

# Content-Based Table Retrieval for Web Queries

Zhao Yan<sup>†\*</sup>, Duyu Tang<sup>‡</sup>, Nan Duan<sup>‡</sup>, Junwei Bao<sup>+\*</sup>,  
Yuanhua Lv<sup>§</sup>, Ming Zhou<sup>‡</sup>, Zhoujun Li<sup>†</sup>

<sup>†</sup>Beihang University      <sup>‡</sup>Microsoft Research, Beijing, China

<sup>+</sup>Harbin Institute of Technology    <sup>§</sup>Microsoft AI and Research, Sunnyvale CA, USA

<sup>†</sup>{yanzhao, lizj}@buaa.edu.cn      <sup>+</sup>baojunwei001@gmail.com

<sup>‡§</sup>{dutang, nanduan, yuanhual, mingzhou}@microsoft.com

## Abstract

Understanding the connections between unstructured text and semi-structured table is an important yet neglected problem in natural language processing. In this work, we focus on content-based table retrieval. Given a query, the task is to find the most relevant table from a collection of tables. Further progress towards improving this area requires powerful models of semantic matching and richer training and evaluation resources. To remedy this, we present a ranking based approach, and implement both carefully designed features and neural network architectures to measure the relevance between a query and the content of a table. Furthermore, we release an open-domain dataset that includes 21,113 web queries for 273,816 tables. We conduct comprehensive experiments on both real world and synthetic datasets. Results verify the effectiveness of our approach and present the challenges for this task.

## 1 Introduction

Table<sup>1</sup> is a special and valuable information that could be found almost everywhere from the Internet. We target at the task of content-based table retrieval in this work. Given a query, the task is to find the most relevant table from a collection of tables. Table retrieval is of great importance for both natural language processing and information retrieval. On one hand, it could improve existing information retrieval systems. The well-organized information from table, such as product comparison from different aspects and flights between two

specific cities, could be used to directly respond to web queries. On the other hand, the retrieved table could be used as the input for question answering (Pasupat and Liang, 2015).

Unlike existing studies in database community (Cafarella et al., 2008; Balakrishnan et al., 2015) that utilize surrounding text of a table or pagerank score of a web page, we focus on making a thorough exploration of table content in this work. We believe that content-based table retrieval has the following challenges. The first challenge is how to effectively represent a table, which is semi-structured and includes many aspects such as headers, cells and caption. The second challenge is how to build a robust model that measures the relevance between an unstructured natural language query and a semi-structured table. Table retrieval could be viewed as a multi-modal task because the query and the table are of different forms. Moreover, to the best of our knowledge, there is no publicly available dataset for table retrieval. Further progress towards improving this area requires richer training and evaluation resources.

To address the aforementioned challenges, we develop a ranking based approach. We separate the approach into two cascaded steps to trade-off between accuracy and efficiency. In the first step, it finds a small set (e.g. 50 or 100) of candidate tables using a basic similarity measurement. In the second step, more sophisticated features are used to measure the relevance between the query and each candidate table. We implement two types of features, including manually designed features inspired by expert knowledge and neural network models jointly learned from data. Both strategies take into account the relevance between query and table at different levels of granularity. We also introduce a new dataset *WebQueryTable* for table retrieval. It includes 21,113 web queries from search

\* Contribution during internship at Microsoft Research.

<sup>1</sup>[https://en.wikipedia.org/wiki/Table\\_\(information\)](https://en.wikipedia.org/wiki/Table_(information))

log, and 273,816 web tables from Wikipedia.


We conduct comprehensive experiments on two datasets, a real world dataset introduced by us, and a synthetic dataset WikiTableQuestions (Pasupat and Liang, 2015) which has been widely used for table-based question answering. Results in various conditions show that neural network models perform comparably with carefully designed features, and combining them both could obtain further improvement. We study the influence of each aspect of table for table retrieval, and show what depth of table understanding is required to do well on this task. Results show the difference between question and web query, and present future challenges for this task.

This paper has the following contributions. We develop both feature-based and neural network based approaches, and conduct thorough experiments on real world and synthetic datasets. We release an open-domain dataset for table retrieval.

## 2 Task Definition

We formulate the task of table retrieval in this section. Given a query  $q$  and a collection of tables  $T = \{t_1, \dots, t_N\}$ , the goal of table search is to find a table  $t_i$  that is most relevant to  $q$ .

major cities of netherlands



Query

Name	Population
Amsterdam , North Holland	741,636
Rotterdam , South Holland	598,199
The Hague , South Holland	474,292
Utrecht , Utrecht	290,529

Biggest Cities Netherlands - GeoNames

Headers

Cells

Caption

Figure 1: A example of query-table pair.

Typically, a query  $q$  is a natural language expression that consists of a list of words, such as “major cities of netherlands”. A table  $t$  is a set of data elements arranged by vertical columns and horizontal rows. Formally, we define a table as a triple  $t = \{\text{headers}, \text{cells}, \text{caption}\}$  that consists of three aspects. A table could have multiple headers, each of which indicates the property of a column and could be used to identify a column. A table could have multiple cells, each of which is a unit where a row and a column intersects. A table could have a caption, which is typically an explanatory text about the table. Figure 1 gives an example to illustrate different aspects of a table.

It is helpful to note that tables from the web are not always “regular”. We regard a table as a “regular” table if it contains header, cell and caption, and the number of cells in each row is equal to the number of header cells. In this work, we make a comprehensive study of table retrieval on regular tables, and would like to release benchmark datasets of good quality. It is trivial to implement heuristic rules so as to convert the irregular tables to regular one, so we leave it to the future work.

## 3 Approach Overview

In this section, we give an overview of the proposed approach. To build a system with high efficiency, we separate the task into two cascaded modules, including candidate table retrieval and table ranking. Candidate table retrieval aims to find a small set of tables, such as 50 or 100. These candidate tables will be further used in the table ranking step, which uses more sophisticated features to measure the relevance between a query and a table. In the following subsections, we will give the work-flow of candidate table retrieval and table ranking. The detailed feature representation will be described in the next section.

### 3.1 Candidate Table Retrieval

Candidate table retrieval aims to get a small candidate table set from the whole table set of large scale, which is hundreds of thousands in our experiment. In order to guarantee the efficiency of the searching process, we calculate the similarity between table and query with Okapi BM25 (Robertson et al., 1995), which is computationally efficient and has been successfully used in information retrieval. Specifically, we represent a query as bag-of-words, and represent table with plain text composed by the words from caption and headers. Given a query  $q = x_1, x_2, \dots, x_n$ , a table  $t$  and the whole table set  $T$ , the BM25 score of query  $q$  and table  $t$  is calculated as follows.

$$BM25(q, t) = \sum_{i=1}^n idf(x_i) \frac{tf(x_i, t) \cdot (k_1 + 1)}{tf(x_i, T) + k_1(1 - b + b \frac{|t|}{avg_{tl}})}$$

where  $tf(x_i, t)$  is the term frequency of word  $x_i$  in  $t$ ,  $idf(x_i)$  is its inverse document frequency,  $avg_{tl}$  is the average sequence length in the whole table set  $T$ , and  $k_1$  and  $b$  are hyper-parameters.

### 3.2 Table Ranking

The goal of table ranking is to rank a short list of candidate tables by measuring the relevance between a query and a table. We develop a feature-based approach and a neural network approach, both of them effectively take into account the structure of table. The details about the features will be described in next section. We use each feature to calculate a relevance score, representing the similarity between a query and a table from some perspective. Afterwards, we use LambdaMART (Borges, 2010), a successful algorithm for solving real world ranking problem, to get the final ranking score of each table.<sup>2</sup> The basic idea of LambdaMART is that it constructs a forest of decision trees, and its output is a linear combination of the results of decision trees. Each binary branch in a decision tree specifies a threshold to apply to a single feature, and each leaf node is real value. Specifically, for a forest of  $N$  trees, the relevance score of a query-table pair is calculated as follow,

$$s(q, t) = \sum_{i=1}^N w_i tr_i(q, t)$$

where  $w_i$  is the weight associated with the  $i$ -th regression tree, and  $tr_i(\cdot)$  is the value of a leaf node obtained by evaluating  $i$ -th tree with features  $[f_1(q, t), \dots, f_K(q, t)]$ . The values of  $w_i$  and the parameters in  $tr_i(\cdot)$  are learned with gradient descent during training.

## 4 Matching between Query and Table

Measuring the relevance between a query and a table is of great importance for table retrieval. In this section, we present carefully designed features and neural network architectures for matching between a query and a table.

### 4.1 Matching with Designed Features

We carefully design a set of features to match query and table from word-level, phrase-level and sentence-level, respectively. The input of a feature function are two strings, one query string  $q$  and one aspect string  $t_a$ . We separately apply each of the following features to each aspect of a table, resulting in a list of feature scores. As described in Section 2, a table has three aspects, including

<sup>2</sup>We also implemented a ranker with linear regression, however, its performance was obviously worse than LambdaMART in our experiment.

headers, cells and caption. We represent each aspect as word sequence in this part.

(1) **Word Level.** We design two word matching features  $f_{wmt}$  and  $f_{wmq}$ . The intuition is that a query is similar to an aspect of table if they have a large amount of word overlap.  $f_{wmt}$  and  $f_{wmq}$  are calculated based on number of words shared by  $q$  and  $t_a$ . They are also normalized with the length of  $q$  and  $t_a$ , calculated as follows,

$$f_{wmt}(t_a, q) = \frac{\sum_{w \in t_a} \delta(w, q) \cdot idf(w)}{\sum_{w' \in t_a} idf(w')}$$

$$f_{wmq}(t_a, q) = \frac{\sum_{w \in t_a} \delta(w, q) \cdot idf(w)}{\sum_{w' \in q} idf(w')}$$

where  $idf(w)$  denotes the inverse document frequency of word  $w$  in  $t_a$ .  $\delta(y_j, q)$  is an indicator function which is equal to 1 if  $y_j$  occurs in  $q$ , and 0 otherwise. Larger values of  $f_{wmt}(\cdot)$  and  $f_{wmq}(\cdot)$  correspond to larger amount of word overlap between  $t_a$  and  $q$ .

(2) **Phrase Level.** We design a paraphrase-based feature  $f_{pp}$  to deal with the case that a query and a table use different expressions to describe the same meaning. In order to learn a strong and domain-independent paraphrase model, we leverage existing statistical machine translation (SMT) phrase tables. A phrase table is defined as a quadruple, namely  $PT = \{\langle src_i, trg_i, p(trg_i|src_i), p(src_i|trg_i) \rangle\}$ , where  $src_i$  (or  $trg_i$ ) denotes a phrase, in source (or target) language,  $p(trg_i|src_i)$  (or  $p(src_i|trg_i)$ ) denotes the translation probability from  $src_i$  (or  $trg_i$ ) to  $trg_i$  (or  $src_i$ ). We use an existing SMT approach (Koehn et al., 2003) to extract a phrase table  $PT$  from a bilingual corpus. Afterwards, we use  $PT$  to calculate the relevance between a query and a table in paraphrase level. The intuition is that, two source phrases that are aligned to the same target phrase tend to be paraphrased. The phrase level score is calculated as follows, where  $N$  is the maximum n-gram order, which is set as 3, and  $src_{i,n}^{at}$  and  $src_{j,n}^q$  are the phrase in  $t_a$  and  $q$  starts from the  $i$ -th and  $j$ -th word with the length of  $n$ , and  $i \in \{1, \dots, |t_a| - n + 1\}$  and  $j \in \{1, \dots, |q| - n + 1\}$ .

$$f_{pp}(t_a, q) = \frac{1}{N} \sum_{n=1}^N \frac{\sum_{i,j} score(src_{i,n}^{at}, src_{j,n}^q)}{|t_a| - N + 1}$$

$$score(src_x; src_y) = \sum_{PT} p(tgt_k|src_x) \cdot p(src_y|tgt_k)$$

(3) **Sentence Level.** We design features to match a query with a table at the sentence level. We use CDSSM (Shen et al., 2014), which has been successfully applied in text retrieval. The basic computational component of CDSSM is sub-word, which makes it very suitable for dealing the misspelling queries in web search. The model composes sentence vector from sub-word embedding via convolutional neural network. We use the same model architecture to get query vector and table aspect vector, and calculate their relevance with cosine function.

$$f_{s1}(t_a, q) = \text{cosine}(\text{cdssm}(t_a), \text{cdssm}(q))$$

We train model parameters on WikiAnswers dataset (Fader et al., 2013), which contains almost 12M question-similar question pairs. In addition, since vector average is an intuitive way to compute sentence vector and does not induce additional parameters, we calculate another relevance score by representing a query and a table aspect with element-wise vector average. We use a publicly available word embedding which is released by Mikolov et al. (2013).

$$f_{s2}(t_a, q) = \text{cosine}(\text{vec\_avg}(t_a), \text{vec\_avg}(q))$$

## 4.2 Matching with Neural Networks

We present neural network models for matching a query with a table. As a table includes different aspects such as headers, cells and caption, we develop different strategies to measure the relevance between a query and a table from different perspectives. In this subsection, we first describe the model to compute query representation, and then present the method that measures the relevance between a query and each aspect.

A desirable query representation should be sensitive to word order as reversing or shuffling the words in a query might result in totally different intention. For example, “*list of flights london to berlin*” and “*list of flights berlin to london*” have different intentions. We use recurrent neural network (RNN) to map a query of variable length to a fixed-length vector. To avoid the problem of gradient vanishing, we use gated recurrent unit (GRU) (Cho et al., 2014) as the basic computation unit, which adaptively forgets the history and remembers the input, and has proven to be effective in sequence modeling (Chung et al., 2014). It recursively transforming current word vector  $e_t^q$  with the output vector of the previous step  $h_{t-1}$ .

$$\begin{aligned} z_i &= \sigma(W_z e_i^q + U_z h_{i-1}) \\ r_i &= \sigma(W_r e_i^q + U_r h_{i-1}) \\ \tilde{h}_i &= \tanh(W_h e_i^q + U_h (r_i \odot h_{i-1})) \\ h_i &= z_i \odot \tilde{h}_i + (1 - z_i) \odot h_{i-1} \end{aligned}$$

where  $z_i$  and  $r_i$  are update and reset gates of GRU. We use a bi-directional RNN to get the meaning of a query from both directions, and use the concatenation of two last hidden states as the final query representation  $v_q = [\vec{h}_n, \overleftarrow{h}_n]$ .

A table has different types of information, including headers, cells and caption. We develop different mechanisms to match the relevance between a query and each aspect of a table. An important property of a table is that randomly exchanging two rows or tow columns will not change the meaning of a table (Vinyals et al., 2015). Therefore, a matching model should ensure that exchanging rows or columns will result in the same output. We first describe the method to deal with headers. To satisfy these conditions, we represent each header as an embedding vector, and regard a set of header embeddings as external memory  $M_h \in \mathbb{R}^{k \times d}$ , where  $d$  is the dimension of word embedding, and  $k$  is the number of header cells. Given a query vector  $v_q$ , the model first assigns a probability  $\alpha_i$  to each memory cell  $m_i$ , which is a header embedding in this case. Afterwards, a query-specific header vector is obtained through weighted average (Bahdanau et al., 2015; Sukhbaatar et al., 2015), namely  $v_{header} = \sum_{i=1}^k \alpha_i m_i$ , where  $\alpha_i \in [0, 1]$  is the weight of  $m_i$  calculated as below and  $\sum_i \alpha_i = 1$ .

$$\alpha_i = \frac{\exp(\tanh(W[m_i; v_q] + b))}{\sum_{j=1}^k \exp(\tanh(W[m_j; v_q] + b))}$$

Similar techniques have been successfully applied in table-based question answering (Yin et al., 2015b; Neelakantan et al., 2015). Afterwards, we feed the concatenation of  $v_q$  and  $v_{header}$  to a linear layer followed by a *softmax* function whose output length is 2. We regard the output of the first category as the relevance between query and header. We use  $NN_1()$  to denote this model.

$$f_{nn}(\text{header}, q) = NN_1(M_h, v_q)$$

Since headers and cells have similar characteristics, we use a similar way to measure the relevance between a query and table cells. Specifically, we derive three memories  $M_{cel}$ ,  $M_{row}$  and



$M_{col}$  from table cells in order to match from cell level, row level and column level. Each memory cell in  $M_{cel}$  represents the embedding of a table cell. Each cell in  $M_{row}$  represent the vector a row, which is computed with weighted average over the embeddings of cells in the same row. We derive the column memory  $M_{col}$  in an analogous way. We use the same module  $NN_1()$  to calculate the relevance scores for these three memories.

$$\begin{aligned} f_{nn}(cell, q) &= NN_1(M_{cel}, v_q) \\ f_{nn}(column, q) &= NN_1(M_{col}, v_q) \\ f_{nn}(row, q) &= NN_1(M_{row}, v_q) \end{aligned}$$

Since a table caption is typically a descriptive word sequence. We model it with bi-directional GRU-RNN, the same strategy we have used for modeling the query. We concatenate the caption vector  $v_{cap}$  with  $v_q$ , and feed the results to a linear layer followed by *softmax*.

$$f_{nn}(caption, q) = NN_2(v_{cap}, v_q)$$

We separately train the parameters for each aspect with back-propagation. We use negative log-likelihood as the loss function.<sup>3</sup>

$$loss = -\frac{1}{|D|} \sum_{(t_a, q) \in D} \log(f_{nn}(t_a, q))$$

## 5 Experiment

We describe the experimental setting and analyze the results in this section.

### 5.1 Dataset and Setting

To the best of our knowledge, there is no publicly available dataset for table retrieval. We introduce **WebQueryTable**, an open-domain dataset consisting of query-table pairs. We use search logs from a commercial search engine to get a list of queries that could be potentially answered by web tables. Each query in query logs is paired with a list of web pages, ordered by the number of user clicks for the query. We select the tables occurred in the top ranked web page, and ask annotators to label whether a table is relevant to a query or not. In this way, we get 21,113 query-table pairs. In the real scenario of table retrieval, a system is required to find a table from a huge collection of tables.

<sup>3</sup>We also implemented a ranking based loss function  $\max(0, 1 - f_{nn}(t_a, q) + f_{nn}(t_a^*, q))$ , but it performed worse than the negative log-likelihood in our experiment.

Therefore, in order to enlarge the search space of our dataset, we extract 252,703 web tables from Wikipedia and regard them as searchable tables as well. Data statistics are given in Table 1.

	WQT dataset	WTQ dataset
# of tables	273,816	2,108
Avg # of columns	4.55	6.38
Max # of columns	52	25
Min # of columns	1	3
Avg # of rows	9.15	28.50
Max # of rows	1,517	754
Min # of rows	2	5
# of questions	21,113	22,033
Avg # of questions	4.61	11.25

Table 1: Statistics of WebQueryTable (WQT) dataset and WikiTableQuestions (WTQ) dataset.

We sampled 200 examples to analyze the distribution of the query types in our dataset. We observe that 69.5% queries are asking about “a list of XXX”, such as “*list of countries and capitals*” and “*major cities in netherlands*”, and about 24.5% queries are asking about an attribute of an object, such as “*density of liquid water temperature*”. We randomly separate the dataset as training, validation, test with a 70:10:20 split.

We also conduct a synthetic experiment for table retrieval on **WikiTableQuestions** (Pasupat and Liang, 2015), which is a widely used dataset for table-based question answering. It contains 2,108 HTML tables extracted from Wikipedia. Workers from Amazon Mechanical Turk are asked to write several relevant questions for each table. Since each query is written for a specific table, we believe that each pair of query-table can also be used as an instance for table retrieval. The difference between WikiTableQuestions and WebQueryTable is that the questions in WikiTableQuestions mainly focus on the local regions, such as cells or columns, of a table while the queries in WebQueryTable mainly focus on the global content of a table. The number of table index in WikiTableQuestions is 2,108, which is smaller than the number of table index in WebQueryTable. We randomly split the 22,033 question-table pairs into training (70%), development (10%) and test (20%).

In the candidate table retrieval phase, we encode a table as bag-of-words to guarantee the efficiency of the approach. Specifically, on WebQueryTable dataset we represent a table with caption and headers. On WikiTableQuestions dataset we represent a table with caption, headers and cells. The re-

calls of the candidate table retrieval step on WikiTableQuestions and WebQueryTable datasets are 56.91% and 69.57%, respectively. The performance of table ranking is evaluated with *Mean Average Precision (MAP)* and *Precision@1 (P@1)* (Manning et al., 2008). When evaluating the performance on table ranking, we filter out the following special cases that only one candidate table is returned or the correct answer is not contained in the retrieved tables in the first step. Hyper parameters are tuned on the validation set.

## 5.2 Results on WebQueryTable

Table 2 shows the performance of different approaches on the WebQueryTable dataset.

Setting	MAP	P@1
BM25	58.23	47.12
Feature	61.02	47.79
NeuralNet	61.94	49.02
Feature + NeuralNet	67.18	54.15

Table 2: Results on the WebQueryTable dataset.

We compare between different features for table ranking. An intuitive baseline is to represent a table as bag-of-words, represent a query with bag-of-words, and calculate their similarity with cosine similarity. Therefore, we use the BM25 score which is calculated in the candidate table retrieval step. This baseline is abbreviated as **BM25**. We also report the results of using designed features (**Feature**) described in Section 4.1 and neural networks (**NeuralNet**) described in Section 4.2. Results from Table 2 show that the neural networks perform comparably with the designed features, and obtain better performance than the BM25 baseline. This results reflect the necessary of taking into account the table structure for table retrieval. Furthermore, we can find that combining designed features and neural networks could achieve further improvement, which indicates the complementation between them.

We further investigate the effects of headers, cells and caption for table retrieval on WebQueryTable. We first use each aspect separately and then increasingly combine different aspects. Results are given in Table 3. We can find that in general the performance of an aspect in designed features is consistent with its performance in neural networks. Caption is the most effective aspect on WebQueryTable. This is reasonable as we find

that majority of the queries are asking about a list of objects, such as “*polish rivers*”, “*world top 5 mountains*” and “*list of american cruise lines*”. These intentions are more likely to be matched in the caption of a table. Combining more aspects could get better results. Using cells, headers and caption simultaneously gets the best results.

Setting	Feature		NeuralNet	
	MAP	P@1	MAP	P@1
Header (H)	22.39	9.76	26.03	13.35
Cell (Cel)	28.85	14.95	27.47	12.92
Caption (Cap)	57.12	56.83	60.16	48.48
H + Cel	31.99	17.08	30.73	16.25
H + Cel + Cap	61.02	47.79	61.94	49.02

Table 3: Performance on WebQueryTable dataset with different aspects.

Moreover, we investigate whether using a higher threshold could obtain a better precision. Therefore, we increasingly use a set of thresholds, and calculate the corresponding precision and recall in different conditions. An instance is considered to be correct if the top ranked table is correct and its ranking score is greater than the threshold. Results of our NeuralNet approach on WebQueryTable are given in 2. We can see that using larger threshold results in lower recall and higher precision. The results are consistent with our intuition.

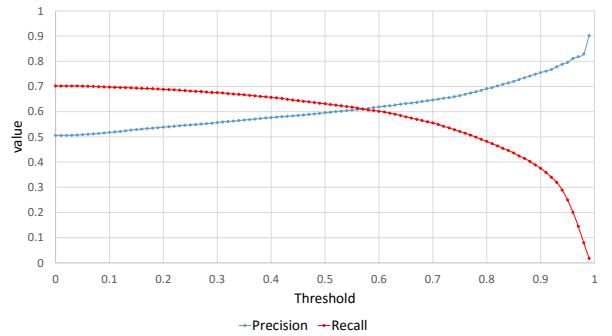


Figure 2: PR Curve on WebQueryTable.

We conduct case study on our NeuralNet approach and find that the performance is sensitive to the length of queries. Therefore, we split the test set to several groups according to the length of queries. Results are given in Figure 4. We can find that the performance of the approach decreases with the increase of query length. When the query length changes from 6 to 7, the performance of P@1 decreases rapidly from 58.12%

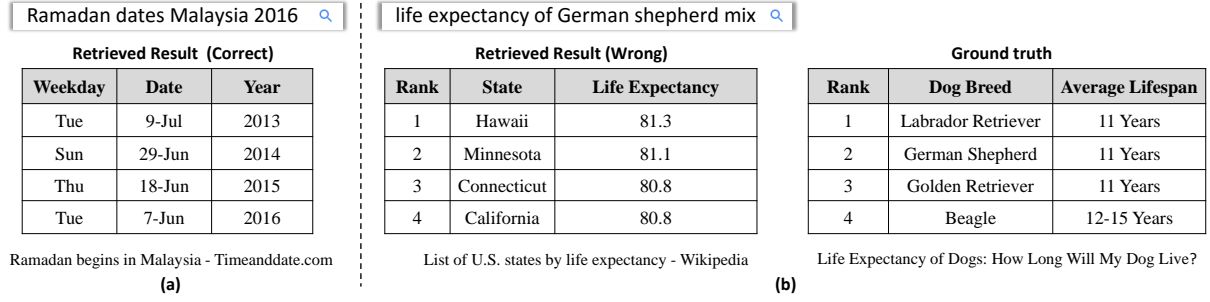


Figure 3: Results generated by NeuralNet on WebQueryTable.

to 50.23%. Through doing case study, we find that long queries contain more word dependencies. Therefore, having a good understanding about the intention of a query requires deep query understanding. Leveraging external knowledge to connect query and table is a potential solution to deal with long queries.

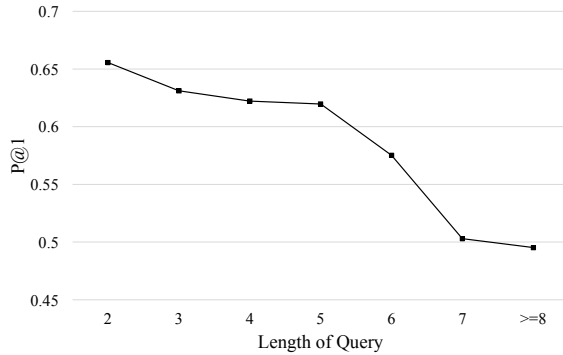


Figure 4: P@1 with different query length on WebQueryTable dataset.

We illustrate two examples generated by our NeuralNet approach in Figure 3. The example in Figure 3(a) is a satisfied case that the top ranked result is the correct answer. We can find that the model uses evidences from different aspects to match between a query and a table. In this example, the supporting evidences come from caption (“ramadan” and “malaysia”), headers (“dates”) and cells (“2016”). The example in Figure 3(b) is a dissatisfied case. We can find that the top ranked result contains “life expectancy” in both caption and header, however, it is talking about the people in U.S. rather than “german shepherd”. Despite the correct table contains a cell whose content is “german shepherd”, it still does not obtain a higher rank than the left table. The reason might be that the weight for header is larger than the weight for cells.

### 5.3 Results on WikiTableQuestions

Table 4 shows the results of table ranking on the WikiTableQuestions dataset.

Setting	MAP	P@1
BM25	51.02	41.02
CDSSM-Header	45.92	31.58
Feature	67.70	56.25
NeuralNet	67.44	54.95
Feature + NeuralNet	72.49	61.50

Table 4: Results on the WikiTableQuestions dataset with different features.

We implement two baselines. The first baseline is BM25, which is the same baseline we have used for comparison on the WebQueryTable dataset. The second baseline is header grounding, which is partly inspired by Venetis et al. (2011) who show the effectiveness of the semantic relationship between query and table header. We implement a CDSSM (Shen et al., 2014) approach to match between a table header and a query. We train the model by minimizing the cross-entropy error, where the ground truth is the header of the answer. Results are given in Table 4. We can find that designed features perform comparably with neural networks, and both of them perform better than BM25 and column grounding baselines. Combining designed features and neural networks obtains further improvement.

We also study the effects of different aspects on the WikiTableQuestions dataset. Results are given in Table 5. We can find that the effects of different aspect in designed features and neural networks are consistent. Using more aspects could achieve better performance. Using all aspects obtains the best performance. We also find that the most effective aspect for WikiTableQuestions is header. This is different from the phenomenon in Web-

Setting	Feature		NeuralNet	
	MAP	P@1	MAP	P@1
Header (H)	46.36	32.52	52.93	36.47
Cell (Cel)	44.33	30.97	43.49	26.41
Caption (Cap)	33.36	24.79	46.83	31.54
H + Cel	60.03	47.42	60.55	45.71
H + Cel + Cap	67.70	56.25	67.44	54.95

Table 5: Results on the WikiTableQuestions dataset with different aspects.

QueryTable that the most effective aspect is caption. We believe that this is because the questions in WikiTableQuestions typically include content constrains from cells or headers. Two randomly sampled questions are “*which country won the 1994 europeans men’s handball championship’s preliminary round?*” and “*what party had 7,115 inactive voters as of october 25, 2005?*”. On the contrary, queries from WebTableQuery usually do not use information from specific headers or cells. Examples include “*polish rivers*”, “*world top 5 mountains*” and “*list of american cruise lines*”. From Table 1, we can also find that the question in WikiTableQuestions are longer than the queries in WebQueryTable. In addition, we observe that not all the questions from WikiTableQuestions are suitable for table retrieval. An example is “*what was the first player to be drafted in this table?*”.

## 6 Related Work

Our work connects to the fields of database and natural language processing.

There exists several works in database community that aims at finding related tables from keyword queries. A representative work is given by Cafarella et al. (2008), which considers table search as a special case of document search task and represent a table with its surrounding text and page title. Limaye et al. (2010) use YAGO ontology to annotate tables with column and relationship labels. Venetis et al. (2011) go one step further and use labels and relationships extracted from the web. Pimplikar and Sarawagi (2012) focus on the queries that describe table columns, and retrieve tables based on column mapping. There also exists table-related studies such as searching related tables from a table (Das Sarma et al., 2012), assembling a table from list in web page (Gupta and Sarawagi, 2009) and extracting tables using tabular structure from web page (Gatter-

bauer et al., 2007). Our work differs from this line of research in that we focus on exploring the content of table to find relevant tables from web queries.

Our work relates to a line of research works that learn continuous representation of structured knowledge with neural network for natural language processing tasks. For example, Neelakantan et al. (2015); Yin et al. (2015b) develop neural operator on the basis of table representation and apply the model to question answering. Yin et al. (2015a) introduce a KB-enhanced sequence-to-sequence approach that generates natural language answers to simple factoid questions based on facts from KB. Mei et al. (2016) develop a LSTM based recurrent neural network to generate natural language weather forecast and sportscasting commentary from database records. Serban et al. (2016) introduce a recurrent neural network approach, which takes fact representation as input and generates factoid question from a fact from Freebase. Lebrete et al. (2016) presented an neural language model that generates biographical sentences from Wikipedia infobox.

Our neural network approach relates to the recent advances of attention mechanism and reasoning over external memory in artificial intelligence (Bahdanau et al., 2015; Sukhbaatar et al., 2015; Graves et al., 2016). Researchers typically represent a memory as a continuous vector or matrix, and develop neural network based controller, reader and writer to reason over the memory. The memory could be addressed by a “soft” attention mechanism trainable by standard back-propagation methods or a “hard” attention mechanism trainable by REINFORCE (Williams, 1992). In this work, we use the soft attention mechanism, which could be easily optimized and has been successfully applied in nlp tasks (Bahdanau et al., 2015; Sukhbaatar et al., 2015).

## 7 Conclusion

In this paper, we give an empirical study of content-based table retrieval for web queries. We implement a feature-based approach and a neural network based approach, and release a new dataset consisting of web queries and web tables. We conduct comprehensive experiments on two datasets. Results not only verify the effectiveness of our approach, but also present future challenges for content-based table retrieval.



## References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations (ICLR)*.
- Sreeram Balakrishnan, Alon Y Halevy, Boulos Harb, Hongrae Lee, Jayant Madhavan, Afshin Ros-tamizadeh, Warren Shen, Kenneth Wilder, Fei Wu, and Cong Yu. 2015. Applying webtables in practice. *Proceedings of Conference on Innovative Data Systems Research (CIDR)*.
- Christopher JC Burges. 2010. From ranknet to lambdarakn to lambdamart: An overview. *Microsoft Research Technical Report MSR-TR-2010-82* 11(23-581):81.
- Michael J Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. 2008. Webtables: exploring the power of tables on the web. *Proceedings of the VLDB Endowment* 1(1):538–549.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using rnn encoder-decoder for statistical machine translation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1724–1734. <http://www.aclweb.org/anthology/D14-1179>.
- Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy, Hongrae Lee, Fei Wu, Reynold Xin, and Cong Yu. 2012. Finding related tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, pages 817–828.
- Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Wolfgang Gatterbauer, Paul Bohunsky, Marcus Herzog, Bernhard Krüpl, and Bernhard Pollak. 2007. Towards domain-independent information extraction from web tables. In *Proceedings of the 16th international conference on World Wide Web (WWW)*. ACM, pages 71–80.
- Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature* 538(7626):471–476.
- Rahul Gupta and Sunita Sarawagi. 2009. Answering table augmentation queries from unstructured lists on the web. *Proceedings of the VLDB Endowment* 2(1):289–300.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* 1:48–54.
- Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language (EMNLP)*.
- Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. 2010. Annotating and searching web tables using entities, types and relationships. *Proceedings of the VLDB Endowment* 3(1-2):1338–1347.
- Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.
- Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. [What to talk about and how? selective generation using lstms with coarse-to-fine alignment](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 720–730. <http://www.aclweb.org/anthology/N16-1086>.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NIPS)*. pages 3111–3119.
- Arvind Neelakantan, Quoc V Le, and Ilya Sutskever. 2015. Neural programmer: Inducing latent programs with gradient descent. *arXiv preprint arXiv:1511.04834*.
- Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Rakesh Pimplikar and Sunita Sarawagi. 2012. Answering table queries on the web using column keywords. *Proceedings of the VLDB Endowment* 5(10):908–919.
- Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP* 109:109.

- Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. [Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 588–598. <http://www.aclweb.org/anthology/P16-1056>.
- Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*. pages 101–110.
- Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems (NIPS)*. pages 2431–2439.
- Petros Venetis, Alon Halevy, Jayant Madhavan, Marius Pasca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu. 2011. Recovering semantics of tables on the web. *Proceedings of the VLDB Endowment* 4(9):528–538.
- Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. 2015. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*.
- Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4):229–256.
- Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2015a. Neural generative question answering. *arXiv preprint arXiv:1512.01337*.
- Pengcheng Yin, Zhengdong Lu, Hang Li, and Ben Kao. 2015b. Neural enquirer: Learning to query tables with natural language. *arXiv preprint arXiv:1512.00965*.