GeoTime Retrieval through Passage-based Learning to Rank *

Xiao-Lin Wang[1,2], Hai Zhao[1,2], Bao-Liang Lu[1,2†]
[1]Center for Brain-Like Computing and Machine Intelligence
Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems
Shanghai Jiao Tong University
800 Dong Chuan Rd., Shanghai 200240, China
arthur.xl.wang@gmail.com, {zhaohai; blu}@cs.sjtu.edu.cn

ABSTRACT

The NTCIR-9 GeoTime task is to retrieve documents to answer such questions as when and where certain events happened. In this paper we propose a Passage-Based Learning to Rank (PGLR) method to address this task. The proposed method recognizes texts both strongly related to the target topics and containing geographic and temporal expressions. The implemented system provides more accurate search results than a system without PGLR. Performance, according to the official evaluation, is average among submitted systems.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Algorithms, Performance, Experimentation

Keywords

GeoTime Retrieval, BM25F, Learning to Rank

1. INTRODUCTION

Queries that request geographic and temporal information, such as "where and when did ... happen?", are very common in information retrieval. Therefore, a customized search method processing geographic and temporal expressions may benefit a large number of users. The NTCIR-9 GeoTime task provides researchers with a platform for testing such methods.

The Center for Brain-like Computing and Machine Intelligence, Shanghai Jiao Tong University (SJTU-BCMI) participates in the NTCIR-9 GeoTime English subtask. We propose a Passage-based GeoTime Learning to Rank method (PGLR) to address this task. PGLR first evaluates the relevance and GeoTime informativeness of each passage in candidate documents, which are retrieved by a BM25F algorithm, and then re-ranks the documents according to the maximum score of their passages.

This paper takes paragraphs as passages, and various methods of feature extraction and ranking are applied. The underlying intuition is that a document could contain both query terms and GeoTime expressions without any semantic relation between them, and such a document would not be a correct answer for GeoTime Retrieval. On the other hand, if the query terms and GeoTime expressions appear in the same passage, they are more likely to have a semantic relation. A document containing such a passage is more likely to be a correct answer for GeoTime Retrieval.

This paper describes our proposed method and discusses its evaluation results. The rest of the paper is organized as follows: the system is presented in Sec. 2; then the submitted runs are described and discussed in Sec. 3; finally this paper in concluded in Sec. 4.

2. SYSTEM DESCRIPTION

The core of our system for NTCIR-9 GeoTime is to employ a passage-based GeoTime Learning to Rank (PGLR) method to improve the relevance related to GeoTime retrieval. Figure 1 presents its framework where PGLR is used as post processing after page retrieval and page rank.

The system works as follows:

1. Parse the GeoTime topics into query terms

2. Find the top-N relevant documents by BM25F similarity;

3. Re-rank with the PGLR algorithm;

The following three subsections give details.

*This work was partially supported by the National Natural Science Foundation of China (Grant No. 60903119, Grant No. 61170114 and Grant No. 90820018), the National Basic Research Program of China (Grant No. 2009CB320901), the Science and Technology Commission of Shanghai Municipality (Grant No. 09511502400), and the European Union Seventh Framework Programme (Grant No. 247619).
†Corresponding author

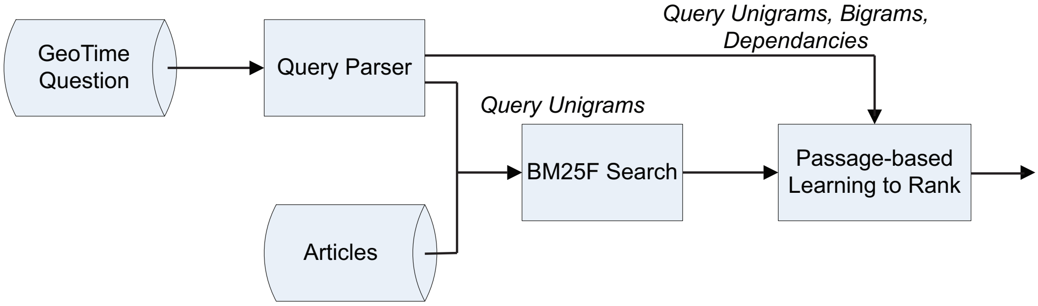

Figure 1: Framework of SJTU-BCMI's system for NTCIR-9 GeoTime task

2.1 GeoTime Topic Parsing

The representation of a GeoTime topic is normally a question that starts with "when" and "where" noted as DESCRIPTION, a small piece of narrowing description noted as NARRATIVE. In this step they are parsed into queries. Three kinds of representations – unigram, bigram and dependency – are employed in order to catch the semantics of a GeoTime topic. Tab. 1 gives four examples in NTCIR9-GeoTime.

The query unigrams, bigrams and dependencies are generated according to the following rules :

- *Unigrams of DESCRIPTION*(UoD) are the unigrams of the DESCRIPTION not in a customized stop word list. This list consists of common English stop words plus words with high frequency in GeoTime topics such as "when" and "where".
- *Bigrams of DESCRIPTION* (BoD) are pairs of UoDs which are neighbors, with stop words skipped.
- *Dependencies of DESCRIPTION* (DoD) are the results of the Stanford parser [1] [3, 1] that have at least one UoD.
- *Unigrams of NARRATIVE* (UoN) are the unigrams which appear in the *bigrams and dependencies of NARRATIVE* (BoN and DoN).
- *Bigrams of NARRATIVE* (BoN) are those neighboring non-stop unigrams in NARRATIVE that have at least one UoD.
- *Dependencies of NARRATIVE* (DoN) are those dependencies in NARRATIVE that have at least one UoD.

2.2 Document Retrieval with BM25F

The document retrieval component is a modified Lucene search engine. Lucene is an open source search engine from the Apache Software Foundation [2]. The default similarity model of Lucene is a combination of a boolean model and a vector space model [3]. We implemented the Standard BM25F similarity formula on Lucene [6, 9], and this slightly raises the accuracy according to our pilot experiments.

The implementation of BM25F follows Eq. 1 presented at [4].

$$BM25F(d,q) = \sum_{t \in q} \frac{tf(t,d)}{k_1 + tf(t,d)} \cdot idf(t) \quad (1)$$

$$tf(t,d) = \sum_{c \in d} w_c * tf_c(t,d) \quad (2)$$

$$tf_c(t,d) = \frac{occur_c(t,d)}{1 - b_c + b_c \frac{l_{d,c}}{l_c}} \quad (3)$$

$$idf(t) = log \frac{N - n(t) + 0.5}{n(t) + 0.5} \quad (4)$$

where d is a document, q is a query, t represents a word, c represents a field contained in d, $occur_c(t,d)$ is the number of occurrences of t in c of d, $l_{d,c}$ is the length of c at d, l_c is the average length of c, N is the number of documents, n(t) is the number of documents containing t, and k_1,w_c and b_c are parameters. $tf_c(t,d)$ is called the field term frequency function, $tf(t,d)$ is called the term frequency function, and $idf(t)$ is called the inverse document frequency.

In our system, the title and body of a document are taken as the two fields in BM25F (please see Sec. 3 for details).

2.3 Passage-based GeoTime Learning to Rank (PGLR)

The intuition of PGLR is that if a document has a small piece of text both strongly related to the GeoTime topic, and containing temporal and geographic expressions, this document is highly likely to be a correct one. By observing many documents of real-world news, the unit of passage is taken as the granule for computation.

Learning to rank [7] is taken to evaluate the rightness of each passage given a GeoTime Topic. The passages of previously retrieved documents are first converted to vector representations based on a group of features; then a trained SVM^{rank} [2] [4] predicts the rightness of these passages. In the end the documents are re-ranked according to the maximum rightness of their passages.

The following nine features are used to represent passages.

- the document's BM25F score
- the TF·IDF similarity of *DESCRIPTION's unigrams*
- the TF·IDF similarity of *DESCRIPTION's bigrams*
- the TF·IDF similarity of *DESCRIPTION's dependencies*

[1] `http://nlp.stanford.edu/software/lex-parser.shtml`
[2] `http://lucene.apache.org/`
[3] `http://lucene.apache.org/java/2_9_0/api/core/org/apache/lucene/search/Similarity.html`

[4] `http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

Table 1: Examples of Parsing GeoTime Questions into Queries

ID	Topic	Query unigram	Query bigram	Query dependency
26	(D)Where and when did the space shuttle Columbia disaster take place? (N)The user wants to know when and in what state the space shuttle Columbia exploded.	space shuttle columbia disaster state space shuttle columbia explode	space_shuttle shuttle_columbia columbia_disaster state_space space_shuttle shuttle_columbia columbia_explode	shuttle_space disaster_columbia shuttle_disaster columbia_space columbia_shuttle explode_columbia
27	(D)When was the last flight of Concorde and where did it land? (N)The user wants to know when was the last time that the supersonic airliner Concorde flew and in what city it landed.	last flight concorde land last time supersonic airliner concorde fly city land	last_flight flight_concorde concorde_land last_time airliner_concorde concorde_fly city_land	flight_last flight_concorde time_last concorde_supersonic concorde_airliner fly_concorde time_land fly_land
28	(D)When and where were the Washington beltway snipers arrested? (N)The user wants to know when and in what state were the Washington snipers arrested who killed ten people and critically injured three more.	washington beltway sniper arrest state washington sniper arrest kill	washington_beltway beltway_sniper sniper_arrest state_washington washington_sniper sniper_arrest arrest_kill	arrest_washington sniper_beltway washington_sniper sniper_state sniper_washington sniper_arrest sniper_kill
29	(D)When was the euro put in circulation and which three member states of the eurozone by that time declined its use? (N)The user wants to know when and in what state were the Washington snipers arrested who killed ten people and critically injured three more.	euro put circulation three member state eurozone time decline use exact euro official currency eurozone enter circulation information 15 member state european union adopt euro require	euro_put put_circulation circulation_three three_member member_state state_eurozone eurozone_time time_decline decline_use exact_euro euro_official currency_eurozone eurozone_enter enter_circulation circulation_information 15_member member_state state_european adopt_euro euro_require	put_euro put_circulation state_three state_member decline_state state_eurozone eurozone_time put_decline decline_use enter_euro euro_currency currency_eurozone enter_circulation state_member adopt_state state_union adopt_euro

- the TF·IDF similarity of *NARRATIVE's unigrams*
- the TF·IDF similarity of *NARRATIVE's bigrams*
- the TF·IDF similarity of *NARRATIVE's dependencies*
- whether it contains temporary named entities recognized by Stanford's parser
- whether it contains geographic named entities recognized by Stanford's parser

3. EXPERIMENTS

3.1 Experimental Setting

Four runs of SJTU-BCMI were submitted to NTCIR-9 GeoTime task. Table 2 summarizes the detailed approach of each run. The parameters of BM25F are tuned to maximum performance on the dataset of the NTCIR-8 GeoTime task. The features are transferred into vectors for classification. SVM^{rank} is used to learn the weights of features, and the results are presented in the last column of Table 2.

3.2 Experimental Results

NTCIR-9 INTENT-DR employs five evaluation matrices: MAP,Q, NDCG@10, NDCG@100 and NDCG@1000. The evaluation results for SJTU-BCMI's runs are presented in Table 3. The runs using the PGLR component provide higher accuracies as expected, since scores of RUN2 and RUN4 are higher than those of RUN1 and RUN3.

4. CONCLUSIONS

In this paper, we proposed a Passage-based GeoTime Learning to Rank (PGLR) method to address the problem of GeoTime retrieval. The implemented system achieves above average performances on the NTCIR-9 GeoTime task. The proposed PGLR obviously improves the GeoTime-related relevance according to the official evaluations of comparison experiments.

The idea of PGLR is actually to recognize the rightness of a piece of text serving as the answer to a GeoTime topic (or question). In this paper we employ a linear model with the input of TF·IDF similarities and presence of temporal and geographical expressions. In the future, we plan to incorporate more technologies from the domain of automatically question answering [5, 8]

5. REFERENCES

[1] M. De Marneffe, B. MacCartney, and C. Manning. Generating typed dependency parses from phrase structure parses. In *LREC 2006*. Citeseer, 2006.

[2] T. Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM, 2006.

[3] D. Klein and C. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003.

[4] J. Pérez-Agüera, J. Arroyo, J. Greenberg, J. Iglesias, and V. Fresno. Using bm25f for semantic search. In *Proceedings of the 3rd International Semantic Search Workshop*, pages 1–8. ACM, 2010.

[5] D. Ravichandran and E. Hovy. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 41–47. Association for Computational Linguistics, 2002.

[6] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. NIST, 1995.

[7] A. Trotman. Learning to rank. *Information Retrieval*, 8(3):359–381, 2005.

[8] E. Voorhees and H. Dang. Overview of the trec 2002 question answering track. *NIST SPECIAL PUBLICATION SP*, pages 57–68, 2003.

[9] H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. Robertson. Microsoft cambridge at trec-13: Web and hard tracks. In *Proceedings of TREC 2004*. Citeseer, 2004.

Table 2: Description of SJTU-BCMI submitted runs to GeoTime English subtask

Runs	NARRATIVE?	Doc. Retrieval	PGLR?(*weight*-s)
SJTUBCMI-EN-EN-01-DN	Yes	BM25F[a]	No
SJTUBCMI-EN-EN-02-DN	Yes	BM25F[a]	Yes(6.8,1.1,0.1,0.0,1.2,0.3,0.0,0.6,0.5)[b]
SJTUBCMI-EN-EN-03-D	No	BM25F[a]	No
SJTUBCMI-EN-EN-04-D	No	BM25F[a]	Yes(6.8,1.1,0.1,0.0,0.6,0.5)[c]

[a] The parameters are k_1=2, b_{title}=0.1, w_{title}=4, b_{main}=0.1,w_{main}=2.
[b] corresponding to the features in Sec. 2.3.
[c] corresponding to the features in Sec. 2.3 with three NARRATIVE's similarities removed.

Table 3: Evaluation of SJTU-BCMI submitted runs to NTCIR-9 Intent-DR Chinese subtask

RunName	MAP	Q	nDCG@10	nDCG@100	nDCG@1000
SJTUBCMI-EN-EN-01-DN	0.3326	0.3557	0.4498	0.4511	0.5772
SJTUBCMI-EN-EN-02-DN	0.3648	0.3884	0.5127	0.4977	0.6045
SJTUBCMI-EN-EN-03-D	0.2885	0.3111	0.4217	0.4085	0.5338
SJTUBCMI-EN-EN-04-D	0.3141	0.3370	0.4429	0.4523	0.5567
MAXIMUM	0.5549	0.5738	0.7379	0.6883	0.7654
AVERAGE	0.3517	0.3710	0.4794	0.4676	0.5684
MEDIAN	0.3326	0.3512	0.4591	0.4563	0.5772
MINIMUM	0.1392	0.1474	0.1712	0.2409	0.3195