IGCESH2014
Universiti Teknologi Malaysia, Johor Bahru, Malaysia 19-21 August 2014

Mining Question-Answer Pairs from Web Forum: A Survey of Challenges and Resolutions

Adekunle I. Obasa[1*], Naomie Salim[2] and Yazan A. Al-khassawneh[3]

[1, 3] Faculty of Computing, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia.
(E-mail: [1]iaobasa@yahoo.com, [3]yakhassawneh@yahoo.com)
[2] Faculty of Computing, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia.
(E-mail: [2]naomie@utm.my)

ABSTRACT

Internet forums, which are also known as discussion boards, are popular web applications. Members of the board discuss issues and share ideas to form a community within the board, and as a result generate huge amount of content on different topics on daily basis. Interest in information extraction and knowledge discovery from such sources has been on the increase in the research community. A number of factors are limiting the potentiality of mining knowledge from forums. Lexical chasm or lexical gap that renders some Natural Language Processing techniques (NLP) less effective, Informal tone that creates noisy data, drifting of discussion topic that prevents focused mining and asynchronous issue that makes it difficult to establish post-reply relationship are some of the problems that need to be addressed. This survey introduces these challenges within the framework of question answering. 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.

KEYWORDS: Internet forum; Text mining; Lexical Chasm; Informal tone; Topic drifting.

1. 0 INTRODUCTION

A forum can be considered as a topic-based document set that has a definite boundary separated by members and non-members. Almost all forums have hierarchical structures. A forum comprises of sub-forums depending on the broad topic categories. 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 [1].

The huge amount of responses and the variations of response context lead to the problems of efficient knowledge accumulation and retrieval [2]. Mining of human generated contents of forums is non-trivial due to its nature. In this paper, four issues that hinder effective mining of knowledge from forums are discussed. Different approaches that researchers consider in overcoming them are explored with a few of presenting the actions

that have been taken so far to resolve them. We also proffer suggestions that can further assist in addressing the problems.

2.0 CHALLENGES AND RESOLUTIONS OF MINING QUESTION- ANSWER PAIRS FROM FORUMS

Predominantly, the content generated in forums are questions / problems and their answers / resolutions. It was empirically confirmed by [3] that 90% of 40 forums investigated contain question-answer knowledge. Mining these question-answer pairs available in different domains will be an asset to 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 [3-6]. Some of the challenges hindering effective Mining of Question-answer pairs are: i. Lexical chasm ii. Informal tones iii. Asynchronous issue and iv. Topic drifting

2.1 Lexical Chasm Issue

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 synonymy (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"), polysemy (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 post of web forum usually includes 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 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.

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 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.

2.1.1 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-content based features shall be reviewed.

a) Query expansion- In mining question and answer from forum, the query question is usually composed with 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.

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 $\rightarrow$ cigarette, why $\rightarrow$ because, URL $\rightarrow$ website and MS $\rightarrow$ Microsoft). These words are added to the query being used for the mapping so as to augment the original query to produce a representation that better reflects the underlying information need.

b)Word sense disambiguation (WSD) - is a method that identifies the meaning of words in a computational manner within the context of their usage [16]. It has been applied successfully in machine translation, information retrieval, information extraction, etc. It is a promising approach for bridging gabs 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* 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.

c) 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 M_D built from D generating Q:

$$sim(Q, D) \approx P(Q|M_D) \tag{1}$$

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:

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

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

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 [17] 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:

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

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. 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.

d) Non-content features –A much more prevalent approach of tackling lexical gaps in web forum question answering is to avoid the use of contextual data. The non-content 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 [4, 18] 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 [1, 19]. A major problem with non-content features is their availability. Some non-content features used by some forums may not be found in others. The degree of availability of some non-content features across forums can be found in [19]. It worth noting that combination of both the contextual and non-contextual is desirable for effective mining of question-answer pairs from forum. The contextual features measure the degree of relevance between question and answer while non-contextual can be used to estimate the quality of answers [20].

2.2 Informal Tone

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.
The very 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 [3, 9, 18, 21, 22].

The noise content of forum can be said to come from two sources. These sources appear to be in line with sources identified by [23] 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 and 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 [23]. In forum, text normalization activities have been concentrated on the second noise source. Categorization of forum noise as contained in [1] is shown in Table 1.

Table 1 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

2.2.1 Informal Tone Resolution Approaches

A number of methods from different research areas have emerged for identifying and correcting words in text. A good work by [24] 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 [25] 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[1] or Hunspell[2] can also be used to fix some of the spelling mistakes found in forum corpora. An empirical result of [26] 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 [27] to fix the four noise classes x-rayed in Table 1.

2.3 Asynchronous Problem

In web forum, multiple questions and answers are often discussed in parallel. Many a time the discussions are interwoven together. It is possible for a post to contain answers to multiple questions. It is also a possibility for one question to have multiple replies. For the post that contains several questions, a scenario often referred to as *complex question*, the answers to these questions may be found in separate replies which in a way will need extra efforts to bring them together. A post containing multiple answers to questions in different

[1] http://aspell.net
[2] http://hunspell.sourceforge.net

posts may lead to question-answer mismatch. It was confirmed empirically by [28] that nearly 50% of post in forums contains two questions and above.

2.3.1 Asynchronous Problem Resolution

Segmentation techniques are often being used to resolve this problem [7, 28]. In [28], six different strategies were adopted for question segmentation. The best of the six strategies recorded 86% accuracy. [7] applied the technique for answer detection. They implemented thread segmentation to reorganize the posts contained in the threads into several fairly independent units, which reduce the influence of asynchrony and preserve the strong relevance for the posts within the same segment. Issue of complex questions have been addressed by a number of researchers using different approaches that can be broadly classified as learning [29, 30] and non-learning [31]. The learning methods are much more promising judging by the results they produce but are expensive.

2.4 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 unique topic is discussed in a thread, different features and aspects of it may be considered in the discussion. There is 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. It has rightly been pointed out by [25, 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 contains many posts, say 6 and above.

2.4 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.

3.0 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) Informal tone that makes forum data to be highly noise. iii) Asynchronous problem that at times do lead to question and answer mismatched and iv) 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.

REFERENCES

[1] A.I. Obasa, N. Salim. Mining Faq From Forum Threads: Theoretical Framework, Journal of Theoretical & Applied Information Technology, 63 (2014).

[2] W.-C. Hu, D.-F. Yu, H.C. Jiau. A FAQ Finding Process in Open Source Project Forums, Fifth International Conference on Software Engineering Advances, (2010) 259-264.

[3] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, Y. Sun. Finding question-answer pairs from online forums, in: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval: ACM, 2008, pp. 467-474.

[4] L. Hong, B.D. Davison. A classification-based approach to question answering in discussion boards, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval: ACM, 2009, pp. 171-178.

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

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

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

[8] E. Brill, S. Dumais, M. Banko. An analysis of the AskMSR question-answering system, in: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10: Association for Computational Linguistics, 2002, pp. 257-264.

[9] B. Wang, B. Liu, C. Sun, X. Wang, L. Sun. Extracting Chinese question-answer pairs from online forums, in: Systems, Man and Cybernetics, 2009 SMC 2009 IEEE International Conference on: IEEE, 2009, pp. 1159-1164.

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

[11] D. Bernhard, I. Gurevych. Combining lexical semantic resources with question & answer archives for translation-based answer finding, in: 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: Volume 2-Volume 2: Association for Computational Linguistics, 2009, pp. 728-736.
[12] Z. Gong, M. Muyeba, J. Guo. Business information query expansion through semantic network, Enterprise Information Systems, 4 (2010) 1-22.
[13] J. Bai, D. Song, P. Bruza, J.-Y. Nie, G. Cao. Query expansion using term relationships in language models for information retrieval, in: Proceedings of the 14th ACM international conference on Information and knowledge management: ACM, 2005, pp. 688-695.
[14] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, Y. Liu. Statistical machine translation for query expansion in answer retrieval, in: ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2007, pp. 464.
[15] J.-T. Lee, S.-B. Kim, Y.-I. Song, H.-C. Rim. Bridging lexical gaps between queries and questions on large online Q&A collections with compact translation models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing: Association for Computational Linguistics, 2008, pp. 410-418.
[16] Z. Zhong, H.T. Ng. Word sense disambiguation improves information retrieval, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1: Association for Computational Linguistics, 2012, pp. 273-282.
[17] A. Berger, J. Lafferty. Information retrieval as statistical translation, in: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval: ACM, 1999, pp. 222-229.
[18] L. Sun, B. Liu, B. Wang, D. Zhang, X. Wang. A study of features on Primary Question detection in Chinese online forums, in: Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on: IEEE, 2010, pp. 2422-2427.
[19] R. Catherine, A. Singh, R. Gangadharaiah, D. Raghu, K. Visweswariah. Does Similarity Matter? The Case of Answer Extraction from Technical Discussion Forums, in: COLING (Posters), 2012, pp. 175-184.
[20] J. Jeon, W.B. Croft, J.H. Lee, S. Park. A framework to predict the quality of answers with non-textual features, in: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval: ACM, 2006, pp. 228-235.
[21] K. Muthmann, A. Löser. Detecting near-duplicate relations in user generated forum content, in: On the Move to Meaningful Internet Systems: OTM 2010 Workshops: Springer, 2010, pp. 698-707.
[22] K. Pattabiraman, P. Sondhi, C. Zhai. Exploiting Forum Thread Structures to Improve Thread Clustering, in: Proceedings of the 2013 Conference on the Theory of Information Retrieval: ACM, 2013, pp. 15.
[23] L.V. Subramaniam, S. Roy, T.A. Faruquie, S. Negi. A survey of types of text noise and techniques to handle noisy text, in: Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data: ACM, 2009, pp. 115-122.
[24] K. Kukich. Techniques for automatically correcting words in text, ACM Computing Surveys (CSUR), 24 (1992) 377-439.
[25] W. Xi, J. Lind, E. Brill. Learning effective ranking functions for newsgroup search, in: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval: ACM, 2004, pp. 394-401.
[26] E. Clark, K. Araki. Text normalization in social media: progress, problems and applications for a pre-processing system of casual English, Procedia-Social and Behavioral Sciences, 27 (2011) 2-11.
[27] Z. Xue, D. Yin, B.D. Davison. Normalizing microtext, in: Proceedings of the AAAI Workshop on Analyzing Microtext, 2011, pp. 74-79.
[28] C.-J. Lin, C.-H. Cho. Question pre-processing in a QA system on internet discussion groups, in: Proceedings of the Workshop on Task-Focused Summarization and Question Answering: Association for Computational Linguistics, 2006, pp. 16-23.
[29] S. Fan, W.W. Ng, X. Wang, Y. Zhang, X. Wang. Semantic chunk annotation for complex questions using conditional random field, in: Coling 2008: Proceedings of the workshop on Knowledge and Reasoning for Answering Questions: Association for Computational Linguistics, 2008, pp. 1-8.
[30] F. Li, X. Zhang, J. Yuan, X. Zhu. Classifying what-type questions by head noun tagging, in: Proceedings of the 22nd International Conference on Computational Linguistics, 2008, pp. 481-488.
[31] N. Tomuro. Interrogative reformulation patterns and acquisition of question paraphrases, in: Proceedings of the second international workshop on Paraphrasing-Volume 16: Association for Computational Linguistics, 2003, pp. 33-40.
[32] J. Seo, W.B. Croft, D.A. Smith. Online community search using thread structure, in: Proceedings of

the 18th ACM conference on Information and knowledge management: ACM, 2009, pp. 1907-1910.
[33] S.N. Kim, L. Wang, T. Baldwin. Tagging and linking web forum posts, in: Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Association for Computational Linguistics, 2010, pp. 192-202.
[34] P.H. Adams, C.H. Martell. Topic detection and extraction in chat, in: Semantic Computing, 2008 IEEE International Conference on: IEEE, 2008, pp. 581-588.
[35] S.H.S. Khandelwal. Automatic Topic Extraction and Classification of Usenet Threads.
[36] D. Shen, Q. Yang, J.-T. Sun, Z. Chen. Thread detection in dynamic text message streams, in: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval: ACM, 2006, pp. 35-42.
[37] L. Shi, B. Sun, L. Kong, Y. Zhang. Web forum Sentiment analysis based on topics, in: Computer and Information Technology, 2009 CIT'09 Ninth IEEE International Conference on: IEEE, 2009, pp. 148-153.
[38] J. Huang, M. Zhou, D. Yang. Extracting Chatbot Knowledge from Online Discussion Forums, in: IJCAI, 2007, pp. 423-428.
[39] J.W. Kim, K.S. Candan, M.E. Dönderler. Topic segmentation of message hierarchies for indexing and navigation support, in: Proceedings of the 14th international conference on World Wide Web: ACM, 2005, pp. 322-331.
[40] A. Labadié, V. Prince. Intended boundaries detection in topic change tracking for text segmentation, International Journal of Speech Technology, 11 (2008) 167-180.
[41] T. Georgiou, M. Karvounis, Y. Ioannidis. Extracting Topics of Debate between Users on Web Discussion Boards, in: ACM SIGMOD Conf, Undergraduate Research Poster Competition, 2010.