MAKARA, TEKNOLOGI, VOL. 16, NO. 2, NOVEMBER 2012: 180-184

KNOWLEDGE DICTIONARY FOR INFORMATION EXTRACTION ON THE ARABIC TEXT DATA

Wahyu Syaifullah Jauharis Saputra[*)], Agus Zainal Arifin, and Anny Yuniarti

Master Program Department of Informatics, Faculty of Information Technology, ITS Surabaya, Keputih Sukolilo Surabaya 60111, Indonesia

[*)]*E-mail: wahyu.s.j.saputra@gmail.com*

Abstract

Information extraction is an early stage of a process of textual data analysis. Information extraction is required to get information from textual data that can be used for process analysis, such as classification and categorization. A textual data is strongly influenced by the language. Arabic is gaining a significant attention in many studies because Arabic language is very different from others, and in contrast to other languages, tools and research on the Arabic language is still lacking. The information extracted using the knowledge dictionary is a concept of expression. A knowledge dictionary is usually constructed manually by an expert and this would take a long time and is specific to a problem only. This paper proposed a method for automatically building a knowledge dictionary. Dictionary knowledge is formed by classifying sentences having the same concept, assuming that they will have a high similarity value. The concept that has been extracted can be used as features for subsequent computational process such as classification or categorization. Dataset used in this paper was the Arabic text dataset. Extraction result was tested by using a decision tree classification engine and the highest precision value obtained was 71.0% while the highest recall value was 75.0%.

Abstrak

***Knowledge Dictionary* untuk Ekstraksi Informasi pada Data Teks Arab**. Ekstraksi informasi merupakan sebuah tahap awal dari proses analisis data tekstual. Ekstraksi informasi diperlukan untuk mendapatkan informasi dari data tekstual sehingga dapat digunakan untuk proses analisis seperti misalnya klasifikasi dan kategorisasi. Data tekstual sangat dipengaruhi oleh bahasa, jika sebuah data tekstual berbahasa Arab maka karakter yang digunakan adalah karakter arab. *Knowledge dictionary* merupakan sebuah kamus yang dapat digunakan untuk mengekstraksi informasi dari data tekstual. Informasi yang diekstraksi menggunakan *knowledge dictionary* adalah konsep. *Knowledge dictionary* biasanya dibangun secara manual oleh seorang pakar yang tentunya membutuhkan waktu yang lama dan spesifik untuk setiap masalah. Pada penelitian ini diusulkan sebuah metode untuk membangun *knowledge dictionary* secara otomatis. Pembentukan *knowledge dictionary* dilakukan dengan cara mengelompokkan kalimat yang memiliki konsep yang sama, dengan asumsi kalimat yang memiliki konsep yang sama akan memiliki nilai similaritas yang tinggi. Konsep yang telah diekstraksi dapat digunakan sebagai fitur untuk proses komputasi berikutnya misalnya klasifikasi ataupun kategorisasi. Dataset yang digunakan dalam penelitian ini adalah dataset teks Arab. Hasil ekstraksi diuji dengan menggunakan mesin klasifikasi *decision tree* dan didapatkan nilai presisi tertinggi 71,0% dan nilai recall tertinggi 75,0%.

Keywords: knowledge dictionary, information extraction, data text, Arabic text

1. Introduction

Information extraction from textual data is a problem that often becomes a major topic in many studies to find the right solution. By using a dictionary of synonyms and prediction rules, each document can be searched for term similarities, thus affecting the value of similarity between the documents, but is limited to documents that have a clear structure [1- 2]. Other studies have been done to extract information pattern from the sentences contained in textual data [3]. By using the term and the frequency of each term, the concept of a document cannot be extracted. Therefore, the recognition of the document is based solely on the term and frequency only, not on the concepts contained in the document. An information extraction method is proposed to use two dictionaries that can help the process of extracting information in the form of the concept of a textual data. The two dictionaries are: proper noun dictionary and knowledge dictionary. By using the knowledge

dictionary, different expressions but having the same meaning can be grouped into a key concept [4]. Thus the document can be identified or classified according to the key concepts that have been extracted from the sentence. The value of similarity of two sentences can then be calculated [5-6].

Knowledge dictionary continues to be developed and a study has proposed a method that automatically acquired the knowledge dictionary [7]. A method for acquiring knowledge in the form of dictionary was proposed using a fuzzy decision tree, and on this acquisition, the knowledge dictionary is created manually by an expert [8]. An application that implements the knowledge of the dictionary is made to classify e-mails on a company's customer center [9]. Knowledge dictionary formed by an expert would take a long time to be produced and depends on the ability of the expert himself/herself. The knowledge dictionary formed would be specific to a particular problem and should be made again for other problems [9]. Each text data is strongly influenced by the language, such as Arabic which is gaining a significant attention in many studies because it is very different from the others. Linguistic character, especially differences in dialect and complex morphology, is a challenge in the research field of (NLP) Natural Language Processing. In contrast to other languages, tools and research on the Arabic language is still lacking [10].

This paper proposed a method for automatically extracting information on the concept of Arabic text data. Extraction is done by clustering sentences that have the same concept, assuming that the sentences having the same concept will have a high similarity value. The results of the extraction process the information obtained will be tested using a decision classification engine. Furthermore, the accuracy will also be calculated using precision, recall and f-measure.

2. Methods

The proposed method of knowledge dictionary is the process of information extraction from textual data automatically by means of recognition of the concept of any expression contained in each document as shown in Figure 1. The concept will be formed into a knowledge dictionary that can be used to perform feature extraction from a document to the classification process. Knowledge creation process automation requires a word dictionary to recognize the main words and recognize the stop words. Expressions that have the same concept are grouped together and are labeled. Clustering of expression that has the same concept is based on the assumption that the expressions having the same concept will have a high similarity value. Then the expression can be grouped based on the value of the

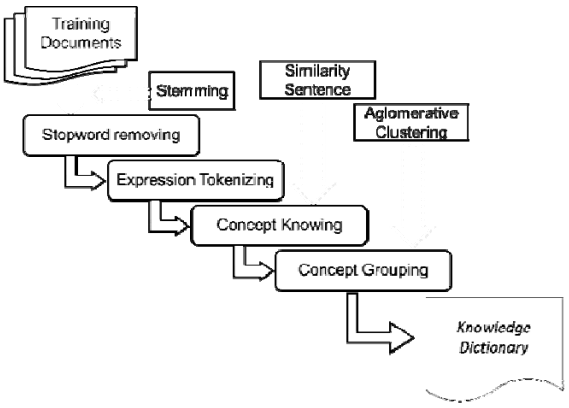

Figure 1. Flow of Knowledge Dictionary Development

expression of the other one. For example, there are two expressions as follows.

- نحن نبحث عنق افكار جديدة للبحث

  Translation: "We're looking for new ideas for research"
- فكرة جديدة من البحث لد ينا لأ يز ال في عملية البحث

  Translation: "a new idea of our research is still in the process of search"

Stopword removal process will eliminate the terms included in the stopwords, the terms which can be categorized as general terms and cannot be used to identify a document, because they exist in almost all documents. Results of stopword removal from two expression before develop the expression as follows.

- نحن بحث عن افكرا جديدة

  Translation: "We're looking for a new idea of research"
- البحثا بحا ثنا من الأفكرالجديدة

  Translation: "The new idea of research we search"

As seen on the first expression there are some words has gone. Stop word removed using stop word dictionary that build manually before. Stop word removal is a process that search each words in whole document and compare to each word in stop word dictionary, if result of comparison is same then the word will be removed.

Next process is expression tokenizing. There are one symbol is represent one single sentence or expression. That symbol is “.” (dot symbol). Expression tokenizing is process that separate document become list of expression. List of expression will be use in next process that compare one expression with another and from this process can get similarity value. Similarity will calculate using cosine similarity [11]. If one document contain two sentences it will create two sentences in list.

The next process is the process of grouping expressions with implementing Hierarcial Aglomerative Clustering. This process begins with the formation of the matrix, each row of the matrix is an expression and each column is a feature of the expression. column in the matrix is a list of words that come from the entire expression in the dataset. Value of the matrix is the weight of each word derived from the calculation of the frequency of the word in each expression is concerned, if the word does not exist then the weight will be 0 (zero). Of the matrix is then processed using a clustering algorithm that is Hierarcial Aglomerative Clustering. Clustering Algorithm of Hierarcial Aglomerative are as follows: 1) For every expression, 2) Calculate the distance of the entire expression in the dataset, 3) Return to step 1, 4) Relationship between the expression of matrix form, 5) For each element in the matrix, 6) Find highest value in the matrix, 7) Make these two expressions that have the highest value in a group, 8) Calculate the matrix again with a new group as a node, 9) Return to step 5.

Once the group is formed expressions, each expression in a group is said to have the same concept. To obtain the relationship between concepts and expressions, where each concept contain 1 (one) or more expressions. This process is the end of the training system. The output of the training process is a knowledge dictionary formed in the relationship between concepts and expressions.

For example, from the two expressions before, two nodes are obtained, where each node represents an expression. All expression nodes are then grouped using Hierarcial Aglomerative Clustering. In the clustering process the features used are terms that make up the expression, and the weights used in each feature is the frequency of the term in an expression. The process of grouping expressions generates a knowledge dictionary, and the form of the knowledge dictionary can be seen in Figure 2.

Knowledge dictionary is used to perform the extraction of information in the test phase. Some of the test process steps are similar to the training process, which is removing stopwords from the test data to be

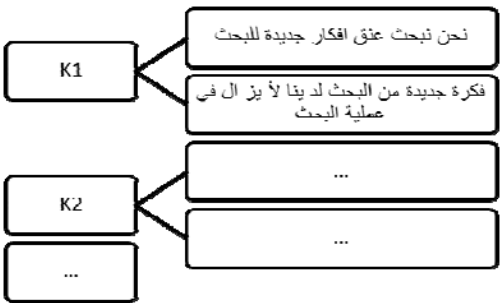

Figure 2. Knowledge Dictionary

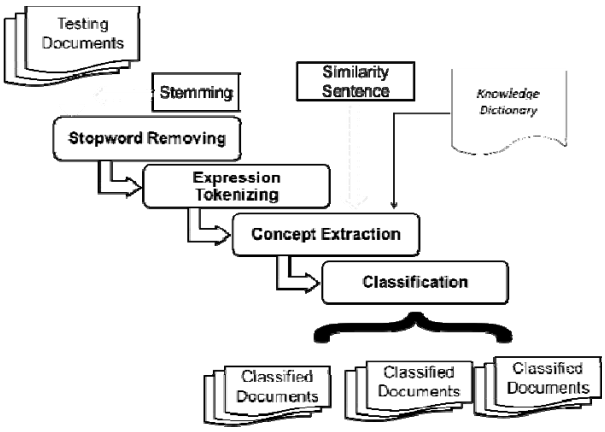

Figure 3. Flow of Knowledge Dictionary Testing

processed by running stopwords removal procedure. The process continues with expression tokenizing, which aims to obtain a list of expressions in a document. Having obtained a list of expression of each document, each expression will be grouped based on the knowledge dictionary that has been established. Grouping of expressions is performed by calculating the similarity of each expression in the test data to every other expression on the dictionary knowledge. Each expression in the test data will be grouped together with the expression of the knowledge dictionary that has the highest similarity value. Step-by-step testing process can be seen in Figure 3.

Concepts extracted from documents are used as features in the classification process, so that each concept has a weight value. The weight of each concept is calculated using formula adopted from the calculation of *tf-idf*. *Tf-idf* weighting calculations are based on term,whereas in this paper weighting calculation are based on concept. Calculcation of weight in this paper using the following formula.

$$Wc = fk.idfc, \quad (1)$$

where: *fk* is the frequency of a concept in a document and idfk is the size of the concept is said to be general or not to all documents and *idfk* can be calculated using the following formula.

$$idfc = \log \frac{|D|}{1+|\{d \in D : k \in d\}|}, \quad (2)$$

where: $|\{d \in D : c \in d\}|$ is the number of documents in which the concept appears, $|D|$ is the total sum of all documents. Letter *c* in each formula means “concept”.

3. Results and Discussion

The design of the input data format of the system is significant be cause it relates to how the implementation of the concept recognition method to form a knowledge

dictionary automatically. In this paper, the input dataset is a collection of magazines in Arabic text, such as Al-Jazirah (http://www.al-jazirah.com/). In this paper, the dataset is divided into two groups of data, namely the training data and test data. Training data is the dataset used for training the system to form a knowledge dictionary. The training data consists of 100 documents which are divided into 2 (two) categories: arts and economy. Each category in the training data contains 50 documents with a size between 9 KB (kilobytes) to 25 KB (kilobytes) with a variety number of expressions (sentences), as shown in Table 1. An example of arabic document is also shown in Figure 4.

Test data is the data dictionary used to test the knowledge generated from the training process. In this research, test data consists of 100 documents divided into 2 (two) categories: arts and economy, and it also applies to the training data. Each document in the test data is a different document from the training data, so that none of the documents in the training data are used as test data. In this paper, both the training data and test data are made to have the same number of categories and the same number of documents, aiming to test the process. At the first trial, the training data is used for training and test data is used for the test, in accordance with their respective functions. At the second trial, the test data will be used for training, and training data will be used for testing. Such trials are intended to determine the ability of the proposed method.

The average value of precision and recall of 7 (seven) tests using 2 (two) categories with 100 training documents and 100 test documents can be seen in Table 2. The number of concepts greatly affects the precision and recall values, as can be seen from Table 2. The art category has the highest precision with the concept number of 250, while the economy has the highest precision when the concept number is 300. The highest average precision value is reached when the concept number is 250. Recall value is the highest when the number of concept is 300 for the art category, and 250 for the economy category. The highest average value of recall is when there is 300 concepts.

بدأ فيأ حلم عرب ختم فيأ مهرج بحر حلم
بكر يواصل سيناريو للفنانين سعد كتب
عبدالرحمن دخيل تماماً كون متوقعاً ضبط
أتي مهرج بحر دول للأغنية صور هزيلاً
وفاشلاً فيأ كلاً شيء نظم سيىء جمل
جمل حدث دراماتيكية يسعني الا ان اقف
مبتسماً أسي سيناريو تراجيدي حدث
فصل منامة. كلاً هتم فني عمل فيأ سعد

Figure 4. Example of an Arabic Document

Table 1. Dataset

Category	Document Number	Total Expression
Training Data:		
Arts	50	1117
Economy	50	834
Test Data:		
Arts	50	667
Economy	50	408

The test is also performed using the general method, which is using the term as a result of information extraction. The term will be used as a feature for the next process, and each feature has a weight that is based on term frequency and inverse document frequency (IDF). The results of trials using the whole terms is 19,302 term results obtained as shown in Table 3.

Table 2 shows that the proposed method can achieve an average precision of 71.0% on the amount of 250 concepts and can achieve an average recall of 75.6% on

Table 2. Accuracy

Concept Count	Arts (%)	Economy (%)	Average (%)
Precision:			
200	38.0	88.0	63.0
250	60.0	83.0	71.0
300	36.0	96.0	66.0
350	42.0	86.0	64.0
400	62.0	36.0	49.0
450	62.0	48.0	55.0
500	44.0	66.0	55.0
Recall:			
200	76.0	58.7	67.0
250	76.2	67.2	72.0
300	90.0	60.0	75.0
350	75.0	59.8	67.4
400	49.2	48.6	59.9
450	54.3	55.8	55.1
500	56.4	54.1	55.2
F-Measure:			
200	50.6	70.4	60.5
250	67.4	73.9	70.6
300	51.4	73.8	62.6
350	53.8	70.5	62.2
400	54.8	41.4	48.2
450	57.9	51.6	54.8
500	49.4	59.4	54.4

Tabel 3. Result Test Using All Terms

Accuracy	Arts (%)	Economy (%)	Average (%)
Precision	81.6	98.0	89.0
Recall	97.6	84.5	91.0
F1	88.9	90.7	89.8

the 300 concepts. From the results of experiments with 7 (seven) variations in the number of concepts, the value of precision, recall, and f-measure is less than those obtained when using the entire term as shown in table 3. This occurs because of the variations in the content of the dataset, as well as variations in the amount of training data used to perform tests [9].

4. Conclusions

From the test results, it can be concluded that the concept of expressions can be recognized automatically and then calculated for their similarity. Similarity in this paper was calculated using the cosine similarity with the maximum value of 1 (one) and a minimum value of 0 (zero). The concept of knowledge is represented in the dictionary which is described in the relationship between concepts and expressions clustered within a concept. Extraction of information from documents using the knowledge dictionary can be achieved by identifying the entire expression in the document. The whole expressions of the documents were then compared to the entire expressions in the dictionary knowledge. Expression of the documents will be grouped together with the expression on the knowledge dictionary that has the most similarity. The result of information extraction using dictionary knowledge is strongly influenced by the number of concepts used as a cluster of extracted expression.

Information extraction process can be performed using the proposed method and obtain the highest precision classification results of 71.0% with 250 concepts and the highest recall of 75.0% on the 300 concepts. The highest F-Measure value obtained is 70.0% with 250 concepts, based on information extracted by the proposed method.

From the research that has been done on methods of knowledge dictionary, it can be seen that there is a point in this method that needs to be developed in future work. It is the determination of the number of concepts required, which is a formula to obtain the optimal number of concepts. This would determine the optimal number of concepts without using a trial and error procedure.

References

[1] R.J. Mooney, U.Y. Nahm, Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa, W. Daelemans, T. du Plessis, C. Snyman, L. Teck (Eds.), Van Schaik Pub., South Africa, 2005, p.141.

[2] N. Kanya, S. Geetha, IET-UK International Conference on Information and Communication Technology in Electrical Sciences (ICTES 2007), Dr. M.G.R. University, Chennai, Tamil Nadu, India, 2007, p.1111.

[3] S. Patwardhan, E. Rillof, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, 2007, p.717.

[4] Y. Ichimura, Y. Nakayama, M. Miyoshi, T. Akahane, T. Sekiguchi, Y. Fujiwara, Proceedings of the 14th Annual Conference of JSAI, Japan, 2000, p.532.

[5] J.-Z. Hu, T. Xu, J.-B. Shu, P. Lu, 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE), Chengdu, China, 2010, p.V4-344. Doi: 10.1109/ICACTE.2010.5579485.

[6] J. Zhang, Y. Sun, H. Wang, Y. He, J. Converg. Inf. Technolo. 6/2 (2011) 22.

[7] S. Sakurai, Y. Ichimura, A. Suyama, R. Orihara, IJCAI 2001 Workshop on Text Learning: Beyond Supervision, 2001, p.45.

[8] S. Sakurai, Y. Ichimura, A. Suyama, R. Orihara, ISMIS 2002, LNAI 2366, Springer-Verlag Berling Heidelberg 2002, p.103.

[9] S. Sakurai, A. Suyama, Apll. Soft Comput. 6 (2005) 62.

[10] D. Mona, H. Kadri, J. Daniel, Proceeding HLT-NAACL-Short '04 Proceedings of HLT-NAACL 2004: Short Papers, Stroudsburg, PA, USA, 2004, p.149.

[11] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison-Wesley, Boston, 2005, p.500.