An Annotation Similarity Model in Passage Ranking for Historical Fact Validation

Jun Araki and Jamie Callan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{junaraki, callan}@cs.cmu.edu

ABSTRACT

State-of-the-art question answering (QA) systems employ passage retrieval based on bag-of-words similarity models with respect to a query and a passage. We propose a combination of a traditional bag-of-words similarity model and an annotation similarity model to improve passage ranking. The proposed annotation similarity model is generic enough to process annotations of arbitrary types. Historical fact validation is a subtask to determine whether a given sentence tells us historically correct information, which is important for a QA task on world history. Experimental results show that the combined model gains up to 7.7% and 4.2% improvements in historical fact validation in terms of precision at rank 1 and mean reciprocal rank, respectively.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.4 [**Information Storage and Retrieval**]: Systems and Software; J.m [**Computer Applications**]: Miscellaneous

Keywords

passage retrieval; question answering; text annotation

1. INTRODUCTION

Passage retrieval is a core component for question answering (QA) systems [18, 5, 6, 10]. Many passage retrieval approaches commonly used in QA systems cannot check linguistic and semantic types annotated in passages at query time [1, 2]. We believe that a main reason for the inability of passage retrieval is the lack of a general ranking scheme to incorporate annotations in retrieval processes along with traditional retrieval models.

As a first step toward the goal to embody such ranking scheme, we propose a three-stage passage ranking approach to effectively integrate a range of annotations in passage retrieval for QA. Our approach has two advantages. First, it can rank passages using those annotations in an unsupervised manner. Second, it can deal with annotations of arbitrary types. The rest of this paper is organized as follows. In Section 2, we discuss past work on passage retrieval for QA. We describe our approach in Section 3, and show and discuss experimental results in Section 4. We make a conclusion and describe future work in Section 5.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609522.

2. RELATED WORK

There is a considerable amount of work on passage retrieval. Passage retrieval was initially explored to overcome the shortcomings of document retrieval [15, 7, 8]. It was further investigated and integrated into QA [14, 4, 18]. This is primarily because a passage is a more appropriate unit to rank for answering questions than a document.

In retrieval processes for QA, researchers often augment information sources with various types of annotations. For instance, Prager et al., 2000 [14] used named entities to improve the performance of passage retrieval. Tiedemann, 2005 [19] integrated dependency relations into a multi-layer index in passage retrieval for Dutch question answering. Bilotti et al. 2007 [2] leveraged named entities and semantic roles to perform sentence-level structural retrieval for QA. Our approach is different from theirs in that we do not pre-annotate any corpora, which is a very expensive process. Shen and Lapata, 2007 [17] examined the effectiveness of semantic roles in factoid QA with a graph matching technique. Our model is different from theirs in that ours can easily incorporate annotations of other types.

More recently, much work showed QA performance improvements using supervised learning models for (re-)ranking passages with linguistic and knowledge-based features [5, 9, 1, 16, 20]. As compared to these learning-based studies, there is much less work on utilizing such linguistic and knowledge-based resources in passage ranking with retrieval models in an unsupervised manner.

3. APPROACH

3.1 World History QA Task

NTCIR-11 set up a shared task called QA Lab[1]. In this task, participants are expected to collaboratively develop module-based QA systems for solving real-world university entrance exam questions. One of the exams is from a standardized test created by the National Center Test for University Admissions in Japan. The original exam corpus was

[1] http://ntcir.nii.ac.jp/QALab/

. . . , most of those who excelled in culture and the arts were those who had passed the Imperial examinations, but in the (2) Ming period, there was a shift toward . . .

Question 2. From 1-4 below, choose the most appropriate sentence concerning events that occurred during the period referred to in the underlined portion (2).
1. Japanese silver circulated in China.
2. A Buddhist sect called Zen was created.
3. The play "The Story of the Western Wing (Xixiangji)" was created.
4. The capital was established in Lin'an (present-day Hangzhou).

Figure 1: An illustrative question in the exam corpus. The correct answer is 1.

created in Japanese, but this work uses an English translation version of the corpus. We use a set of 26 true-false questions from the 2009 exam on world history. All the questions are multiple-choice questions with four answer choices, and one of them is the correct answer. Each answer choice is given in a single sentence.

Figure 1 shows an example of the true-false questions in the corpus, and the correct answer to this question is 1. Examinees are instructed to read introductory text with some contextual information, and to solve questions following the text. The questions are strongly or weakly dependent on their corresponding introductory text. A degree of the dependency varies among individual questions, but in any case the correct answer does not appear anywhere in that text or in the entire corpus. Therefore, examinees must rely solely on their knowledge on world history in their brains to answer the questions. In the case of Figure 1, for example, the introductory text has a strong dependency on Question 2 since the underlined portion (2) provides an indispensable piece of temporal information for examinees to solve the question. There is no answer-bearing sentence in the introductory text, including the ". . . " portions that we omitted to save space in Figure 1.

3.2 Historical Fact Validation

The focus of this paper is not on QA but rather on passage ranking as a system module for QA. To evaluate our passage ranking system, we first collect historical facts from the exam corpus. In this work, we define a *historical fact* as a sentence that tells us historically correct information. We ensure the historical correctness by a reference to information sources that we rely on.

The process of collecting historical facts is a sequence of the following manual steps. We first select the 26 true-false questions from the corpus. We then divide the 26 into two groups: (A) a subset of questions whose answer choices comprise one historically correct sentence (the correct answer) and three historically incorrect sentences, and (B) the other subset of questions whose answer choices are one historically incorrect sentence (the correct answer) and three historically correct sentences. Each question in group (A) produces one historical fact, and each in group (B) produces three. Since the 26 questions consist of 21 questions in group (A) and 5 questions in group (B), we collect 36 historical facts in total.

Table 1: Wikipedia dump statistics.

Article type	# Articles
All	14,226,207
Non-redirect	7,741,191
Non-redirect & main namespace	4,514,662

In collecting historical facts, we possibly make a modification to raw answer-choice sentences when a question is strongly dependent on its introductory text. For instance, Question 2 in Figure 1 is strongly dependent on the introductory text, since solving the question requires the temporal phrase specified with the underlined portion (2), as described in Section 3.1. In such cases, we append such phrase to answer-choice sentences in order to make the sentences as historically specific as possible. As a result, from Question 2 we create a history fact "Japanese silver circulated in China during the Ming period." We observe that we do not need to make such modification when questions are weakly dependent on their introductory text, because answer-choice sentences of these questions are historically specific enough to be a complete historical fact by themselves.

We define *historical fact validation* as a subtask to determine whether or not a given sentence is a historical fact and output a binary value (i.e., true or false) about it. It is clear that a system component to perform the subtask is directly useful to QA on the 26 true-false questions. For historical fact validation, we employ passage retrieval. The basic idea is that if a system taking a given sentence as a query (historical hypothesis) retrieves and ranks a passage (historical evidence) with a reasonably high score, then the system regards the sentence as a historical fact.

3.3 Indexing Wikipedia Articles

We choose Wikipedia as information sources for two reasons. First, it is abundant of historical facts and highly likely to cover historical topics of posed questions in the exam corpus. Second, Wikipedia is text-based, and thus makes it easier for us to incrementally add desirable annotations using natural language processing or knowledge base tools than structured resources such as DBpedia. We use the '2014-02-03' dump of English Wikpedia articles[2] to construct an index. Table 1 shows the number of articles in the dump. We index only non-redirect articles with a Main namespace[3] (the third row of Table 1), excluding the other articles because they do not have any textual contents to be effectively retrievable for historical fact validation. As another preprocess to obtain plain text, we clean up Wikipedia markups in the dump using Bliki[4]. After the preprocess, we build up the index using Apache Solr[5].

3.4 Passage Ranking

For historical fact validation, we propose a three-stage passage ranking approach shown in Figure 2. All the re-

[2] http://dumps.wikimedia.org/enwiki/20140203/enwiki-20140203-pages-articles.xml.bz2
[3] http://en.wikipedia.org/wiki/Wikipedia:Namespace
[4] https://code.google.com/p/gwtwiki/
[5] http://lucene.apache.org/solr/

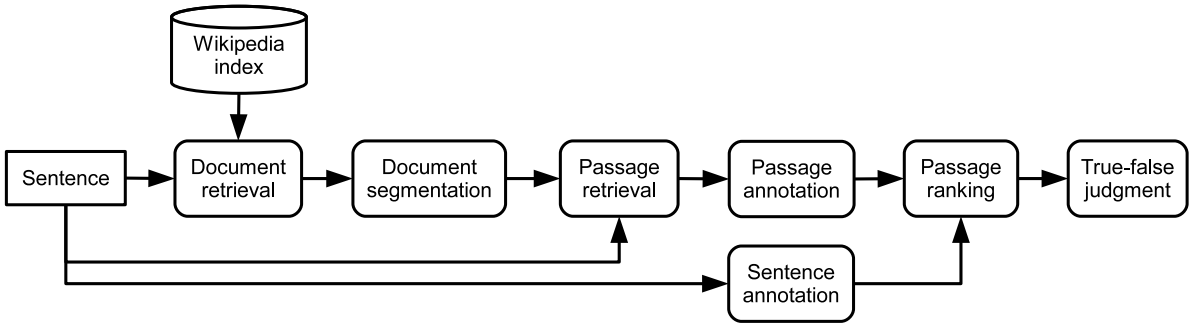

Figure 2: Passage ranking architecture for historical fact validation.

trieval processes are done at a query time. The first stage is document retrieval. We assume that a query comprising raw words in an answer-choice sentence can retrieve documents with substantially high recall. Therefore, we take a simple approach to formulate a query using just all raw words in the sentence. We restrict the number of retrieved documents to N_d for further processing. Our another assumption here is that a properly tuned N_d gives us a set of retrieved documents with relatively high recall.

The second stage is passage retrieval without involving any annotations except sentence segmentation. In this stage, we segment a retrieved document into passages by a fixed-length windows [8, 3]. The effectiveness of fixed-length arbitrary passages is not particularly sensitive to passage length in the standard information retrieval setting [8, 11]. Although it is not clear that this is the case with passage retrieval for historical fact validation, we follow the approach and use a sliding window of N_s sentences to obtain passages. We use Stanford CoreNLP[6] for sentence segmentation. We also restrict the number of retrieved passages to N_p for subsequent processes. Our assumption here is that we can also tune N_p properly so it gives us a set of retrieved passages with moderately high recall. We use TF-IDF [12] for retrieval in the both first and second stage.

Algorithm 1 Annotation Similarity Model.

Input: $G_1 = (E_1 = (te_1), R_1 = (tr_1), T)$
Input: $G_2 = (E_2 = (te_2), R_2 = (tr_2), T)$
Input: $T_c \in T$
1: $E'_1 \leftarrow (te_1)$ where $te_1 \in T_c$
2: $E'_2 \leftarrow (te_2)$ where $te_2 \in T_c$
3: $R'_1 \leftarrow (tr_1)$ where $tr_1 \in T_c$
4: $R'_2 \leftarrow (tr_2)$ where $tr_2 \in T_c$
5: $sim(G_1, G_2, T_c) \leftarrow 2\frac{|E'_1 \bigcap E'_2| + |R'_1 \bigcap R'_2|}{|E'_1| + |E'_2| + |R'_1| + |R'_2|}$
Output: $sim(G_1, G_2, T_c)$

The third stage is passage ranking with a range of annotations. The underlying idea of this stage is to improve passage ranking by combining a bag-of-words similarity model (sim_{BOW}) and an annotation similarity model (sim_{ANN}). Following [1], we represent a set of anntations as an anntation graph. More specifically, an annotation graph G is represented as $G = (E, R, T)$ where E is a set of elemental annotations and R is a set of relational annotations specified under a type system T. Algorithm 1 shows our algorithm to output the annotation similarity using vertex/edge overlap [13]. T_c is a subset of T used for the similarity calculation. In this work, we determine the final score with respect to sentence s and passage p by a multiplication of the TF-IDF score and the annotation similarity score as follows.

$$\begin{aligned} score(s,p) &= sim_{BOW}(s,p) \times sim_{ANN}(s,p) \\ &= \textit{TF-IDF}(s,p) \times (1 + \alpha\ sim(G_s, G_p, T_c)) \end{aligned}$$

where α is a weight on the annotation similarity model, G_s and G_p are annotation graphs of s and p, respectively. We intend that sim_{ANN} gives a small amount of similarity adjustment to sim_{BOW}. Thus, we seek relatively small values when tuning α. With respect to annotations, we used Stanford CoreNLP for named entities and dependencies and ClearNLP[7] for semantic roles.

3.5 Evaluation

Since there is no gold standard for a ranked list of passages as output of passage ranking, we judge the relevance of ranked passages manually by ourselves. More specifically, the judgment process consists of the following manual steps. We first figure out a full set of semantic components (e.g., key entities, and temporal and geographical information) in a given sentence to constitute a historical fact. We then examine whether a ranked passage contains the necessary pieces of information to determine whether it is an answer-bearing passage, i.e., whether it can validate the correctness of a historical fact. To measure the performance of passage ranking, we use precision at rank 1 (P@1) and mean reciprocal rank (MRR). P@1 is the percentage of historical facts where an answer-bearing passage is ranked at the first position. MRR is computed as follows: $MRR = \frac{1}{|Q|}\sum_{q=1}^{|Q|}\frac{1}{rank(q)}$, where Q is a set of queries, q is a query in Q, and $rank(q)$ is the rank of the first answer-bearing passage in ranked passages retrieved from q.

4. EXPERIMENTAL RESULTS

We conducted an experiment to investigate the impact of the annotation similarity model in passage ranking using the 26 true-false questions. Table 2 shows our experimental results. The first row of this table shows the performance of a baseline where we run the system up to the second stage. We found out that named entities of a person type is the most useful annotation. The annotation obtained 7.7% and 4.2% gains in terms of P@1 and MRR, respectively. This

[6] `http://nlp.stanford.edu/software/corenlp.shtml`

[7] `http://clearnlp.com`

Table 2: Comparison in the performance of passage ranking between passage annotations. In this experiment, the maximum number of documents is 1000, the maximum number of passages is 10, the window size is 3 sentences, and the weighting parameter α is 0.1.

T_c	P@1	MRR
(Baseline)	0.3611	0.4609
Named entity (person)	0.3889	0.4801
Dependency (nsubj, dobj)	0.3889	0.4755
Semantic argument (A0, A1)	0.3611	0.4639

result indicates that names of historical figures are a key element to amplify TF-IDF effects well, and the named entity annotation of persons effectively boost the performance of passage ranking.

For dependency relations, we used two relations 'nsubj' and 'dobj'. The former means a predicate-subject relation, and the latter a predicate-object relation. We observed that these relations also gave us performance gains comparable to named entities. We also examined two semantic arguments 'A0' and 'A1', which mean an agent and a patient, respectively. They achieved a slightly better performance than the baseline, but the performance improvement was quite small. This is mainly due to a sparseness problem of semantic arguments. The system produced a relatively small number of semantic role annotations. Consequently, it is rather rare that a sentence and a passage exhibit the same argument structure over the same tokens.

5. CONCLUSIONS

We proposed a three-stage passage ranking algorithm for historical fact validation, which is an important subtask for on world history. To our knowledge, this is the first work on an annotation similarity model that can process annotations of any type, along with traditional bag-of-words models, in an unsupervised manner. The model showed performance gains in passage ranking in terms of both P@1 and MRR.

Our future work is to refine the model so it can benefit from a combination of different annotations, including WordNet synsets and temporal relations. We also plan to implement a true-false judgment component to be integrated with our passage ranking component for an end-to-end evaluation of a world history QA system.

6. ACKNOWLEDGMENTS

We would like to thank the three anonymous reviewers for their valuable comments. The first author is supported by the Funai Overseas Scholarship.

7. REFERENCES

[1] M. W. Bilotti, J. Elsas, J. Carbonell, and E. Nyberg. Rank Learning for Factoid Question Answering with Linguistic and Semantic Constraints. In *Proceedings of CIKM 2010*, pages 459–468, 2010.

[2] M. W. Bilotti, P. Ogilvie, J. Callan, and E. Nyberg. Structured Retrieval for Question Answering. In *Proceedings of SIGIR 2007*, pages 351–358, 2007.

[3] J. P. Callan. Passage-Level Evidence in Document Retrieval. In *Proceedings of SIGIR 1994*, pages 302–310, 1994.

[4] C. L. A. Clarke, G. V. Cormack, and T. R. Lynam. Exploiting Redundancy in Question Answering. In *Proceedings of SIGIR 2001*, pages 358–365, 2001.

[5] H. Cui, R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua. Question Answering Passage Retrieval Using Dependency Relations. In *Proceedings of SIGIR 2005*, pages 400–407, 2005.

[6] D. A. Ferrucci, E. W. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. M. Prager, N. Schlaefer, and C. A. Welty. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3):59–79, 2010.

[7] M. A. Hearst and C. Plaunt. Subtopic Structuring for Full-length Document Access. In *Proceedings of SIGIR 1993*, pages 59–68, 1993.

[8] M. Kaszkiel and J. Zobel. Passage Retrieval Revisited. In *Proceedings of SIGIR 1997*, pages 178–185, 1997.

[9] J. Ko, E. Nyberg, and L. Si. A Probabilistic Graphical Model for Joint Answer Ranking in Question Answering. In *Proceedings of SIGIR 2007*, pages 343–350, 2007.

[10] E. Krikon, D. Carmel, and O. Kurland. Predicting the Performance of Passage Retrieval for Question Answering. In *Proceedings of CIKM 2012*, pages 2451–2454, 2012.

[11] X. Liu and W. B. Croft. Passage Retrieval Based on Language Models. In *Proceedings of CIKM 2002*, pages 375–382, 2002.

[12] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[13] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina. Web Graph Similarity for Anomaly Detection. *Journal of Internet Services and Applications*, Volume 1(1):19–30, 2010.

[14] J. Prager, E. Brown, A. Coden, and D. Radev. Question-Answering by Predictive Annotation. In *Proceedings of SIGIR 2000*, pages 184–191, 2000.

[15] G. Salton, J. Allan, and C. Buckley. Approaches to Passage Retrieval in Full Text Information Systems. In *Proceedings of SIGIR 1993*, pages 49–58, 1993.

[16] A. Severyn, M. Nicosia, and A. Moschitti. Building Structures from Classifiers for Passage Reranking. In *Proceedings of CIKM 2013*, pages 969–978, 2013.

[17] D. Shen and M. Lapata. Using Semantic Roles to Improve Question Answering. In *Proceedings of EMNLP-CoNLL 2007*, pages 12–21, 2007.

[18] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton. Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. In *Proceedings of SIGIR 2003*, pages 41–47, 2003.

[19] J. Tiedemann. Integrating Linguistic Knowledge in Passage Retrieval for Question Answering. In *Proceedings of HLT/EMNLP 2005*, pages 939–946, 2005.

[20] W. Yih, M. Chang, C. Meek, and A. Pastusiak. Question Answering Using Enhanced Lexical Semantic Models. In *Proceedings of ACL 2013*, pages 1744–1753, 2013.