \right) \right) \right) \quad (7)$$

where, f_t^{dom} and f_t^{ref} represent frequency of term (t) in domain corpus and reference corpus respectively. Expected frequency of a term (t) in domain $(f\_exp_t^{dom})$ and reference corpora $(f\_exp_t^{ref})$ were calculated as follows:

$$f\_exp_t^{dom} = s_{dom} \times \left(\frac{f_t^{dom} + f_t^{ref}}{s_{dom} + s_{ref}} \right) \quad (8)$$

$$f\_exp_t^{ref} = s_{ref} \times \left(\frac{f_t^{dom} + f_t^{ref}}{s_{dom} + s_{ref}} \right) \quad (9)$$

where, s_{dom} and s_{ref} represent total number of tokens in domain corpus and reference corpus respectively. Next, we can calculate the weight of a triple ($\langle$subject, predicate, object$\rangle$) by summing up the weight assigned to each term of the triple

2.3 Domain Corpus

The domain corpus is a collection of text related to the domain of the question being considered. However, finding a corpus which belongs to the same domain as the question is challenge in its own. To overcome this, we have utilized a unsupervised domain corpus creation based on a web snippet extraction with the input as extracted key phrases from questions and answers.

2.4 Reference Corpus

The reference corpus is an additional resource utilized for the LLD based contextual information selection. In essence, to facilitate the LLD calculation to determine whether a term is important for particular domain, a balanced corpus is needed. The reference corpus represents a balanced corpus which contains text from different genres. We have used the British National Corpus (BNC) as the reference corpus.

2.5 Triple retrieval

The model employs the Jena RDF framework for the triple retrieval. We have implemented the Java library to query and automatically download necessary RDF files from DBpedia.

2.6 Threshold based selection

After associating each triple with calculated weight, we then have to limit the selection based on a particular cut-off point as the threshold (θ). Due to absence of knowledge to measure the θ at this stage, it is considered as a factor that needs to be tuned based on experiments. Further discussion on selecting the θ can be found in Section 4.

3 Experimental framework

3.1 Dataset

We used the QALD-2 training and test datasets and removed the invalid questions. The invalid questions include the questions marked as "out of scope" by dataset providers and questions where DBpedia triples do not exist. Table 1 provides the statistics of the dataset, including the distribution of questions in two different question categories, single entity and multiple entity questions.

We have also built a gold triple collection for each question. These gold triples were selected by analysing community provided answers for the questions in our dataset. Using this gold triples in the evaluation and statistics will be discussed in Section 3.2.

Table 1. Statistics related question dataset. Invalid questions are which are marked by dataset providers or questions where for which triples cannot be retrieved from DBpedia

	Training set	Test set
All questions	100	100
Invalid questions	5	10
Single entity questions	47	42
Multiple entity questions	48	48

3.2 Results and discussion

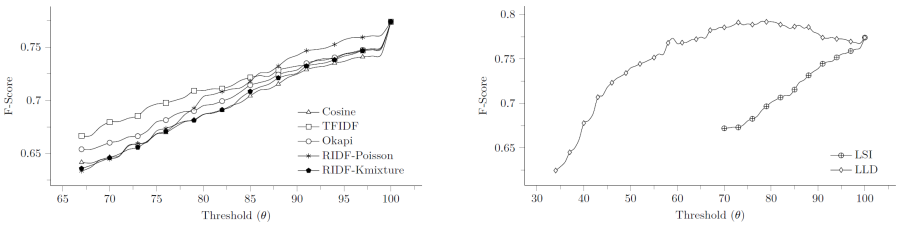
The gold standard evaluation method is utilized for the task [6, 7]. The training question set is used to measure the threshold (θ) which need to be used as the cut-off point for the ranked triples. The idea behind using this threshold value

is that the accurate model should rank all relevant triples higher compared to the irrelevant triples. Therefore, with the increase in θ for an accurate model (a model that rank all relevant triples higher than irrelevant), the precision will remain constant until it starts selecting the irrelevant triples and then it will gradually decrease. The recall will increase with θ and will be constant after it starts selecting irrelevant triples. Therefore, the θ which gives the highest F-score will be the turning point for both precision and recall.

Using the θ identified from training set, we can then test the model using testing dataset. When measuring the θ based on training dataset it is also important to measure the percentage of gold triples from the total triples. This is because if the percentage deviates from mean significantly, then it is hard to find a threshold value that can satisfy the entire question set. A set of statistics related to this calculation is shown in Table 4.

Table 2. Statistics related to the gold triple percentage in total triple collection in training dataset

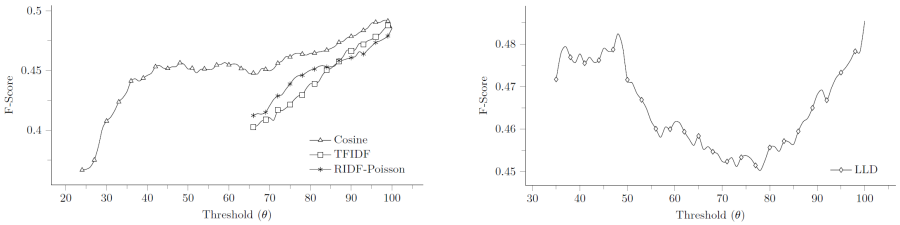
	μ	σ	Max%	Min%
Single entity type	68.89	4.28	78.79	63.58
Multiple entity type	30.43	3.88	37.06	22.93



(a) Bag of Words models (b) Bag of Concepts model

Fig. 1. F-score gained for single entity type questions using Bag of Words models(left) and Bag of Concepts models (right)

According to statistics shown in Table 2 it is clear that there is a possibility to find threshold values for both question sets. Fig. 1(a) and Fig. 1(b) depicts the evaluation performed on the single entity question category from training dataset, for both Bag of Words and Bag of Concepts models.Fig. 2(a) and Fig. 2(b) depicts the evaluation performed on the multiple entity question category from training dataset, for both Bag of Words and Bag of Concepts models.



(a) Bag of Words models (b) Bag of Concepts model

Fig. 2. F-score gained for multiple entity type questions using Bag of Words models(left) and Bag of Concepts models (right)

4 Related work

Benamara and Dizier [8] present the cooperative question answering approach which generates natural language responses for given questions. In essence, a cooperative QA system moves a few steps further from ordinary question answering systems by providing an explanation of the answer.

Bosma [9] incorporates the summarization as a method of presenting additional information in QA systems. He coins the term, an intensive answer to refer to the answer generated from the system. The process of generating intensive answer is based on summarization using rhetorical structures. Several other summarization based methods for QA such as Demner-Fushman and Lin [10], Yu et al. [11], and Cao et al. [12] also exist with different methods. However, the common drawback that they all shares is the inability to select cohesive information units (e.g., triples).

Vargas-Vera and Motta [13] present an ontology based QA system, AQUA. Although AQUA is primarily aimed at extracting answers from a given ontology, it also contributes to answer presentation by providing an enriched answer. The AQUA system extracts ontology concepts from the entities mentioned in the question and present those concepts in aggregated natural language.

5 Conclusion

This study has examined the role of syntactic and semantic models in contextual information selection for answer presentation. The results of this investigation show that although some semantic models (e.g., LLD) performs well for single entity based questions, in general, pragmatic aspects become more important for this task. However, as of our knowledge this is the first study that investigated the syntactic and semantic models in the contextual information selection to enrich answers as a method of presentation. In future, we expect to extend the work by integrating other possible methods to select contextual information. In addition to these extensions, the contextual information selection will be integrated to our Natural Language Generation (NLG) project [14–16] as the content selection module.

References

1. Perera, R.: Ipedagogy: Question answering system based on web information clustering. In: T4E-2012. (2012)
2. Perera, R.: Scholar: Cognitive Computing Approach for Question Answering. Honours thesis, University of Westminster (2012)
3. Perera, R., Nand, P.: Interaction history based answer formulation for question answering. In: KESW-2014. (2014) 128–139
4. Mendes, A.C., Coheur, L.: When the answer comes into question in question-answering: survey and open issues. Natural Language Engineering **19**(01) (January 2013) 1–32
5. Gelbukh, A., Sidorov, G., Lavin-Villa, Eduardo Chanona-Hernandez, L.: Automatic Term Extraction Using Log-Likelihood Based Comparison with General Reference Corpus. In: Natural Language Processing and Information Systems. Springer Berlin Heidelberg (2010) 248–255
6. Perera, R., Nand, P.: Real text-cs - corpus based domain independent content selection model. In: ICTAI-2014. (2014) 599–606
7. Perera, R., Nand, P.: The role of linked data in content selection. In: PRICAI-2014. (2014) 573–586
8. Benamara, F., Dizier, P.S.: Dynamic generation of cooperative natural language responses in webcoop. In: 9th European Workshop on Natural Language Generation, Budapest, Hungary, ACL (2003)
9. Bosma, W.: Extending answers using discourse structure. In: Recent Advances in Natural Language Processing, Borovets, Bulgaria, Association for Computational Linguistics (2005)
10. Demner-Fushman, D., Lin, J.: Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06, Morristown, NJ, USA, Association for Computational Linguistics (July 2006) 841–848
11. Yu, H., Lee, M., Kaufman, D., Ely, J., Osheroff, J.A., Hripcsak, G., Cimino, J.: Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. Journal of Biomedical Informatics **40** (2007) 236–251
12. Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J.J., Ely, J., Yu, H.: AskHERMES: An online question answering system for complex clinical questions. Journal of Biomedical Informatics **44** (2011) 277–288
13. Vargas-Vera, M., Motta, E.: Aqua-ontology-based question answering system. In: Mexican International Conference on Artificial Intelligence, Mexico City, Mexico, Springer-Verlag (2004)
14. Perera, R., Nand, P.: A multi-strategy approach for lexicalizing linked open data. In: CICLing-2015. (2015) 348–363
15. Perera, R., Nand, P.: Realtext-lex: A lexicalization framework for linked open data. In: ISWC-2015 Demonstration. (2015)
16. Perera, R., Nand, P.: Generating lexicalization patterns for linked open data. In: Second Workshop on Natural Language Processing and Linked Open Data collocated with 10th Recent Advances in Natural Language Processing (RANLP). (2015)