The Role of Linked Data in Content Selection

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. This paper explores the appropriateness of utilizing Linked Data as a knowledge source for content selection. Content Selection is a crucial subtask in Natural Language Generation which has the function of determining the relevancy of contents from a knowledge source based on a communicative goal. The recent online era has enabled us to accumulate extensive amounts of generic online knowledge some of which has been made available as structured knowledge sources for computational natural language processing purposes. This paper proposes a model for content selection by utilizing a generic structured knowledge source, DBpedia, which is a replica of the unstructured counterpart, Wikipedia. The proposed model uses log likelihood to rank the contents from DBpedia Linked Data for relevance to a communicative goal. We performed experiments using DBpedia as the Linked Data resource using two keyword datasets as communicative goals. To optimize parameters we used keywords extracted from QALD-2 training dataset and QALD-2 testing dataset is used for the testing. The results was evaluated against the verbatim based selection strategy. The results showed that our model can perform 18.03% better than verbatim selection.

Keywords: Linked Data, Content selection, Log likelihood distance, Text mining

1 Introduction

One of the fundamental tasks in Natural Language Generation (NLG) is deciding the content which is optimally relevant for a purpose. The main goal of content selection is to retrieve relevant content given a communicative goal and a given knowledge source [1]. The communicative goal specifies the objective that final content must satisfy. The knowledge source is the one that content selection will use to retrieve the required content. Therefore, knowledge source must contain general knowledge about different domains. Early content selection models utilized purposely built special knowledge bases to perform content selection. These knowledge bases are very domain specific, since the underlying process of creating a knowledge base is usually application specific in order to save the required effort. Compared to traditional knowledge bases, web today has accumulated vast amounts of knowledge with a vastly wide domain coverage. However, a large proportion of the online knowledge is unstructured which presents challenges for machine processing.

Recently, there has been a rise of the concept of Linked Data, which is archiving of the unstructured knowledge from the web into a more structured form. Linked Data is represented in the form of Resource Description Framework (RDF) which utilizes the triple data structure. A triple consists of a subject, a predicate and an object. A Linked Data resource is essentially a collection of such triples which may also be organized in a hierarchy. Linked Data statistics reports [2] show that Linked Data cloud has grown from 28 billion triples to 31 billion triples within 6 months. This shows the rapidly evolving nature of Linked Data cloud which is increasingly making it an attractive choice as a knowledge source for NLP applications.

This paper explores the opportunity of utilizing ever growing Linked Data cloud for the content selection task. The model we propose in this paper aligns well with the dynamic and evolving nature of the Linked Data cloud. The model is being developed as a part of a larger project which uses NLG in Question Answering (QA). This is the motivation for using keywords extracted from question dataset as the representation of the communicative goal. The selection of content from the Linked Data is carried out using a ranking function which utilizes log likelihood distance calculated based on term frequencies from two corpora, a domain corpus and a general reference corpus. The ranking function assigns a weight to each triple and then a threshold based selection is performed to select the final content. The model also has utility functions to retrieve, verbalize and filter the triples.

The remainder of the paper is structured as follows. In Section 2 we explain the Linked Data resource utilized for this research. Section 3 describes the proposed model in detail. Section 4 describes the results from experiments. Section 5 evaluates the proposed approach against other similar work from the literature and finally Section 6 concludes the paper with an outlook on future work.

2 DBpedia as a Linked Data resource

DBpedia[1] is one of the main Linked Data resources that extracts structured information from Wikipedia data. It is based on an automatic information extraction framework which makes the extracted information available to the public via a free SPARQL Protocol and RDF Query Language (SPARQL) based endpoint to access the RDF data. Currently DBpedia contains knowledge for about 4.58 million entities and categorized under 685 entity classes (e.g., person, organization, place). It hosts 3 billion triples out of which 580 million triples are extracted from Wikipedia English version. Table 1 gives various statistics about

[1] http://dbpedia.org/About

DBpedia as well as two other similar sized Linked Data resources, Freebase[2] and Yago[3].

Table 1. Comparison of DBpedia statistics with Freebase and Yago

Triple store	Entities	Triples	Ontology classes	Query language
DBpedia	4.58 million	3 billion	685	SPARQL
Freebase	44 million	2.4 billion	40616	MQL
YAGO	10 million	120 million	451708	SPARQL

Compared to other Linked Data resources, DBpedia contains more triples categorized under strong class hierarchy which makes it a strong choice as a knowledge source. Although, Freebase has higher number of topics (entities), it does not have as many triples raising the issue of data sparsity. In addition, Yago is still in its initial phases of archiving Linked Data compared to DBpedia.

Table 2 shows the growth rate of DBpedia over 5 releases up to 2014, based on DBpedia release notes[4]. It shows a rapid growth in volume which was another factor which was considered when choosing a knowledge source for content selection. The combination of these factors led us to select DBpedia as a Linked Data resource for content selection as well as for the wider the NLG project.

Table 2. DBpedia growth rate in last 5 releases. Number of entities, triples and ontology classes are considered.

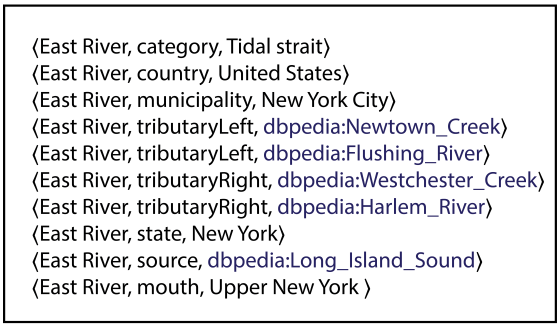
Release version	Entities	Triples	Ontology classes
2014	4.58 million	3 billion	685
3.9	4.26 million	2.46 billion	529
3.8	3.77 million	1.89 billion	359
3.7	3.64 million	1 billion	320
3.6	3.5 million	672 million	272

Each triple in DBpedia has an entity as the subject, a property as the predicate and a Uniform resource Indicator (URI) or a literal value as the object. A sample set of triples is shown in Fig. 1 which were taken from DBpedia entity for *East River*.

[2] `https://www.freebase.com/`

[3] `https://www.mpi-inf.mpg.de/yago-naga/yago/`

[4] `http://blog.dbpedia.org/category/dataset-releases/`



⟨East River, category, Tidal strait⟩
⟨East River, country, United States⟩
⟨East River, municipality, New York City⟩
⟨East River, tributaryLeft, dbpedia:Newtown_Creek⟩
⟨East River, tributaryLeft, dbpedia:Flushing_River⟩
⟨East River, tributaryRight, dbpedia:Westchester_Creek⟩
⟨East River, tributaryRight, dbpedia:Harlem_River⟩
⟨East River, state, New York⟩
⟨East River, source, dbpedia:Long_Island_Sound⟩
⟨East River, mouth, Upper New York ⟩

Fig. 1. Sample triples for *East River* retrieved from DBpedia

3 Content selection model for Linked Data

Since Linked Data cloud grows rapidly with community effort, systems that consume this data must also be dynamic and should work with minimum preprocessing. Rule based approaches that were used in early content selection models are not suitable for models based on such as evolving Linked Data resource.

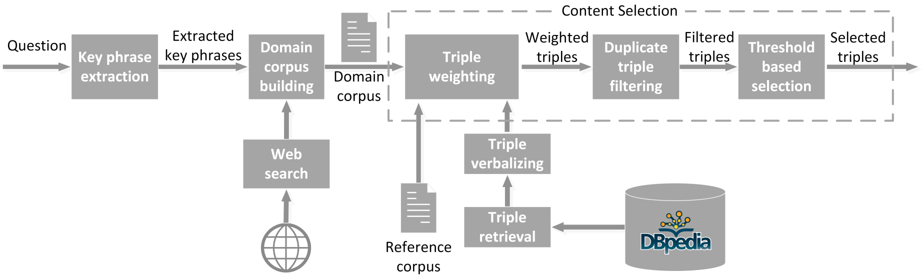


Fig. 2. Schematic representation of the model

The proposed model incorporated the evolving nature of the knowledge source which works with minimum preprocessing. The whole process of content selection is designed as an automatic pipeline model. The schematic representation of the proposed model is depicted in Fig. 2. As the proposed model is designed as a part of a QA system, it takes questions as the initial input. The key phrases from the question are extracted using a predetermined rule set. The resulting key phrases are then treated as the communicative goal for the content selection model. Based on these key phrases content selection model should then be able to select the relevant content. The selection process uses a weighting function to apply weight to each retrieved triple. The weighting function utilizes a combination of general reference corpus and a domain corpus to calculate the frequencies of terms in a triple. If a term has high frequency in domain corpus, but low frequency in general corpus, then it implies that the term is an important term for the domain being considered and thus the weighting function applies high weight for that term. This is performed for terms that appear in the triples retrieved. Finally, all weights assigned for the terms in the triple are summed up to calculate the final weight for the triple.

3.1 Content selection unit

Content selection unit is comprised of three modules:

- *Triple weighting* module assigns a weight for each triple based on their importance to the domain represented by extracted set of key phrases
- *Duplicate triple filtering* module is responsible for eliminating duplicate triples appearing in the retrieved content
- *Threshold based selection* module uses an experimentally determined threshold value to select the finalized content

The following sections describe the process of aforementioned modules in detail.

Triple weighting Triple weighting step assigns the weight for each triple retrieved. This weight represents how important the triple is to the domain represented by the set of key phrases used. Basically, the process first tokenizes the triple into terms and then applies weight value to each term. These values are then summed up to get the total value for the whole triple. Prior to this calculation all stop words are removed from the triple, because existence of stop words can disturb the term weighting. The proposed model utilizes the log likelihood distance [3–5] to calculate weights based on two corpora; a domain corpus and a general reference corpus. The domain corpus represents the knowledge specific to the domain represented by the key phrases. The domain corpus changes with set of key phrases being processed. Therefore, the model uses dynamic process to build a domain corpus relevant to each set of key phrases. Section 3.3 provides the detailed look into the domain corpus building task. The general reference corpus contains general knowledge about various domains. The selection and the usage of the general reference corpus is described in Section 3.3.

The following is the Formula used to calculate the weight (W_t) for a term (t) in a triple (T):

$$W_t = 2 \times \left(\left(f_t^{dom} \times \log \left(\frac{f_t^{dom}}{f\_exp_t^{dom}} \right) \right) + \left(f_t^{ref} \times \log \left(\frac{f_t^{ref}}{f\_exp_t^{ref}} \right) \right) \right) \quad (1)$$

where, f_t^{dom} and f_t^{ref} represent frequency of t in domain corpus and general reference corpus respectively. To calculate the expected frequency of a term in domain corpus ($f\_exp_t^{dom}$) and reference corpus ($f\_exp_t^{ref}$) following Formulas were used:

$$f\_exp_t^{dom} = s_{dom} \times \left(\frac{f_t^{dom} + f_t^{ref}}{s_{dom} + s_{ref}} \right) \tag{2}$$

$$f\_exp_t^{ref} = s_{ref} \times \left(\frac{f_t^{dom} + f_t^{ref}}{s_{dom} + s_{ref}} \right) \tag{3}$$

where, s_{dom} and s_{ref} represent total number of tokens in domain corpus and reference corpus respectively. All of the above calculations were based on stop words removed from the text. A rule based string filtering process was used in stop words removal using the 429 extended list of stop words mentioned in Onix text retrieval reference sheet[5].

Once the weight for each term in triple is calculated, the final weight for the triple is assigned as follows:

$$W_T = \sum_{t \in T} W_t \tag{4}$$

where, W_T is the weight of triple T.

The triples are then sorted based on weight from highest to the lowest after calculation of weights of the triples. The triples at the higher weighting represent the most relevant content and the ones at the bottom represent more general content.

Duplicate triple filtering One of the challenges that needs to be overcome when using Linked Data is the process of managing duplicate knowledge. Since Linked Data concept emphasizes the interlinking between different data sources and even within the same data source, number of duplicate knowledge that can appear in content selection is considerably higher than using a traditional knowledge base. For instance, DBpedia resource for *George W. Bush* contains the triple ⟨George W. Bush, spouse, Laura Bush⟩, while DBpedia resource for *Laura Bush* contains the triple ⟨Laura Bush, spouse, George W. Bush⟩. Essentially, these two triples have the same knowledge and are actually duplicates. However, there can be scenarios which are more complex than this. Triples like ⟨Microsoft, founded by, Bill Gates⟩ and ⟨Bill Gates, founder, Microsoft⟩ are also considered as duplicate knowledge.

To address this challenge, proposed model uses a two-step process to filter duplicates. In first step, we consider triples with exact knowledge where triples have exactly the same content. In these scenarios, we remove all others keeping only one triple. Since all triples have the same weight (as same set of terms are present), it is not significant which one is kept.

Next, the model considers partial matches where subject and object contains similar knowledge, but predicates are different. In such cases, WordNet similarity [6] is calculated for the predicate. Through, experimentation we have found

[5] `http://www.lextek.com/manuals/onix/stopwords1.html`

that a value of 0.25 threshold WordNet similarity factor. Therefore, if triples have similarity factor greater than 0.25 between predicates, we remove all others keeping the triple with the highest weight.

Threshold based selection In this step, a set of triples need to be selected as the final content. Though, all triples are assigned a weight value and subsequently sorted, there is no way to determine the optimum threshold value as the cut-off point for selection. We left this as a factor to be determined experimentally. The experiment and the detailed overview of the optimum threshold value determined was based on the training dataset as described in Section 4.

3.2 Key phrase extraction

Key phrases are the communicative goal used by the content selection model. These key phrases were extracted from a question taken as an input. The extraction process was based on predetermined set of rules. First, the question was Part-Of-Speech (POS) tagged using Stanford POS tagger. Then the following rules were used to identify key phrases based on assigned POS tags:

- Adjacent noun (NN), singular proper noun (NNP) or plural noun (NNS) phrases
- Adjective (JJ) with a noun phrase (NN/NNP/NNS)
- Comparative adjective (JJR) with a noun phrase (NN/NNP/NNS)
- Superlative adjective (JJS) with a noun phrase (NN/NNP/NNS)

3.3 Corpora selection

The triple weighting process was based on two corpora; a domain corpus and a reference corpus. In this section we describe the process of preparing these corpora to use with the triple weighting process.

Domain corpus Domain corpus is the one that represents domain knowledge related to the extracted key phrases. Due to this fact a separate domain corpus needs to be used for each set of key phrases extracted from the questions. Finding this type of corpora which contains a corpus related to key phrases is difficult. This issue is addressed by building a domain corpus dynamically which is related to the set of key phrases under consideration.

First, extracted key phrases were used to search the web using Bing[6] search Application Programming Interface (API) . The main reason that we used the Bing Search API is its flexibility in querying which provides 5000 transaction for a month compared to Google search API which is now deprecated. This search process returns three facts for each search result; website Uniform Resource Locator (URL), title of the web page, and a snippet of text which is an extracted

[6] http://www.bing.com/toolbox/bingsearchapi

passage of text containing the key phrases used to search. The domain corpus was built using the snippets of text returned through this search. Essentially, the model aggregates these snippets and builds the domain corpus for each set of key words.

The corpus built based on aforementioned process generally has a limited textual content. Therefore, it raised the issue that whether this text need to be enriched, because this text is used in the term weighing process which searches terms within domain corpus. Natural Language Processing (NLP) defines several ways of enriching text. Since, the domain corpus is analysed in token level during the weighting process, we decided to look into the token level text enrichment methods. Two widely used such methods are lemmatizing based methods, and associating words with near-synonyms using Synsets derived from lexical resources such as WordNet [7].

Lemmatization is the process of finding the base form of word from inflected forms of words. For an example, the word *develop* has inflected forms; *develops*, *developed*, *developing*, and *developer*. All inflected forms of a word are based on lemma, the base form of the word. However, in this context we consider adding all inflected forms as well as the base form of the word to the text as a method of enriching the text.

Associating words with near-synonyms using Synsets is a widely discussed in topic areas such as lexical choice. The applicability of this model to our approach is to enrich the textual content with related words. Recent literature has shown that textual enrichment can be further improved using background information [8]. For simplicity, in this paper we explore only base level textual enrichment methods.

However, both lemmatization and Synset based textual enrichments were already present in the employed web search module. Hence, no further effort was expended towards further enriching the domain corpus.

General reference corpus General reference corpus should contain general knowledge covering range of domains. Such corpora are referred as balanced corpora in corpus linguistics. Examples of such balanced corpora are, the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), and Sinica Corpus.[7] The proposed model uses BNC which is a 100 million word collection of text from different genres. However, all stop words are filtered from the corpus as they introduce noise in the model. The resulting corpus contains 52.3 million words.

However, the model only needs word frequencies of the BNC. Therefore, we performed a unigram analysis on BNC and extracted all words with their related frequencies. The resulting list contains 207406 unique tokens and these tokens together with their frequencies were stored in an indexed embedded database for efficient access.

[7] `http://www.natcorp.ox.ac.uk/`, `http://corpus.byu.edu/full-text/intro.asp`, `http://rocling.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm`

3.4 Triple retrieval and verbalizing

The model uses Jena framework[8] to retrieve triples from the DBpedia. The DBpedia utilizes two types of specifications to group triples; DBpedia properties (dbprop) and mapped ontology web properties (owl). Our model only uses DBpedia properties specification which provides broad range of knowledge compared to mapped properties.

4 Experimental framework

The experimental framework was designed to achieve two objectives. Firstly, a specific threshold value needs to be identified to limit the triple selection as described in Section 3.1. Secondly, the model needs to be evaluated to check whether it can select the necessary content for a given question.

4.1 QALD dataset

Since the framework described here forms part of a larger framework for NLG in QA, we used a question dataset as the input to the model. There are variety of question dataset series such as TREC QA [9] , QALD [10], and Wikipedia QA [11]. Among these datasets, QALD is the only dataset series that is based on Linked Data and use DBpedia as a source for retrieving answers. Due to this, the experiments used the QALD-2 datasets which contains training and testing datasets each containing 100 questions. However, we have eliminated erroneous questions (empty questions and out of scope questions) marked by dataset providers. The resulting training and testing sets contained 93 and 96 questions respectively.

4.2 Experimental settings and results

The evaluation was based on gold standard process that is introduced for content selection through various early researches [12, 13]. The selection of gold triples was accomplished by analysing triples mentioned in the community provided answers for the QALD questions. We crawled four different question answering sites to acquire these answers: Yahoo! Answers, Answers.com, WikiAnswers, and AnswerBag.[9] The precision (P), recall (R), and F-measure (F*) were used as metrics for the evaluation. The descriptions of these metrics related to the context of evaluation is given below:

$$P = \frac{|triples_{selected} \cap triples_{gold}|}{|triples_{selected}|} \tag{5}$$

[8] https://jena.apache.org/

[9] https://answers.yahoo.com/, http://www.answers.com/, http://wiki.answers.com/, http://www.answerbag.com/

$$R = \frac{|triples_{selected} \cap triples_{gold}|}{|triples_{gold}|} \tag{6}$$

$$F^{*} = \frac{2PR}{P+R} \tag{7}$$

QALD-2 training set was used to tune the threshold parameter and QALD-2 test set was used to evaluate the results. Furthermore, we used verbatim selection process as a method to compare our results with. In essence, verbatim selection determines content to be selected by analysing what triples are mentioned in the domain corpus as verbatim. For instance, if ⟨Bill Gates, founder, Microsoft⟩ triple will be selected if domain corpus contains text "Bill Gates is the founder of Microsoft and served as the CEO for 10 years".

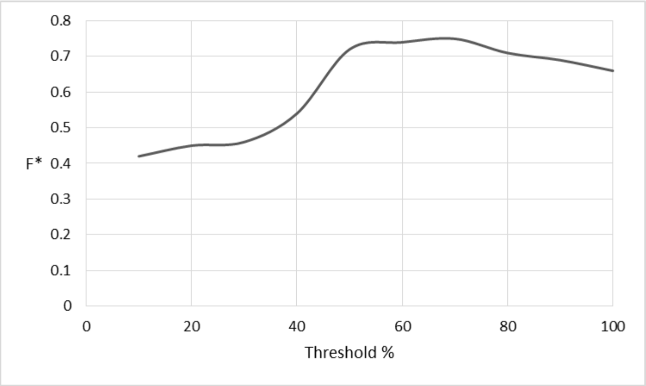


Fig. 3. Average F* vs threshold for QALD-2 training dataset

Fig. 3 shows the results of the experiment carried put to determine the best threshold value for limit triple selection. In this experiment average F-measure was assessed against the threshold values using QALD-2 training dataset and average domain corpus was 1361 words.

Table 3. Statistics about DBpedia resources and triples processed for both training and testing phases

	Training dataset	**Testing dataset**
Number of DBpedia resources	458	482
Similar triples identified	78	91
Invalid triples (Predetermined)	1827	1920
Invalid triples (Selected)	41	28

Table 3 summarizes the statistics collected about DBpedia resources and triples processed for both QALD-2 training and testing datasets. The number of DBpedia resources gives an idea about the amount of RDF files that were processed for the entire question set. The triples identified as similar by the WordNet similarity function are also shown for both datasets. Apart from similar triples, the number of invalid triples were also filtered and recorded by the triple retrieval process. The triple retrieval process used a predetermined rule set to filter invalid triples, which was based on prior observations.

Table 4. Comparison of F* between Verbatim based selection and our approach using QALD-2 test dataset

	Verbatim based selection	**Our approach**
Average F*	0.61	0.72

Table 4 shows the evaluation results of our approach compared to the verbatim based selection. This comparison was based on QALD-2 test dataset and threshold value was set to 68%. The average domain corpus size for this evaluation was 1393 words.

4.3 Observations and discussions

The first experiment (see Fig. 3) showed that the best F-measure value can be achieved in 60-76% threshold range. Due to this, the latter experiments were based on an average threshold value of 68%. We observed that with an increase in threshold value after 60-76% value range, F-measure starts to do down due to decreasing precision. This is because with the increased threshold value, triples that are not included in the gold triple set start to get selected as the relevant content.

Statistics collected during processing DBpedia resources show that both question datasets utilized multiple resources. This is mainly due to the nature of QALD-2 question dataset which utilizes linked data for the answers to the questions. It was also observed that compared to the size of the resources utilized, the number of similar triples identified were very few. Further, there were no duplicate triples in the finalized content. All identified duplicate triples were true positives which were removed. This shows that the duplicate filtering using WordNet similarity function was effective in removing duplicates. The results also show that the threshold value used was effective in raising the overall accuracy.

Based on comparisons in Table 4, the proposed method performed 18.03% better than the verbatim based selection. Verbatim based selection selects content by finding what triples are mentioned as text in domain corpus, compared to the proposed model that selects the content by allocating weight for each triple. The results show that the proposed model outperforms the verbatim model by a significant percentage. This can be clearly attributed to the exploitation of the combined knowledge based on domain specific information as well as general information.

5 Related work

In the past several years, content selection has been tried in various NLG application with a range of approaches. Some of them have used Linked Data as a knowledge source and some have used different knowledge sources. In this section, we specifically discuss approaches that use Linked Data as a knowledge source, but also discuss whether other approaches can work with Linked Data as a knowledge source.

The shared task organized in European Workshop on Natural Language Generation (ENLG) [13] was a recently held significant event that emphasized the use of Linked Data in content selection. The two systems participated in this event were common ground principal based model by Kutlak et al. [14] and heuristic based approach by Venigalle and Eugenio [15]. The basis for the model proposed by Kutlak et al. [14] was that selected content must contain knowledge that is common. They have used Google custom web search hit count as a measurement in analysing commonality of knowledge. The heuristic model by Venigalle and Eugenio [15] is based on inducing rules based on co-occurrence of predicates. Both approaches used FreeBase as the knowledge source and they also mentioned that the main challenge they faced was with verbalizing FreeBase triples. Compared to these two, proposed model in this paper utilizes a more advanced weighting function and filtering mechanism rather than simple heuristics. Furthermore, we employ DBpedia as the knowledge source which is more organized, has much more richer content and does not suffer from data sparsity.

The approach presented by Doube and McKeown [16] uses language modelling (cross entropy) and statistical analysis to retrieve content from Linked Data. The model is based on determining rules for a specific domain and then selecting content using induced rules. However, the main drawback in this model is that for each domain needs a separate set of rules need to be collected. Bouayas-Agha and Wanner [17] also present a rule based method which selects the content from an ontology, with a relevance criteria which can partially automate the process of selection. Still, the domain adaptability remains as an issue in this model. In comparison the model proposed in this paper selects the content on the fly without pre-processing. This fact makes this model suitable for applications which work with range of domains such as open domain QA.

6 Conclusions and future work

This paper explored the suitability of Linked Data used as a knowledge source for content selection task. We presented the results for the use of DBpedia as a Linked Data resource for content selection task as a part of a higher level QA application. The paper proposed a model which assigns a weight for each triple in DBpedia and select content for a set of key phrases derived from a question. The weight calculation was achieved through a log likelihood computation using a domain and a general reference corpus. The model was evaluated using gold standard based evaluation and the results was compared to the verbatim based selection. The evaluation results showed that the model performs 18.03% percent better than verbatim based selection.

In future experiments, we intend to examine linking multiple triple stores to perform content selection. In such a scenario we will need to filter duplicates with advanced functionalities. An immediate future work is to develop more comprehensive techniques to identify such duplicate knowledge present in Linked Data.

Further, the next phases of this research will also explore other types of triple weighting functions that can jointly calculate weight for triples with even better accuracy. For this, we will consider variations of existing content selection methods as well as different term weighting approaches.

References

1. Reiter, E., Dale, R.: Building natural language generation systems. Cambridge University Press (January 2000)
2. Jentzsch, A., Cyganiak, R., Bizer, C.: State of the LOD Cloud. Technical report, Hasso-Plattner-Institut, Potsdam-Babelsberg (2011)
3. Paul Rayson, Damon Berridge, B.F.: Extending the Cochran rule for the comparison of word frequencies between corpora. In: 7th International Conference on Statistical analysis of textual data. (2004)
4. He, T., Zhang, X., Xinghuo, Y.: An Approach to Automatically Constructing Domain Ontology. In: Pacific Asia Computational Linguistics, Wuhan (2006) 150–157
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. Pedersen, P.: WordNet::Similarity - Measuring the Relatedness of Concepts. In: Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics, Boston (2004) 38–41
7. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM **38**(11) (1995) 39–41
8. Penas, A., Hovy, E.: Semantic enrichment of text with background knowledge. In: NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, Los Angeles, Association for Computational Linguistics (June 2010) 15–23
9. Voorhees, E., Tice, D.: Building a Question Answering Test Collection. In: ACM Special Interest Group on Information Retrieval Conference, Athens, Greece, ACM Press (2000)
10. Unger, C.: Question Answering Over Linked Data. Technical report, Bielefeld University, Heraklion, Greece (2012)

11. Smith, N., Heilman, M., Hwa, R., Cohen, S., Gimpel, K.: Question-Answer Dataset. Technical report, Carnegie Mellon University, Pennsylvania, USA (2013)
12. Bouayad-Agha, N., Casamayor, G., Wanner, L., Mellish, C.: Content selection from semantic web data. In: Seventh International Natural Language Generation Conference, Utica IL, USA, Association for Computational Linguistics (May 2012) 146–149
13. Bouayad-Agha, N., Casamayor, G., Wanner, L., Mellish, C.: Overview of the First Content Selection Challenge from Open Semantic Web Data. In: Proceedings of the 14th European Workshop on Natural Language Generation, Sofia, Bulgaria, Association for Computational Linguistics (August 2013) 98–102
14. Kutlak, R., Mellish, C., van Deemter, K.: Content Selection Challenge - University of Aberdeen Entry. In: Fourteenth European Workshop on Natural Language Generation, Sofia, Bulgaria, Association for Computational Linguistics (August 2013) 208–209
15. Venigalla, H., Eugenio, B.D.: UIC-CSC: The Content Selection Challenge Entry from the University of Illinois at Chicago. In: Proceedings of the 14th European Workshop on Natural Language Generation, Sofia, Bulgaria, Association for Computational Linguistics (August 2013) 210–211
16. Duboue, P.A., McKeown, K.R.: Statistical acquisition of content selection rules for natural language generation. In: Proceedings of the 2003 conference on Empirical methods in natural language processing. Volume 10., Morristown, NJ, USA, Association for Computational Linguistics (July 2003) 121–128
17. Bouayad-Agha, N., Casamayor, G., Wanner, L.: Content selection from an ontology-based knowledge base for the generation of football summaries. In: Thirtheenth European Workshop on Natural Language Generation, Nancy, France, Association for Computational Linguistics (September 2011) 72–81