IJE TRANSACTIONS C: Aspects Vol. 26, No. 3, (March 2013) 241-244

International Journal of Engineering

Journal Homepage: www.ije.ir

Use of Generalized Language Model for Question Matching

S. Izadi*, M. Ghasemzadeh

Electrical and Computer Engineering Department, Yazd University, Yazd, Iran

PAPER INFO

Paper history:
Received 22 September 2012
Received in revised form 7 October 2012
Accepted 18 October 2012

Keywords:
Question Matching
Natural Language Processing
Statistical Language Model
Q&A Services

ABSTRACT

Question and answering service is one of the popular services in the World Wide Web. The main goal of these services is to find the best answer for user's input question as quick as possible. In order to achieve this aim, most of these use new techniques foe question matching. We have a lot of question and answering services in Persian web, so it seems that developing a question matching model might be useful. This paper introduces a new question matching model for Persian. This model is based on statistical language model and employs generalized bigram and trigram model. We also describe some results regarding the employment of natural language processing in question matching model. Most of the Q&A services have large number of questions and answers; hence we considered an optimized implementation for the model. We evaluated our model with Rasekhoon question and answering archive which contains about 18000 pairs of questions and answers. The results showed the improvement of precision and recall measures through using this model.

doi: *10.5829/idosi.ije.2013.26.03c.03*

Archive of SID

1. INTRODUCTION

Today many people use web to satisfy their need for information in all over the world. Question and answering services help people to find the answer of their question in acceptable time. *Wondir* and *Google Answer* are two great sites that provide question and answering service. Some of these services use question matching techniques to increase the response rate. However, measuring syntactic similarity singly is not good enough to find similar question. Sometimes two questions have close meaning but the terms that have been used in them are different.

Three different types of approaches have been developed in the literature to solve the word mismatch problem among questions. The first approach uses knowledge databases such as machine readable dictionaries. However, the quality and structure of current knowledge databases are, based on the results of previous experiments, not good enough for reliable performance. The second approach employs manual rules or templates. These methods are expensive and hard to scale for large size collection. The third approach is to use statistical techniques developed in information retrieval and natural language processing. We believe the last approach is the most promising if we have enough training data. Wang et al. [1] showed that a question matching model based on translation probabilities learned from the archive significantly outperforms other approaches in terms of finding similar question despite a considerable amount of lexical mismatch. They used nave question and answering archive as knowledge base. FAQ finder is natural language question-answering system that uses files of frequently asked questions as its knowledge base. This system uses a combination of statistical and natural language processing techniques to match over users' questions against known question-answer pair from FAQ files [2-5]. A new interval framework based on syntactic tree structure for question matching was proposed [1, 6]. We have some question and answering services in Persian[1] . Most of the Persian services are about religion and consultation in social field. Hassanpour [7] made an initial attempt to investigate the reuse of facts contained in the archive of previous questions to help and gain performance in answering future related factoid questions. This paper introduces a new question matching model based on a generalized language model. The remainder of this paper is structured as follow. In next section we discus about this model. Implementation has been briefly described in

*Corresponding Author Email: s.izadi65@yahoo.com (S. Izadi)

[1] www.pezeshkonline.ir ; www.rasekhoon.net

www.SID.ir

section 3. In section 4 we explain our results. Section 5 is conclusion of this paper.

2. QUESTION MATCHING MODEL

In this section we introduce our model. This model contains online and offline processing parts. At first we build some dataset using question and answering archives. Then, online processing part uses them to compute similarity between questions. These two main sections also contain some subsections.

2. 1. Offline Processing This section contains 3 subsections:

Preprocessing
Computing DF measures
Producing bigram and trigram datasets.

2. 1. 1. Preprocessing Each pair of questions and answers was considered as a document in this paper. At the beginning we eliminate writing marks such as "؟","!", "،","."," ",… from documents. Also, stopword have been deleted in preprocessing. Stopwords are some words that have no semantic valence. In Persian conjunctives like "در", "به", "از" and "که" and some verbs like "باشد", "است" and "بود" are considered as stopwords [8].

Eliminating these frequent terms could reduce the computation and space needed for storage.

2. 1 . 2. Compute DF Measure We used vector space model in our approach. So computing DF in offline process reduces online computation. DF means frequent of each term in whole documents.

2. 1. 3. Producing Bigram and Trigram Datasets In this model, we consider the possibility of occurrence the word w_n after each of the words w_{n-1}, w_{n-2} and w_{n-3} , separately. This means that we have to consider the distance between words in our relations. It is the main difference between our approach and the standard language model. The distance between two words in same document is given by:

$$dist_e(w_i, w_j) = \frac{1}{\min(|p(w_i), p(w_j)|)} \quad (1)$$

where $p(w_i)$ is index of w_i and $p(w_j)$ is index of w_j in document e. Also, $dist_e(w_i, w_j)$ is the distance between w_i and w_j in document e.

The main measure is sum of the distance between two words in whole documents and is given by:

$$Weight(w_i, w_j) = \sum_{i=1}^{n} dist_e(w_i, w_j) \quad (2)$$

where n is the total number of documents which contains both w_i and w_j.

The above method and relations have been used to compute bigram dataset. We use the same method to compute trigram dataset.

2. 2. Online Processing This section also contains 3 subsections:

- ✓ Compare input question and all existing questions based on vector space model.
- ✓ Compare input question and all existing questions based on the generalized language model.
- ✓ Computing total similarity score for the input question and each existing question.

- Vector space model

The main benefit of using vector space model is independence of questions length. Final score is sum of the obtained score for same words in two questions.

- Generalized language model

In this section we extract words from offline bigram and trigram datasets that have relation with words in input question. Then expanded question compare with the entire question in archive.

- Total score

Finally, we combine all of the scores: the vector space model (s_a), the bigram similarity score (s_b) and the trigram similarity score (s_c):

$$Total = \frac{T * s_a + U * s_b + V * s_c}{T + U + V} \quad (3)$$

T, U and V are constant weights associated with vector space model, bigram similarity and trigram similarity.

3. IMPLEMENTATION

Implementing this system is written in 2500 lines of code in visual C#. Because of the huge number of question and answers in Q&A services, it is necessery to consider efficient programming for implementation of this model. Thus, we designed algorithms that run in liner time on average. For example, Table 1 shows the pseudo-code to extract bigram from documents.

4. EVALUATION

Rasekhoon questions and answerings dataset is employed to evaluate question matching model. Most of questions and answerings in this service is about religion and had been answered by experts. Table 2 contains a few examples of bigram dataset. For each single word, bigram dataset return some related words based on the average distance between them. It makes processing easier; however, they lose the semantics of the text.

Archive of SID

www.SID.ir

TABLE 1. Pseudo-code to extract bigram from documents

```
procedure bigram(String[] AllQA, int Threshold)
Begin
 for i = 1 to  Number of AllQA
 Begin
  FirstWord = nextword(AllQA[i,1]);
  distance=1;
  for j = 2 to  Number of Words in AllQA[i]
  Begin
    SecondWord = nextword(FirstWord, distance);
    if Dist (FirstWord, SecondWord < Threshold)
    Then
      DataSet Bigram (FirstWord, SecondWord) =
      DataSet Bigram (FirstWord, SecondWord)
      + Dist (FirstWord, SecondWord);
      distance++;
    Else
      distance=1;
      FirstWord = nextword(FirstWord, distance);
    End
  End
 End
End
```

TABLE 2. Examples of bigram dataset

Main word	Relation words
نماز (pray)	زکات (Charity), روزه (Fast), رکعت (Knees), جمعه (Friday), شب (Night)
امام (Imam)	خمینی (Khomeini), صادق (Sadegh), حضرت (Holiness), زمان (Zaman), السلام (Alsalam)
قرآن (Quran)	آیات (Signs), کریم (Holy), مجید (Glorious), تفسیر (Interpretation), خداوند (God)
زهرا (Zahra)	فاطمه (Fatima), حضرت (Holiness), علی (Ali), پیامبر (Prophet), فاطمة (Fatimah)
تشیع (Funeral)	مذهب (Religion), آئین (Faith), تاریخ (History), تسنن (Sunni), مکتب (School)

TABLE 3. Examples of trigram dataset

Main word	Relation words
ژان پل (john Paul)	روحانیون (clergy), واتیکان (Vatican), دوم (Second),
مبارزه طالبان (Fighting The Taliban)	شوروی (Soviet), پاکستان (Pakistan)
غسل تعمید (Baptism)	آداب (Customs), مادر (Mother), نوزاد (Baby),
هوی نفس (Self-fad)	محافظت (Protection), هوس (Lust), دین (Religion),
یازدهمین پیشوا (Eleventh PISHVA)	شیعه (Funeral), حسن (Hassan), عسکری (Askari),

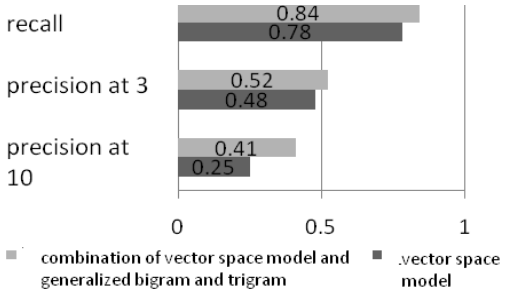

Figure 1. Comparing generalized language model with vector space model regardless of the type of questions

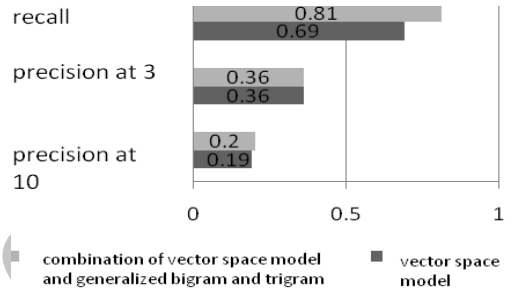

Figure 2. Comparing generalized language model with vector space model based on question type

Some examples of trigram dataset are shown in Table 3. The trigram dataset could save semantices of the text more than bigram dataset. Combining each double words with their related words may yeild to a meaningful sentence. Online processing section evaluated two different situations. At first, we considered all questions for measuring precision and recall for our model regardless of the type of question. The results show that these two measures are improved by employing generalized bigram and trigram in question matching model. Figure 1 shows a comparison between question matching model when it is based on vector space model singly and based on a combination of vector space model and generalized language model. Figure 1 shows this test results. In the second type of testing we considered three types of question sets based on their subject as dataset. We assumed that user specified the type of his/her question at first. The result of this evaluation is shown in Figure 2.

5. CONCLUSION

In this paper, we introduced a new question matching model for Persian language based on generalized language model. Also, we discussed about implementation and evaluated our model for two different situations. Our results showed that use of generalized language model yield to improve the

Archive of SID

www.SID.ir

question matching model in Persian. It seems that by employing generalized language model we could improve our results as well as question matching model for other languages that use advance natural language processing tools like wordnet.

6. ACKNOWLEDGEMENT

This work was partially sponsored by Research Institute for ICT-ITRC.

7. REFERENCES

1. Wang, K., Ming, Z. and Chua, T. S., "A syntactic tree matching approach to finding similar questions in community-based qa services", in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, (2009), 187-194.
2. Burke, R., Hammond, K., Kulyukin, V., Lytinen, S., Tomuro, N., and Schoenberg, S., "Natural language processing in the faq finder system: Results and prospects", in Working Notes from AAAI Spring Symposium on NLP on the WWW., (1997), 17-26.
3. Lytinen, S. and Tomuro, N., "The use of question types to match questions in FAQFinder", in AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, (2002), 46-53.
4. Ceeper, E., "Improving FAQfinder's Performance: Setting Parameters by Genetic Programming", (1996).
5. Burke, R. D., Hammond, K. J., Kulyukin, V., Lytinen, S. L., Tomuro, N., and Schoenberg, S., "Question answering from frequently asked question files: Experiences with the faq finder system", ***AI magazine***, Vol. 18, No. 2, (1997), 57.
6. Moschitti, A., "Making tree kernels practical for natural language learning", in Proceedings of EACL. Vol. 6, No. Issue, (2006), 113-120.
7. Hassanpour, H., "Boosting passage retrieval through reuse in question answering", ***International Journal of Engineering-Transactions C: Aspects***, Vol. 25, No. 3, (2012), 187-196.
8. Taghva, K., Beckley, R. and Sadeh, M., "A list of farsi stopwords", ***Retrieved September***, Vol. 7, No., (2003).

Use of Generalized Language Model for Question Matching

S. Izadi, M. Ghasemzadeh

Electrical and Computer Engineering Department, Yazd University, Yazd, Iran

PAPER INFO

Paper history:
Received 22 September 2012
Received in revised form 7 October 2012
Accepted 18 October 2012

Keywords:
Question Matching
Natural Language Processing
Statistical Language Model
Q&A Services

چکیده

از جنبه‌های مهم در فناوری اطلاعات امکان یافتن پاسخ سوالات از بسترهای آن می‌باشد. فضای اینترنت شامل حجم عظیمی از اطلاعات و از آن جمله جفت‌های پرسش و پاسخ است. لذا این قابلیت که بتوانیم سوال معادل و یا سوال مشابه با سوال کاربر را به سرعت یافته و پاسخ مربوطه را ارائه دهیم اهمیت ویژه‌ای یافته است. در این زمینه کوشش‌هایی برای سایر زبان‌ها صورت پذیرفته و انجام آن برای زبان فارسی نیز الزامی می‌باشد. در این مقاله روشی مبتنی بر ترکیب فضای برداری و تعمیمی از مدل‌های زبانی یونی‌گرم و بای‌گرم برای تطابق سوال فارسی ارائه می‌گردد. روش مورد نظر پیاده‌سازی و بر روی داده‌های محک انبوه ارزیابی شده‌اند. داده‌های محک شامل بایگانی سرویس پرسش و پاسخ برخط راسخون، که حاوی بیش از هجده هزار جفت پرسش و پاسخ است، می‌باشد. حجم پردازش و سایز ورودی لزوم بکارگیری الگوریتم-های کارآمد با درجه پیچیدگی زمانی و همچنین درجه پیچیدگی حافظه پایین‌تری را ملزم می‌داشت که از جمله نتایج این تحقیق می‌باشند. از آنجایی که تمرکز اصلی در این تحقیق، ارزیابی کارایی مدل‌های زبانی است، میزان بهبود تطابق سوال نسبت به روشی که تنها فضای برداری استفاده شود نیز مقایسه شده است. نتایج این مقایسه نشان از بهبود معیارهای دقت و فراخوانی با استفاده از مدل‌های زبانی ارائه شده است. همچنین این مدل، در مقایسه با مدل‌های تطبیق سوال ارائه شده برای سایر زبان‌ها که از روش‌های پیچیده‌تری مانند هستان‌شناسی در تطبیق سوال استفاده کرده‌اند نیز پاسخ بهتری ارائه می‌دهد.

doi: *10.5829/idosi.ije.2013.26.03c.03*

Archive of SID

www.SID.ir