Mohamed Vall Mohamed Salem
Khaled Shaalan   Farhad Oroumchian
Azadeh Shakery   Halim Khelalfa (Eds.)

LNCS 7097

Information Retrieval Technology

7th Asia Information Retrieval Societies Conference, AIRS 2011
Dubai, United Arab Emirates, December 2011
Proceedings

Springer

Lecture Notes in Computer Science 7097

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Mohamed Vall Mohamed Salem
Khaled Shaalan Farhad Oroumchian
Azadeh Shakery Halim Khelalfa (Eds.)

Information Retrieval Technology

7th Asia Information Retrieval
Societies Conference, AIRS 2011
Dubai, United Arab Emirates, December 18-20, 2011
Proceedings

Springer

Volume Editors

Mohamed Vall Mohamed Salem
University of Wollongong in Dubai, Faculty of Computer Science and Engineering
Dubai Knowledge Village, P.O. Box 20183, Dubai, United Arab Emirates
E-mail: mohamedsalem@uowdubai.ac.ae

Khaled Shaalan
British University in Dubai, Faculty of Engineering and IT
Dubai International Academic City, Block 11, 1st and 2nd Floor
P.O. Box 345015, Dubai, United Arab Emirates
E-mail: khaled.shaalan@buid.ac.ae

Farhad Oroumchian
University of Wollongong in Dubai, Faculty of Computer Science and Engineering
Dubai Knowledge Village, P.O. Box 20183, Dubai, United Arab Emirates
E-mail: farhadoroumchian@uowdubai.ac.ae

Azadeh Shakery
University of Tehran, Faculty of Engineering
Department of Electrical and Computer Engineering
North Kargar Street, P.O. Box 14395-515, Tehran, Iran
E-mail: shakery@ut.ac.ir

Halim Khelalfa
University of Wollongong in Dubai, Faculty of Computer Science and Engineering
Dubai Knowledge Village, P.O. Box 20183, Dubai, United Arab Emirates
E-mail: halimkehlalfa@uowdubai.ac.ae

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-25630-1 e-ISBN 978-3-642-25631-8
DOI 10.1007/978-3-642-25631-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011941661

CR Subject Classification (1998): H.3, H.4, F.2.2, I.4-5, E.1, H.2.8

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The Asian Information Retrieval Societies Conference (AIRS) is one of the most established and competitive information retrieval conferences; the seventh edition of this conference (AIRS 2011) aimed to bring together international researches and developers to exchange new ideas and the latest results in information retrieval (IR). The scope of the conference encompassed theory and practice of all aspects of IR in text, audio, image, video and multimedia data. The call for papers invited submissions to the following areas of research:

- Arabic Script Text Processing and Retrieval
- IR Models and Theories
- Multimedia IR
- User Study, IR Evaluation, and Interactive IR
- Web IR, Scalability and Adversarial IR
- IR Applications
- Machine Learning for IR
- Natural Language Processing for IR

AIRS 2011 was the first edition to be organized in the western part of the Asian continent with a growing interest to foster IR research and communalities in natural language processing. A new track on Arabic Script Text Processing and Retrieval was added for the first time to the main areas of research in the conference.

Historically, AIRS 2011 is a continuation of the series of conferences that grew from the Information Retrieval with Asian Languages (IRAL) workshop series back in 1996. It has become a mature venue of IR work, finding support from the ACM Special Interest Group and Information Retrieval (SIGIR) and many other associations.

The Organizing Committee was very pleased with the quality and level of interest received to our call for contributions from the research community in the IR field. We received a total of 132 papers representing work by academics and practitioners from all over the world and we would like to thank all of them. The Program Committee used a double-blind reviewing process and as result 31 articles (23.5%) were accepted as full papers and 25 (19%) were accepted as short (poster) papers.

The success of this conference was only possible with the support of the extremely active Program Committee members without whom the present proceedings would not have been possible. We would like to acknowledge the contributions of Ali Farghaly (Oracle, USA), Minjie Zhang (University of Wollongong, Australia), Joemon M. Jose (University of Glasgow, UK), Tetsuya Sakai (Microsoft Research Asia), Min Zhang Tsinghua (University, China), Wang Bin (Chinese Academy of Sciences, China), Tie-Yan Liu (Microsoft Research Asia) and Chia-Hui Chang (National Central University, Taiwan).

For a conference to run smoothly, much behind-the-scene work is necessary, most of which is largely unseen by the authors and delegates. We would like to thank our Publication Chairs (Azadeh Shakery and Halim Khelalfa) who painstakingly worked with each individual author to ensure formatting, spelling, dictation and grammar were completely error-free.

October 2011

Khaled Shaalan
Farhad Oroumchian
Mohamed Vall Mohamed Salem

Organization

AIRS 2011 was organized by the Faculty of Computer Science and Engineering, University of Wollongong in Dubai in cooperation with ACM/SIGIR.

Executive Committee

Conference Co-chairs

Mohamed Val Salem	University of Wollongong in Dubai, UAE
Farhad Oroumchian	University of Wollongong in Dubai, UAE

Program Co-chairs

Khaled Shaalan	British University in Dubai, UAE
Farhad Oroumchian	University of Wollongong in Dubai, UAE

Publication Co-chairs

Azadeh Shakery	University of Tehran, Iran
Halim Khelalfa	University of Wollongong in Dubai

Publicity Chair

Asma Damankesh	University of Wollongong in Dubai
Abolfazl AleAhmad	University of Tehran, Iran

Program Committee

Program Co-chairs

Khaled Shaalan	British University in Dubai, UAE
Farhad Oroumchian	University of Wollongong in Dubai, UAE

Area Chairs

User Study, IR Evaluation, and Interactive IR	Tetsuya Sakai, Microsoft Research Asia
IR Models and Theories	Minjie Zhang, University of Wollongong, Australia
Multimedia IR	Joemon M Jose, University of Glasgow, UK

IR Applications	Kazunari Sugiyama, National University of Singapore, Singapore
Web IR, Scalability and Adversarial IR	Min Zhang, Tsinghua University, China Wang Bin, Chinese Academy of Sciences (CAS), China
Machine Learning for IR	Tie-Yan Liu, Microsoft Research Asia
Natural Language Processing for IR	Chia-Hui Chang, National Central University, Taiwan
Arabic Script Text Processing and Retrieval	Ali Farghaly, Oracle, USA

Referees

I. Sengor Altingovde
I. Abu El-Khair
M. Adriani
A. Aizawa
Z. Al Aghbari
H. Aliane
H. Amiri
M. Asadpour
E. Ashoori
C. Man Au Yeung
S. Bandyopadhyay
L.H. Belguith
P. Bhattacharyya
W. Bin
M. Blondel
R. Cai
T.W. Cai
F. Can
S. Carberry
V. Cavalli-Sforz
Chia-Hui Chang
Chien Chin Chen
Hsin-Hsi Chen
Sung-Kwon Choi
G. Dupret
M. Diab
W. Ding
K. Eguchi
A. Elnagar
M. Evens
H. Faili
C. Gurrin
L. Gao
R. Girju
Bo Gong
G. Grefenstette
T. Gungor
Jiafeng Guo
N. Habash
A. Hamdi-Cherif
Y. Harada
F. Harrag
K. Hatano
Ben He
Sam Yin He
Yulan He
Xuanjing Huang
S. Hussain
A. Jatowt
A.M. Jaoua
K. Jarvelin
Wei Jin
Hideo Joho
G. Jones
Hanmin Jung
Noriko Kando
In-Su Kang
E. Kanoulas
Jaana Kekalainen
S. Khadivi
Jungi Kim
F. Kimura
K. Kishida
P. Kolari
K. Kuriyama
M. Lalmas
A. Lampert
F. Lazarinis
Seungwoo Lee
Yeong Su Lee
Hyowon Lee
Guoliang Li
Peng Li
Sujian Li
Xiaojun Lin

Yiqun Liu
Zhiyuan Liu
Wen-Hsiang Lu
Yuanhua Lv
Qiang Ma
A. Maeda
M. Melucci
M.R. Mustaffa
C. Mahlow
M. Maragoudakis
K. Megerdoomian
D. Metzler
R. Mitkov
M.F. Moens
A. Moffat
T. Mori
Seung-Hoon Na
P. Nakov
H. Nanba
Kenta Oku
M. Okumura
F. Oroumchian
I. Ounis
W. Radford
F. Radlinski
A. Rafaa
R. Ren
K. Riaz
Hae-Chang Rim
I. Ruthven
T. Sakai
M. Sanderson
J. Savoy
F. Scholer
K. Seki
S. Sekine
A. Shakery
M. Shamsfard
H. Shima
M. Shokouhi
T. Shumizu
Sa-Kwang Song
Young-In Song
A. Soudi
K. Sugiyama
A. Sun
L. Sun
Y. Suzuki
Sim-Hui Tee
U.S. Tiwary
M. Vilares-Ferro
J. Vilares-Ferro
Jingdong Wang
Mingwen Wang
Taifeng Wang
W. Webber
R. White
Kam-Fai Wong
Yunqing Xia
Zheng Ye
E. Yilmaz
M. Yoshioka
N. C. Zamorani
A. ZarehBidoki
Lanbo Zhang
Yin Zhang
Jun Zhao
I. Zitouni

Sponsoring Institutions

Dubai Knowledge Village
Micorsoft Middle East

Table of Contents

User Study, Information Retrieval Evaluation and Interactive Information Retrieval

Web Information Retrieval, Scalability and Adversarial Information Retrieval

Machine Learning for Information Retrieval

Natural Language Processing for Information Retrieval

Arabic Script Text Processing and Retrieval

Maintaining Passage Retrieval Information Need Using Analogical Reasoning in a Question Answering Task

Hapnes Toba, Mirna Adriani, and Ruli Manurung

Faculty of Computer Science, University of Indonesia
Depok 16424, Indonesia
hapnes.toba@ui.ac.id, {mirna,maruli}@cs.ui.ac.id

Abstract. In this paper we study whether a question and its answer can be related using analogical reasoning by using various kinds of textual occurrences in a question answering (QA) task. We argue that in a QA passage retrieval context, low cost language features can contribute some positive influence in the representation of the information need that also appears in other passages, which have some analogical features. We attempt to leverage this through query expansion and query stopwords exchange strategies among analogical question answer pairs, which are modeled by a Bayesian Analogical Reasoning framework. Our study by using ResPubliQA 2009 and 2010 dataset shows that the predicted analogical relation between question answer pairs can be used to maintain the information need of the QA passage retrieval task, but has a poor performance in determining the question type. Our best accuracy score was achieved by using*'bigram occurrences by using stemmer and TF-IDF weighting completed with named-entity'* feature set for the query expansion approach, and *'bigram occurrences by using stemmer and TF-IDF weighting'* feature set for the stopwords exchanged approach.

Keywords: Bayesian Analogical Reasoning, Question Answering System, Passage Retrieval, Query Expansion, ResPubliQA.

1 Introduction

Question Answering System (QAS) is a form information retrieval that tries to produce an exact answer given a natural language question. Despite its natural task to find a single answer, a QAS needs supporting textual context from one or more document collections, in the size of a sentence, a passage, a paragraph or even the whole document [1]. This is one of the reasons why question answering also considered as a challenging field in information retrieval.

Most typical QAS pipeline architectures consist of four main components, i.e.: question analyzer, query formulation, information retrieval and answer validation. The task of a question analyzer is to classify a question into one or more question types, which will be used as the expected answer type during the answer validation phase. In the query formulation component, a question will be formulated into a

M.V.M. Salem et al. (Eds.): AIRS 2011, LNCS 7097, pp. 489–498, 2011.

specific keyword-based query, for instance by using bag-of-words (BOW) approach after removing stopwords, or by using WordNet for term expansion [2]. In the information retrieval component, usually by using third-party search engines, such as: Indri, or Lucene, appropriate top-*n* textual candidates will be retrieved. Finally, the answer validation component needs to validate whether a retrieved answer candidate reflects some information need, with respect to the expected answer type, and produce a final single answer. The difficult task of constructing a final answer will be made easier if the final answer is already included in a limited set of passage retrieval results [3, 4, 5]. In this context, the performance of an underlying information retrieval system is important to retrieve relevant passages.

Recent works in information retrieval strategies that are specific to the question answering (QA) task are mostly focused on: linguistic and semantic constraints [4, 6], relevance feedback [7], semantic role labeling [8] or by topic indexing [9]. Despite these recent approaches, performing QA passage retrieval in a more conventional information retrieval way, i.e. by using textual features consisting of appropriate question terms, could be preferable if important search terms are already stated in the question. Recently, a new approach has been developed that focuses in the relational data between existing questions answer pairs [10]. By assuming that answers are related to their questions through certain types of implicit links, it is theoretically possible to learn these links from existing data, and to apply the learned model for relating unseen questions to their appropriate answers.

Table 1. Example of overlapping information need between QA pairs collections

QA Pairs Collections	Question	Passage Gold Standard
ResPubliQA 2010 (#91, question type: Factoid)	In which country will the **2010 FIFA World Cup** be held?	Repeats its demand that the Mugabe regime … value from either the run-up to the **2010 World Cup** or the tournament itself; in this regard, calls on [South Africa], the host nation, and on **FIFA** to exclude Zimbabwe … in pre-World Cup matches, … national teams involved in the event;
ResPubliQA 2009 (#7, question type: Factoid, feature: 'unigram occurrences')	**In** which **areas** will **objective information** be provided on **drugs and drugs addiction**?	The Centre's **objective** is to provide, [**in** the **areas** referred to **in** Article 4], the Community and … with **objective**, reliable and comparable **information** … … **drugs and drug addiction** and their consequences.
Overlapping unigram between QA pairs	in, which, will, be, on, and, the	

Inspired by this work, we study whether a question and its answer can be related using low cost language features in a QA passage retrieval scenario. We argue that in a QA passage retrieval context, low cost language features, such as *n*-gram, can actually contribute some positive influence to represent the information need that also appear in other passages, which have some analogical or related features. Table 1 gives an example of such a case, the words in **bold** show the overlapping words between the question and the answer, underlined words show the overlapping words between both question answer pairs, and the text surrounded by '[ ]' are the exact answers.

In Table 1, seems that the two questions have different question types. The first question could be classified into *'COUNTRY'* question type and the second one into *'LOCATION'* question type. But if we consider the QA pairs as a relation, the answers of both question are directing into something in common, i.e. a kind of named-entity, by using the question word *'WHICH'*, either it is about 'location of an event' or 'location of a section in a regulation document'. In this way, we could define an analogy as measure of similarity between structures of related objects.

2 Question Answering Approach

As stated in [2, 3], most typical QA architectures consider questions and answers as independent elements. The consequence of this kind of architecture is that question type and the related expected answer type cannot be learned in a single learning mechanism framework. To compensate for the independence of a question and its answer pair, we exclude the question type component in our approach, and use a single learning mechanism framework to learn the relations of a question and its answer pair. We propose to use the related features of a question and its answer as a means to recognize the information need of a question, which at the same time could also give an indication of how a question should be answered. The related features are learnt from a collection of question answer pairs, in which the answer is given in the form of a passage. The complete approach can be seen in Fig. 1.

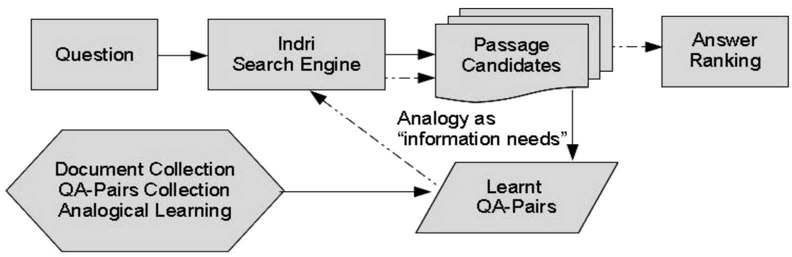

Fig. 1. Proposed QAS Approach. Question analysis component is excluded and the learnt question answer pairs are used as a means to recognize the information need.

In our approach, we assume that all words that appear in a question are important and have influences during the information retrieval phase. Each passage candidate and its related question will be compared to a set of question answer pairs which relation has been learnt by using an analogical learning mechanism. This comparison value will produce a new score that shows how information need from a passage candidate is related to the set of learnt question answer pairs, which also depends on the feature set during the analogical learning. We hypothesized that it is theoretically possible to exchange those related features or to enhance the original question, and use them to re-run the query in a second phase (shows as dotted-line in Fig. 1), to form the final passage answer.

3 Bayesian Analogical Reasoning

Bayesian Analogical Reasoning (BAR) was originally introduced in [11, 12], which basic idea is to learn model parameters and priors from related objects, and update it during the comparison process of a query to obtain marginal probability that relates the query with the objects that have been learnt.

Assume there is a space of unseen functions $Q \times A \rightarrow \{0, 1\}$. If two objects, a question Q and an answer A are members of a set S, which are related by an unknown function $f(Q,A) = 1$, what needs to be quantified is how similar the function $f(Q,A)$ is to another unseen function $g(. , .)$, that classifies all pairs of $(Q^i, A^j) \in S$ as being linked where $g(Q^i,A^j) = 1$. The functions $f(., .)$ and $g(. , .)$ are unseen, and thus we need a set of priors that will be used to integrate them over the function space.

Suppose for each pair $(Q^i \in Q, A^j \in A)$, there exists a feature vector: $X^{ij} = [\Phi_1(Q^i, A^j) \ldots \Phi_k(Q^i, A^j)]^T$, defined by the mapping: $\Phi : Q \times A \rightarrow \Re^K$, as a single point of link representation on the feature space Φ.

This feature space mapping computes a K-dimensional vector of features of the question answer pairs, which is hoped to have a relevant link prediction between the objects in the pairs. The feature vector X^{ij}, for each pair of question and answer consists of the same number of features, and thus we can define a measure as the link representation between such pair. In this case we use the cosine distance.

If there is an unseen label L^{ij}, with $L^{ij} \in \{0,1\}$ as a predicted indicator of the existence of a relation between Q^i and A^j, then we will have a model parameters vector $\Theta = [\Theta_1 \ldots \Theta_K]^T$ which models the presence or absence of interaction between objects, and could be learnt by performing the logistic regression model:

$$P\left(L^{ij} = 1 \mid X^{ij}, \Theta\right) = \ell ogistic\left(\Theta^T X^{ij}\right) \tag{1}$$

where *logistic(x)* is defined as: $1/(1 + e^{-x})$.

The priors are learnt by using:

$$P(\Theta) = N(\tilde{\Theta}, (c\tilde{T})^{-1}) \tag{2}$$

where $\tilde{\Theta}$ is the Maximum Likelihood Estimator (MLE) of Θ $N(m, v)$ is a normal of mean m and variance V. Matrix $\tilde{T}$ is the empirical second moment's matrix of the link object features, and c is a smoothing parameter, which is set to the number of links that exist in the trained set.

During the retrieval process of linked pairs, a query is compared by the functions for links prediction by marginalizing over the parameters of the functions. If we have L^S as the vector of link predictions for S, then each $L \in S$ has the value $L = 1$, indicating that every pair of objects in S is linked. The final score of a retrieval process indicating the order of predicted links between the query and the related objects that has been learnt, and is compute as follows:

$$\mathrm{score}\left(Q^i, A^j\right) = \log P\left(L^{ij} = 1 \mid X^{ij}, S, L^S = 1\right) - \log P\left(L^{ij} = 1 \mid X^{ij}\right) \tag{3}$$

Silva et al. in [11, 12] use the variational logistic regression to compute this scoring function.

4 Experimental Setting

To evaluate the influence of analogical reasoning in our question answering scenario, we use Indri [13, 14], to retrieve the top-*5* passage candidates of the testing question set, in a bag-of-words approach. We evaluate the performance in terms of accuracy at the top-1 answer against the gold standard. The accuracy is computed according to the formula:

$$C@1 = tp/(tp + fp) \tag{4}$$

where *tp* is the number of 'true' answer at the top-1, and *fp* is the number of 'wrong' answer.

We use the question answer pairs from ResPubliQA 2009 paragraph selection gold standard as our training set, and ResPubliQA 2010 as our testing data. To maintain the question type's equality between the two question collection sets, we use the: *Definition* (95 questions), *Factoid* (139), *Reason-Purpose* (187), and *Procedure* (79) question types during the experiments, and exclude the Opinion and Other question types. In total, we have 500 question answer pairs from ResPubliQA 2009 collection, and 133 questions from ResPubliQA 2010, which consist of: *Definition* (32 questions), *Factoid* (35), *Reason-Purpose* (33), and *Procedure* (33) .

The document collections that were used during this study are JRC-ACQUIS and EUROPARL [15]. We created an index that was based on paragraph segmentation with Indri indexing tools. In total, we have about 1.5 million paragraphs indexed that were considered as documents. Indri is a search engine that is specially designed for passage retrieval, thus will be fitted to the retrieval task in this study [14]. The works in [15] showed that paragraph selection is a challenging task, and one of the successful methods is to improve paragraph retrieval by using overlapping uni- and bigram occurrences' as contextual information. This is the main motivation that we explored the following textual feature sets during the experiments:

- unigram occurrences;
- unigram occurrences after using (Porter) stemmer;
- unigram occurrences by removing stopwords and using a stemmer;
- unigram occurrences by using TF-IDF weighting;
- unigram occurrences by using stemmer and TF-IDF weighting;
- unigram occurrences completed with named-entity;
- bigram occurrences;
- bigram occurrences by using a stemmer and TF-IDF weighting;
- bigram occurrences by using stemmer and TF-IDF weighting completed with named-entity;
- named-entity occurrences, by using Stanford NER and dictionary-based NER [2].

Finally, we decomposed the feature sets into SVD 25-dimension, as the main feature dataset, in order to reduce the word features dimensionality.

During the passage retrieval phase, we made two types of query enhancement. The first one is to add overlapping non-stopwords word occurrences, which appear in the top-*5* retrieved analogous question answer pairs, to the original question, as a kind of query expansion method. The second one is to use the stopwords that appear in the best analogous question answer pair, to complete the stopwords that was removed from the original question. Each query is considered as a BOW model. To evaluate our expansion method, we also run some experiments of the original question by using Indri pseudo-relevance feedback, which were set to some configuration of smoothing parameters and weights for the original query [14, 16].

5 Results and Discussions

We present our results in the following aspects: the accuracy performance of the query expansion strategies, the comparison of the performance with respect to Indri relevance feedback, and the influence of the retrieved analogous pairs to the passage retrieval ranking performance.

5.1 Performance of Query Expansion

The result of the re-run scenario by using overlapping of non-stopwords which occur in the top-*5* of analogical pairs is given in Table 2. The best accuracy score was achieved by the *'bigram occurrences by using stemmer and TF-IDF weighting completed with named-entity'* feature set, i.e. 0.31. For our example in Table 1, by using the *'bigram-stem-TFIDF-ne'* feature set, the question will be enriched by some other non-stopwords term(s) that occur in the top-*5* analogical pairs, as follows : *"In which country will the 2010 FIFA World Cup be held + European"*.

The low accuracy performance is mainly influenced by out-of-topic terms. Such cases are mostly occurred when the analogical pairs come from different topic or when the semantics of the question and answer are too far to be captured in the analogical model. For example the question: *"what is maladministration?" (Q#6-2010: Definition type)*, the analogical model only considered the word *"what is"*, as related important features, and thus fail to enrich the query with something that is related to "*maladministration"*.

To create another view of retrieval performance, a running result of the original questions that were expanded using WordNet is included in Table 2. This expansion strategy is performed for every verb and noun from the original questions [2, 17]. The result is mostly below the accuracy of our approach. The original questions were mostly enriched with out-of-topic terms which decreased the retrieval accuracy. For our example in Table 1, the query would be expanded as follows: *"In which (country OR commonwealth OR state OR land OR nation OR "res publica" OR "body politic") will the 2010 FIFA World Cup be (held OR maintained OR kept)"*.

Table 3 shows the result of the re-run scenario by using the stopwords that appear in the best analogous question answer pair. Again, the retrieval result for the WordNet query expansion is also included, which performance is lower than our approach. The best performed feature set is the '*bigram occurrences by using stemmer and TF-IDF weighting',* with 0.34 accuracy. For our example in Table 1, the question, by using the *'bigram occurrences by using stemmer and TF-IDF weighting',* will be reformulated as the following bag-of-words query: *"country FIFA World Cup held + What is the of A top side to of which is"*.

The stopwords exchanged have the same failure analysis as the non-stopwords enhancement. In overall the stopwords exchanged performance is slightly better in terms of accuracy than the non-stopwords expansion.

Table 2. Overlapping non-stopwords performance in decreasing order

"Overlap-of-Terms-Top5" Features	C@1
Indri BOW	0.35
Bigram-stem-TFIDF-NE	0.31
Bigram-stem-TFIDF	0.30
Named- entity	0.30
Bigram	0.29
Unigram	0.26
Unigram-stem-remove stopwords	0.26
Unigram-stem-remove stopwords-NE	0.24
Unigram-TFIDF	0.24
Unigram-stem	0.23
WordNet	0.23
Unigram-TFIDF-stem	0.21

Table 3. Stopwords exchanged performance in decreasing order

"Stopwords-Exchanged" Features	C@1
Indri BOW	0.35
Bigram-stem-TFIDF	0.34
Bigram-stem-TFIDF-NE	0.33
Unigram-TFIDF-stem	0.33
Unigram-stem	0.31
Bigram	0.29
Unigram	0.29
Unigram-stem-remove stopwords	0.28
Unigram-stem-remove stopwords-NE	0.28
Unigram-TFIDF	0.27
Named-entity	0.26
WordNet	0.23

5.2 Indri Pseudo-relevance Feedback

To evaluate the performance of our best feature set expansion approach, we compare our results to the Indri pseudo-relevance feedback of the original questions, with various parameter settings. The first parameter setting is regarding the document smoothing to overcome data sparseness problem [14]. We use Dirichlet smoothing, and experimenting with: μ *= 2500 (default),* and μ *= 2000 (optimum for the query and document length).* Those values were chosen based on the work in [16]. Another parameter setting is the Indri feedback smoothing *(fbMu = 0.0 (default),* and *0.5)*, the query word weighting *(fbOrigWeight = 0.5, 0.8 and 1.0),* the number of terms for the feedback *(fbTerms=10)*, and the number of documents for the feedback *(fbDocs=5)*. The comparison is presented in Table 4.

Table 4. Best BAR feature set (*Bigram-stem-TFIDF-NE*) of non-stopwords expansion vs. Indri pseudo-relevance feedback in decreasing order

Parameter Setting	C@1	MRR@5
Indri BOW (mu = 2500, no relevance feedback)	0.35	0.45
Indri BOW (mu = 2500, fbOrigWeight = 1, fbMu = 0)	0.35	0.45
Indri BOW (mu = 2000, fbOrigWeight = 0.8, fbMu = 0.5)	0.32	0.43
BAR expansion with *'Bigram-stem-TFIDF-NE'* (mu=2500; no rel. feedb.)	0.31	0.43
Indri BOW (mu = 2000, fbOrigWeight = 0.5, fbMu = 0.5)	0.31	0.43
Indri BOW (mu = 2500, fbOrigWeight = 0.5, fbMu = 0.5)	0.29	0.40

From Table 4, we can observe that the accuracy of our expansion approach (0.31) is quite similar to the accuracy of Indri pseudo relevance feedback (0.32). This indicates that the expanded terms of the analogical question answer pairs can maintain the information need of the original query. Further analysis on the top-*5* retrieval results, in terms of Mean Reciprocal Rank (MRR) performance; give us promising results for answer validation strategy, which is beyond this study.

5.3 Question Type and Retrieval Performance Issues

Table 5 gives a number of analogous pairs examples, relating to the question type, of the *'Bigram-TFIDF-NE'* feature set in Table 2. The 'question type' classification accuracy from this feature set is 0.31. In our opinion, one of the problems is due to the term variations in the training and testing sets. The ResPubliQA collection [15] is characterized by its wide scope of questions and documents coverage in parliamentary domains. On the other hand, the BAR framework assumes that the feature space should provides a reasonable classifier to predict the existence of links. Such case is not in general decomposable as similarities between only the textual features in the question part, but also the presence or absence of the features in the answer part of related pair.

Table 5. Examples of question analogical pairs with respect to the question type (QT)

No.	2010	Question	QT	2009	Best Analogous Question	QT
1.	#73	What actions does the competent authority for maritime security of a port carry out?	PR	#338	What should be done in the case of epizootic?	PR
2.	#91	In which country will the 2010 FIFA World Cup be held?	F	#216	Who will be involved in radiotherapeutic practices?	F
3.	#105	What is the WTO Agreement?	D	#58	Why is the increase of the weight of the 50 cent coin from 7 g to 7,8 g necessary?	R
4.	#188	What was the purpose of EU states in establishing new permanent political and military bodies?	R	#418	What is the main objective of producing electricity in public thermal plants?	R

If we inspect further into each question type, for instance, from the *'Bigram-TFIDF-NE'* feature set in Table 2, this will gives us a distribution of question type accuracy that can be seen in Table 6. Such typical distribution also occurs in other feature sets in Table 2 and Table 3. The *Reason-Purpose* question type always has the best accuracy, and *Definition* question type always has the lowest one.

Table 6. Question Type Accuracy Distribution *'Bigram-TFIDF-NE'* Feature Set in Table 2

Question Type	Accuracy
Reason-Purpose	0.45
Factoid	0.37
Procedure	0.21
Definition	0.19

This result gives us an indication that the analogical relations among common bigram terms, such as: *'in order'*, *'order to'* or *'objective to'* in the *Reason-Purpose* type, could provide us much better expanded terms, in contrast to the relations of quite specific terms in the *Definition* question types, such as: *'define as'*, or *'the meaning*.

Table 7 presents some cases of expanded queries with their influence to the retrieval ranking performance.

Table 7. Some examples (as in Table 5), of expanded queries with their influence to the retrieval ranking (we only consider the top-5 retrieval)

No.	2010	Question	Expanded Q. Terms	Baseline Retrieval	After Expansion
1.	#73	What actions does the competent authority for maritime security of a port carry out?	application, competent	5	2
2.	#91	In which country will the 2010 FIFA World Cup be held?	european	1	1
3.	#105	What is the WTO Agreement?	order	> 5	> 5
4.	#188	What was the purpose of EU states in establishing new permanent political and military bodies?	account, competent, decide, order	> 5	4

The result presented in Table 7 indicates that a simple question has in fact more term variations in the answer, as for example in the *Definition* type. In contrast, a more complex question with numerous term occurrences' in the answer part, has the tendency to be more related to their analogous pair, and hence could achieved a better retrieval performance, as in the *Reason-Purpose* type.

6 Conclusions and Future Work

In general we conclude that the predicted analogical relation between question answer pairs can be used to maintain the information need of the QA passage retrieval task (c.f. section 5.2), but in the case of our experiments, analogical reasoning does a very poor job of classifying the expected answer type of a question (c.f. section 5.3). The overall passage accuracy in this study is much below the best performance of the ResPubliQA 2010 baseline [15], which is 0.73. It seems that the feature sets which were explored during the experiments are not enough to bridge the semantic gap between question and answer pairs. The choice of feature set is a crucial step in our study, which give significant influence to the retrieval results. In our study the best performed feature set is *'bigram occurrences by using stemmer and TF-IDF weighting completed with named-entity'* for the query expansion approach, and *'bigram occurrences by using stemmer and TF-IDF weighting'* for the stopwords exchanged approach (c.f. section 5.1).

Considering that we cannot always have all possibilities of question answer pairs during the training, it might valuable to aggregate patterns from *n*-most analogous question answer pairs, as recurring patterns, would seem to specify an indicative feature of the information need. Such automatic pattern generation strategy will be useful to expose question type analysis and its expected answer type in a question answering system. To address these issues we plan to conduct study in feature selection mechanism to be fitted in the analogical model as our future work.

References

1. Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., Karger, D.R.: The Role of Context in Question Answering Systems. In: Extended Abstracts on Human Factors in Computing Systems, Fort Lauderdale, Florida, pp. 1006–1007 (2003)
2. Schlaefer, N., Gieselmann, P., Schaaf, T., Waibel, A.: A Pattern Learning Approach to Question Answering Within the Ephyra Framework. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 687–694. Springer, Heidelberg (2006)
3. Bilotti, M.W.: Linguistic and Semantic Passage Retrieval Strategies for Question Answering. Dissertation Thesis. Language Technologies Institute. School of Computer Science, Carnegie Mellon University (2009)
4. Aktolga, E., Allan, J., Smith, D.A.: Passage Reranking for Question Answering Using Syntactic Structures and Answer Types. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 617–628. Springer, Heidelberg (2011)
5. Cui, H., Sun, R., Li, K., Kan, M.-Y., Chua, T.-S.: Question answering passage retrieval using dependency relations. In: SIGIR 2005, pp. 400–407. ACM, New York (2005)
6. Bilotti, M.W., Elsas, J., Carbonell, J., Nyberg, E.: Rank Learning for Factoid Question Answering with Linguistic and Semantic Constraints. In: Proceedings of CIKM (2010)
7. Pizzato, L.A., Molla, D., Paris, C.: Pseudo-Relevance Feedback using Named Entities for Question Answering. In: Australasian Language Technology Workshop, ALTW (2006)
8. Pizzato, L.A., Molla, D.: Indexing on Semantic Roles for Question Answering. In: Proceedings of the 2nd Workshop on Information Retrieval for Question Answering (2008)
9. Ahn, K., Webber, B.: Topic Indexing and Information Retrieval for Factoid QA. In: Proceedings of the 2nd ACL Workshop on Information Retrieval for Question Answering (2008)
10. Wang, X.-J., Tu, X., Feng, D., Zhang, L.: Ranking Community Answers by Modeling Question-Answer Relationship via Analogical Reasoning. In: Proceedings of SIGIR Conference (2009)
11. Silva, R., Heller, K., Ghahramani, Z.: Analogical Reasoning with Relational Bayesian-sets. In: Proceedings of AISTATS (2007)
12. Silva, R., Heller, K., Ghahramani, Z., Airoldi, E.: Ranking Relations Using Analogies in Biological and Information Networks. The Annals of Applied Statistics 4(2), 615–644 (2010)
13. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language-model based search engine for complex queries, `http://ciir.cs.umass.edu`
14. Metzler, D.: Indri Retrieval Model Overview, `http://ciir.cs.umass.edu/~metzler/indriretmodel.html`
15. Peñas, A., Forner, P., Rodrigo, A., Sutcliffe, R., Forascu, C., Mota, C.: Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation. In: Working Notes of CLEF ResPubliQA (2010)
16. Fang, H., Tao, T., Zhai, C.X.: A Formal Study of information Retrieval Heuristics. In: Proceeding of SIGIR. ACM, New York (2005)
17. Sun, R., Jiang, J., Tan, Y.F., Cui, H., Chua, T.S., Kan, M.Y.: Using syntactic and semantic relation analysis in question answering. In: Proceedings of the 14th Text REtrieval Conference (2005)