Building a Multi-lingual Field Association Terms Dictionary

Mahmoud Rokaya † **and Abdallah Nahla** ††,

University of Taif, University of Taif, KSA

Summary
Field association terms (FATs) dictionary concept was studied and tested for English. All algorithms were developed considering the English rules. power link algorithm presented a new rules to improve the quality of field association terms (FATs) dictionary in English. This method is independent of the Language then it can be applied to any language. Arabic language has many differences from the English language so it needs special techniques for preprocessing before applying the power link algorithm. In this paper depending on available FATs dictionary in English a multilingual FATs dictionary in English and Arabic is presented. The results showed that the precession and recall for the Arabic dictionary were less than the English one. Recall and precision for the Arabic FATs dictionary were 87.7% and 83.6%. In our simulation the power link algorithm increased average precision for the Arabic FATs were by 18.6%. Also the complete dynamic system to build a multi lingual FATs dictionary was built and tested.

Key words:
FATs, Co-word Analysis, Power Link, Concentration Ratio, Precision.

1. Introduction

Field association terms (FATs) are the terms that indicate each subject matter category in the classification scheme [10]. Field means a basic and common knowledge that can be used in human communication [9] and [16].
A document field can be ranked as: a super-field, a sub-field or terminal field. FATs are grouped according to how well they indicate specific fields. FATs have different rank to associate with document fields, so five field precision levels can be used to classify FATs according to document fields [3].
Fuketa et al. (2000) presented an early field understanding method using FATs. Rules had been defined and an efficient method of determining compound FATs had been presented [7].
Elmarhomy et al. (2006) present two criteria: deleting unnecessary terms with low frequencies, and deleting unnecessary terms using category information [6].Uddin et al. (2007) present a method for improvement of building FATs dictionary using passage retrieval They proposed a new approach for extracting FATs using passage (portions of a document text) technique rather than extracting them from the whole documents [17].

*Faculty of computer & information systems

Manuscript received March 5, 2011
Manuscript revised March 20, 2011

All these previous methods depend on the absolute frequency of the individual terms or the position of the term inside the documents.
Using the absolute frequency means that the frequency of a term will reflect the length of the documents rather than the weights of terms. Also depending on the frequency only, as a tool to decide the fields of the terms, wastes valuable information that can be derived from the co-occurrence relations between terms. [14].`
Co-word analysis considers the dynamics of science as a result of actor strategies. Changes in the content of a subject area are the combined effect of a large number of individual strategies. This technique should allow us in principle to identity the actors and explain the global dynamic [5].
Based on co-word analysis, Rokaya & Atlam (2009) presented a modification of the rules used to determine the FAts level. The new rule depends on concept of links [15]. In this new rule, they presented the power link Algorithm to calculate the power of link between terms as well as power of a link between terms and fields.
The algorithm depends on building a base FATs dictionary, then the algorithm uses this base FATs dictionary to produce a modified FATs dictionary that has a better recall and precision. The system continues in a loop till a certain termination condition is met.
In this paper pre-prepared FATs dictionary in English will be translated into Arabic.
The intersection between the translated and an extracted list of Arabic candidates will be considered as the base dictionary to produce a dynamic FATs dictionary in Arabic.
This paper is organized as follows. Section 2 presents the basic definitions and concepts used in the FATs literature. Section 3 presents Arabic language and preprocessing methods, computing algorithm for producing a multi lingual Field Association terms dictionary is presented in section 4 . Section 5 presents the details of the experiments and system evaluation. Section 6 analysis the achieved results.

2. Field Association terms.

It is natural for people to identify the field of document when they notice specific words. These specific words are referred as Field Association Terms (FATs); specifically, they are words that allow us to recognize intuitively a field of text or field –coherent passage. Therefore, FATs can be used to identify the field of a passage, and can be also used to classify different fields among passages. For these reasons FATs can be used as a clue to identify a passage field. FATs can be either words or phrases [17]. This section presents the concepts and common notion that are used in the literature of FATs like "concentration value", "field", "subfield" etc.

Field association terms (FATs) are the words that indicate each subject matter category in the classification scheme [10].

A minimum term , or a word, is defined as one which cannot be further divided without losing its semantic meaning, as a single FAT. Compound FAT are defined to consist of two or more single FAT. Both terms are expressed by enclosing them within quotation mark. A compound FAT is regarded as being single if it loses its field information when divided. Compound FATs (e.g. حزب الله" Hizb Allah" or الدفء العالمي "global worming") are considered to be simple FATs because document field information is easily lost when those compound terms are divided. So, proper nouns (e.g. قناة الجزيرة "Aljazera TV", أوليمبياد القاهرة "Cairo Olympic" and ساحل العاج "Ivory Coast") are considered to be simple FATs. Personal names (e.g. أحمد زويل "Ahmed Zweel" and نجيب محفوظ "Nagib Mahfoz") are considered to be simple FATs, but proper noun containing a title (e.g. المدرب حسن شحاته "Coach Hassan Shehata") are divided into two single terms: ''المدرب'' (coach) and ''حسن شحاته'' (Hassan Shehata) which belong to the same document field Soccer, but are on different levels. FAT could be a word (e.g., "هدف" : goal), a phrase (e.g., "الحرب و السلام" war and peace) or right truncated word (e.g., "ظهير" Back), which represents that matches the letters to the FAT of the colon [2].

Field means a basic and common knowledge that can be used in human communication. And for convenience; hierarchical fields are categorized as Sub-Fields and Super-Fields [9] and [16]. Therefore, "ظهير" (back) can relate to sub-field Soccer of super-field SPORTS and "ظهير" (back) may be classified SPORTS/Soccer. Selecting useful FATs requires consideration of relationships between simple and compound FATs and field classification [16].

3. Arabic language

Arabic and other Semitic languages, in general, present a somewhat idiosyncratic morphology. What sets these languages aside is the predominantly nonlinear or non-concatenation character of their word structure. Unlike in English or French where they are appended one after the other, morphemes in Arabic are typically intertwined with each other in a way that blurs morpheme boundaries.

For example, while the English words writer and written are more or less easily analyzable into a discrete base, presumably the verb write, immediately followed by a suffix (-er and -en, respectively), their Arabic counterparts kaatib and maktuub resist such a linear analysis [8].

Many algorithms had been developed to solve the problem of finding all approximate pattern of a given string. Seller (1980) presented the approximate search algorithm which relies on dynamic programming [13].

Alnajem (2005) presented an algorithm that considers the Arabic syllabic and use it to handle variation in Arabic orthography. The main disadvantage of this algorithm is that the algorithm covers only a small part of the Arabic rules, namly the Arabic rules that covers verb-initial glottal stop and verb final glides [1].

In this paper an algorithm that relies on Soundex algorithm handle spelling errors in Arabic.

The algorithm suggested two coding systems of the words. The first coding depend on homo-phonology and the other coding system depends on the similarity in letters shape. This algorithm was developed to handle miss-spellings in the Arabic language, the practical results showed that this algorithms has high accuracy. This algorithm will be considered to filter both corps and FATs candidates from speling wrongs.

Filtering stop list words and stop list phrases: Stop-list terms are referred to as noise words. These words are those that do not contribute to the meaning of a sentence or a document, yet they help in forming a proper sentence [12]. These terms like ‘where’, ’when’, ‘numbers’, ‘it’, ‘ever’, etc. Stop list words are determined according to predefined lists.

Similarly to stop-list terms, stop-list phrases are sentences that occur within a document, yet they do not contribute to the meaning of the document. For example, a document may contain the phrase “Ladies and Gentlemen” , yet it tells nothing about Ladies or Gentlemen. Stop-list phrases are detected by comparing the first word of the phrase with a certain set of words that holds the starting words of commonly used stop-list phrases. If a matching word is detected, then the rest of the phrase is compared with a set of stop-list phrases that begin with the same word [11].

Parts of speech analysis: In this step we determine whether a given term is a noun or a verb, whether a noun is in its singular form or plural form, whether a verb is in its past, present or future tense, and whether pronouns are attached to a word. Furthermore we determine the category of the noun which it is a name of human, animal, location, plant or other things.

Mansour et al. (2008) presented a method to determine whether a given Arabic word is a noun or a verb. In this

paper this approach will be adopted to determine whether a given term is a noun or a verb since the approach is designed for Arabic language and it achieves a good precision and recall scores[11].

4. Algorithm for building Arabic FATs dictionary

Rokaya & Atlam presented a modification of the rules used to determine the FATs. These new rules depend on concept of links [15]. They presented a new rule to calculate the power of link between terms and new rules to measure the power of a link between terms and fields. Based on these new concepts, they presented a new algorithm to decide the fields of FATs. The algorithm depend on building a base FATs dictionary, then the algorithm uses this FATs dictionary to produce a modified FATs dictionary that has a better recall and precision. The system continues in a loop till a certain termination condition is met. The results that they got reflect a slight improvement in precision [14].
In what follows we will use this algorithm to build a FATs in Arabic.
To be able to deal with Arabic texts. Figure 1 illustrates the steps of this algorithm. We can summarize these steps as follows:

- An English FATs will be translated using a machine translation into Arabic.
- The resulting candidates will be filtered through the filtering steps (spelling wrongs filter, stop list words filter and parts of speech filtering)
- The FATs candidate will be extracted from the Arabic corps and both of the corps and the extracted candidates will go through the three filtering processes.
- The intersection between the translated candidates and the extracted FATs candidates will be calculated and it will represent the base FATs core of the FATs algorithm based on links. The remaining parts out of the intersection from the two sets, translated and extracted candidates will be the candidate inputs beside the Arabic corps to the power link algorithm.
- The power link algorithm will produce a refined FATs dictionary in Arabic.

5. Evaluation experiments

In this paper the scope of a field in [2], [3] and [4] is used.
In this experiments, the collection of the data sample relied on the data sample that was used in [14] will be used to represent the English corps. For the Arabic corps, A corps with size 40.03 megabyte and 1953 documents of data was extracted; about 90345 candidate terms were extracted. Data was collected using search engine to extract the candidate terms and the candidate documents. Only text from documents was retrieved. This corps was chosen to be distributed over 16 sub-fields and fields. A folder for each sub-field was established and the documents for this sub-field are stored in that folder.
The greatest number of terms came from the field <Sports>, 41397 terms, and the smallest number of terms came from the subfield <Skiing>, 9867 terms.

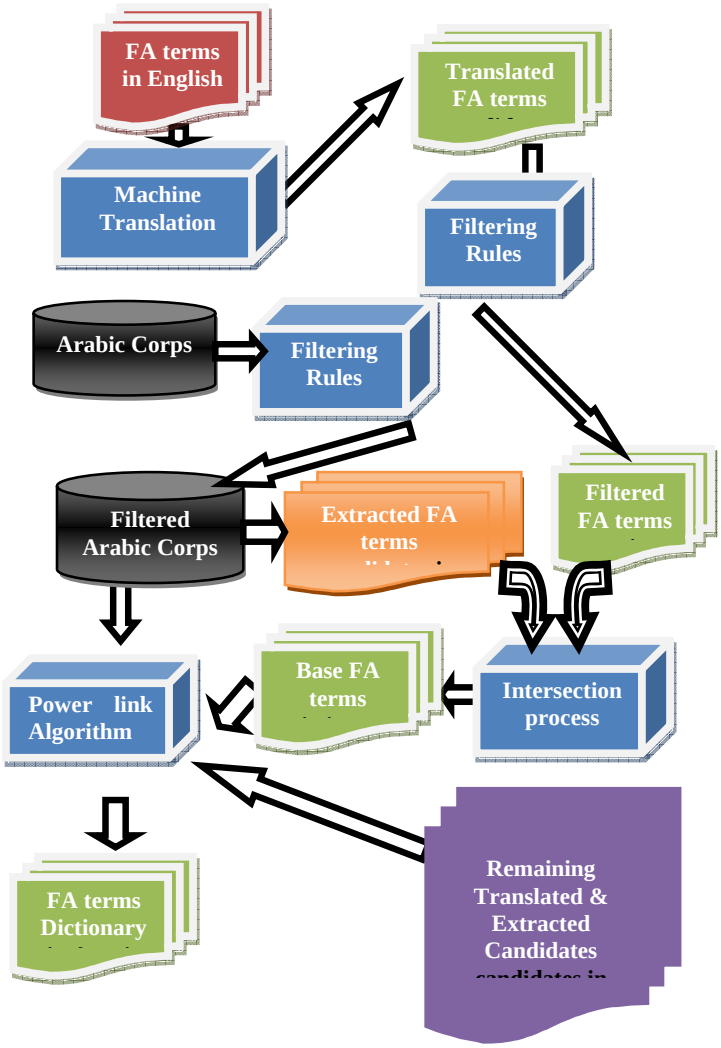

Fig. 1: Steps to get FA terms Dictionary in Arabic

In the experiments of this paper, the resulting FATs dictionary in English in the work of Rokaya and Atlam (2009) was used as the dictionary that was translated into Arabic .
For purpose of symmetry, data in the same 16 subfields collected under the condition that only Arabic documents will be retrieved. The steps of filtering spelling wrongs and stop-words cancelling as well as the mapping each

term to its category (verb or noun) were performed for both the extracted corps as well as the translated candidates.

The power link algorithm was applied to the refined data.

For example, among 18 random sample of terms the system failed to map 5 terms to their right fields. Namely, the system mapped the terms "الديمقراطية" (Democracy) , "ثنائية" (Double), "حارس مرمى" (Goal keeper), "إتحاد" (Association) and "كأس العالم" (World cup) to the subfields <Constitution>, <baseball>, <Soccer>, <Tennis> and <Soccer> respectively, whereas the right fields should be <Policy>, <Ball-games>, <Sports>, <Sports> and <Sports> respectively. We note that the system mapped the terms to subfields of the right field.

For example, the system mapped the terms "الديمقراطية" to the subfield <Constitution> instead of its super-field <Policy>. This means that, in this random example, the system has efficiency 100% to map the terms while the efficiency decreased when the system tried to map the terms to their correct sub-fields. We conclude from that our system has a high efficiency to map the terms till level 3. But in levels 1 and 2 the system will begin to produce some wrongs[14].

Rokay & Atlam (2009) implemented the power link algorithm as follows: experiments were run using values for ε_1 and ε_2 equal 0.001 [14]. Atlam et al. (2003) showed that the best threshold, for of concentration values, is 0.9 [3]. So they used this threshold as a fixed threshold for the concentration values in all loops. In the first trial of the system they noticed that the system stopped after the first loop, when they checked the value of R_1, they found that $R_1 < R_0$. This means that the termination condition was hold, but this gives no sense so they cancelled the termination condition related to the recall value $R_i - R_i < \varepsilon_2$ and kept the termination condition related to the precision value $P_i - P_{i-1}$ [14].

In the experiments of this paper the values for $\varepsilon_1$and ε_2 (0.001) and the concentration value ,(0.9) will be used

6. Results

Figure 2 shows number of extracted FATs and number of relevant FATs in each loop. The system was stopped after the ninth loop. In Figure 2, the number of extracted FATs decreases with increase of the number of loops and the number of extracted FATs which are relevant FATs decreases also. But the rate of decreasing the number of extracted FATs which are relevant is slower than the rate of decreasing the number of extracted FATs.

P values increase with increase in number of loops. P at first loop is $P_0 = 0.65$, P_0 represents the precision of the base FATs Arabic dictionary after applying a pure version of [3] to decide the precision levels for the FATs Arabic dictionary that, the candidates here represents the union of the translated candidates from the English corps and the extracted candidates from the Arabic corps. The resulting precision value, 0.65, seems to be low. This may explained as a result of the used tools for the filtering processes. This shows that these methods may need to be extensively reviwed.

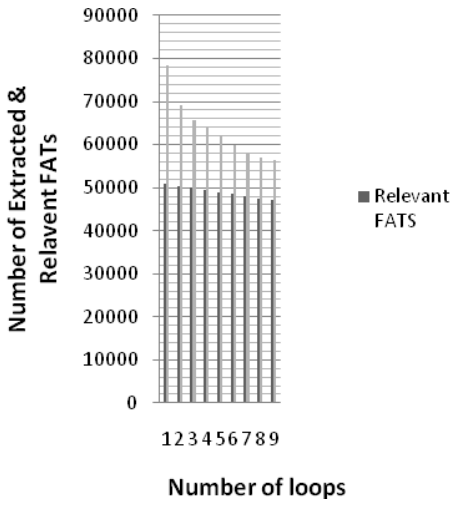

Fig. 2 Relevant and extracted FATs with increasing number of loops for FATs dictionary in Arabic

This leads to the fact that the precision increases and the recall decreases with increase of the number of loops. This result agrees what Rokaya & Atlam (2009) got in their work with the difference that the number of loops is increased [14].

Despite the results of Rokaya and Atlam (2009) we notice that. In Figure 3, increasing the number of loops causes a significant change in R but the decrease in R values is slower than the increasing in P values.

As shown before the base FATs dictionary was established to be the intersection between the translated FATs candidates and the extracted FATs candidates from the Arabic corps. We notice that a significant increasing in the precision value after the first loop, the precision increased by 8%. During the successive loops the rate of increasing in precision values is decreased but it comes to another relatively significant increasing in 5th and 6th loops. Then the rate of change in P becomes slower till we reach the ninth loop where the termination condition was hold ($P_9 - P_8 = 0.00015 < 0.001$).

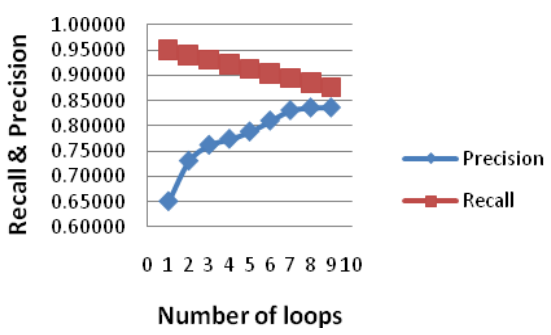

Fig. 3 Recall and Precision with increasing number of loops for FATs dictionary in English

7. Conclousion

In this paper we presented a method to produce a multilingual FATs dictionary. We built FATs dictionary in Arabic based on a new algorithm. This algorithm is independent of the language except in the preprocessing and filtering steps which are inherently depended on the languages. We generalized application of the power link algorithm to be applied to other languages. The results of recall and precision showed that the system was able to improve the precision values. The precision achieved using the traditional algorithm was 0.65, after the 9th loop the precision becomes 0.84. Hence the system succeeded to improve the precision values by 18.6%. 2. Tables, Figures and Equations

In the case that you would like to paragraph your manuscript, please make use of the specified style "paragraph" from the drop-down menu of style categories

Acknowledgments

The authors gratefully acknowledge the Taif University for the financial support (project No: 1-431-788)

References

[1] Alnajem, S, A, Computational Approach to the Variations in Arabic Verbal Orthography,Computer Speech and Language, Vol. 19, 2005, pp. 275-299

[2] Atlam E., Morita, K., Fuketa, M., & Aoe, J., A new method for selecting English field association terms of compound words and its knowledge representation. Information Processing and Management, Vol. 38, 2002, pp. 807-821.

[3] Atlam E., Morita, K., Fuketa, M., & Aoe, J., Documents similarity measurement using field association terms. Information Processing and Management, Vol. 39, 2003, pp. 809-824.

[4] Atlam, E., Elmarhomy G., Morita K., Fuketa, M., Aoe J., Automatic Building of New Field Association Word Candidates Using Search Engine. Information Processing & Management Journal, Vol. 42(4), 2006, pp. 951-962.

[5] Callon, M., Courtid J., & Ladle, F., Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. Science metrics, Vol. 22(1), 1991, pp. 155-205.

[6] Elmarhomy, G., Atlam, E., Fuketa, M., Morita K., Sumitomo, T., Aoe J., Automatic Deletion of Unnecessary Field Association Word Using Morphological Analysis. Journal of Computer and Mathematics, Vol. 83(3), 2006, pp. 247-262.

[7] Fuketa, M., Lee, S., Tsuji, T., Okada, M., & Aoe, J., A document classification method by using field association words. Information Science, Vol. 126, 2000, pp. 57-70

[8] Idrissi, A., Kehayia, E., Morphological Units in Arabic Mental Lexicon: Evidence from an individual with deep Dyslexia, Brain and Language, Vol. 90, 2004, pp. 183-197

[9] Kawabe, K., & Matsumoto, Y., Acquisition of normal lexical knowledge based on basic level category. Information Processing Society of Japan, SIG note, Vol. 125(9), 1998, pp. 87–92

[10] Lee, S., Shishibori, M., Sumitomo, S., Aoe, J., Extraction of field-coherent passages. Information Processing and Management, Vol. 38, 2002. pp. 173-207.

[11] Mansour, N., Haraty, A., Daher, W., Houri, M., An Auto-indexing for Arabic Text, Information Processing & Management Journal, Vol. 44(4), 2008 pp. 1538-1545

[12] McNamee, P., Mayfield, J., Indexing Using Both N-grams and Words. In Proceeding of 7th Text Retrieval conference, 1998, pp. 419-424

[13] Navarro, G., Raffinot, M., Flexible Pattern Matching in Strings: Practical on line search Algorithms for Text and Biological Sequences. Cambridge University Press, Cambridge, 2002.

[14] Rokaya, M., Atlam, E., Building of Field Association Terms Based on Links, Journal of Computer Applications in Technology (IJCAT) Special Issue on: "Intelligent Text Processing with its Applications and Computational Linguistics", Vol. 38, No. 4, 2010, pp. 298-305.

[15] Rokaya, M., Atlam, E., Fuketa, M., Dorji, C., & Aoe, J.. Ranking of field association terms using Co-word analysis, Information Processing & Management,. Vol. 44 No. 2, 2008, pp. 738-755.

[16] Tsuji, T., Nigazawa, H., Okada, M., & Aoe, J., Early field recognition using field association words. In Proceedings of 18th international conference on computer processing of oriental languages, ICCPOL'99, 1999, pp. 301–304.

[17] Uddin, S. Elmarhomy G., Atlam, E., Fuketa, M., Morita K., Aoe J., Improvement of Automatic Building Field Association Term Dictionary Using Passage Retrieval. Information Processing & Management Journal, Vol. 43, 2007, pp. 1793–1807

Mahmoud Rokaya received B.Sc. and M. Sc. Degrees in Mathematics from, Faculty of Science, Tanta University, Egypt, in 1997 and 2003 respectively, and the Ph.D. degree in information science and Intelligent systems from University of Tokushima, Japan, in 2009. He is currently an assistant professor in Dept. of Statistical and Computer science, Tanta University, Egypt. Mahmoud is a member in Egyptian Mathematical Association (EMA). His research interests include information retrieval, natural language processing and document processing.

Abdallah Nahla received B.Sc. and M. Sc. Degrees in Mathematics from, Faculty of Science, Tanta University, Egypt, in 1997 and 2003 respectively, and the Ph.D. degree in nuclear reactors from University of Tanta, Egypt, in 2006. He is currently an assistant professor in Dept. of Mathematics, Taif University, Saudi. Abdallah is a member in Egyptian Mathematical Association (EMA). His research interests include partial differential equations, nuclear reactors and information systems.