2011 International Conference on

Advanced Computer Science and Information Systems

(Proceedings)

Mercure Convention Centre, Jakarta

December 17th-18th , 2011

Published by:

Faculty of Computer Science

Universitas Indonesia

ISBN: 978-979-1421-11-9

Welcome Message from
General Chairs

On behalf of the Organizing Committee of this International Conference on Advanced Computer Science and Information Systems 2011 (ICACSIS 2011), we would like to extend our warm welcome to all of the presenter and participants, and in particular, we would like to express our sincere gratitude to our plenary and invited speakers.

This international conference is organized by the Faculty of Computer Science, Universitas Indonesia, and is intended to be the first step towards a top class conference on Computer Science and Information Systems. We believe that this international conference will give opportunities for sharing and exchanging original research ideas and opinions, gaining inspiration for future research, and broadening knowledge about various fields in advanced computer science and information systems, amongst members of Indonesian research communities, together with researchers from Germany, United Kingdom, Rusia, Australia, Japan, South Korea, Malaysia, Thailand, Vietnam and other countries.

This conference focuses on the development of computer science and information systems. Along with 5 plenary and 3 invited speeches, the proceedings of this conference contains 66 papers which have been selected from a total of 134 papers from fourteen different countries. These selected papers will be presented during the conference.

We also want to express our sincere appreciation to the members of the Program Commitee for their critical review of the submitted papers, as well as the Organizing Commitee for the time and energy they have devoted to editing the proceedings and arranging the logistics of holding this conference. We would also like to give appreciation to the authors who have submitted their excellent works to this conference. Last but not least, we would like to extend our gratitude to the Ministry of Education of the Republic of Indonesia, the Rector of Universitas Indonesia, and the Dean of the Faculty of Computer Science for their continued support towards the the ICACSIS 2011 conference.

Sincerely yours,
General Chairs
Ito Wasito

ISBN: 978-979-1421-11-9

 ISBN: 978-979-1421-11-9

Welcome Message from the Dean of Faculty of Computer Science, Universitas Indonesia

On behalf of all the academic staff and students of the Faculty of Computer Science, Universitas Indonesia, I would like to extend our warmest welcome to all the participants to the Mercure Convention Centre in Ancol, Jakarta on the occasion of the 2011 International Conference on Advanced Computer Science and Information Systems (ICACSIS).

Just like the previous two events in this series (2009 in Depok and 2010 in Bali), I am confident that ICASIS 2011 will play an important role in encouraging activities in research and development of computer science and information technology in Indonesia, and give an excellent opportunity to forge collaborations between research institutions both within the country and with international partners. The broad scope of this event, which includes both theoretical aspects of computer science and practical, applied experience of developing information systems, provides a unique meeting ground for researchers spanning the whole spectrum of our discipline. I hope that over the next two days, some fruitful collaborations can be established.

I also hope that the special attention devoted this year to the field of pervasive computing, including the very exciting area of wireless sensor networks, will ignite the development of applications in this area to address the various needs of Indonesia's development.

I would like to express my sincere gratitude to the distinguished invited speakers for their presence and contributions to the conference. I also thank all the program committee members for their efforts in ensuring a rigorous review process to select high quality papers.

Finally, I sincerely hope that all the participants will benefit from the technical contents of this conference, and wish you a very successful conference and an enjoyable stay in Jakarta.

Sincerely,
Professor T. Basaruddin Ph.D
Dean of the Faculty of Computer Science
Universitas Indonesia

 ISBN: 978-979-1421-11-9

Welcome Message from Vice Minister of Education Ministry of Education and Culture of the Republic of Indonesia

Ladies and Gentlemen, speakers and guests of the 2011 International Conference on Advanced Computer Science and Information Systems, or simply ICACSIS 2011, Good Day, Assalamu'alaikum Wr. Wb.

Allow me to first express my gratitude towards our honorary chairs and our honored speakers from all around the world, who have spared their valuable time to contribute to this conference along with all the other distinguished participants who have assembled here in Jakarta, over the next two days, for academic discussions on advanced computer science and information systems.

In today's information age, it seems that there is no longer an aspect of life that is unaffected by the advances of information and communication technology, or ICT. The Ministry of Education and Culture of the Republic of Indonesia recognizes that ICT has a huge role to play in addressing national issues and is committed to supporting research on how ICT can further solve these problems.

In recognition of the importance of ICT in national development, the Indonesian government's recently unveiled Master Plan for the Acceleration and Expansion of Indonesia's Economic Development (MP3EI) includes ICT as a crucial component of its 22 primary activities. This master plan is a bold initiative which aims to make Indonesia one of the world's 10 biggest economies by 2025, taking GDP to $4.5 trillion and increasing the per capita income from $3000 now to $15,000. One of the strategic initiatives of this Master Plan is to encourage large scale ICT investment, including the provision of essential infrastructure such as affordable and usable broadband throughout the archipelago.

Such initiatives will be expected to serve as an enabling technology, and the government sees the national education sector – particularly higher education – as one of the catalysts to leverage this technology to directly impact Indonesia's national competitiveness. To that end, the Ministry encourages researchers and academics to improve national competitiveness through outstanding research achievements in the field of ICT. There are many research areas which can improve Indonesia's competitiveness, ranging from e-Government solutions that improve efficiency and effectiveness of public services, to information retrieval systems that are able to support information requirements at lightning speed through various online media, to the state-of-the-art discoveries in fields such as nano technology and pervasive computing, which are expanding the horizons of what can be achieved with ICT.

 ISBN: 978-979-1421-11-9

The Ministry appreciates the efforts conducted by the organizing committee that has worked hard through this conference to achieve two important objectives towards the development of advanced computer science and information systems. Firstly, it is to disseminate the state of the art of research and development in ICT, cognizant of its significant value for Indonesia's future. Secondly, it is intended to provide a media for exchanging ideas and information concerning ICT. I am convinced that the scholars who have gathered here at this conference will bring valuable contributions to this discipline.

Finally, I want to convey my deep appreciation and gratitude to the Faculty of Computer Science, Universitas Indonesia, and all of our distinguished plenary and invited speakers. I hope this conference will be enlightening for all of us, and I hope also that we will be able to continuously collaborate to push the frontiers of science and solve the problems of our nation.

Sincerely,
Professor Musliar Kasim
Vice Minister of Education
Ministry of Education and Culture of the Republic of Indonesia

 ISBN: 978-979-1421-11-9

COMMITTEES

Honorary Chairs:

- T. Basaruddin, Dean of Faculty of Computer Science, Universitas Indonesia, ID
- Z. Bien, Fellow IEEE, KR
- B. Mirkin, University of London, UK / University Moscow, RU
- T. Fukuda, Director IEEE Regional X, JP

General Chairs:

- I. Wasito, Universitas Indonesia, ID
- Z.A. Hasibuan, Universitas Indonesia, ID
- H. Suhartanto, Universitas Indonesia, ID

Program Chairs:

- P. Mursanto, Universitas Indonesia, ID
- S. Yazid, Universitas Indonesia, ID
- W. Jatmiko, Universitas Indonesia, ID

Financial Chairs :

- R. Y. K. Isal, Universitas Indonesia, ID
- S. Louvan, Universitas Indonesia, ID

Program Committee:

- A. Tiu, Australia's National University, AU
- A. Tucker, Brunel University, UK
- A. Litchfield, University of Technology Sydney, AU
- A. Buono, Bogor Agricultural University, ID
- A.Z. Arifin, SepuluhNovember Institute of Technology, ID
- A. Murni, Universitas Indonesia, ID
- A. Srivihok, Kasetsart University, TH
- A. Purwarianti, Bandung Institute of Technology, ID
- B. H. Wijaya, Universitas Indonesia, ID
- B. Yuwono, Universitas Indonesia, ID
- C. Pang, CommonwealthScience and Industrial Research Organization, AU
- D. I. Sensuse, Universitas Indonesia, ID
- D. Jana, Institute of Electronics &Telecommunications Engineers, IN
- E. Suh, Ulsan National Institute of Science and Technology, KR
- H. Sayama, State University of New York, US
- H. Kim, Gwangju Institute of Science andTechnology, KR
- K. Sekiyama, Nagoya University, JP
- K. Abdullah, International Islamic University of Malaysia, MY
- M. Adriani, Universitas Indonesia, ID
- M. Ishak, Universiti Teknologi Malaysia, MY
- M. Nakajima, Nagoya University, JP

ISBN: 978-979-1421-11-9

- Marimin, Bogor Agricultural University, ID
- M. Pechenizkiy, EindhovenUniversity of Technology, NL
- N. Salim, Universiti Teknologi Malaysia, MY
- P. Hitzler, Wright State University, US
- Q. Zhang, CommonwealthScience and Industrial Research Organization, AU
- R. Setiono, National University of Singapore, SG
- S. Bressan, National University of Singapore, SG
- T. Usagawa, Kumamoto, University, JP
- W. Prasetya, Universiteit Utrecht, NL
- W. Chutimaskul, King Mongkut University Thailand, TH
- X. Liu, Brunel University, UK
- X. Li, University of Queensland, AU
- Y. S. Ho, Gwangju Institute of Science and Technology, KR

Table of Contents

Information Retrieval

 ISBN: 978-979-1421-11-9

Expected Answer Type Construction using Analogical Reasoning in a Question Answering Task

Hapnes Toba, Mirna Adriani, and Ruli Manurung
Information Retrieval Laboratory
Faculty of Computer Science, Universitas Indonesia
Email: hapnes.toba@ui.ac.id, mirna@cs.ui.ac.id, maruli@cs.ui.ac.id

***Abstract*—In a question answering system (QAS), question analysis component has an important task to determine the expected answer type (EAT) of a given question. Recently, there are a number of studies which investigates the influence of statistical relational framework to learn question-answer pairs in particular component of a QAS. In this study, we propose an approach that utilizes the strength of statistical learning of question-answer pairs as a means to develop EAT patterns. In a question analysis experiment setting by using factoid testing questions from QA@CLEF 2008, our result outperforms the accuracy of manually constructed patterns of Open Ephyra (OE)[1], with 84.17% against 81.67%.**

I. Introduction

A QUESTION ANSWERING SYSTEM (QAS) is a system that attempts to return a specific answer given a natural language question. Most of successful systems that took part in question answering evaluation forums, such as Text Retrieval Conference[2] (TREC) or Cross Language Evaluation Forum[3] (CLEF), were in general using a pipeline that is composed of: question analysis, query generation, information retrieval and answer validation components [1]. The task of a question analysis component is to determine the expected answer type (EAT) of a given question. For example the question: *"Where was the Volkswagen Polo Playa built?"* would expect a '*location'* as the answer. This EAT information will be convenient to validate the final answer.

Current research in question analysis could be grouped into two main approaches, i.e.: pattern-based and machine learning. In a pattern-based approach, the question analyzer tries to match sequences of word in a question with a set of predefined patterns, mostly in the form of regular expressions [2, 3]. In a machine learning approach, the performance of the question analyzer depends on the accuracy of the question classifiers. The classifiers are mostly learnt based on lexical or linguistic feature sets, by using machine learning algorithms, such as: SVM or Maximum Entropy [4].

The main advantage of the pattern-based approach is that it has high precision, but rather low recall, due to incomplete patterns. On the other hand, the machine learning approach has high recall, but rather low precision, if the learnt feature sets are not fit enough during the classification process. The main drawbacks of both approaches is that either they rely on human annotation, for example during the pattern constructions, or highly depends on complex natural language tools, such as syntactic parser, during the feature set extractions.

A number of more advanced techniques, which is beyond matching or learning the 'question part' of question-answer gold standards, have been recently investigated as well in [4, 5, and 6]. These studies investigated the influence of statistical relation learning of question-answer pairs. Lita and Carbonell [5] used the question-answer pairs to train the answer model, query content model, and extraction model to shape the strategies for answering new questions by using a clustering-based strategy. Huang et al. [4] showed that the quality of question analyzer improved if it knows how to classify the feature set of the 'question part' to its 'answer part'. Aktolga et al. [6] studied a translation model to notice which relation paths are more probable to be seen together in true question answer pairs, and the results showed that such relational paths could improve the passage reranking during retrieval. Another inspiring research has been conducted in [7] which showed that statistical relation learning of question-answer pairs is effective to retrieve answer in a community-based question answering.

In this study, we propose an approach that combine the strength of statistical learning approach of question-answer pairs and the expected high precision of pattern-based question analyzer, as a means to

[1] http://sourceforge.net/projects/openephyra/
[2] http://trec.nist.gov/
[3] http://celct.isti.cnr.it/ResPubliQA/

determine the EAT of a given question. We argue that if we can relate some positive influence of specific word occurrences in the 'question part', with some true responses of EAT in the 'answer part', then we can construct a pattern set of EAT that eventually outperforms the completeness of manually constructed patterns.

Consider a number of question-answer pair's examples, taken from QA@CLEF 2008, as follows:

Q#	Question	Answer
# 76	Where was Irish politician Willie O'Dea born?	Limerick
#105	Where is the car manufacturer Morgan Motor Company based?	Malvern
#108	Where was the Volkswagen Polo Playa built?	South Africa

Each of the questions above use the question word 'where', and the answer of the questions are pointed to a certain location, respectively: a city, a province, and a country.

Based on these examples, it is obvious useful to relate a certain question word that appears in a question to a certain type of answer. This kind of relation will provide us to some kind of 'information need' that lead us to a certain 'answer validation' strategy, further in a QAS pipeline, which is beyond this study. As an initial step of a complete pipeline, we conduct some experiments which objective is to develop an EAT pattern set based on the occurrences' of question words in the 'question part' and their related named-entity types that appear in the 'answer part'.

II. Pattern Construction Strategy

Most of QAS classify new questions according to static ontologies, such as: dictionaries or pattern sets [2]. These ontologies incorporate human knowledge about the EAT (such as: location, person, organization), the EAT granularity (location.country, location.state, location.city), and very often semantic information about the question type (location of an event). Considerable manual effort is invested into building and maintaining accurate ontologies even though EAT does arguably not always disjoint and hierarchical in nature (e.g. *"Where was Irish politician Willie O'Dea born?"*, expects an answer in the form of a location, either a country or a city).

In this research, we try to construct EAT pattern set automatically by learning the similarity of related pairs, i.e.: the question word and the named-entity of the answer. Our pattern construction strategy is based on the overlapping of certain question words and their related named-entity appearances' in a set of question-answer pairs. Related question-answer pairs are studied by using a recently developed statistical relation framework, called the Bayesian Analogical Reasoning (BAR) [8, 9]. The detail of this framework is given in subsequent section. Our complete approach can be seen in Fig. 1.

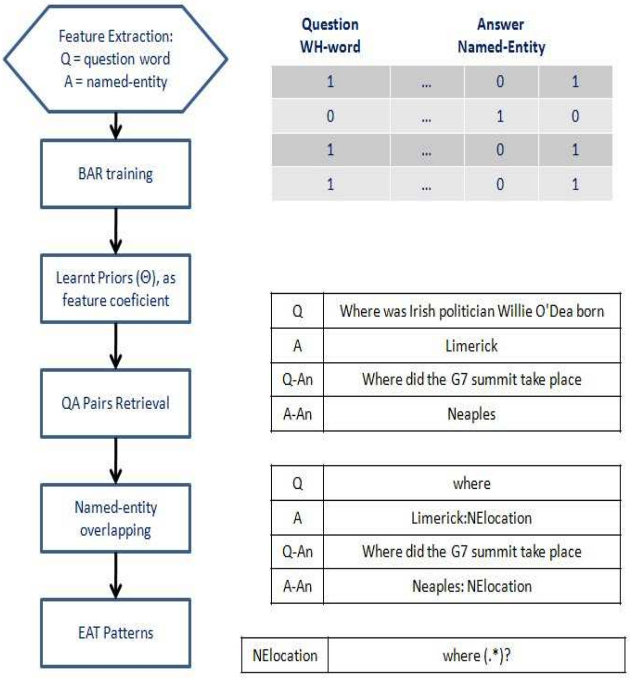

Fig. 1. The Pattern Construction Approach is mainly based on the overlapping of certain question words and their related named-entity appearances' in question-answer pairs.

The first step of the approach is to extract the feature set from question-answer pairs in the training and testing set, i.e. the question word and the named-entity. The features are binary, which indicate the presence (1) or absence (0) of a certain question word or named-entity. After the feature extraction are completed, the BAR framework learns the related features from the training set, and computes estimated priors for each feature. The next step is to perform question-answer pair's retrieval from the testing set, to identify which question-answer pairs of the testing set analogous to those in the training set. This retrieval step will produce *'pairs-of question-answer pairs'* that in some measure indicates how the pairs are related. Those pairs will be used as the foundation to identify the pattern of a particular named-entity.

Since we are interested in the overlapped features, the next step is to identify which *'pairs-of question-answer pairs'* have identical question word and named-entity appearances. From the list of the overlapping pairs, we grouped the pairs according to their named-entity appearances, and extract the question word for each named-entity group. For example in Fig. 1, we grouped the pairs according to the *'location'* named-entity, and extract the question word *'where'*, which indicates a close relation between the question word and the EAT. Our patterns are rather general, in the sense, that we only take the appearance of question word into account, and ignore all other words that appear in a question.

III. BAYESIAN ANALOGICAL REASONING

Bayesian Analogical Reasoning (BAR) is a method for ranking relations based on the Bayesian similarity criterion [10]. The underlying idea of BAR is to learn model parameters and priors from related objects (question and answer pairs in our case), and update the priors during the retrieval process of a query. The objective is to obtain marginal probability that relates the query with the objects that have been learnt.

Most methods of classification or similarity measures focus on the similarity between the *features* of objects in a candidate pairs and the features of objects in the query pairs. Further than that BAR focus on the similarity between *functions* that map pairs to links.

Consider there is a space of unseen mapping 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 $f(Q,A) = 1$, we need to measure 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 $g(Q^i,A^j)$ is related. The functions $f(., .)$ and $g(. , .)$ are unseen, and thus we need a prior set that will be used to integrate them over the function space.

Consider for each pair of object $(Q^i \in Q, A^j \in A)$, there is a feature vector X^{ij}, with:

$$X^{ij} = \left[\Phi_1\left(Q^i, A^j\right) \ldots \Phi_k\left(Q^i, A^j\right)\right]^T$$

which is 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. In our experiment, the objects of question and answer pairs have the same type of features set, and therefore dimensionality. 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 employ the 'cosine distance' to quantify a link representation between linked feature vectors.

Consider we have an unseen label L^{ij}, where $L^{ij} \in \{0,1\}$ as an expected indicator of the existence of a relation between related objects, then we will have model parameters in a vector of:

$$\Theta = \left[\Theta_1 \cdots \Theta_K\right]^T$$

This vector models the presence or absence of interaction between objects. These model parameters could be learnt by performing logistic regression estimation:

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

where *logistic(x)* is defined as:

$$\frac{1}{1 + e^{-x}} \quad (2)$$

The priors in the model parameter are learnt by performing the following equation:

$$P(\Theta) = N(\tilde{\Theta}, (c\tilde{T})^{-1}) \quad (3)$$

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 training set.

During comparison process, 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 have been learnt, and is compute as follows:

$$\begin{aligned} &score\left(Q^i, A^j\right) = \\ &\log P\left(L^{ij} = 1 \mid X^{ij}, S, L^S = 1\right) - \log P\left(L^{ij} = 1 \mid X^{ij}\right) \end{aligned} \quad (4)$$

The works in [8, 9] use the variational logistic regression approach [11] to compute the scoring function in equation 4.

IV. EXPERIMENTAL SETTING

In our experiments, we use the QA@CLEF 2003-2007 factoid English monolingual question-answer pairs gold standards as the training set, which consists of 1,888 pairs. The question type distribution can be seen in Table 1.

The testing set comes from QA@CLEF 2008 gold standard that consists of 120 factoid question-answer pairs. We exclude the 'OTHER', OBJECT' and 'MANNER' answer type. This is to maintain the equality of the manually developed patterns in Open Ephyra (OE) which is specifically designed to cover factoid-typed questions. The reason that we choose those sets due to the EAT equality which could be categorized into the following named-entity: person, organization, location, time, measure, and count. Each type consists of 20 question-answer pairs.

Table 1. Question type distribution of the training set

EAT	Frequency
PERSON	423
LOCATION	315
MEASURE	279
ORGANIZATION	273
OTHER	250
TIME	215
OBJECT	86
MANNER	26
COUNT	21
TOTAL	1888

We use two types of named-entity recognizer (NER) tools to extract the named-entity in the 'answer part' of each pair. The first one is based on statistical model, i.e.: the Stanford NER. This is used to extract the person, organization and location named-entity. The second NER tool is a dictionary-based NER as described in [2]. The dictionary consists of 46 named-entities in the type of 'number' and 'date', such as: 'length', 'size', 'month', 'year', etc.; and 95 in the type of nouns, such as: acronym, actor, animal, etc. It is possible that a word is categorized into more than one named-entity, for example the word *'Rome'* is recognized as 'capital' and 'location' entities.

To extract the 'question word' from a question, we use a simple identification approach. First we build a list of used question words from the training set, detected by Stanford[4] Part-of-Speech (POS) tagger. Then for each question, we identify a particular question word at the beginning of a question to form the feature set.

In order to reduce the dimensionality of the feature set, we decompose the original feature set matrix into 25-dimension singular value decomposition (SVD). SVD has been proved to be effective in many information retrieval experiments [12].

During the BAR training step, we captured two kinds of 'negative examples' random sampling proportions, i.e.: 1 and 10 'negative answers' for each 'positive' one. These settings are important to evaluate the proportion of linked to unlinked pairs in the set, and how it would influence the resulted pairs in the retrieval process.

To evaluate the BAR retrieval performance, we also experimented with 'cosine distance' (COS) measure, a common distance measure widely used in information retrieval. Comparison between BAR and COS performance indicates how effective the BAR algorithm. The COS generated patterns are constructed with the same strategy as described in Section II, but this time the question-answer pairs retrieval is measured by using COS distance.

To evaluate the coverage of the 'analogous pairs', we experimented with the 'top-1' and 'top-5' retrieval results. All constructed patterns, will finally be used in a 'question analysis' setting to evaluate the EAT accuracy.

V. Results and Discussions

In this section we present the results of our experiments during this study in following point of observations: the gold standards evaluation in terms of EAT accuracy, the retrieval results of analogous pairs, the identified named-entity, and the results of our pattern construction.

A. *Gold Standard Evaluation*

For the purpose of gold standards evaluation, we run some experiments in the 'question analysis' component of the OE framework by changing the manually developed pattern sets with our approach. Each question analysis result from the test question will be evaluated against the 2008 QA@CLEF EAT gold standard. During the evaluation judgment, we consider that if at least one named-entity occurs in the final EAT then it is assumed to be a true judgment (similar to the logical 'or-rule'). For instance the question (#53 in QA@CLEF 2008): "*In which university is Esperanto used?*" could be classified into: 'NEorganization, NElocation or NEcapital'. The EAT will be judged as true, because it is expected in the gold standard that the EAT is 'NEorganization'.

Table 2 presents the gold standard EAT evaluation. The columns represent the following experiment settings:

- *BAR-neg1-top5*: evaluation using BAR generated patterns with 1:1 positive:negative random sampling proportion at top-5 retrieval.
- *OE-full*: evaluation using all of the 154 detailed EAT patterns in Open Ephyra [2, 3].
- *BAR-neg1-top1*: evaluation using generated BAR patterns with 1:1 positive:negative random sampling proportion at top-1 retrieval.
- *OE-top level*: evaluation using only the 44 top levels EAT manual patterns in Open Ephyra [2, 3].
- *COS-top5*: evaluation using COS-measure generated patterns at top-5 retrieval.
- *BAR-neg10-top1*: evaluation using BAR generated patterns with 1:10 positive:negative random sampling proportion at top-1 retrieval.
- *COS-top1*: evaluation using COS-measure generated patterns at top-1 retrieval.

In the overall accuracy, the *'BAR-neg1-top5'* pattern set achieved the highest accuracy score, 84.17%. The *'OE-full'* pattern set achieved 81.67% accuracy, and the *'OE top-level'* pattern set achieved 72.50%, which

[4] http://nlp.stanford.edu/software/tagger.shtml

is to our surprise lower than the accuracy of the *'BAR-neg1-top1'* pattern set, 75.00%.

All pattern sets achieved high performance in the *'COUNT'* and *'PERSON'* answer type. Our constructed patterns performed worst (35%) in the *'TIME'* answer type, in contrast to the manual patterns of OE, which can achieved 80% (with full patterns) and 85% (top-level patterns) accuracy.

Table 2. Gold standard EAT evaluation.
The constructed pattern set from *'BAR-1:1proportion-top5'*outperforms the manually generated pattern sets of Open Ephyra.

	BAR-neg1-top5	OE-full	BAR-neg1-top1	OE-top level	COS-top5	BAR-neg10-top1	COS-top1
Over. Acc.	84.17%	81.67%	75.00%	72.50%	68.33%	56.67%	44.17%
Count	100.00%	100.00%	100.00%	100.00%	90.00%	100.00%	90.00%
Loc.	100.00%	85.00%	100.00%	60.00%	100.00%	75.00%	10.00%
Meas.	75.00%	65.00%	75.00%	35.00%	55.00%	75.00%	50.00%
Org.	95.00%	70.00%	90.00%	65.00%	30.00%	20.00%	0.00%
Pers.	100.00%	90.00%	85.00%	90.00%	95.00%	70.00%	75.00%
Time	35.00%	80.00%	0.00%	85.00%	40.00%	0.00%	40.00%

Table 3. Examples of Retrieved Analogous Pairs

Question (#166 QA@CLEF 2008, EAT = ORGANIZATION)	
What company does Neil Sweig work for?	
Landenburg, Thalmann & Co.	
Retrieval Results	
BAR-neg1-top5	**COS-top5**
What party did Hitler belong to?	What is the abbreviation for mad cow disease?
Nazi party	BSE
In what year was the Statue of Liberty built	What does the abbreviation EEC mean?
1886	European Economic Community
Of what political party is Ian Paisley the leader?	What does the abbreviation DNA stand for?
Democratic Unionist Party	Deoxyribose Nucleic Acid
Where is the Leaning Tower?	What is the capital of the Republic of South Africa?
Pisa	Pretoria
Where is Kishinev located?	What is the name of the Swiss national airline?
Moldavia	Swissair

Table 4. Examples of identified named-entity

Question (#166 QA@CLEF 2008, EAT = ORGANIZATION)	
what	
Landenburg , Thalmann & Co:NEorganization	
Retrieval Results	
BAR-neg1-top5	**COS-top5**
what	what
Nazi party:NEorganization	BSE:NEacronym
what	what
1886:NEdate 1886:NEnumber 1886:NEyear 1886:NEnumber	European Economic Community:NEorganization
what	what
Democratic Unionist Party:NEorganization	Acid:NEdrug Acid:NEnarcotic
where	what
Pisa:NElocation	Pretoria:NEcapital Pretoria:NElocation
where	what
Moldavia:NElocation	Swissair:NEorganization

B. Retrieval of Analogous Pairs

The analogous pairs from the testing set are computed by using BAR and COS-measure during question-answer pair's retrieval. Examples of the retrieved analogous pairs can be seen in Table 3.

From the examples in Table 3, we can observe that BAR framework does not always retrieve the same feature appearances. For instance, the question word *'what'* and *'where'* in Table 3.

As a comparison, we could also see in Table 3 that COS distance has retrieved more pairs that have *'what'* as the question word features. On the other hand, the COS distance does not guarantee that the retrieved pairs have a close features (named-entity) relation.

This result indicates that BAR returns more analogous pairs. The similarity between the pairs is defined in terms of the similarity of the function (i.e. the logistic regression function) which mapped the pairs as being linked.

Table 5. Identified named-entity in various experiment settings (The X's means the appearance of the named-entity)

Named-entity	BAR-neg10-top1	BAR-neg1-top1	BAR-neg1-top5	COS-top1	COS-top5
acronym	X	X	X	X	X
animal			X		
book			X		
capital			X		X
causeofdeath			X	X	X
color			X		
country	X	X	X	X	X
date			X	X	X
day	X	X	X	X	X
ethnicgroup					X
film	X	X			
firstname			X	X	X
hour			X		
length	X	X	X	X	X
location	X	X	X	X	X
month					X
musicalinstrument				X	X
musictype					X
nationality		X			
newspaper			X		X
number	X	X	X	X	X
organization	X	X	X		X
person	X	X	X	X	X
profession	X	X	X		
province		X	X		X
range	X	X	X	X	X
scientist			X	X	X
show			X		
size	X	X	X	X	X
team					X
uspresident					X
socialtitle	X	X	X		
sport			X		
year			X	X	X
# of. identified NE	13	15	26	15	24

C. Identified Named-Entity and EAT Patterns

Further inspection of the retrieval results can be observed in terms of the identified named-entity among the pairs. For instance in Table 4, we examine the identified named-entities for the example in Table 3. As we observe in Table 4, a particular question word does not always relate to the same named-entity in the answer. But it seems in this example, that there is a close relation between the question words 'what', and the 'organization' named-entity. In this way, we can generalized that the regular expression *'what (*.)?'* is related to 'NEorganization'.

Table 5 gives the total number of identified named-entity in various experiment settings as described in Section IV. It can be seen in Table 5 that *'BAR-neg1-top-5'*, has the highest total number of identified named-entities, i.e. 26, and *'BAR-neg10-top1'* has the lowest number of identified named-entities, i.e. 13. It seems that by using BAR framework, a less number of negative samples lead to fewer noise data, and thus can give a better retrieval performance. It can also be seen in Table 5, that a higher number of retrieval results (the top-5), gives a better coverage of named-entities identification.

If we look further into the identified named-entity, than we could see that the 'person' and 'location' named-entities - that recognized by the model based NER-tool (Stanford NER) - are all presence in each experiment setting. Except for the 'organization' named-entity, this could not be identified in the *'COS-top1'* setting. This result suggests that BAR has better 'awareness for related named-entity with respect to the appearance of a particular question word, than a common distance function, such as COS measure.

Table 6. Example of constructed patterns for the 'person' NE

BAR-neg1-top5	Open Ephyra
name (.*)?	who(m|se)?
what (.*)?	(what|which|name|give|tell) (.*)?(name at birth|nickname|(full|original|real) name)
which (.*)?	(what|which|name|give|tell) (.*)?(abolitionist|actor|actress|adventurer|apostle|architect|artist|assassin|astronaut|aunt|author|(arch)?bishop|boxer|boy|brother|builder|candidate|captain|CEO|chairman|champion|chancellor|character|chief executive|chief justice|child|choreographer|coach|comedian|commander|c
(who|whose) (.*)?	

An example of constructed patterns for the 'person' named-entity by utilizing the BAR framework, and its comparison to the manual constructed patterns in OE can be seen in Table 6. As we can see in Table 6, the constructed patterns are rather general, in contrast to the detailed manual patterns in OE.

D. Evaluation in a QAS Pipeline

To evaluate the ability of our constructed patterns, we performed two types of answer validation experiments in OE framework. The first one is by using the gold standard snippets as the source of answer extraction process. The second one is by using Indri passage retrieval as the source of answer extraction process.

We used two metric for this purpose. The first one is precision, which shows whether a question could be correctly answered. The second one is mean reciprocal rank (MRR), which shows in which rank a correct answer is retrieved. We compare our results with the original OE manually constructed pattern set. The result can be seen in Table 7 and 8.

Table 7. Answer validation performance from gold standard snippets

Pattern-set	Precision	MRR
BAR-neg1-top5	0.525	0.372
Open Ephyra Full	0.475	0.389

In Table 7, we can see that our patterns set outperforms the manually constructed patterns of OE in terms of precision, but has a slightly lower MRR.

Table 8. Answer validation performance from Indri passage retrieval

Pattern-set	Precision	MRR
BAR-neg1-top5	0.475	0.048
Open Ephyra Full	0.475	0.077

As we can see in Table 8, by performing an information retrieval strategy, our pattern set has a comparable precision against OE, but has a lower MRR. In our opinion, the lower MRR is caused by the high recall of our patterns during the question analysis phase which will give a high coverage of pattern, but on the other hand can reduce the rank in which the true answer is retrieved.

E. Discussions

There are a number of interesting findings during this study that could be further investigated in the field of question answering, such as:

1. Statistical learning framework, such as BAR could be preferable above the commonly used distance function, such as COS, to measure the relation between pairs of question and answer feature space. However, distance function still will be important to measure the 'closeness' of feature set, for instance for evaluation purposes.
2. By using simple feature set, such as question word and named-entity in this research, the task of determining an exact EAT is still remain a hard decision problem. Neither, the manually developed patterns nor the automatic approach can be directly used to suggest the best EAT during question analysis. A more linguistic-based feature set, such as: sequences of POS-tags or statistics correlated information of syntactic parsed-tree, might produce better analogous pairs.

VI. CONCLUSION AND FUTURE WORK

A number of conclusions from this study could be summarized as follows:

1. We have showed that our constructed EAT pattern set – by utilizing statistical relational learning – outperforms the manually constructed patterns (c.f. section V. A).
2. We have demonstrated that simple semantic features, i.e. 'question word' and 'named-entity', would be fit enough to relate positive question-answer pairs (c.f. section V. B).
3. We have demonstrated that relational pairs of question and answering could be utilized to construct EAT patterns (c.f. section V. C).
4. Our approach produces a high recall and high precision pattern set. A high recall means that our patterns give a wide area of EAT coverage (c.f. section V. A). A high precision means that our patterns give a high accuracy of predicted EAT which lead to a high precision in answer validation evaluation (c.f. section V. D). The trade-off of our approach in the answer validation evaluation is that we achieved a lower MRR in comparison to the manually constructed patterns in OE.

Which kind of pattern is preferable: the general patterns, such as in our approach, or detailed patterns, such as in OE, is still open to be answered. It seems that each has its own advantages. A general pattern set, could be preferable if structure of the question is not much varied, such as in the *'COUNT'* answer type (with the typical '*how (many|much)'-pattern*) or *'PERSON'* answer type (with the typical *'who'*-pattern). If the question structure of an EAT is rich, such as in the '*DATE (temporal)*' answer type, for instance: *'when is …?'*, *'in what year …?'*, *'during which period …?'*, manual human observation might be preferable.

As our future work, we plan to investigate the influence of statistical learning algorithm, such as BAR, to find related 'answer strategy' among question answer pairs, with much richer linguistic features, for instance: the sequences of POS-tags, and statistics correlated information of syntactic parsed-tree.

REFERENCES

[1] M.W. Bilotti, "Linguistic and Semantic Passage Retrieval Strategies for Question Answering," Ph.D. dissertation, Language Technologies Institute, School of Com. Sci., Carnegie Mellon Univ., 2009.

[2] N. Schlaefer, P. Gieselmann, T. Schaaf, and A. Waibel, "A Pattern Learning Approach to Question Answering within the Ephyra Framework," *LNAI* 4188, pp. 687–694, Springer-Verlag, 2006.

[3] N. Schlaefer, J. Ko, J. Betteridge, G. Sautter, M. Pathak, and E. Nyberg, "Semantic Extensions of the Ephyra QA System for TREC 2007," In *Proc. Text Retrieval Conference*, 2007.

[4] Z. Huang, M. Thint, and A. Celikyilmaz, "Investigation of Question Classifier in Question Answering," *Proc. Conf. on Empirical Methods in Natural Language Processing*, Singapore, 2009, pp. 543–550.

[5] L.V. Lita and J. Carbonell. (2004, January). "Instance-based Question Answering: A Data-Driven Approach". *Institute for*

Software Research. Paper 432. Available: http://repository.cmu.edu/isr/432

[6] E. Aktolga, J. Allan, and D. A. Smith, "Passage Reranking for Question Answering Using Syntactic Structures and Answer Types," *LNCS* 6611, pp. 617–628, Springer-Verlag, 2011.

[7] X-J. Wang, X. Tu, D. Feng, and L. Zhang, "Ranking Community Answers by Modeling Question-Answer Relationship via Analogical Reasoning," *Proc. of SIGIR*, 2009.

[8] R. Silva, K. Heller, and Z. Ghahramani, "Analogical Reasoning with Relational Bayesian-sets", *Proc. AISTATS*, 2007.

[9] R. Silva, K. Heller, and Z. Ghahramani, and E. Airoldi, "Ranking Relations Using Analogies in Biological and Information Networks," *The Annals of Applied Statistics* 4 (2), 2010.

[10] Z. Ghahramani, and K. Heller, "Bayesian Sets," *Advances in Neural Information Processing Systems* 18, pp. 435-442, 2005.

[11] T. Jaakkola, and M. Jordan, "Bayesian Parameter Estimation via Variational Bounds," *Statistics and Computing* 10, pp. 25-37, 2000.

[12] C.D. Manning, P. Raghavan, H. Schutze, *Introduction to Information Retrieval*. New York, USA: Cambridge University Press, 2008, pp. 369-384.