ARABIC DATASET FOR AUTOMATIC KEYPHRASE EXTRACTION

Mohammed Al Logmani[1] and Husni Al Muhtaseb[2]

[1]Information Technology, Saudi Aramco, Dhahran, Saudi Arabia
Mohammed.logmani@aramco.com
[2]Information & Computer Science Department, King Fahd University for Petroleum & Minerals, Dhahran, Saudi Arabia
muhtaseb@kfupm.edu.sa

ABSTRACT

We propose a dataset in Arabic language for automatic keyphrase extraction algorithms. Our Arabic dataset contains 400 documents along with their keyphrases. The dataset covers eighteen different categories. An evaluation using a state-of-the-art algorithm demonstrates the accuracy of our dataset is similar to that of English datasets.

KEYWORDS

Keyphrase extraction, Arabic, dataset

1. INTRODUCTION

Automatic keyphrase extraction aims to extract a set of phrases that are highly relevant and most descriptive phrases for the input text. The process of extracting keyphrases is achieved systematically and with no or minimal human interference. Keyphrase extraction algorithms mine the data corpus to extract important phrases and label the documents with these phrases. To verify the accuracy of the extraction algorithms, datasets are used. These datasets contain training and test documents along with the keyphrases representing the content of each document.

For the English language, there are many verified datasets. To illustrate some examples, there is Reuters Dataset [1] which contains more than 20,000 of documents with focus on text classification field. Also, the work presented in [2] prepared a large dataset consists of 2000 text from scientific papers. Additionally and on the same field of scientific papers, there are datasets submitted for the Workshop on Semantic Evaluation 2010 (Sem-Eval 2010) [3]. These datasets are tailored for machine-learning automatic keyphrase extraction algorithms.

For the Arabic language, we didn't find any published datasets that target the area of Arabic keyphrase extraction. However, we did find some related research including the human annotated Arabic dataset which provides annotation on books reviews [4]. Also, the work in [5] describes a corpus that classifies newspaper text into seven domains. Whereas the work in [6] explains a dataset with more than 17,000 texts with focus on text classification problem.

David C. Wyld et al. (Eds) : CSITA, ISPR, ARIN, DMAP, CCSIT, AISC, SIPP, PDCTA, SOEN - 2017
pp. 217– 222, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70121

In the few proposed algorithms for keyphrase extraction in Arabic, the number of documents used to experiment with the algorithm was small as mentioned in the respective publications. For example, KP-Miner [7] used a set of 100 Wikipedia documents where the algorithm proposed by El-Shishtawy [8] used a dataset consisted of 50 documents.

The main contribution of this work is the new Arabic dataset we have prepared. Below we explain the sources for the articles and the methodology used to create the dataset.

2. ARABIC DATASET

This section explains the sources for the articles and the methodology used to create the dataset. Our dataset contains 400 documents distributed on 18 different categories.

2.1. Sources

We have collected the documents from two sources: Arabic Wikipedia [9] and King Abdullah Initiative for Arabic Content [10].

- Arabic Wikipedia: is the main source of articles used in creating this dataset. Arabic Wikipedia contains more than 198,349 pages. For our goal, we obtained 365 articles from Wikipedia. Out of the 365 articles, approximately 200 articles were obtained from a previous work by Shaaban [11] where the rest were collected using BzReader [12]. BzReader is an application that allows offline browsing of the Wikipedia dump files and displays the text-only version of Wikipedia pages

- King Abdullah Initiative for Arabic Content: is an initiative aims to enrich the Arabic content on the internet after noticing the small percentage of Arabic content. According to this initiative, the percentage of Arabic digital content does not exceed 0.3% out of the world content composed of other languages. For our goal here, we obtained 35 articles with focus on medical topics.

2.2. Selection and organization

The corpus covers different knowledge areas like religion, history, geography, technology, sciences, sports…etc. Selecting the documents from different fields would help future automatic keyphrase extraction algorithm to cover general domains and not be tied to a specific domain like scientific papers. These documents vary also in size from 1 to 30 pages. The total number of words in these documents ranges approximately from 172 to 17,589 words. The number of words in the whole dataset is 1,708,168 words distributed on 288,191 lines. The documents are saved in text files with the extension (.txt) as Unicode format UTF-8.

The largest category with regard to number of documents is the people category with 59 documents where the smallest one is the food category with 3 documents. We also calculated the density percentage defined as the average number of words per file in each category. This measure shows the richness of a certain category based on the longest files they have and not based on the number of documents under that category. When calculating the density score, countries category scored the highest with 7,366. The next highest category is religion with 5,960 average words per file. This category contains 16 files. In this measure, food category scored last

with 1,233. For the health and medicine category, the density score is 1,950, which is very small comparing to the number of files (51).

The largest file in the Arabic dataset is from the history category and it is about the Ottoman Empire (الدولة العثمانية) with 17,589 words and 3,094 lines. The smallest file belongs to the environment category and it discusses radioactive pollution (التلوث الإشعاعي) with 172 words and 32 lines.

Table 1 shows the 18 categories we have chosen to use for the categorization of the files in our dataset. It also shows the sub-categories, the number of files, the total number of words, the total number of lines, and the density percentage in each category.

Table 1. Distribution of the documents in the Arabic dataset.

Category	**Sub-Category**	**Number of Files**	**Number of lines**	**Number of words**	**Density**
History	History	39	32,976	4,991	49.9%
Culture	Culture, Social, Cloths, Language, Buildings, palace, Festival, Flags	22	15,615	4,172	41.7%
Countries	Country, City	58	73,757	7,366	73.7%
Aviation	Airplane, Airport, Air Machine	5	1,954	2,450	24.5%
Health & Medicine	Health, Medicine, Medical	51	16,605	1,950	19.5%
Animals	Animal, Dinosaur, Zoology	29	26,606	5,459	54.6%
War	Battles, War Machines	21	8,460	2,459	24.6%
Technology	Technology, Software Engineering	12	8,631	4,237	42.4%
Sciences	Chemistry, Electricity, Energy, physics, Law	11	6,136	3,165	31.6%
Economy	Company, Economy	10	4,310	2,556	25.6%

Environment	Environmental Issues, Pollution	12	4,904	2,353	23.5%
Space	Space	20	10,621	3,258	32.6%
Entertainment	Fiction, Movie, Music	12	7,418	3,766	37.7%
Food	Fruit	3	624	1,233	12.3%
Geography	Geography, Mountain	8	2,696	1,989	19.9%
People	People	59	45,400	4,698	47.0%
Religion	Religion	16	16,400	5,960	59.6%
Sports	Sports	12	5,078	2,577	25.8%

2.3. Cleaning Up

When converting from Wikipedia pages to text format using BzReader, some clean-up for the format was needed. The clean-up process included removing some text that may confuse the readers or make the articles hard to read. Text that was generated due to the conversion from HTML/ Rich text format was eliminated. This includes place holders of graphics, sounds, and videos. To illustrate this point by an example, if the article contains several images, then the word (png) or (jpg) will be repeated several times in the text version of the article. Hence, this will increase the chance of selecting (png) or (jpg) as a keyword. This is because many of Keyphrase Extraction algorithms e.g. KEA [13] and KP-Miner [7] use term frequency as a factor when selecting candidates for keyphrases. The clean-up process included Wikipedia tables, side images captions, some references ...etc.

2.4. Manual Keyphrase Extraction

All documents were assigned to six readers to read and extract 10 keyphrases from each file. These keyphrases are stored in separate files with '.key' extension. This format is the one used by KEA and some other Automatic Keyphrase Extraction tools. In the '.key' files, each row represents a Keyphrase. They are sorted based on their importance in the article from high importance to low importance. The '.key' file name is matching exactly the '.txt' file. This is done to help the algorithm to locate the files in the training phases and help in organizing the dataset.

2.5. Keyphrase Verification

The final step in the methodology of preparing the Arabic dataset is the verification step. It included proofreading the articles and adjusting or concurring with the extracted keyphrases. This step also included reviewing and correcting the spelling mistakes, the number of keyphrases, and the '.key' files format.

As part of this work, the dataset was made available for future work related to Arabic keyphrase extraction. The dataset can be found at this link: https://github.com/logmani/ArabicDataset.

3. EVALUATION OF THE DATASET

To evaluate of the quality of our dataset, we conducted an evaluation using KEA, which is one of the most worked on algorithms in the area of keyphrase extraction. In our evaluation, we trained KEA using 300 documents, and our test set contained 100 documents. The F-measure (F-score) on our dataset was 19% which is very close to the results reported on [2] which was 19.08%. Furthermore, we ran KEA using the English dataset prepared in [3] and we found out the F-score was 15.3%. Hence, the quality of our dataset can be used reliably in future work related to keyphrase extraction algorithms. Table 2 shows a summary of our experiment on the Arabic datasets including the values of exact matching measure, precision, and recall.

Table 2. Summary of results obtained for the Arabic dataset.

Measure	Results on Arabic Dataset
Exact Match	189
Precision	18.9%
Recall	19.1%
F-Score	15.3%

4. CONCLUSIONS

In this research work, we prepared and presented an Arabic dataset for automatic keyphrase extraction. The dataset contains 400 documents along with their correspondence 400 keyphrases files. The dataset is publicly available for future automatic keyphrase extraction research on Arabic language. The evaluation showed that the quality of the dataset is reliable to be used in the future.

ACKNOWLEDGEMENTS

The authors would like to thank Saudi Aramco and King Fahd University of Petroleum and Minerals for supporting this research and providing the computing facilities.

REFERENCES

[1] Lewis, David. "Reuters-21578." Test Collections 1 (1987).

[2] Krapivin, Mikalai, Aliaksandr Autaeu, and Maurizio Marchese. "Large dataset for keyphrases extraction." (2009).

[3] Kim, S. N., Medelyan, O., Kan, M. Y., Baldwin, T.: Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 21-26. Association for Computational Linguistics. (2010)

[4] AL-Smadi, M., Qawasmeh, O., Talafha, B., Quwaider, M.: Human Annotated Arabic Dataset of Book Reviews for Aspect Based Sentiment Analysis. Proceedings of 3rd International Conference on Future Internet of Things and Cloud (FiCloud 2015), Rome, Italy (2015)

[5] Goweder, A., De Roeck, A.: Assessment of a significant Arabic corpus. Presented at the Arabic NLP Workshop at ACL/EACL 2001, Toulouse, France, (2001)

[6] Khorsheed, M., Al-Thubaity, A.: "Comparative evaluation of text classification techniques using a large diverse Arabic dataset." Language resources and evaluation vol.47, no. 2, pp.513-538 (2013).

[7] El-Beltagy, S., Rafea, A.: KP-Miner: A keyphrase extraction system for English and Arabic documents. Inf. Syst, vol. 34, no. 1, pp. 132–144. (2009)

[8] El-Shishtawy, T., Al-Sammak, A.: Arabic keyphrase extraction using linguistic knowledge and machine learning techniques. arXiv preprint arXiv:1203.4605. (2012)

[9] Wikipedia, “Arabic Wikipedia.” [Online]. Available: http://ar.wikipedia.org/wiki. [Accessed: 01-Jul-2012].

[10] King Abdullah Initiative for Arabic Content, “King Abdullah Initiative for Arabic Content.” [Online]. Available: http://www.econtent.org.sa. [Accessed: 07-Oct-2012].

[11] Shaaban, O.: Automatic Diacritics Restoration for Arabic Text. King Fahd University of Petroleum and Minerals. (2013)

[12] Tymchenko, V.: BzReader, an application to browse Wikipedia compressed dumps offline. http://code.google.com/p/bzreader/ (2012)

[13] Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., Nevill-Manning, C. G.: KEA: Practical automatic keyphrase extraction. In Proceedings of the fourth ACM conference on Digital libraries, pp. 254-255. ACM (1999).

AUTHORS

Mohammed Al Logmani works as Systems Analyst at Saudi Aramco. He obtained his M.S. degree in information & computer science from King zzFahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, in 2013 and the B.S. degree from King Abdul-Aziz University, Saudi Arabia in 2002. His research interests include software development and cyber security.

Husni Al-Muhtaseb Assistant Professor, Computer Science Department. He Obtained a PhD degree from the University of Bradford, UK in 2010. He received his M.S. degree in computer science and engineering from King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, in 1988 and the B.E. degree from Yarmouk University, Irbid, Jordan in 1984. His research interests include software development, Arabic Computing, computer Arabization, and Arabic OCR.