Jurnal Teknologi

Full paper

A Survey Of Challenges And Resolutions Of Mining Question-Answer Pairs From Internet Forum

Adekunle Isiaka Obasa*, Naomie Salim, Yazan A. Al-Khassawneh

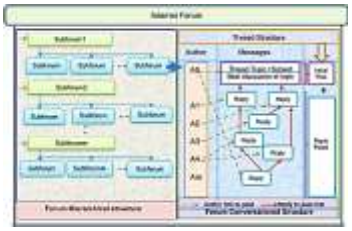
SCRG Lab, Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor Malaysia

*Corresponding author: iaobasa@yahoo.com

Article history

Received : 31 July 2014
Received in revised form :
23 November 2014
Accepted : 1 December 2014

Graphical abstract



Abstract

Internet forum is a web community that brings people in different geographical locations together. Members of the forum exchange ideas and expertise and as a result generate huge amount of content on different topics on daily basis. A good percentage of human generated content of Internet forums have been found to be question-answer (QA) pairs. These QA pairs are useful for automating question answering system. Mining these QA pairs has become a hot issue in the research community. Effective mining of the QA pairs is being hindered by a number of factors. Lexical chasm that renders some Information Retrieval (IR) techniques less effective, casual language that creates noisy data; multiple authors that bring about unfocused topics are some of the issues that need to be addressed. In this paper, an extensive overview of the strategies and findings relevant to these three challenges are addressed. The survey revealed that researchers are adopting non-lexical features as against lexical to resolve the issue of data sparseness. Noise level is mostly controlled using conventional dictionary rather than using domain-specific dictionary.

Keywords: Internet forum; question-answer pairs; lexical chasm; casual language

Abstrak

Internet forum is a web community that brings people in different geographical locations together. Members of the forum exchange ideas and expertise and as a result generate huge amount of content on different topics on daily basis. A good percentage of human generated content of Internet forums have been found to be question-answer (QA) pairs. These QA pairs are useful for automating question answering system. Mining these QA pairs has become a hot issue in the research community. Effective mining of the QA pairs is being hindered by a number of factors. Lexical chasm that renders some Information Retrieval (IR) techniques less effective, casual language that creates noisy data; multiple authors that bring about unfocused topics are some of the issues that need to be addressed. In this paper, an extensive overview of the strategies and findings relevant to these three challenges are addressed. The survey revealed that researchers are adopting non-lexical features as against lexical to resolve the issue of data sparseness. Noise level is mostly controlled using conventional dictionary rather than using domain-specific dictionary.

Kata kunci: Internet forum; question-answer pairs; lexical chasm; casual language

■1.0 INTRODUCTION

Internet forum is a web application that is becoming more and more popular. Its popularity may be attributed to the fact that it provides customer support for business enterprises that use it. Both technical and less technical issues are discussed in forums. Forum brings together experts from all walks of life. Members of a forum can make their contribution at the comfort of their homes without geographical and time zone barriers. Forums have both hierarchical and conversational structures. The hierarchical structure has to do with sub-forums emanating from the main forum, depending on the broadness of the category. For example, a computer technology forum can have hardware and software as sub-forums. The hardware sub-forum may also have mother-boards, input devices and output devices as sub-forums. The conversational structure takes place within a sub-forum. A sub-forum is made up of threads. A thread is the minimal topical unit that addresses a specific topic. A thread is usually initiated by an author's post (usually called initial post), which constitute the topic of discussion. Members who are interested in the topic send reply posts. Figure 1 shows the structure of an Internet forum.

Interaction within the forum community is naturally through question and answer scenario. It was empirically confirmed by [1] that 90% of 40 forums investigated contain question-answer the various domains. This is because different business enterprise, which sells on the Internet need to provide customer call-centres to address customers' queries. Mined question-answer pairs can be archived to serve this purpose. This will not only reduce the cost of operating call centres but also enhance response time. Benefits of question-answer pairs are x-rayed in [1-4]. Some of the challenges hindering effective Mining of Question-answer pairs are: Lexical chasm, Informal tone and Unfocused Topic mining.

In this paper, we carry out an extensive overview of these three challenges that are limiting the potentiality of mining knowledge from Internet forum. Different approaches that researchers consider in overcoming them are explored with actions that have been taken so far to resolve them. We also proffer suggestions that can further assist in addressing the problems. Mining of human generated contents of forums is non-trivial due to its nature. The huge amount of responses and the variations of response context lead to the problems of efficient knowledge accumulation and retrieval [5]. Table 1 shows different forums that are serving different purposes with volume of human generated content they contain.

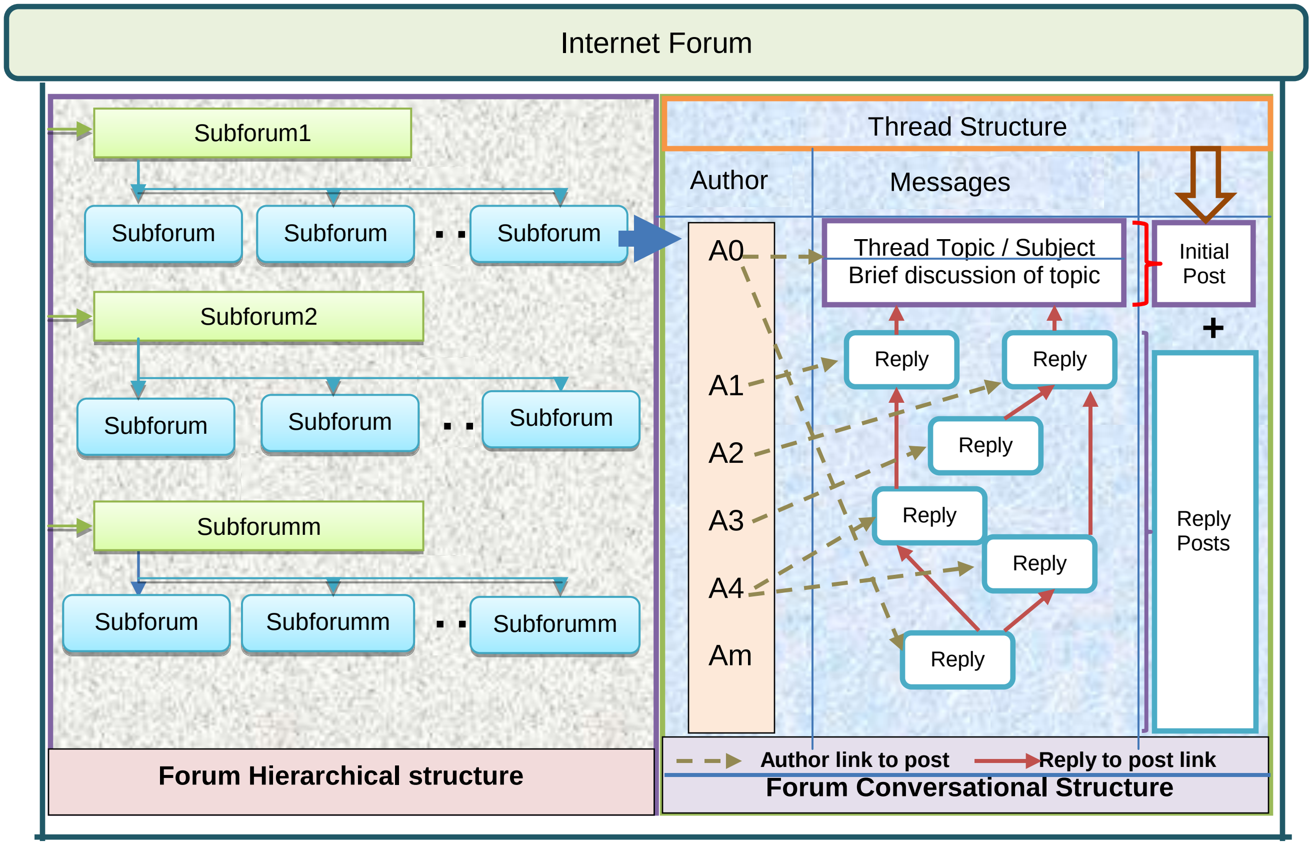


Figure 1 Hierarchical and conversational structure of internet forum (modified from [6])

The research activities in this domain is focusing on how to use the human generated contents reported in column 3 under the heading "Statistics" for the benefits of mankind. A good number of research activities are going on in the forum domain. Some of these research activities include retrieving relevant forum threads, clustering forum threads, finding similar threads, evaluating threads quality and mining question-answer pairs.

Another type of discussion board that is becoming popular is the Community Question Answering (CQA). Some good examples of CQA are Yahoo! Answers, Stackoverflow, and Baidu a popular Chinese CQA. The CQA renders purely question answering services which are similar to that of the Internet forum. The CQA's are highly restrictive. A number of CQA's welcome purely objective contributions that do not call for too much debate from members. Members that wish to seek for subjective opinion may have to turn to Internet forum.

A number of commercial question answering services like telephone answering system, chat bot, speaktoit, etc. are systems that benefit directly from automatic mining of QA pairs from CQA and Internet forums. These systems are products of Artificial Intelligence (AI). It should also be noted that the AI researchers are using the mined QA pairs to conduct Machine Learning (ML) training and testing while producing the systems. There are many other uses of QA pairs that can be found in the literature. It is on this premise that we decided to survey some of the issues that hinder effective mining of these QA pairs.

Table 1 Examples of internet forum with volume of human generated contents

Forum	Genre	Statistics
Ubuntu Forums	Official Ubuntu Forums	1.8 million threads, 1.7 million users, started in 2004
Lowyat.net	Malaysia’s Largest Online Community	0.3 million threads, 0.5 million users, started in 2003
Body Building.com	Body building related content	4.8 million threads, 5.2 million users, started in 2000
Breast Cancer	Breast Cancer dedicated forums	0.95 million threads, 0.11 million users, started prior to 2008
Christian Forums	Faith and beliefs related content	7.69 million threads, 0.31 million users, started on 2003

■2.0 LEXICAL CHASM IN MINING QA PAIRS

Lexical chasm, also known as lexical gap, is one of the issues hindering effective mining of knowledge from forums [7-9]. A lexical Chasm occurs whenever a language expresses a concept with a lexical unit whereas the other language expresses the same concept with a free combination of words [10]. Lexical gap problem can be attributed to different ways of writing that calls for the use of polysemy (same word with different meanings, such as “book” as in the following examples: “The book is on the table” and “I will book my flight tomorrow”), synonym (different words with the same or similar meanings, such as “agree” and “approve” as in “I agree with his going to London” and “I approve his going to London”) and the use of paraphrasing. The problem is more severe when retrieving shorter documents such as sentence, question and answer retrieval in QA archives [11].

Human generated posts of web forum usually include a very short content, which always have much fewer sentences than that of web pages. The implication of this is that some useful models for similarity computing such as Cosine similarity, Kullback Leibler (KL) divergence and even Query Language that have yielded useful results in information retrieval become less powerful when faced with forum contents. The short contents cannot also provide enough semantic or logical information for deep language processing [9].

In forum’s question-answer detection system, it will be difficult to expect a great match between the lexical contents of question and its corresponding answer. In fact, there is often very little similarity between the tokens in a question and the one appearing in its answer. For example, a good answer to the question *“Which hotel in Skudai is pet friendly?”* might be *“No Man’s Land at Sri Pulai”*. The two statements have no tokens in common. Even at times the answers provided may be just a single word. For example the answer to the question *“Where can I get a good clipper to buy?”* can just be given as *“Jusco”*. The relevance models that are stated above use common tokens to establish similarity. Hence, they failed to yield good results in forums.

The established vocabularies for questions and answers are the same, but the probability distributions over those vocabularies are different for questions and their answers. The vocabulary mismatch and non-linkage between query and response vocabularies is often referred to as a *lexical chasm.* This problem between queries and documents or questions and answers has been identified as a common problem to both information retrieval and question answering [11]. It is even more pronounced in question answering because of the prevailing data sparseness in the domain. Bridging the lexical chasm between questions and their answers will require techniques that will move from lexical level toward semantic level.

The lexical chasm problem has made it difficult to establish a good similarity between questions and answers posts. As a result of this, researchers have to find alternative approaches to relevance modelling in getting answers in forum threads. Some of these approaches and some relevant suggestions are given in the next section.

2.1 Lexical Chasm Resolution Approaches

Several techniques have been used by researchers to resolve problem of lexical chasm. In this section, four of these resolution measures, namely, query expansion, word sense disambiguation, machine translation and non-lexical based features shall be reviewed.

2.1.1 Query Expansion

In mining QA pairs from forum, the query question is usually composed from relevant tokens with some of the context dropped. This scenario is a contributory factor to the problem of lexical chasm. For this reason, there has been much interest in query expansion techniques [12-15]. The basic query expansion technique involves adding words to the query; the words may likely be synonyms or somehow related words in the original query. The techniques used in query expansion can be classified as i) getting synonyms of words by searching for them ii) determining various morphological forms of words by stemming words in the search query iii) correcting spelling errors automatically by searching for the corrected form iv) re-weighting the terms in the original query [16].

A more focused expansion can be generated using question-answer pairs’ training set. All it requires is to learn a mapping between words in the query (that is, the question) and their corresponding responses (such as smoking → cigarette, why →

because, URL → website and MS → Microsoft). These words are added to the query being used for the mapping to augment the original query to produce a representation that better reflects the underlying information need.

2.1.2 Word Sense Disambiguation (WSD)

This is a method that identifies the meaning of words in a computational manner within the context of their usage [17]. It has been applied successfully in machine translation, information retrieval, information extraction, etc. It is a promising approach for bridging gaps between question and answer pairs of web forum. It is mostly being implemented using WordNet in the domain. WSD approaches are classified based on the sense primary source. Dictionary-based or knowledge-based WSD uses dictionaries, thesauri, and lexical knowledge bases without using any corpus evidence. Other approaches are unsupervised, supervised or semi-supervised. These approaches use unannotated corpora, annotated corpora or seed data in a bootstrapping process for training purposes.

2.1.3 Machine Translation

The basic language modelling structure for retrieval which establishes similarity between a query Q and a document D may be modelled as the probability of the document language model MD built from D generating Q:

$$sim(Q,D) \approx P\,(Q|M_D) \qquad (1)$$

Query words are often considered to occur independently in a particular document language model, as such, the query-likelihood $P\,(Q|M_D)$is calculated as:

$$P\,(Q|M_D) = \prod_{q\in Q} P\,(q|M_D) \qquad (2)$$

where q is a query word. The probability $P\,(q|M_D)$ is usually calculated using maximum likelihood estimation [15].

It should be noted that this basic language model structure does not address lexical gaps issue between queries and question. Information retrieval was viewed by [18] as statistical document-query translation and as such added translation models to map query words to document words. The established translation-based retrieval model obtained by modelling $P\,(q|M_D)$ in equation (2) above is:

$$P\,(q|M_D) = \sum_{w\in D} T(q|w)P\,(w|M_D) \qquad (3)$$

where *w* represents document word. The translation probability *T(q|w)* fundamentally represents the level of association between query word q and document word w captured using different machine translation setting [15]. The use of translation models judging from traditional information retrieval perspective, produce an implicit query expansion effect, since query words that are not found in a document are mapped to associated words in the document. A positive impact could only be made by this translation-based retrieval models if only the pre-constructed translation models have consistent translation probability distributions.

2.1.4 Non-Lexical Features

A much more prevalent approach of tackling lexical gaps in web forum question answering is to avoid the use of lexical data. The non-lexical features are at times referred to as structural features. Forum meta data such as authorship, answer length, normalized position of post, etc. are used in determining questions and answers. In [2, 19] total number of posts and authorship were used to mine questions with a reasonable performance. A host of these features with detailed descriptions for mining questions and answers are contained in [6, 20]. A major problem with non-lexical features is their availability. Some non-lexical features used by some forums may not be found in others. The degree of availability of some non-lexical features across forums can be found in [20]. It is worth noting that combination of both the lexical and non-lexical is desirable for effective mining of question-answer pairs from forum. The lexical features measure the degree of relevance between question and answer while non-lexical can be used to estimate the quality of answers [21].

■3.0 CASUAL LANGUAGE

Forum content generation is at times done with some laxity. Members initializing or replying a post tends to use an informal tone / language which is more closed to his/her oral habit. The informal tone is often considered in literature as unstructured casual language [22]. The useful information is concealed inside majority of trivial, heterogeneous, and sometimes irrelevant, text data of different quality. This attitude usually make forum content to be highly noisy [1, 9, 19, 23, 24].

The noise content of forum can be said to come from two sources. These sources appear to be in line with sources identified by [25] for text generally: 1) noise can occur during the conversion process, when a textual representation of information is produced from some other form. For example, web pages, printed/handwritten documents, camera-captured images, spontaneous speech are all intended for human use. Their conversion into some other forms may results in noisy text. 2) Noise can also be introduced when text is generated in digital form. Most especially in informal settings such as SMS (Short Messaging Service or Texting), online chat, emails, web pages and message boards, the text produced is inherently noisy. This type of text contains spelling errors, special characters, grammar mistakes, non-standard word forms, usage of multilingual words and so on [25]. In forum, text normalization activities have been concentrated on the second noise source. Categorization of forum noise as contained in [6] is shown in Table 2.

Table 2 Classes of Noise with examples

Class of Noise	Example

Orthographic	Msg= Message, befour =before
	Positon=position

Phonetic	Rite=right, gooood= good
	Smokin= smoking

Contextual	In other to = in order to I can here you= I can hear you
Acronym	Asap = as soon as possible Lol = laughs out loudly

3.1 Casual Language Resolution Approaches

A number of methods from different research areas have emerged for identifying and correcting words in text. A good work by [26] described in details various methods for correcting spelling mistakes. A common measure for rectifying spelling errors is edit distance or Levenshtein distance. For any two character strings t_1 and t_2, the edit distance between them is considered as the minimum number of edit operations needed to transform t_1 into t_2. The expected edit operations are: (i) insertion of a character into a string; (ii) deletion of a character from a string and (iii) replacement of a character of a string by another character. For example, the edit distance between dog and rat is 3. The edit distance model is at times being augmented by a Language Model (LM) from the corpus of Web queries. This is based on the notion of distributional similarity [27] between two terms, which is high between a frequently occurring misspelling and its correction, and low between two irrelevant terms only with similar spellings.

Open source dictionaries such as Aspell or Hunspell can also be used to fix some of the spelling mistakes found in forum corpora. An empirical result of [22] confirms the effectiveness of these open source dictionaries in correcting words in text. However, dictionaries can only correct spelling mistakes with some being able to fix phonetic errors. Noise is often modelled depending on the application. Four different noise channels, namely, Grapheme Channel, Phoneme Channel, Context Channel and Acronym Channel are proposed by [28] to fix the four noise classes x-rayed in Table 2. The noise channels are described in the following four paragraphs.

The grapheme channel is responsible for the spelling distortion. A way of modelling this channel is to consider it as being directly proportional to the similarity between a corrupted token and its normalization. The more similar a normalization candidate is to the corrupted token, the more likely it is the correct substitution for it.

The phoneme channel is responsible for distortion in pronunciations. It is similar to the grapheme channel; the probability of a correct string being transformed into an incorrect string is proportional to the similarity between the two terms, area of difference being that the similarity in this case is measured on the phonetic representations instead of orthographic forms. A major step in phoneme is Letter-to-Phoneme (L2P) conversion, which estimates the pronunciation of a term, represented as a sequence of letters. A lot of research is going on in this area of letter-to-phoneme conversion. Some notable ones are the work of [29-31]. After the L2P conversion, the similarity measure between two phoneme sequences becomes the same as the similarity measure implemented in the grapheme channel, the only difference is that a uniform weight Levenshtein distance is considered instead of weighted Levenshtein distance.

Context channel - a context-based correction procedure would not only handle the problem of real-word errors, i.e., errors that result in another valid word, like *form* instead of *from,* but it would also be good in correcting those non-word errors that have more than one possible correction. A good example of such is the string *ehre.* Without context there is little reasoning one could make, some possible options to considered as the intended correction among others are *here, ere, ether, where, there.* Developing context-based correction procedures has become a notable challenge for automatic word recognition and error correction in text [26]. Correct normalization using context is often determined by considering the n-gram probability. The n-gram language model is normally trained on a large Web corpus to return probability score for a query word or phrase.

Acronym Channel - the three channel models considered so far deal with word-to-word normalization. There exist a number of acronyms such as “fyi” (for your information), “asap” (as soon as possible) and “lol” (laugh out loudly) that are commonly used and involve word-to-phrase mappings. The acronym channel can then be considered as a model of one-to-many mapping.

■4.0 TOPIC DRIFT

Threads in Internet forum are composed by many authors. As a result, they are less coherent and more susceptible to sudden jumps in topics. The existence of several topics in a thread is something very common in popular discussions. Even if a unique topic is discussed in a thread, different features and aspects of it may be considered in the discussion. There is a need to uncover the content structure of threads so as to establish post-to-post discourse structure. Specifically, it will be better to establish which earlier post(s) a given post responds to. It has rightly been pointed out by [27, 32] that post-to-post discourse structure will enhance information retrieval. A good illustration of this problem is contained in [33]. Topic drift is mostly found in threads that contain many posts, say 6 and above.

4.1 Topic Drift Resolution Strategies

The usage of term frequency (TF- IDF) and text similarity methods is a very common approach for extracting topic of discussion [34-37]. Quotation within post is often being used to establish context coherence. It indicates the relevance between a reply and the root message if root message is quoted. Drift resolution is implemented in [38] using two quotation features: a reply quoting root message and a reply quoting other replies. A reply quoting root message indicates that the reply is relevant to the message. In contrast, a reply quoting other replies may not be relevant to the root message hence it can be considered as topic drift. A blended quoting technique that utilizes some special features offered from the structure of web forums is proposed by [39] to cluster the posts of a discussion with the same topic. In their work, an algorithm that uses temporal information such as time and date of posts, the post authors etc. is implemented to create posting chains that uses topic similarity algorithm augmented with the utilization of the quoting system.

An exciting method to track topic drifting in a discussion is proposed by [40]. They use lexical similarity and thematic distance to identify topic boundaries in a discussion and fragmented it into topic related clusters. An algorithm proposed by [41] that isolates parts of a discussion in order to extracts the topics using just these parts and not the entire thread is good approach to tackle problem of topic drift in forums. Utilization of term weights and domain technical words will probably enhance performance.

Some other popular approaches are the use of dialogue act tagging (DAT) and discourse disentanglement. Dialogue act tagging helps in capturing the purpose of a given utterance in relation to an encompassing discourse. Discourse disentanglement is being implemented to automatically identify coherent sub-discourses in a single thread. The two concepts are implemented in [33] to establish post-to-post relationship. Three categories of features, namely, structural features, post context features and semantic features were considered in the work. The use of topic modelling such as Latent Dirichlet Allocation may be necessary for long threads that contain tens of posts.

■5.0 SUMMARY AND CONCLUSION

In this paper, a review of four challenges and resolutions militating against effective mining of questions and their answers from web forums is presented. We specifically focused the review on: i) Lexical chasm problem that renders good similarity computing algorithm like cosine to be less effective with forum data. ii) Casual language that makes forum data to be highly noise. iii) Topic drift that makes discussion to be less coherent. We explored relevant materials in the fields of information retrieval, information extraction, data mine and text mining to address the issues. The survey provides description of the problems, cites and explores useful publications to the reader for further examination, provides an overview of resolution strategies and findings relevant to the challenges. We also proffer suggestions that can further assist in addressing the problems.

Acknowledgement

This work was supported by the ministry of higher education (MOHE) and research management centre (RMC) at Universiti Teknologi Malaysia (UTM) under research university grant category (Vot:Q.J130000.2528.07H89).

References

[1] Cong G, Wang L, Lin C-Y, Song Y-I, Sun Y, 2008. Finding question-answer pairs from online forums. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval: ACM. 467–74*.

[2] Hong L, Davison B.D. 2009. A classification-based approach to question answering in discussion boards. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval: ACM*. 171–8.

[3] Raghavan P, Catherine R, Ikbal S, Kambhatla N, Majumdar D. 2010. Extracting problem and resolution information from online discussion forums. *Management of Data*. 77.

[4] Sumit B, Prakhar B, Prasenjit M. 2012. Classifying User Messages For Managing Web Forum Data. *Fifteenth International Workshop on the Web and Databases (WebDB 2012), Scottsdale, AZ, USA*.

[5] Hu W-C, Yu D-F, Jiau HC. 2010. A FAQ Finding Process in Open Source Project Forums. *Fifth International Conference on Software Engineering Advances*. 259–64.

[6] Obasa AI, Salim N. 2014. Mining FAQ From Forum Threads: Theoretical Framework. *Journal of Theoretical & Applied Information Technology*. 63.

[7] Wang B-X, Liu B-Q, Sun C-J, Wang X-L, Sun L. 2013. Thread Segmentation Based Answer Detection in Chinese Online Forums. *Acta Automatica Sinica*. 39:11–20.

[8] Brill E, Dumais S, Banko M. 2002. An analysis of the AskMSR question-answering system. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10: Association for Computational Linguistics*. 257–64.

[9] Wang B, Liu B, Sun C, Wang X, Sun L. 2009. Extracting Chinese question-answer pairs from online forums. *Systems, Man and Cybernetics, 2009 SMC 2009 IEEE International Conference on: IEEE*. 1159–64.

[10] Bentivogli L, Pianta E. Looking for lexical gaps. 2000. *Proceedings of the ninth EURALEX International Congress: Citeseer*. 8–12.

[11] Bernhard D, Gurevych I. 2009. Combining lexical semantic resources with question & answer archives for translation-based answer finding. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Association for Computational Linguistics* 2: 728-36.

[12] Gong Z, Muyeba M, Guo J. 2010. Business information query expansion through semantic network. *Enterprise Information Systems*. 4:1–22.

[13] Bai J, Song D, Bruza P, Nie J-Y, Cao G. 2005. Query expansion using term relationships in language models for information retrieval. *Proceedings of the 14th ACM international conference on Information and knowledge management: ACM*. 688–95.

[14] Riezler S, Vasserman A, Tsochantaridis I, Mittal V, Liu Y. 2007. Statistical machine translation for query expansion in answer retrieval. *Annual Meeting-Association For Computational Linguistics*. 464.

[15] Lee J-T, Kim S-B, Song Y-I, Rim H-C. 2008. Bridging lexical gaps between queries and questions on large online Q&A collections with compact translation models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Association for Computational Linguistics*. 410–8.

[16] Carpineto C, Romano G. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*. 44:1.

[17] Zhong Z, Ng HT. 2012. Word sense disambiguation improves information retrieval. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Association for Computational Linguistics*.1: 273–82.

[18] Berger A, Lafferty J. 1999. Information retrieval as statistical translation. *Proceedings Of The 22nd Annual International ACM SIGIR Conference On Research And Development In Information Retrieval: ACM*. 222–9.

[19] Sun L, Liu B, Wang B, Zhang D, Wang X. 2010. A study of features on Primary Question detection in Chinese online forums. *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on: IEEE*. 2422–7.

[20] Catherine R, Singh A, Gangadharaiah R, Raghu D, Visweswariah K 2012. Does Similarity Matter? The Case of Answer Extraction from Technical Discussion Forums. *COLING (Posters)*. 175–84.

[21] Jeon J, Croft WB, Lee JH, Park S. 2006. A framework to predict the quality of answers with non-textual features. *Proceedings Of The 29th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval: ACM*. 228–35

[22] Clark E, Araki K. 2011. Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. *Procedia-Social and Behavioral Sciences*. 27:2–11.

[23] Muthmann K, Löser A. 2010. Detecting near-duplicate relations in user generated forum content. *On the Move to Meaningful Internet Systems: OTM 2010 Workshops: Springer*. 698–707.

[24] Pattabiraman K, Sondhi P, Zhai C. 2013. Exploiting Forum Thread Structures to Improve Thread Clustering. *Proceedings of the 2013 Conference on the Theory of Information Retrieval: ACM*. 15.

[25] Subramaniam LV, Roy S, Faruquie TA, Negi S. 2009. A survey of types of text noise and techniques to handle noisy text. *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data: ACM*. 115–22.

[26] Kukich K. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*. 24:377–439.

[27] Xi W, Lind J, Brill E. 2004. Learning effective ranking functions for newsgroup search. *Proceedings Of The 27th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval: ACM*. 394–401.

[28] Xue Z, Yin D, Davison BD. 2011. Normalizing microtext. *Proceedings of the AAAI Workshop on Analyzing Microtext*. 74–9.

[29] Rama T, Singh AK, Kolachina S. 2009. Modeling letter-to-phoneme conversion as a phrase based statistical machine translation problem with minimum error rate training. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium: Association for Computational Linguistics*. 90–5.

[30] Dou Q, Bergsma S, Jiampojamarn S, Kondrak G. 2009. A ranking approach to stress prediction for letter-to-phoneme conversion. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Association for Computational Linguistics*. 1: 118–26.

[31] Bartlett S, Kondrak G, Cherry C. 2008. Automatic Syllabification with Structured SVMs for Letter-to-Phoneme Conversion. *ACL* . 568–76.

[32] Seo J, Croft WB, Smith DA. 2009. Online community search using thread structure. *Proceedings Of The 18th ACM Conference On Information And Knowledge Management: ACM*. 1907–10.

[33] Kim SN, Wang L, Baldwin T. 2010. Tagging and linking web forum posts. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Association for Computational Linguistics*. 192–202.

[34] Adams PH, Martell CH. 2008. Topic detection and extraction in chat. *Semantic Computing, IEEE International Conference on: IEEE*. 581–8.

[35] Khandelwal SHS. 2004. Automatic Topic Extraction and Classification of Usenet Threads.

[36] Shen D, Yang Q, Sun J-T, Chen Z. 2006. Thread detection in dynamic text message streams. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval: ACM*. 35–42.

[37] Shi L, Sun B, Kong L, Zhang Y. 2009. Web forum Sentiment analysis based on topics. *Computer and Information Technology, 2009 CIT'09 Ninth IEEE International Conference on: IEEE*. 148–53.

[38] Huang J, Zhou M, Yang D. 2007. Extracting Chatbot Knowledge from Online Discussion Forums. *IJCAI*. 423–8.

[39] Kim JW, Candan KS, Dönderler ME. 2005. Topic segmentation of message hierarchies for indexing and navigation support. *Proceedings of the 14th international conference on World Wide Web: ACM*. 322–31.

[40] Labadié A, Prince V. 2008. Intended boundaries detection in topic change tracking for text segmentation. *International Journal of Speech Technology*. 11: 167–80.

[41] Georgiou T, Karvounis M, Ioannidis Y. 2010. Extracting Topics of Debate between Users on Web Discussion Boards. *ACM SIGMOD Conf, Undergraduate Research Poster Competition*.