Automatic Building an Extensive Arabic FA Terms Dictionary

El-Sayed Atlam, Masao Fuketa, Kazuhiro Morita, and Jun-ichi Aoe

Abstract—Field Association (FA) terms are a limited set of discriminating terms that give us the knowledge to identify document fields which are effective in document classification, similar file retrieval and passage retrieval. But the problem lies in the lack of an effective method to extract automatically relevant Arabic FA Terms to build a comprehensive dictionary. Moreover, all previous studies are based on FA terms in English and Japanese, and the extension of FA terms to other language such Arabic could be definitely strengthen further researches. This paper presents a new method to extract, Arabic FA Terms from domain-specific corpora using part-of-speech (POS) pattern rules and corpora comparison. Experimental evaluation is carried out for 14 different fields using 251 MB of domain-specific corpora obtained from Arabic Wikipedia dumps and Alhyah news selected average of 2,825 FA Terms (single and compound) per field. From the experimental results, recall and precision are 84% and 79% respectively. Therefore, this method selects higher number of relevant Arabic FA Terms at high precision and recall.

Keywords**—**Arabic Field Association Terms, information extraction, document classification, information retrieval.

I. Introduction

IN recent years, the amount of data of all kinds available electronically has increased dramatically. However, collection often remained a challenge to retrieve and process into useful information and into actionable knowledge due to a lack of infrastructure for data**.** Novel techniques based on Field Association (FA) Terms [2][9][19][26] which is more suitable than the traditional feature selection methods such as vector space models [16][18] and probabilistic methods [10] have been introduced to be very effective in document classification.

The concept of Field Association (FA) Terms is based on the fact that the subject of a text (document field) can usually be identified by looking at certain specific words or phrases in that text. People can recognize the field of a document when they extract these specific words called FA Terms [1][9]. For example, “homerun” indicates the subfield <Baseball> of super-field <Sports>, and “election” indicates sub-field <Election> of super-field <Politics>.

Moreover, Atlam, Morita, Fuketa and Aoe [3] have proposed a method to select compound FA Terms from a pool of single FA Terms only. This method has a drawback because it does not extract compound FA Terms directly from a document even compound FA Terms form a majority of the relevant FA Terms in a given field. Other main problem lies in the selection of FA Terms do not use POS information and rely too heavily on the term frequency. Dorji, Atlam, Yata, Morita, Fuketa and Aoe [6] have proposed a method for selection of FA Terms using POS pattern rules for automatically building FA Terms dictionary.

Presently, all the previous studies are based on FA terms in English and Japanese, and the extension of FA terms to other language such Arabic could be definitely strengthen further researches. Therefore, this paper presents a new methodology for building extensive Arabic dictionary uses linguistic methods to extract relevant compound as well as single FA Terms from domain-specific corpora using Arabic POS.

Section 2 in this paper presents the overview and FA tree. Section 3 presents the related works. Section 4 presents our new methodology. Section 5 presents the experimental evaluation. Finally, Section 6 presents the conclusion and future work.

II. OVERVIEW

A. FA Term

FA Term: FA Terms are words or phrases that allow humans to recognize intuitively the field to which a text belongs. Technically, a FA Term is defined as the minimum word or phrase that can identify a field in a document field representation scheme called field tree.

For example, “tournament” is a proper FA Term of super-field <Sports> and “tennis tournament” is a proper FA Term of subfield <Tennis> under super-field <Sports>. In this case, the addition of the word “tennis” has added new field information to the word “tournament”.

B. Field Tree

In this study, we use a field tree based on Imidas’99[7] containing 14 super-fields, 50 median fields and 393 terminal fields (sub-fields). For example, in Figure 1, the path <Sports/Soccer/Egyptian Goalkeeper> describes super-field <Sports> having median-field <Soccer> and terminal field < Egyptian Goalkeeper> and this path can be represented by field code 13.1.5.

Each FA Term is connected to a particular field inside a hierarchical field tree like the one shown in Fig. 1. Since a

E. Atlam, M. Fuketa, K. Morita and J. Aoe are with the Department of information science and Intelligent systems, the University of Tokushima, Japan, phone: 0081-886-24-4599; fax: 0081-886-24-4599; e-mail: atlam@ is.tokushima-u.ac.jp).

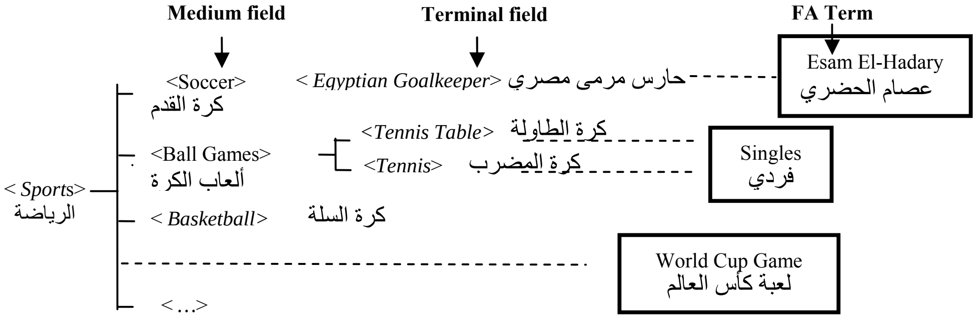

Fig. 1 sample of field tree structure

FA Term may belong to more than one field, it is possible that the same FA Term may be connected to the field tree at more than one node. In the FA Terms dictionary, whether a FA Terms belongs to more than one field or not is represented by its level.

C. FA Term Levels

Some FA Terms can uniquely identify a certain field, while some FA terms may belong to two or more fields. Thus each FA Term has a different scope to associate with a field. In order to take this into consideration, FA Terms are classified into five different levels [9][3] based on how well they indicate specific fields. The FA Term levels are defined as follows:

Level 1 Perfect FA terms: Associate with only one terminal field. For example, (Esam El-Hadary) is associated with one terminal field <*Soccer*>.

Level 2 Semi-perfect FA terms: Associate with more than one terminal field. For example, (singles) is associated with more than one terminal fields, <*Tennis*> and < *Table Tennis* > in the same medium field <*Ball Games*>.

Level 3 Ordinary FA terms: Associate with one top field. For example, (World cup) is associated with one top field <*Sports*>.

Level 4 Cross fields FA terms: Associate with more than one fields in different top fields. For example, (victory) is associated with the top field <*Sports*> and the terminal field <*Humanities and Social Sciences /Government> and <Politics/Election*>.

Level 5 Non-specific FA terms: Do not specify any fields. For example, (size) and (method) are some conjunctions and nouns lack of distinct fields.

III. Related Research Study

A. Important Aspects

The extraction of domain terminology from textual data is an important step for creating a specialized dictionary of terminologies [27]. Krauthammer and Nenadic [13] provide an overview of a number of approaches used for term identification. Co-occurring words are usually extracted as compound term candidates. Approaches used to identify co-occurrences consist of dictionary-based, syntactic rule-based using POS Patterns and machine-learning. For instance[4] has developed syntactic pattern rules for extracting noun phrases while [21] use syntactic patterns as well as bigrams to extract terminology candidates from log files.

Statistical methods are generally used with syntactic methods for evaluating the adequacy of terminological candidates. Under the framework established by traditional terminology extraction methods, we use specially developed POS patterns to extract FA Term candidates from domain-specific corpora using a sliding window of ten words. Relevant FA Terms are then selected by corpora comparison and using a unique series of statistical formulae based on *tf-idf*.

The selected FA terms are then added to FA Terms dictionary under their relevant fields in the field tree. Currently we use a hierarchical field tree based on Imidas'99 [7]. This field tree is a classification system similar to those used in knowledge organization systems.

B. Arabic Part of Speech (APOS)

Much work has been done on addressing different specific natural language processing tasks for Arabic, such as tokenization, diacritization, morphological disambiguation, part-of-speech (POS) tagging, stemming and lemmatization. (The papers cited below contain a discussion of relevant work.) The MADA system along with TOKAN provides one solution to all of these different problems. Our approach distinguishes between the problems of morphological analysis (*what are the different readings of a word out-of-context*) and morphological disambiguation (*what is the correct reading in a specific context*). Once a morphological analysis is chosen in context, we can determine its full POS tag, lemma and diacritization. Morphological analysis and disambiguation are handled in the MADA component of our system. Knowing the morphological analysis also allows us to tokenize and stem deterministically. Since there are many different ways to tokenize Arabic (tokenization is a convention adopted by researchers), the TOKAN component allows the user to specify any tokenization scheme that can be generated from disambiguated analyses.

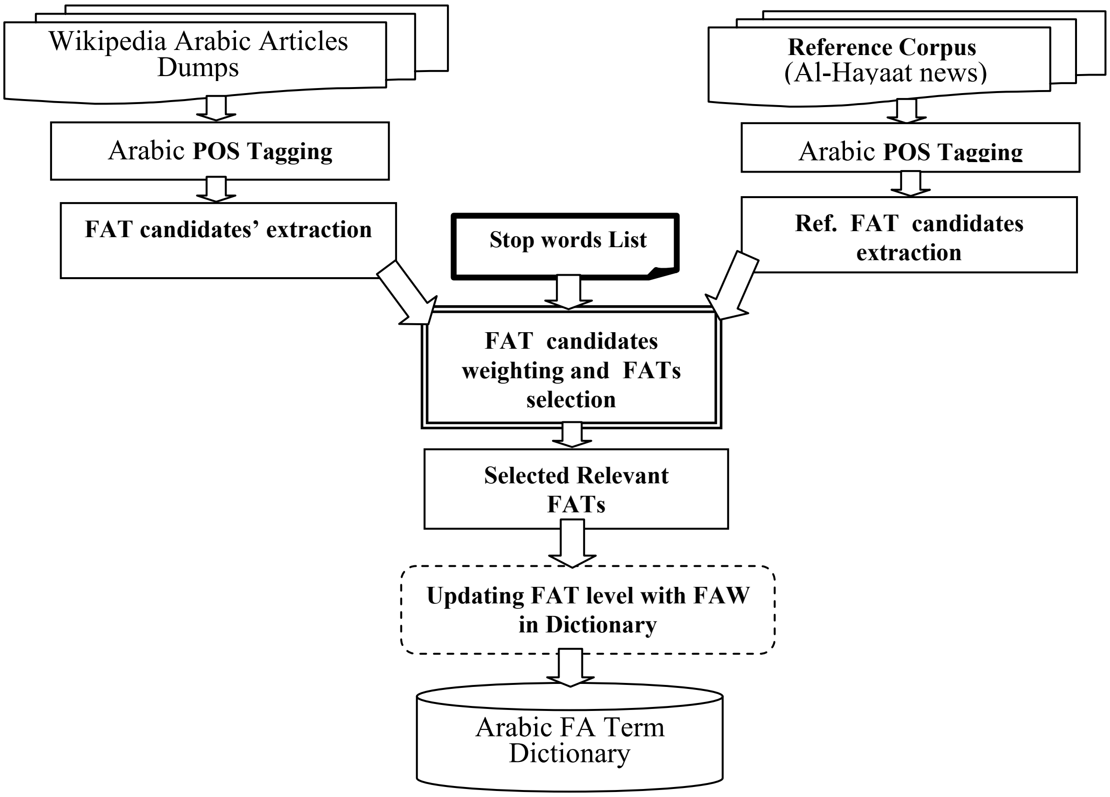
Wikipedia Arabic Articles Dumps
Arabic POS Tagging
FAT candidates' extraction
Reference Corpus (Al-Hayaat news)
Arabic POS Tagging
Ref. FAT candidates extraction
Stop words List
FAT candidates weighting and FATs selection
Selected Relevant FATs
Updating FAT level with FAW in Dictionary
Arabic FA Term Dictionary

Fig. 2 System outline of the proposed FA Terms selection methodology

TABLE III SAMPLE ARABIC OUTPUT OF TREE TAGGER

Token	POS	English	Token	POS	English	Token	POS	English
طورت	VVN	developed	في	IN	at	مرئية	JJ	visual
انتل	NP	Intel	انتل	NP	Intel	جديدة	JJ	new
معالجا	NN	processor	الشرق	NP	East	،	،	،
من	IN	of	الأوسط	NP	Middle	مع	IN	with
أسرة	NN	family	:	:	:	النظم	NNS	systems
بينتيوم	NP	Pentium	ستتاح	JJ	available	مجهزة	VVN	equipped
بسرعة	NNS	Speeds	للمستخدمين	NNS	Users	بمعالجة	NNS	processor
333	CD	333	في	IN	in	انتل	NP	Intel
ميغاهرتز	NN	MHz	قطاع	NN	sector	الجديدة	JJ	new
و	CC	and	الأعمال	NN	business	،	،	،
قال	VVD	said	و	CC	and	بينتيوم	NP	Pentium
نديم	NP	Nadem	المستخدمين	NNS	Users	الثاني	NP	II
جارودي	NP	Jaroudi	العاديين	JJ	casual	.	.	.
مدير	NN	manger	،	،	،			
تطوير	NN	development	قوة	NN	power			
الأعمال	NN	business	حوسبة	NN	computing			

The tokenized version is produced using the ARAGEN generator [11]. MADA (Morphological Analysis and Disambiguation for Arabic) makes use of 19 orthogonal features to select, for each word, a proper analysis from a list of potential analyses provided by the Buckwalter Arabic Morphological Analyzer [5]. The BAMA analysis which matches the most of the predicted features wins. These 19 features consist of the 14 morphological features, e.g. number, gender, case, mood, etc., which MADA predicts using 14 distinct Support Vector Machines trained on the PATB. In addition, MADA uses five features capturing spelling variations and n-grams statistics among others. Since MADA selects a complete analysis from BAMA, all decisions regarding morphological ambiguity, lexical ambiguity, tokenization, diacritization and POS tagging in any possible POS tagset are made in one fell swoop [11].

IV. The new methodology

A. System outline

The outline of the new system is shown in Figure 2. We use as inputs domain-specific corpora for the various fields of interest from Wikipedia Arabic Articles Dumps and reference corpora from Al-Hayaat news for comparison. The system consists of a part-of-speech (POS) tagger, a module for FA Terms candidate extraction, a module for weighting candidate terms and selecting the relevant FA Terms, and lastly a appending them to the FA Terms dictionary.

Firstly, documents in a domain-specific corpus are POS tagged using Tree Tagger [5]. The tagged corpus is then fed as input to the FA Term candidate extractor module. This module extracts FA Term candidates that match predefined POS pattern rules. The extracted FA term candidates are then weighted and ranked by comparing with term candidates from a reference corpus and using formulae based on *tf-idf* and at same time stop words list to filter out term candidates is used during the selection. Candidate terms that have final weights are automatically selected as new FA Terms. The selected FA Terms are then manually checked to confirm their relevance. Finally, the selected FA Terms are compared with all other FA Terms in the dictionary. Then the selected FA Terms are appended to the FA Terms dictionary under their relevant fields. All these procedures involved in the extraction and selection of FA Terms are described in detail in the following sub-sections.

B. Arabic POS tagging

The documents in the domain-specific corpora and the reference corpora are POS-tagged using Tree Tagger Arabic-tagger trained on the train part of the ATB as split for the 2005 JHU Summer Workshop, using Bies tags (data distributed by Mona Diab). Tree tagger is reported to achieve accuracy of 96.72% on dev portion according to Diab split (77.49% on unknown words)[5][11].

Example 1

The results of tagging the following sentence in Table 2 using the Tree Tagger are shown in Table 3. The sentence was taken from a document in the domain-specific corpus of the 'Operating system' field.

C.FA Term Candidates Extraction

1. Single FA Term candidates extraction

Single words like common nouns, proper nouns, adjectives or gerunds are extracted as candidates for Arabic single FA Terms. The words that belong to these parts-of-speech are the most likely candidates for single FA Terms

.Example 2

Let us consider extracting single FA Term candidates from the sentence given in Example 1. All words, the POS of which have been identified as nouns, adjectives or gerunds in Table 2 would be extracted as candidates. That would include the following words:

انتل, معالجاً, أسرة, بينتيوم, بسرعة, ميغاهرتز, نديم, جارودي, مدير, تطوير,
الأعمال, انتل, الشرق, الأوسط, قطاع, الأعمال, المستخدمين, العاديين، قوة,
حوسبة, مرئية, جديدة, النظم, المجهزة, معالج, انتل, بينتيوم, الثاني.

Table II Sample of Arabic Text and English translation from "Operating systems" filed

English translated text	Arabic Text
Intel has developed a family of Pentium processors at speeds of 333 MHz, "said Nadem Jaroudi business development manager at Intel Middle East said: be made available to business users and casual users, the power of new visual computing, dealing with systems equipped with Intel's new Pentium II .	طورت انتل معالجاً من أسرة بينتيوم بسرعة 333 ميغاهرتز وقال نديم جارودي مدير تطوير الأعمال في انتل الشرق الأوسط: ستتاح للمستخدمين في قطاع الأعمال والمستخدمين العاديين، قوة حوسبة مرئية جديدة، مع النظم المجهزة بمعالجة انتل الجديدة، بينتيوم الثاني.

2. Compound FA Term candidate's extraction

Arabic Compound FA Terms are formed by collocations. (Smadja, 1993) identified three types of collocations: rigid noun phrases, predicative relations and phrasal templates. Compound FA Terms consist of an uninterrupted sequence of words such as "Human rights-حقوق الإنسان", "President Hosni Mubarak- الرئيس حسني مبارك", "The Council of Ministers- مجلس الوزراء", "Islamic party of Justice-حزب العدالة الإسلامي", "Doctor of Philosophy-دكتوراة في الفلسفة", etc. and fall under the category of "rigid noun phrases".

Bennet and Schatz [4] have developed syntactic rules for extracting noun phrases in general, but they can not be applied directly for our purpose as we are interested only in some special noun phrases that are candidates for compound FA Terms. All noun phrases cannot be candidates for FA Terms. Based on previous studies [4] and on our own study of FA Terms, we developed the following sequence of POS patterns for a maximum length of ten words and minimum of two words, as rules for determining compound FA Term candidates:

1. [Noun] – [Noun] – [up to 8more nouns]
2. [Noun] – [Preposition] – [Noun] – [up to 7 more nouns]
3. [Noun] – [Preposition] – [Article] – [Noun] – [up to 6 more nouns]
4. [Adjective] – [Noun/Gerund] – [up to 8 more nouns]
5. [Adjective] – [Adjective] – [Noun] – [up to 7 more nouns]
6. [Gerund] – [Noun] – [up to 8 more nouns]

These rules are applied to the tagged documents from the corpora using a sliding window of ten words. The window is placed on the words such that the word at the beginning of the window is a noun, adjective or a gerund as per the POS pattern for a compound FA Term candidate identified above. The

Arabic POS pattern rule is then applied to the window contents. The window will be truncated when a word that does not conform to the identified Arabic POS pattern is encountered or a punctuation mark other than the hyphen is encountered. Whether a candidate term is located or not, the window then slides over to the next word that matches the starting Arabic POS for a FA Term candidate following the word where the previous window was truncated. If the previous window was not truncated, the window moves to the word that matches the starting Arabic POS for a FA Term candidate next to where the previous window ended. The process is repeated until the end of the file is reached.

Example 3

Let us look at the extraction of compound FA Term candidates from the sentence given in Example 1. The window is truncated after identifying a possible FA Term candidate. Firstly, the underlined phrases are extracted as compound FA Term candidates.

طورت انتل معالجاً من أسرة بينتيوم بسرعة 333 ميغاهرتز وقال نديم
جارودي مدير تطوير الأعمال فى انتل الشرق الأوسط: ستتاح للمستخدمين
في قطاع الأعمال والمستخدمين العاديين، قوة حوسبة مرئية جديدة، مع
النظم المجهزة بمعالجة انتل الجديدة، بينتيوم الثانى.

Secondly, some of the FA Term candidates can furnish other smaller FA Term candidates since some of the Arabic POS pattern rules are subsets of other rules.

Candidate terms made up of three or more words have the potential to yield smaller FA Term candidates since some of the Arabic POS pattern rules are subsets of other rules. In Example 3, the FA Term candidate "انتل معالجاً من أسرة بينتيوم بسرعة 333 ميغاهرتز" was extracted t first. Then this FA Term candidate also yielded two smaller FA Terms انتل معالجاً من أسرة بينتيوم" and "بينتيوم بسرعة 333 ميغاهرتز".

D.FA Terms weighting and selection

1. Corpora comparison

Comparing the occurrence of a FA Term candidate in the domain-specific corpus with its occurrence in the reference corpus Drouin [8] and Jiang [12] is an effective method to find FA Terms with high field specificity. The reference corpus is chosen in such a way that it would help us discriminate FA Terms in the domain-specific corpus more distinctly.

For each candidate term, we measure the regional term frequency (the frequency of term within the given document), the universal term frequency (term frequency within the whole of domain-specific corpus) and the document frequency (number of documents in the domain-specific corpus that contain the term). Likewise, we also measure the universal term frequency and the document frequency of the term in the reference corpus. The more relevant a term is to the field, the higher will be its term frequency and document frequency in the domain-specific corpus, while its term frequency and the document frequency in the reference corpus would be lower.

2. Weighting formula

The weighting formula is based on a modified version of *tf-idf.* Lan and Sung [14] have made a comparative study of various versions of *tf-idf* term weighting schemes.

We make the calculations at two levels: one at the document level and the other at the corpus level. "Regional" refers to the calculation at the level of the document, while "Universal" refers to the calculation at the level of the whole corpus of a particular field. The final selection of FA Terms is based on the universal weight.

wt_regional_tf refers to weighted regional term frequency of a FA Term candidate. *regional_tf* refers to regional term frequency and *regional_avg_tf* to average frequency of all FA term candidates in a document.

$$wt\_regional\_tf = \frac{1+\dfrac{df1/n1}{df1/n1+df2/n2}\times\beta\times\log(regional\_tf)}{1+\dfrac{df2/n2}{df1/n1+df2/n2}\times\beta\times\log(regional\_avg\_tf)} \tag{1}$$

where n1 is the number of documents in the domain-specific corpus, *df1* is the number of documents containing the term in the domain-specific corpus, *n2* is the number of documents in the reference corpus, *df2* is the number of documents containing the term in the reference corpus and β is an adjustment factor.

By using equation (1), then, *regional_weight* is calculated as follows:

$$regional\_weight = wt\_regional\_tf \times \left(\frac{itf_2}{itf_1}\times\frac{idf_2}{idf_1}\right)^2 + \alpha \tag{2}$$

Where itf_1, itf_2, idf_1 and idf_2 are calculated as $log_{10}(nt_1/tf_1)$, $log_{10}(nt_2/tf_2)$, $log_{10}(n_1/df_1)$ and $log_{10}(n_2/df_2)$ respectively. nt_1 and tf_1 are the total number of candidate terms and the total frequency of a particular term in the domain-specific corpus, while nt_2 and tf_2 are the total number of candidate terms and the total frequency of a particular term in the reference corpus. Moreover, α is the additional weight given to compound FA Term candidates if they contain a single FA Term.

Finally Universal term weights are calculated as in equation (3) by taking the average of the regional term weights that remain.

$$universal\_weight = \frac{\sum_{i=1}^{n} regional\_weight_i}{N}$$

TABLE IV RECALL AND PRECISION USING ARABIC SINGLE AND COMPOUND FA TERMS

Field (Size in B)	AFAT Type	FA Term Candidates	Automatically Retrieved Arabic FA Term	Total Relevant Arabic FA Terms	Relevant Arabic FA Terms selected automatically	Recall %	Precision %
Science	ASFAT	99.780	970	913	807	0.88	0.83
(1.64 MB)	ACFAT	155.650	1.559	1.404	1.066	0.75	0.68
Economic	ASFAT	205.659	2.566	2.338	2.028	0.86	0.79
(15.4 MB)	ACFAT	405.325	4.530	4.327	3.868	0.89	0.85
General	ASFAT	455.867	4.558	4.377	3.862	0.88	0.84
(36.7 MB)	ACFAT	549.300	5.293	5.211	4.552	0.87	0.86
Cars	ASFAT	101.500	852	782	614	0.78	0.72
(868 KB)	ACFAT	121.345	1.899	1.645	1.447	0.87	0.76
Computer	ASFAT	177.452	1.412	1.294	1.036	0.80	0.73
(1.64 MB)	ACFAT	199.402	1.930	1.858	1.216	0.65	0.63
News	ASFAT	307.849	4.631	4.480	3.842	0.85	0.82
(44.8 MB)	ACFAT	567.834	6.498	6.382	5.543	0.86	0.85
Sports	ASFAT	98.678	1.255	1.184	1.060	0.89	0.84
(18.9 MB)	ACFAT	289.456	3.447	3.313	2.843	0.85	0.82
Politics	ASFAT	260.453	2.576	2.398	2.020	0.84	0.78
(40.6 MB)	ACFAT	398.564	4.661	4.086	3.651	0.89	0.78
Religions	ASFAT	198.534	2.477	2.371	2.019	0.85	0.81
(30.89 MB)	ACFAT	378,694	4,295	3,908	3,451	0.88	0.80

ASFAT=Arabic Single FA Term, ACFAT = Arabic Compound FA Term.

Where N is the number of documents in which the term appears in the domain-specific corpus.

V. EXPERIMENTAL EVALUATION

A. Data Collection

The domain-specific corpora used in this research were collected from the English Wikipedia dumps (Wikipedia Foundation, Inc.) downloaded on 5 November 2009. As Wang [28] have shown that a thesaurus of concepts built from Wikipedia is effective in enhancing previous approaches for text classification, Wikipedia dumps is a good source of corpora for extracting FA Terms. From the downloaded Wikipedia dumps, the individual documents (articles) are extracted and Arabic POS-tagged with Tree tagger. We then use these tagged corpora as the source of our Alhyah corpora. We divide the documents (articles) into different fields based on their Wikipedia category and title [29] using a computer program. Some manual checking was required to get rid of garbage or empty files. The size of domain-specific corpora for different fields is shown in Table 4.

Experimental evaluation is carried out for 14 different fields using 251 MB of Alhyah corpora. The fields are: <Science>, <Economic>, <General>, <Cars>, <Computer>, <News>, <Sports>, etc. Reference corpora for comparison were also collected from the Wikipedia dumps.

B. Evaluation of FA Terms selection results

Once the domain-specific corpora and reference corpora are ready, we extract the Arabic single FA Terms and compound FA Terms as described before. The number of candidates selected in different fields is shown in Table 4. The extracted FA Term candidates are then given weights using the method described in the previous sections. The FA Term candidates with $universal\_weight$ above a heuristic threshold value are selected as FA Terms. Based on our experimental observations and after many trying the best selected values for α and β were α = 5, β = 10.

Table 4 shows the results of FA Terms selection for only 9 fields out of the 14 used in the experimental evaluation. Hence, precision and recall are calculated as follows:

$$precision = \frac{\mathrm{Re}\,lavant\ ArabicFATerms\ Selected\ Automatically}{Automatically\ \mathrm{Re}\,treived\ Arabic\ FATerms}$$

$$\mathrm{Re}\,call = \frac{\mathrm{Re}\,lavant\ ArabicFATerms\ Selected\ Automatically}{Total\ \mathrm{Re}\,lavant\ Arabic\ FATerms}$$

VI. CONCLUSION

This paper has presented a new methodology for building extensive Arabic dictionary uses linguistic methods to extract relevant compound as well as single FA Terms from domain-specific corpora using Arabic POS. The new method has extracted Arabic FA Terms from domain-specific corpora using part-of-speech (POS) pattern rules, corpora comparison and modified *tf-idf* weighting. Experimental evaluation is carried out for 14 different fields using 251 MB of domain-specific corpora obtained from Arabic Wikipedia dumps and Alhyah news selected average of 2,825 FA Terms (single and compound) per field. From the experimental results, recall and precision are 84% and 79% respectively. The results show that the proposed methodology is effective for building a comprehensive Arabic dictionary of FA Terms.

Future studies will further improve the proposed methodology by adding a document classification module so that documents can be classified automatically and FA Term candidates extracted from them. Moreover, text summarization using Arabic filed association dictionary knowledge will be presented.

REFERENCES

[1] Atlam, E., Fuketa, M., Morita, K., Aoe, J. (2003). Documents Similarity Measurement using Field Association Terms, *Information Processing & Management*, 39(6): 809-824.

[2] Atlam, E., Ghada, E., Morita, K., Fuketa, M., Aoe, J. (2006). Automatic building of new field association word candidates using search engine, *Information Processing & Management*, 42(4): 951-962.

[3] Atlam, E., Morita, K., Fuketa, M., Aoe, J. (2002). A new method for selecting English field association terms of compound words and its knowledge representation, *Information Processing & Management*, 38(6): 807-821.

[4] Bennet N.A., He, Q., Powell K., Schatz, B.R. (1999). Extracting noun phrases for all of MEDLINE, In *Proceedings of the AMIA Symposium*, pp. 671-5.

[5] Diab M., Kadri Hacioglu (2004), and Daniel Jurafsky. Automatic tagging of Arabic text: From raw text to base phrase chunks. In *Proceedings of the 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLTNAACL04)*, Boston, MA, 2004.

[6] Dorji, T., Atlam, E., Yata, S., Fuketa, M., Morita, K., Aoe, J. (2009) Building a Dynamic and Comprehensive Field Association Terms Dictionary from Domain-specific Corpora using Linguistic Knowledge, *In Proceedings of the fifth Corpus Linguistics Conference*, Liverpool, UK.

[7] Dozawa, T. (1999). Innovative multi information dictionary Imidas'99. Annual Series. Japan: Zueisha Publication Co. [in Japanese].

[8] Drouin, P. (2004). Detection of domain specific terminology using corpora comparison, In *Proceedings of the 4th International conference on Language resources and evaluation (CLREC)*, pp. 79-82.

[9] Fuketa, M., Lee, S., Tsuji, T., Okada, M., Aoe, J. (2000). A Document Classification Method by using Field Association Words, *International Journal of Information Sciences* 126: 57-70.

[10] Graham-Cumming, J. (2005) Naive Bayesian Text Classification: Fast, accurate, and easy to implement, *Dr. Dobb's Journal*, http://www.ddj.com/development-tools/184406064, [Accessed 3 September 2009].

[11] Habash, Nizar and Owen Rambow (2005). Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop. In Proceedings of the Conference of American Association for Computational Linguistics (ACL05)

[12] Jiang, G., Sato, H., Endoh, A., Ogasawara, K., Sakurai, T. (2005). Extraction of Specific Nursing Terms Using Corpora Comparison, In *Proceedings of the AMIA Annual Symposium*, 2005: 997.

[13] Krauthammer, M., Nenadic, G. (2004). Term identification in the biomedical literature, *Journal of Biomedical Information*, 37(6): 512–526.

[14] Lan M., Tan C., Low H., Sung S. (2005). A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *Posters Proc. 14th International World Wide Web Conference*, pp. 1032–1033.

[15] Lee, S., Shishibori, M., Sumitomo, T., Aoe, J. (2002). Extraction of Field-coherent Passages, *Information Processing & Management*, 38(2): 173-207.

[16] Pang, S., Kasabov, N. (2009) Encoding and decoding the knowledge of association rules over SVM classification trees, *Knowledge and Information Systems*, 19(1): 79-105.

[17] Patry, A., Langlais, P., (2005) Corpus-based terminology extraction. *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering*, Copenhagen, Denmark, pp. 313–321.

[18] Peng, T., Zuo, W., He, F. (2008) SVM based adaptive learning method for text classification from positive and unlabeled documents, *Knowledge and Information Systems*, Springer London, 16(3): 281-301.

[19] Rokaya, M., Atlam, E., Fuketa, M., Dorji, T., Aoe, J. (2008) Ranking of Field Association Terms using co-word analysis, *Information Processing and Management*, 44(2): 738-755.

[20] Salton, G., Allan, J., Buckley, C. (1993) Approaches to passage retrieval in full text information systems. *Proceedings of the 16th annual international ACM/SIGIR conference on research and development in information retrieval*, pp. 49–58.

[21] Saneifar, H., Bonniol, S., Laurent, A., Poncelet, P., Roche, M. (2009) Terminology Extraction from Log Files, *Database and Expert Systems Applications, Lecture Notes in Computer Science*, 5690: 769 - 776.

[22] Sharif, U. M., Ghada, E., Atlam, E., Fuketa, M., Morita, K., Aoe, J. (2007). Improvement of building field association term dictionary using passage retrieval, *Information Processing and Management*, 43(2): 1793-1807.

[23] Shereen Khoja. 2001. *APT: Arabic Part-of-speech Tagger.*, Proc. of the Student Workshop at NAACL 2001Smadja, F. (1993) Retrieving collocations form text: Xtract, *Computational Linguistics*, 19(1): 143–177.

[24] Srinivasan, P., Pant, G., Menczer, F. (2005) A general evaluation framework for regional crawlers. *Information Retrieval*, 8(3):417–447.

[25] Stanford TreeTagger – a Language-Independent Part-of-speech Tagger, *http://nlp.stanford.edu/software/tagger.shtml* [Downloaded 5 November 2009]

[26] Tsuji, T., Nigazawa, H., Okada, M., Aoe, J. (1999) Early Field Recognition by Using Field Association Words, In *Proceedings of the 18th International Conference on Computer Processing of Oriental Languages*, pp. 301-304.

[27] Velardi, P., Navigli, R., D'Amadio, P. (2008) Mining the Web to Create Specialized Glossaries, *IEEE Intelligent Systems*, 23(5): 18-25.

[28] Wang, P., Hu, J., Zeng, H., Chen, Z. (2008) Using Wikipedia knowledge to improve text classification, *Knowledge and Information Systems*, 19(3): 265–394.

[29] Wikipedia Foundation, Inc., English Wikipedia Dumps, *http://dumps.wikimedia.org/arwiki/* [Downloaded 5 November 2009]

Dr. El-Sayed Atlam: Received B.Sc. and M. Sc. Degrees in Mathematics from, Faculty of Science, Tanta University, Egypt, in 1990 and 1994, respectively, and the Ph.D. degree in information science and Intelligent systems from University of Tokushima, Japan, in 2002. He has been awarded by a ***Japan Society of the Promotion of Science*** (*JSPS*) postdoctoral Fellow from 2003 to 2005 in Department of Information Science & Intelligent Systems, Tokushima University; He is currently assistant professor at the Department of information science and Intelligent systems from University of Tokushima, Japan. He is also Associate professor at the Department of Statistical and Computer science, Tanta University, Egypt. Dr. Atlam is a member in the Computer Algorithm Series of the IEEE computer society Press (CAS) and the Egyptian Mathematical Association (EMA). His research interests include information retrieval, natural language processing and document processing.

Prof. Jun-ichi Aoe received B.Sc. and M.Sc. Degrees in electronic engineering from the University of Tokushima, Japan, in 1974 and 1976, respectively, and the Ph.D. degree in communication engineering from the University of Osaka, Japan 2980. Since 1976 he has been with the University of Tokushima. He is currently a Professor in the department of Information Science & Intelligent Systems, Tokushima University, Japan. His research interest include Design of an automatic selection method of key search algorithms based on expert knowledge bases, natural language processing, a shift-search strategy for interleaved LR parsing, robust method for understanding NL interface commands in an intelligent command interpreter, and trie compaction algorithms for large key sets. He is the editor of the computer Algorithm Series of the IEEE computer Society Press. He is a member in the association for computing machinery, the association for the natural language processing of Japan.