2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing

A Method for Recommending the Most Appropriate Expansion of Acronyms using Wikipedia

Dongjin Choi
Dept. of Computer Engineering
Chosun University
Gwangju, Republic of Korea
dongjin.choi84@gmail.com

Juhyun Shin
Dept. of Leaders in INdustry-university Cooperation
Chosun University
Gwangju, Republic of Korea
jhshinkr@chosun.ac.kr

Eunji Lee
Dept. of Computer Engineering
Chosun University
Gwangju, Republic of Korea
eunbesu@gmail.com

Pankoo Kim
Dept. of Computer Engineering
Chosun University
Gwangju, Republic of Korea
pkkim@chosun.ac.kr

***Abstract*—Over the years, many researchers have been studied to detect expansions of acronyms in texts by using linguistic and syntactical approaches in order to overcome disambiguation problems. Acronym is an abbreviation formed which is composed of initial components of single or multiple words. These initial components bring huge mistakes when a machine conducts experiments to find meaning from given texts. Detecting expansions of acronyms is not a big issue now days. The problem is that a polysemous acronym. In order to solve this problem, this paper proposes a method to recommend the most related expansion of acronym through analyzing co-occurrence words by using Wikipedia. Our goal is not finding acronym definition or expansion but recommending the most appropriate expansion of given acronyms.**

Keywords—Acronyms; Wikipedia; Acronym expansion; Information extraction; Text mining

I. INTRODUCTION

Acronym is an abbreviation formed which is composed of initial components of multiple words to indicate words in a short format. The acronyms are similar to abbreviation but the number of original word has to be bigger than two. The abbreviation is a format to omit when the length of a word is normally too long such as "ad" for advertisement. In case of acronyms, it is a format to shorten multiple words. For instance, "NLP" indicates not only *Natural Language Processing* in computer science or artificial intelligence field but also *Natural Law Party* which was a transnational party founded on the principles of Transcendental Meditation. This polysemous acronym (an acronym has multiple expansions) brings a big obstacle when a machine tried to distinguish what "NLP" precisely means. Over the years, many researchers have been studied to detect what expansions of acronyms in given text data by using linguistic and syntactical approaches [1, 2, 3, 12] in order to enhance the performance of Information Retrieval, Information Processing, and more. These works focused on the issue to detect and extract the precise expansion of acronyms in given texts. Because of these researches, the precision rate for detecting expansions or definitions of acronyms reached to over 90%. However, it is still challenging matter to solve disambiguation words such as polysemous acronyms. In order to overcome Word Sense Disambiguation (WSD) problem, people have been applied Knowledge Base (KB) such as WordNet [4, 5] which was developed and maintained by the Cognitive Science Laboratory of Princeton University. Moreover, [6] proposed a method to apply context information in Wikipedia to enrich concept networks in WordNet. These related works proved that Wikipedia has valuable information to overcome WSD problems. However, the weakness of these works is the fact that they did not focused on the acronyms. They only aimed to enrich noun and verb types of word. Although the acronyms indicate and describe technical or specific terms, they were commonly considered as a noisy data.

In this paper, we are aimed to recommend the most appropriate expansion acronyms after building acronym data set consists of co-occurrence words by analyzing Wikipedia by using simple syntactic patterns.

The reminder of this paper is organized as follows: Section 2 describes the related works of this paper; Section 3 explains a method to detect acronyms in Wikipedia and a method to recommend the most related expansion acronyms for test sentences with experiments; finally, we conclude our research with future works in section 4.

II. RELATED WORKS

In this section, we explain why we used Wikipedia data to find a clue to recommend the most appropriate expansions of acronyms in given texts. Besides, we describe the related works such as acronym detection and expansion, WSD, document classification, and etc.

978-0-7695-4974-3/13 $26.00 © 2013 IEEE
DOI 10.1109/IMIS.2013.44

A. Wikipedia

Wikipedia[1] is an open and free internet encyclopedia which is collaboratively edited and maintained by experts around the world. This Collective Intelligence contains diverse kinds of information concerning our real world from physical entities to philosophical theories which consist of more than 3,770,000 articles in English. Wikipedia is written in almost perfect grammatical form due to collaboration of people. Moreover, this valuable data is freely provided by DBpedia [2] which contains diverse structured information about Wikipedia such as title, short abstract, extended abstract, infobox, categories, external links, disambiguation links, and more. Due to the strong strength of Wikipedia, many researchers have been applied Wikipedia data to overcome WSD problems or Knowledge Base Enrichment [7]. The human natural language is so complex for computer to understand, there are many challenging issue needed to be solved. For example, a noun "bat" has five kinds of senses defined in WordNet. These senses are chiropteran, baseball bat, racket, cricket bat, and club. In order to overcome this ambiguity problem, there were researches [8, 9] to find the most related annotations for given images based on WUP similarity in WordNet. However, these researches did not consider acronyms as an important data to find context words from given texts.

Over the years, people have been gave great efforts to discover expansions and definitions of acronyms based on linguistic and syntactic patterns. Therefore the performance of discovering definitions and expansions of acronyms reached higher than 90% currently. However, the problem is that they only focus on detecting acronyms not solving disambiguation problem. Acronym is a combination of initial components of words so it is likely to be the same as other even though the expansions are different from each other. The following table 1 shows the examples of polysemous acronyms.

TABLE I. EXAMPLES OF POLYSEMOUS ACRONYMS

Acronyms	Expansions
ABC	Australian Broadcasting Corporation
ABC	American Broadcasting Company
ABC	Associated British Corporation
ABC	Australian Broadcasting Commission
ABC	African-American Businesswomen CEOs
ABC	Alcoholic Beverage Control
ABC	Association for Business Communication
…	…

As described in table 1, the number of expansions of acronym "ABC" is 347 defined in World Wide Web Acronym and Abbreviation Server (WWWAAS)[3]. This is a huge obstacle when we conduct experiments to find concept networks or sense disambiguation in NLP area. These polysemous acronyms have to be distinguished by somehow. Therefore, we propose a method for finding the most adequate expansions of acronyms by using co-occurrence words in Wikipedia.

[1] http://www.wikipedia.org
[2] http://dbpedia.org
[3] http://www.acronymfinder.com

III. A METHOD FOR RECOMMENDING THE MOST APPROPRIATE EXPANSIONS OF ACRONYMS

In order to recommend the most appropriate expansions of acronyms, we firstly need to make acronym data set which contains acronyms, expansions, titles, and co-occurrence words by analyzing Wikipedia extended abstracts. The reason why we analyzed extended abstracts is the fact that the extended abstracts describe a core point of articles corresponding to given titles. In order to extract these kinds of information, we are simply based on a linguistic approach to discover acronyms and expansions from given extended abstracts in Wikipedia. The following figure 1 indicates not only process to extract acronyms and their expansions from Wikipedia but also the method to recommend the most adequate expansions by using co-occurrence words.

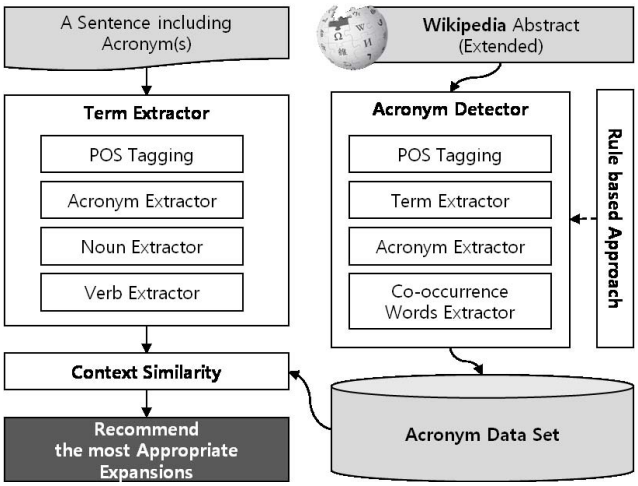

Fig. 1. The proposed system processes.

A. A Methodology for Extracting Acronyms and their Expansions from Wikipedia Extended Abstracts

In this paper, we applied a simple linguistic approach to detect acronyms and their expansion from Wikipedia extended abstracts due to the fact that the major goal of this paper is not extracting acronyms but recommending the most appropriate expansions for given test sentences. The following table 2 gives examples of extended abstracts in Wikipedia.

TABLE II. EXAMPLES OF EXTENDED ABSTRACTS IN WIKIPEDIA

Extended Abstracts in Wikipedia
<http://.../Autism> <http://.../abstract> "Autism is a disorder of neural development characterized by impaired social interaction … @en .
<http://.../Albedo> <http://.../abstract> "Albedo, or reflection coefficient, is the diffuse reflectivity or reflecting power of a surface. … @en .
<http://.../Anarchism> <http://.../abstract> "Anarchism is a political philosophy which considers the state undesirable, … @en .
<http://.../A> <http://.../abstract> "A is the first letter and a vowel in the basic modern Latin alphabet. It is similar to the Ancient Greek letter Alpha, from which it derives."@en .
<http://.../Achilles> <http://.../abstract> "In Greek mythology, Achilles was a Greek hero of the Trojan War, … @en .
…

As we can see in table 2, the extended abstracts contain *url*s to distinguish each articles and sentences to explain given titles in detail. After removing *url*s from the abstracts, acronyms will be extracted based on following linguistic rules.

- Acronyms in Wikipedia are likely to be located between parentheses and commonly represented by a capital letter. Therefore, if a term satisfies both conditions, the term will be considered as an acronym candidate at first. Afterward, we decide the term as an acronym when the term has more than two small letters.
- After the above step, we developed an algorithm to find expansion words for an acronym based on the rule that the first character of successive words has to be equal to the given acronym.
- We omitted words such as or, of, and, and more for discovering expansion words.

After applying simply linguistic rules, we can obtain 244,144 candidate acronyms from Wikipedia abstracts shown in table 3.

TABLE III. EXAMPLES OF EXTRACTED ACRONYM CANDIDATES

Acronyms	Expansions
PDDNOS	Pervasive Developmental Disorder-Not Otherwise Specified
CDC	Centers for Disease Control
BRDF	bidirectional reflectance distribution function
AMPAS	Academy of Motion Picture Arts and Sciences
IANA	Internet Assigned Numbers Authority
ASA	American Standards Association
AFC	American Football Conference
…	…

The weakness of this simple process is the fact that it is not able to detect expansions if the order of the capital letter of successive words was not the same as the given acronym. For instance, the expansion words for acronym "UTC" is Coordinated Universal Time. Moreover, it is hard to find expansions such as name of elemental symbols; Sodium (Na), Potassium (K), Rubidium (Rb), and more. We have to emphasize our goal again that the purpose of this paper is not finding expansions of acronyms but recommending the most semantically related expansion words for given acronyms.

We finally obtained 110,432 acronyms after removing duplicated ones which the expansions were the same as each other. As we described at the beginning of section 3, the acronym data set consists of acronyms, expansions, titles, and co-occurrence words. The co-occurrence words will be determined and calculated their Term Frequency (TF) values only if the word is not a stopword, number and special character. So, finally the acronym data set was extracted as described in following table 4.

TABLE IV. EXAMPLES OF EXTRACTED ACRONYM DATA SET

Acronyms	Expansions	Titles	Co-occurrence Words
PDDNOS	Pervasive Developmental …	Autism	Autism, children, autistic, diagnosed, disorder, rare, signs, …
CDC	Centers for Disease …	Autism	
BRDF	Bidirectional Reflectance …	Albedo	Albedo, surface, radiation, distribution, average, computer, …
AMPAS	Academy of Motion …	Academy Award	Academy, awards, ceremony, film, ampas, held, help, hollywood …
IANA	Internet Assigned …	ASCII	Ascii, characters, code, encoding, graphic, space, text, American, devices, printing, standard, …
ASA	American Standards …	ASCII	
…	…		

B. A Methodology for Recommending the Most Appropriate Expansion words

In order to find the most appropriate expansion words for given acronyms, we applied WUP similarity (Wu and Palmer, 1994) in WordNet. The WUP similarity measurement is a function to compute the path length from the least common subsume (LCS) of the two given concepts C_1 and C_2 followed by formula 1.

$$Sim_{wup} = \frac{2 \times depth(LCS(C_1, C_2))}{depth(C_1) + depth(C_2)} \quad (1)$$

where *depth*(*C*) is the depth of concept *C* in WordNet hierarchy so the value of this approach goes to high if two concepts shared an ancestor with long depth.

Let us assume that there is a sentence "The ABC is an American commercial broadcasting television network." The terms were not in stopword list will be considered as a context word. So the extracted context words will be "American," "commercial," "broadcasting," "television," and "network." An acronym "ABC" has 347 kinds of different expansions so we compared each of expansions of acronym "ABC" with given context words by using WUP similarity and find their average value as described in table 5.

TABLE V. EXPERIMENT RESULT TO FIND THE MOST APPROPRIATE EXPANSIONS FOR GIVEN ACRONYM "ABC"

"American," "commercial," "broadcasting," "television," "network"			
Acronym	**Expansions**	**WUP**	
		Sum	**Ave.**
ABC	Australian Broadcasting Corporation	1.490	0.298
ABC	**American Broadcasting Company**	**1.657**	**0.331**
ABC	African Bird Club	1.396	0.279
ABC	African American Business woman CEO	1.533	0.307
ABC	Asahi Basotho Convention	1.250	0.250
ABC	Achieve Baby Care	0.934	0.187
ABC	Activity Based Costing	0.961	0.192
ABC	Acorn Business Computer	1.313	0.263
ABC	Arab Banking Corporation	1.302	0.260
ABC	American Bowling Congress	1.508	0.302
ABC	Alcoholic Beverage Control	1.312	0.262
ABC	Australian Bird Count	1.391	0.278
ABC	Approximate Bayesian computation	0.472	0.094
ABC	Artificial Bee Colony	0.888	0.178
ABC	Associated British Corporation	1.145	0.229
…	…	…	…
"number," "commercial," "television," "companies," "established," "United," "Kingdom"			
Acronym	**Expansion**	**WUP**	
		Sum	**Ave.**
ABC	Australian Broadcasting Corporation	2.341	0.167
ABC	American Broadcasting Company	2.449	0.175
ABC	African Bird Club	2.052	0.147
ABC	African American Business woman CEO	2.045	0.146

"American," "commercial," "broadcasting," "television," "network"			
ABC	Asahi Basotho Convention	3.833	0.274
ABC	Achieve Baby Care	1.465	0.105
ABC	Activity Based Costing	1.836	0.131
ABC	Acorn Business Computer	2.103	0.150
ABC	Arab Banking Corporation	2.302	0.164
ABC	American Bowling Congress	2.442	0.174
ABC	Alcoholic Beverage Control	2.804	0.200
ABC	Australian Bird Count	2.522	0.180
ABC	Approximate Bayesian computation	0.905	0.065
ABC	Artificial Bee Colony	1.499	0.107
ABC	**Associated British Corporation**	**4.119**	**0.294**
…	…		…

As described in table 5, the most appropriate expansion words for acronym "ABS" corresponding to given context words is "American Broadcasting Company" by using WUP similarity measurement. The strength of our proposed method is the fact that it can distinguish if expansions are semantically differ from each other. However, our proposed method is weak in if expansions have the same terms such as "American Broadcasting Company" and "African American Business woman CEO."

IV. Conclusion and Future Works

This research deals with a basic approach for discovering acronyms in Wikipedia extended abstract and recommending the most relevant expansion words for given acronyms by using WUP similarity measurements. Many people have been studied to detect and discover an acronym ant its expansion words from text data based on linguistic patterns so its performance reached to higher than 90% currently. However, it is hard to solve acronym disambiguation problem. An acronym is a combination of initial components of words so there is a high possibility to have the same acronyms even though the expansion words are different. In order to distinguish which expansions are the most appropriate to given texts, we built acronym data set which contains acronyms, expansion words, titles and co-occurrence words from Wikipedia extended abstract. The performance for detecting acronyms in this paper is not superior to related works. However, we introduce a methodology to recommend the most appropriate expansion words. The goal of this paper is not detecting acronyms but giving an idea to prove that polysemous acronyms can be discriminated by their co-occurrence words semantically. Although we have applied simple syntactic patterns to detect acronyms and WUP measure, the expansion words for polysemous acronyms were successfully determined. However, there is a weakness to recommend the expansions if the expansions are similar to each others. Hence, we have a plan to amend our approach to overcome the weakness in the future work through applying a method proposed in research [6] and n-gram statistics which is the powerful method to reveal context information in human text data [10, 11]. Moreover, we have to conduct diverse experiments by using different similarity measures such as Latent semantic analysis (LSA), Normalized Google distance (NGD), Maximum Entropy method [12], and more to justify our proposed method.

Acknowledgment

This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation.

References

[1] M. Zahariev, "A Linguistic Approach to Extracting Acronym Expansions from Text," Journal of Knowledge and Information Systems, vol. 6, no. 3, pp. 366-373, May 2004.

[2] X. Ji, G. Xu, J. Bailey, and H. Li, "Mining, Ranking, and Using Acronym Patterns," Lecture Notes in Computer Science, vol. 4976, pp. 371-382, 2008.

[3] D. Sanchez and D. Isern, "Automatic extraction of acronym definitions from the Web," Journal of Applied Intelligence, vol 34, no. 2, pp. 377-327, April, 2011.

[4] H. Kong, M. Hwang, and P. Kim, "A new methodology for merging the heterogeneous domain ontologies based on the WordNet," International Conference on Next Generation Web Services Practices, pp. 22-26, August, 2005.

[5] M. Hwang, C. Choi, and P. Kim, "Automatic Enrichment of Semantic Relation Network and Its Application to Word Sense Disambiguation," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, pp. 845-858, June, 2011.

[6] M. Hwang, D. Choi, B. Ko, J. Choi, and P. Kim, "An Automatic Method for WordNet Concept Enrichment using Wikipedia Titles," Reliable and Autonomous Computational Science, pp. 347-365, 2010.

[7] M. Hwang, D. Choi, and P. Kim, "A Method for Knowledge Base Enrichment using Wikipedia Document Information," An International Interdisciplinary Journal, vol. 13, no. 5, pp. 1599-1612, September, 2010.

[8] D. Choi, J. Kim, H. Kim, M. Hwang, and P. Kim, "A Method for Enhancing Image Retrieval based on Annotation using Modified WUP Similarity in WordNet," Proc. of 11th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, pp. 83-87, Feb., 2012.

[9] D. Choi and P. Kim, "Automatic Image Annotation using Semantic Text Anlaysis," Lecture Notes in Computer Science, vol. 7465, pp. 479-487, 2012.

[10] D. Choi, B. Ko, E. Lee, M. Hwang, and P. Kim, "Automatic Evaluation of Document Classification using N-Gram Statistics," 15th International Conference on Network-Based Information Systems, pp. 739-742, Sept., 2012.

[11] D. Choi, M. Hwang, B. Ko, and P. Kim, "Solving English Questions through Applying Collective Intelligence," Communications in Computer and Information Science, vol. 184, pp. 37-46, 2011.

[12] S. Pakhomov, "Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts," Association for Computational Linguistics, pp. 160-167, 2002.