Multimed Tools Appl (2014) 68:401–412
DOI 10.1007/s11042-012-1338-z

Contextual keyword extraction by building sentences with crowdsourcing

Soon Gill Hong · Sungho Shin · Mun Yong Yi

Published online: 3 January 2013

Abstract Automatic keyword extraction from documents has long been used and proven its usefulness in various areas. Crowdsourced tagging for multimedia resources has emerged and looks promising to a certain extent. Automatic approaches for unstructured data, automatic keyword extraction and crowdsourced tagging are efficient but they all suffer from the lack of contextual understanding. In this paper, we propose a new model of extracting key contextual terms from unstructured data, especially from documents, with crowdsourcing. The model consists of four sequential processes: (1) term selection by frequency, (2) sentence building, (3) revised term selection reflecting the newly built sentences, and (4) sentence voting. Online workers read only a fraction of a document and participated in sentence building and sentence voting processes, and key sentences were generated as a result. We compared the generated sentences to the keywords entered by the author and to the sentences generated by offline workers who read the whole document. The results support the idea that sentence building process can help selecting terms with more contextual meaning, closing the gap between keywords from automated approaches and contextual understanding required by humans.

Keywords Crowdsourcing · Keyword extraction · Document summary · Content extraction · Sentence building · Contextual term extraction

1 Introduction

As the number of documents and multimedia resources is continuously increasing, our need to automatically classify and extract knowledge from them grows to a greater extent every

S. G. Hong · M. Y. Yi (✉)
Department of Knowledge Service Engineering, KAIST, Daejeon, Republic of Korea
e-mail: munyi@kaist.ac.kr

S. G. Hong
e-mail: hsoongil@kaist.ac.kr

S. Shin
Korea Institute of Science and Technology Information, Daejeon, Republic of Korea
e-mail: maximus74@kisti.re.kr

day. Consequently, the role of keywords in information search has become important. If documents are given with proper keywords, searching and retrieving of information by users can be vastly improved. Automatic keyword assignment seems to be practical in that it is efficient and cost-effective compared to using human indexers [7]. Tagging for multimedia resources by anonymous users behind the screen also has emerged as one of the automatic approaches and shows promising result for multimedia resources, which cannot be classified effectively by automatic approach [14]. The concept of crowdsourcing is similar to that of outsourcing except that traditional knowledge workers are replaced with public members in the form of an open call [4]. Workers receive assignments and submit the results through the Internet. Requesters post tasks online and approve or reject the results through the Internet as well.

In this paper, we propose a new model of extracting key contextual terms from documents by building sentences with crowdsourcing. The new model consists of four sequential processes: (1) term selection by frequency, (2) sentence building, (3) revised term selection reflecting the newly built sentences, and (4) sentence voting. Online workers read only a fraction of a document and participate in sentence building and sentence voting processes. The model allows human intelligence to intervene in an automatic keyword extraction process and make major decisions in selecting key contextual terms. Further, by asking anonymous online users to build sentences with given words in the form of drop down lists, human responses are facilitated. Forming simple statements can attract online users to perform time consuming and potentially dull tasks with ease and joy. In short, the goal of this research is to propose and validate a model of key contextual term extraction in which automatic processes are augmented by human intelligence operated through crowdsourcing.

2 Related work

The basic premise of automatic keyword extraction from a single document is largely based on the idea that frequently appearing words better represent the contents of the document than less frequently used words do. In a single document, the measure of importance is the co-occurrences of a specific word with a particular subset of frequent words [9], the occurrences of a word in a specific location such as the head of noun phrases, and the length of the words [1]. In multiple documents, more frequently used words in one document but less frequently repeated in other documents can be selected as keywords for that document, which is known as tf*idf (Term Frequency — Inverse Document Frequency).

Machine learning algorithms analyze documents and build structural models. When a new document comes in, the models process it and extract keywords based on their past experiences. Turney's algorithm is based on Naive Bayes, and it uses term frequency (tf), collection frequency (idf), and the position of a term [13]. Frank et al. trained Turney's algorithm on domain-specific documents and showed that the results were comparable to the state of the art [3]. Some additional external sources such as thesauruses for a specific domain can be used to find more appropriate words that do not appear in an original document [7].

Although some of these approaches have shown improved outcomes in some areas, they also pose some limitations [15]. For example, a keyword cannot have more than three tokens [3, 13], or external resources should constantly be updated to remain effective [6]. In order to find the co-occurrence of words, the order of the words has to be ignored so that the modification relationship of sentences, which is important to understanding the meaning of the sentences, is not taken into account [9]. Although applying recent developments in natural language processing to keyword extraction shows improvement of the extraction accuracy [6], the inherent insufficiency of automatic keywords extraction or texts retrieval originates from the fact that software has the difficulty of understanding the contents of a document because software relies on syntactic information or past experience rather than on the context of the document. The best way for keyword assignment would be to utilize human indexers or people who have real needs of using particular information. However, human indexers are expensive, and it is not easy to timely reach the people who may need the information.

With the advent of crowdsourcing, however, this situation has changed. The concept of crowdsourcing is similar to the idea of outsourcing except that traditional human indexers are substituted with unspecified individuals who use the Internet in the form of an open call [4]. The tasks covered by crowdsourcing range from instantaneous image tagging to natural language processing, with which traditional artificial intelligent systems have difficulty in accomplishing their goals. One of the typical crowdsourcing applications is tagging images in the form of a game [14]. Snow et al. perform human linguistic annotation with crowdsourcing and show promising results [12]. Another subjective task done with crowdsourcing is Soylent, a word processing interface that allows writers to ask for help from workers through the Internet [2].

One of the biggest beneficiaries of crowdsourcing can be multimedia resources. Automatic processing doesn't perform well for recognizing image patterns, identifying voice signals, or understanding the context of documents. However, with crowdsourcing in alternative operationalization possibilities (i.e., gaming, social interaction), tagging conducted by humans has become a prevalent approach for the self-organization or retrieval purpose of the resources. However, as manual tagging allows arbitrary, idiosyncratic words to be used, it poses quality problems related to redundant, inaccurate, or ambiguous tags [5]. Also, because users do not consider their tags as indexing terms or search terms when they are created, tags are not so practically helpful for enhancing search efficiency [10].

3 Key contextual term extraction with crowdsourcing

In order to perform key contextual term extraction with crowdsourcing, we propose a process model with four sequential steps, as shown in Fig. 1. In the first step, we select frequent noun phrases and verbs based on frequencies. In the second step, we ask anonymous online users to build sentences with those frequent noun phrases and verbs. In the third step, we revisit the most frequent noun phrases by aggregating their frequency rates in the newly built sentences and in the original document. We then choose a set of sentences that have the revisited noun phrases as subjects and put those sentences to vote. In the fourth step, we ask a different set of anonymous online users to vote for the sentences that best

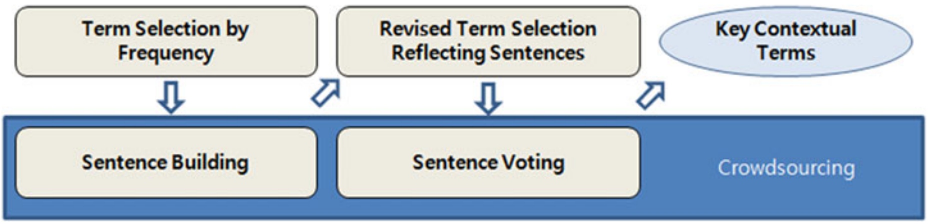


Fig. 1 Model of key contextual term extraction with crowdsourcing

describe the document. Finally, words or phrases from the sentences with a majority vote are regarded as key contextual terms. This model provides a built-in quality control structure as the work results of one group of online users are reviewed later by another group of users [11]. Also, the model supports efficiency by not requiring the users to read the whole document.

For our research, we selected a document entitled "Knowledge Sharing and Yahoo Answers: Everyone Knows Something," which describes Yahoo Answer, one of the largest question-answer forums. We selected this paper as it was deemed to be a sound, representative document for key contextual term extraction. The paper's keywords are consisted of noun phrases and verbs. Verbs are essential to capture the contextual meaning of nouns or noun phrases by specifying how they are related. In traditional keyword extraction approaches, only nouns and noun phrases regarded as keywords for a document.

3.1 Term selection by frequency

We preprocessed the chosen document by using the Stanford CoreNLP suite. This software reads a document and returns pairs of original token, tag, and lemma. Lemma is a base or dictionary form of a word from which inflectional ending is removed. For example, 'goes' and 'went' have the same lemma 'go.' After the preprocessing, we used two patterns for extracting noun phrases and verbs, and any set of nouns and verbs complying with one of these patterns were extracted. Those patterns were expressed in regular expressions: ((VBG)? (NNPS)+(NNPS)+) or ((JJ)+(NNPS)+). VBG stands for verb, gerund, or present participle; JJ for adjective; NNPS for singular/plural noun; '?' for zero or more; and '+' for one or more. Any set of words matching any of these patterns was extracted as candidate keywords for further processing. This processing generated 575 unique noun phrases, including one noun phrase 'social network' that appeared 23 times. It also generated verbs from words used in present or past tense, gerund or present participle, and past participle, resulting in 194 unique verbs extracted from the document, including one verb 'answer' that appeared 34 times.

As we had too many noun phrases and verbs for building sentences, we reduced the set by selecting the most frequently appearing ones. Specifically, we selected the most frequently appearing 5 % noun phrases (i.e., 25) out of 575 unique noun phrases. As for verbs, they were less distinctive (194 verb vs. 575 noun phrase), but their average frequency rates were higher than that of noun phrases (4.27 for verbs vs. 1.4 for noun phrases). This shows that verbs were used more frequently than noun phrases. To adjust the bias of noun phrase and verb distributions, we applied the following rules to select the most frequently appearing terms in their own category.

$$\mathrm{F_{np}} = (\mathrm{Total\,\#\,of\,NP})^{*}(\mathrm{Avg.\,freq.\,of\,NP})^{*}(\mathrm{Sampling\,Rate}) = 575^{*}1.4^{*}0.05 \approx 25$$
$$\mathrm{F_{verb}} = (\mathrm{Total\,\#\,of\,Verb})^{*}(\mathrm{Avg.\,freq.\,of\,Verb})^{*}(\mathrm{Sampling\,Rate}) = 194^{*}4.27^{*}0.05 \approx 40$$

Accordingly, the most frequent 25 noun phrases and 40 verbs were chosen as base terms in this step. At this point, the nouns phrases of the base terms could be regarded as keywords for the document in traditional keyword extraction approaches, but we went further to employ crowdsourcing.

3.2 Sentence building

One of the biggest challenges with crowdsourcing is how to ensure job quality. If tasks are objective and require low cognitive load or little expertise, online workers are willing to perform the tasks even for little money or fun with free of charge. However, if tasks require a certain amount of knowledge or creativity, the workers tend to skip those tasks or regard them as dull assignments and provide random answers [8]. As keyword extraction is a subjective task and requires cognitive load or creativity, we need to prevent workers' possible negligent attitude toward the given task, thereby ensuring good work quality. We employed two quality control processes: the first process was used in this step of sentence building. We asked workers to construct sentences with given base terms through the interface shown in Fig. 2. We informed them that their work would be reviewed later by other people. In the second process, which was used in the fourth step, we asked other workers, for a validation purpose, to vote for the sentences that best described the meaning of the presented document.

Specifically for this study, we posted the sentence building task for 2 days (from December 4, 2011 to December 5, 2011) on the Amazon Mechanical Turk website. Workers whose past approval rate was more than 90 % were allowed to participate while each worker was allowed to build the maximum of 3 sentences. Participants were told that they would be paid $0.05 upon approval and were given the maximum of 45 minutes to complete the task. Sentences to be built had a simple *subject-predicate-object* syntax. Workers were provided with the abstract of the document, and 3 sets of subject-predicate-object drop down lists (see Fig. 2). Workers built a sentence by selecting one subject, one predicate, and one object from the list, and repeated this procedure three times. A total of 30 workers participated in the task, spending about 3 minutes and 59 seconds on average per task, resulting in 90 sentences in total. Among the 90 sentences, 87 sentences were unique as 3 sentences were built two times each and 84 sentences were built once each

3.3 Revised term selection reflecting sentences

As discussed before, it might be inefficient to use the sentences obtained through the second step without any validation process because online workers work anonymously under little

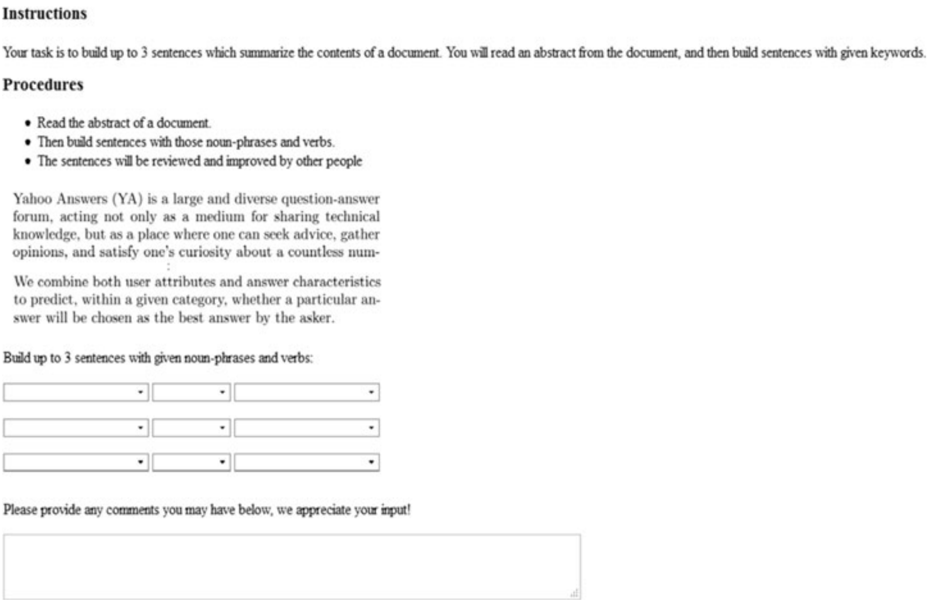


Fig. 2 Sentence building interface

surveillance [11]. We extracted the most frequent top 10 noun phrases used as subjects and objects from the sentences and selected the most frequent top 10 noun phrases from the document. We scored them by their frequencies and selected the top 5 most frequent noun phrases from the total 20 noun phrases. Also, we selected non-frequent noun phrase *active user* on purpose for validation and checked later whether sentences with this noun phrase have a majority vote, which was not the case. Based on these top 5 frequent noun phrases and 1 non-frequent noun phrase, we extracted 51 out of 90 sentences, which used one of those six noun phrases as a subject, and subsequently they were put to the vote.

3.4 Voting for sentences

We posted a voting task on the Amazon Mechanical Turk for 1 day on December 7, 2011.

Twenty workers participated in the voting task (Fig. 3), and voted 130 times in total. Theoretically, because 20 workers could vote for each sentence, one sentence could have up to 20 votes. The sentence with the largest number of votes was “online interaction help knowledge sharing,” which had 8 votes. Each worker voted 6.5 times on average (i.e., each worker selected 6.5 sentences), and a total 44 out of 51 sentences had at least one vote.

3.5 User responses

During the online experiment, we asked users to leave comments on the assigned task and some did. Two of them said the task was “very” interesting, one person said that “This type of sentence forming is very useful to us.” Another person considered the assigned task as a “good mind game,” which we did not anticipate. Collectively, the comments show that the participants thought that the assigned task was useful and interesting.

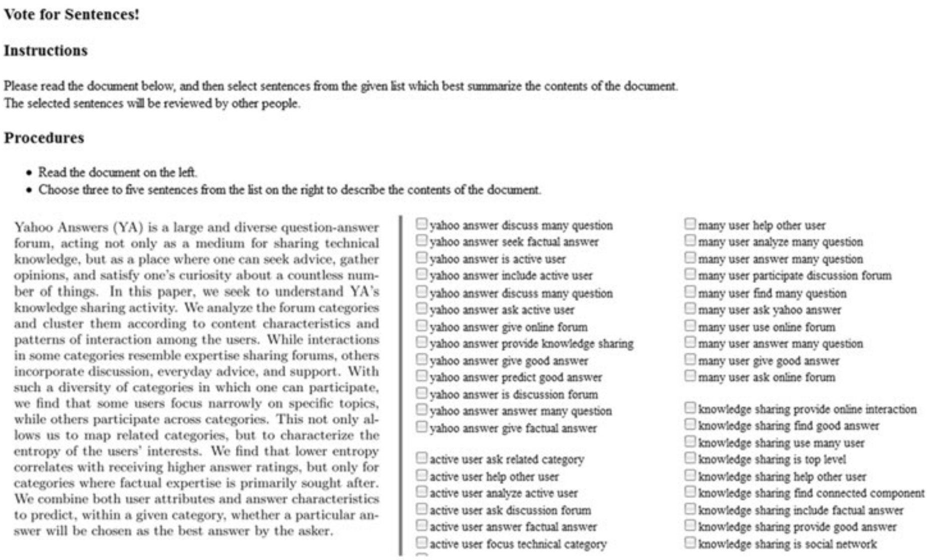


Fig. 3 Interface for voting sentences

4 Evaluation

We evaluated our proposed model from two perspectives. First, we compared precision and recall of online users' outputs to those of the frequency-based approach. This comparison shows how our new model performs compared to an automated (machine only) approach. Second, we calculated similarity scores between online users' outputs and offline users' outputs. This comparison shows how our model performs compared to human only approach.

4.1 Comparison to automated approach

For evaluation, one on one (1:1) matching of the terms from online workers (Online Group) and the terms from the frequency-based approach, which is commonly used for automatic keyword extraction, was conducted. To make the comparison process straightforward, we compared the precision and recall of the Online Group with those of the frequency-based approach. We used distinct 13 words from the 6 keywords entered by the author in the original document as gold standard.

As there were distinct 13 words from the 6 keywords originally entered, we selected the Online Group's top 4 sentences of which the distinct number of words was also 13. Accordingly, to equalize the number of words, we selected 7 most frequent noun phrases, identified by the frequency-based approach, of which the distinct number of words was also 13.

When we compared the words from the frequency-based approach to the words from the author, precision and recall were each 46 % (6 out of 13). When we compared the words from the Online Group with the words from the author, precision and recall were each 62 % (8 out of 13). The combined results (Fig. 4) show that the Online Groups outperforms the frequency-based approach substantially. When the two approaches (i.e., Frequency-based vs. Online Group) were directly compared, precision and recall for the keywords from the Online Group were each higher by about 33 % than the frequency-based approach.

4.2 Comparison to human readers

The previous experiment was designed to assess how much our proposed approach improves an automated approach. We conducted another experiment to assess how much the sentences built by the online workers by going through the proposed process (i.e., sentence building and sentence voting) but without reading the whole document are similar to the sentences

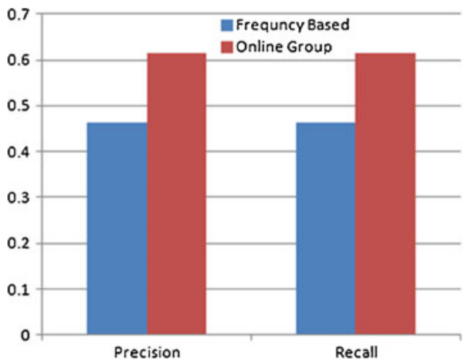


Fig. 4 Precision and recall for frequency-based approach vs. online group

people produce in a normal reading setting by going through careful digestion of the document and creating a summary using a few sentences.

The second experiment involved 13 graduate students in a research university in South Korea, who are majoring in information science related fields. The subjects voluntarily participated in the experiment and received about $20 monetary reward for participation. In the experiment, the subjects were asked to build 7 sentences (of which the number of words doubles the number of words from the keywords by the author to cover more broader range of ideas) in order to best describe the document content using the same set of base words Online Group used before (in the sentence building step). All the participants were given 2 hours in a lab setting, dedicated to reading the whole paper and creating a summary.

To measure similarities, we treated each set of keywords and sentences as a document and calculated the cosine similarity between them using the formula shown in Fig. 5. Cosine similarity is commonly used to measure similarities between term vectors in information retrieval. The similarity scores were normalized so that it is free from the bias associated with the document length.

We calculated the similarity between the keywords entered by the author and the keywords found in those sentences generated by each treatment group (Offline Group and Online Group). To maintain consistency across the groups, we used top 7 frequently selected sentences as key sentences from Online Group. Top 7 frequently selected sentences (13.7 %) had 43 votes (33.1 %). Figure 6 shows those sentences from online workers and the keywords entered by the author.

For Offline Group, we computed the similarity score between the keywords entered by the author and the keywords derived from each offline worker's sentences, and then computed the overall average of those scores, which was 0.46. For Online Group, following the same procedure, we calculated the similarity score between the keywords entered by the author and the keywords derived from the sentences generated by Online Group, and then computed the overall average of those scores, which was 0.67. Figure 7 summarizes the obtained similarity results.

The results show that the online approach provides a higher number of similar key terms than the offline approach. This seems reasonable because online users were provided just the abstract of the document and the keywords provided by the author while offline users were provided the whole document that includes the two components. As a result, it was likely that online workers paid more close attention on utilizing the keywords provided by the author while trying to cover the general context of the whole paper. However, offline workers were required to read the whole document. Consequently, they could pay attention to the details of the document and obtain diverse contextual information from the document.

We also calculated similarities between the sentences from the online workers and the sentences from the offline workers to see how comparable the results from the two approaches are. We computed the similarity score between each online worker's set of sentences and each offline worker's set of sentences, and then derived the overall average of the scores. The overall average similarity score between the sentences produced by the two alternative approaches (i.e., crowdsourcing involving online workers vs. detailed reading by offline human readers) was 0.42. In addition, we computed the similarity scores between the

Fig. 5 Cosine similarity

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Sentences from online workers	Freq.
online interaction help knowledge sharing	8
yahoo answer provide knowledge sharing	7
social network help knowledge sharing	7
yahoo answer discuss many question	6
many user participate discussion forum	6
knowledge sharing provide good answer	5
online interaction examine many question	4

Keywords from the author
Online
Community
Question
Answer
Social
Network
Analysis
Expertise
Find
Help
Seek
Knowledge
Share

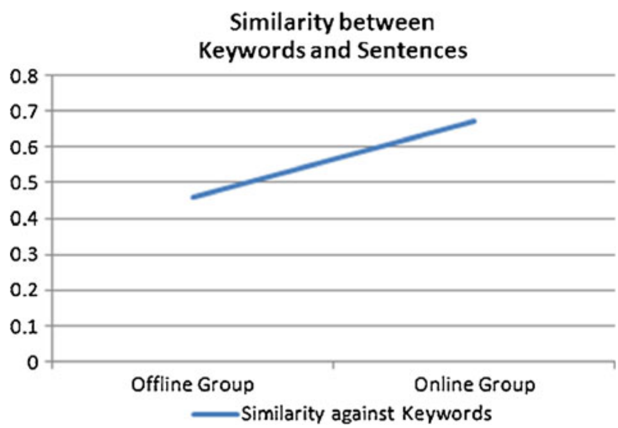
Fig. 6 Sentences from online workers and keywords from the author

set of sentences produced by each offline worker in the detailed reading condition, involving 78 (=13×12÷2) similarity comparisons. The similarity scores between the sentences produced by the offline human readers ranged from 0.29 to 0.81, with the average similarity score of 0.57. Collectively, the combined results show that the sentences produced by the proposed approach is acceptable, as it is within the similarity range produced by the human readers and close to its average.

5 Summary and conclusion

The study results show that combining crowdsourcing with an automated approach is a viable alternative for keyword extraction. The higher precision and recall show that crowdsourcing adds a value to the automated keyword extraction process. The simple activities of sentence building and sentence voting performed by the online workers vastly improve the keyword extraction process, resulting in sentences more inclusive of the keywords the author of the paper intended to use to represent the essence of the paper. Further the comparison with the sentences produced by human readers who spent two hours to read the whole document shows that the proposed approach, which took much less time, produces an acceptable set of sentences that are similar to the sentences human readers produced. While the proposed crowdsourcing approach

Fig. 7 Sentence similarity against keywords



produced sentences better utilizing the keywords entered by the author of the paper, the offline approach produced sentences more reflective of detailed contextual information embedded in the body of the document using diverse words and expressions. The crowdsourcing approach produces sentences comparable to the sentences produced by detailed human readers without requiring the same amount of effort or time commitment.

It seems that by asking users to build simple sentences, users seem to be oriented not to select words randomly. One of the limitations of our online approach is that it depends on the quality of the abstract or other fraction of the document that is given to online workers. If a document does not have an abstraction or introduction, we should decide which part to be given to online workers, and this is not trivial. While our approach needs further validation with many other documents and alternative forms of descriptions, the study findings clearly indicate that it can be applied to augment existing automatic keyword extraction processes. For multimedia applications, the proposed approach can be used to improve the multimedia tagging process. Instead of just asking users to tag, we can ask them to build simple sentences using a predefined set of noun phrases and verbs before tagging. Then they would be more careful to choose tags, possibly reflecting more faithfully the ideas original producer of the multimedia content intended to convey, instead of selecting tag words instantaneously. The exercise can direct users to consider the context of the tags by combining verbs and by seeing whether the sentences make sense.

We proposed a new key contextual term extraction model that combines both machine and human power. Following the proposed model, we first extracted a set of frequent noun phrases and verbs from a document by using the Stanford CoreNLP suite. Next, human workers read an abstract of the document and composed sentences of subject-predicate-object syntax to best describe the contents of the document while using the provided noun phrases and verbs selected from the previous step. Then, we revisited the selection of the keywords to reflect the frequently used words in the newly built sentences, and selected a number of sentences based on the revisited noun phrases. Those sentences were put to the vote. Finally, another group of online users read the abstract of the same document and voted for the sentences that best represent the document. The sentences with a majority vote were selected as a final set from which key contextual terms were derived for comparison.

We compared precision and recall between online workers' results and frequency-based results and found a significant improvement by the online approach, which combines crowdsourcing with the traditional keyword extraction process. Further, we compared the sentences produced by the online workers and offline workers and found that they are similar albeit not identical. The sentences produced by the online workers are more reflective of the keywords entered by the author to capture the essence of the paper while the sentences produced by the offline workers are more reflective of contextual details embedded in the body of the document. The overall results show that the proposed approach sits in between machine only approach and human only approach, with the tendency to remain faithful to the key ideas the original author of the paper intended to express. The proposed approach vastly improves the performance of the automatic keyword extraction process while producing a set of sentences similar to the sentences human readers normally produce after spending a considerable amount of time in digesting the document. Thus our study opens the possibility of combining crowdsourcing with the traditional text summarization and knowledge extraction processes, and enhancing the overall performance of those processes. Also, it taps into the possibility of creating an effective summary without requiring extensive human efforts, but still producing an acceptable solution. More broadly speaking, the proposed idea of building sentences with crowdsourcing should be able to solve many problems, which only humans or only machines have tried to solve but had major difficulties in solving the problems by themselves.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0029185).

References

1. Barker K, Cornacchia N (2000) Using noun phrase heads to extract document keyphrases. Adv Artif Intell 40–52. Springer
2. Bernstein MS, Little G, Miller RC, Hartmann B, Ackerman MS, Karger DR, Crowell D, Panovich K (2010) Soylent: a word processor with a crowd inside. Proceedings of the 23nd annual ACM symposium on User interface software and technology, 313–322. ACM
3. Frank E, Paynter G, Witten I, Gutwin C, Nevill-Manning CG (1999) Domain-specific keyphrase extraction. International Joint Conference on Artificial Intelligence 16:668–673
4. Howe J (2006) The rise of crowdsourcing. Wired magazine 14(6):1–4
5. Hsu W, Mei T, Yan R (2008) Knowledge discovery over community-sharing media: from signal to intelligence. Proceedings of the 17th international conference on World Wide Web, 665-674
6. Hulth A (2003) Improved automatic keyword extraction given more linguistic knowledge. Proceedings of the 2003 conference on Empirical methods in natural language processing, 216-223. Morristown, NJ, USA: Association for Computational Linguistics
7. Hulth A, Karlgren J, Jonsson A, Boström H, Asker L (2001) Automatic keyword extraction using domain knowledge. Comput Linguist Intell Text Process 472–482. Springer
8. Kittur BA (2010) CrowdSourcing, collaboration and creativity. XRDS: Crossroads The ACM Magazine for Students 17(2):22–26
9. Matsuo Y, Ishizuka M (2004) Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 13:157–170
10. Riste Gligorov (2012) User generated metadata in audio-visual collections. Proceedings of the 21st international conference, 139–143
11. Shaw AD, Horton JJ, Chen DL (2011) Designing Incentives for Inexpert Human Raters. Proceedings of the ACM conference on Computer supported cooperative work
12. Snow R, Connor BO, Jurafsky D, Ng AY, Labs D, St C (2008) Cheap and fast — but is it good? Evaluating non-expert annotations for natural language tasks. Comput Linguist 254–263
13. Turney P (2000) Learning algorithms for keyphrase extraction. Information Retrieval 2(4):303–336
14. Von Ahn L, Dabbish L (2008) Designing games with a purpose. Communications of the ACM 51(8):57
15. Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG (1999) KEA: practical automatic keyphrase extraction. Proceedings of the fourth ACM conference on Digital libraries, 254–255

Soon Hong is a doctoral candidate at Korea Advanced Institute of Science and Technology (KAIST), Korea. He obtained his B.A. and M.S. from Sungkyunkwan University and Baruch College, The City University of New York, respectively. Before joining KAIST, he worked as a senior consultant of business intelligence at Teradata Corp., U.S. His current research interests include knowledge extraction, knowledge representation, and semantic web.

Sungho Shin received his B.S. degree from Kyungbook National University, Korea, in 2000. He received his M.S. degree in Business Administration (Management Information Systems major) from Kyungbook National University, Korea in 2002. Currently, he works as a senior researcher at Korea Institute of Science and Technology Information (KISTI), Korea. Previously, He studied and developed a data quality management system. And recently he has researched and developed linguistic resource management system for Semantic service. His current research interests include information retrieval and discovery of technology opportunities based on Text Mining.

Mun Yi is an Associate Professor in the Department of Knowledge Service Engineering and the director of Knowledge Systems Lab at KAIST. He is participating in the Linked Open Data (http://lod2.edu) project, in which several European research institutions and KAIST collaborate on the development of tools for the Semantic Web. He earned his Ph.D. in Information Systems from University of Maryland, College Park. Before joining KAIST in 2009, he taught at University of South Carolina as an Assistant Professor (1998–2004) and (tenured) Associate Professor (2005–2009). His current research interests include semantic information retrieval, recommender systems, knowledge structure engineering, computer skill acquisition, and technology adoption. His work has been published in a number of journals including *Information Systems Research, Decision Sciences, Information & Management, International Journal of Human-Computer Studies,* and *Journal of Applied Psychology*. He is a former editorial member of *MIS Quarterly* and a current Associate Editor for *International Journal of Human-Computer Studies* and a Senior Editor for *AIS Transactions on Human-Computer Interaction*.