Journal of Theoretical and Applied Information Technology
30th September 2014. Vol. 67 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A HYBRID METHOD OF LINGUISTIC APPROACH AND STATISTICAL METHOD FOR NESTED NOUN COMPOUND EXTRACTION

[1] HAMED AL-BALUSHI, [2] MOHD JUZIADDIN AB AZIZ

[1,2]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi Selangor, MALAYSIA

E-mail: [1]almukles7@gmail.com, [2]din@ftsm.ukm.my

ABSTRACT

Arabic noun compound extraction has become a challenging issue in the field of NLP. Several approaches have been proposed in terms of extracting Arabic noun compounds. Some of them have used linguistic-based approach, other have used statistical methods and the rest have used a hybrid between them. This research proposed a hybrid method of linguistic-based approach and statistical method in order to solve the extraction of nested Arabic noun compound. The dataset has been collected from online Arabic newspaper archive from Aljazeara.net and Almotamar.net. Several pre-processing steps have been carried out on the data including transformation, normalization, stemming and POS tagging. After that, the n-gram model has generated bi-gram, tri-gram, 4-gram, and 5-gram candidates of noun compound. Then three association measures which are NC-value, PMI and LLR have been used in order to rank the candidates. The evaluation has been performed using the n-best method with a human annotation (manual selection by expertise). NC-value has outperformed PMI and LLR in terms of extracting nested noun compounds.

Keywords: *Multi-Word Expressions, Noun Compound, Nested Noun Compound, Statistical Method, POS Tagging*

1. INTRODUCTION

Noun compound (NC) is the phrase that has two or more nouns with a space (e.g. orange juice) or without a space (e.g. headquarter) or with a hyphen (e.g. son-in-law) which refers to a single concept [1]. It consists of two parts; the first one is a noun which called head and the rest are called modifier (Kim 2008). Various conditions have been proposed by the researchers in order to identify NCs for instance, Lauer [2] suggested that the head should be a noun and the modifier could be nouns, verbs, adjective or adverb, whereas Buckeridge et al. [3] suggested that the head and the modifier should be nouns. Noun compound seems to be very challenging issue in the field of NLP due to several reasons. Syntactically, sometimes NCs bring an ambiguity to the sentence such as, "glass window cleaner" which refers to several meanings. It could be "(glass (window cleaner)", "window cleaner made of glass" or "((glass window) cleaner)". Semantically, it is difficult to interpret the semantic relations between the head and the modifier. Therefore, several approaches have been proposed in terms of disambiguate the semantic and syntax of NCs [2-4].

Nested noun compound NNC are substring of other longer noun compounds, it may contain two words, three words, four words, five words, etc. This type of noun compound characterized by its arbitrary (it is not clear why to "fail through" means to "fail"), and also it is a domain independent (mobile application or stock market) and it is a co-occurrence-based which means the presence of part of it may suggest the rest of it (e.g. Sultanate of could be imply with Oman or Brunei) [5] The problem of NNC lies on extracting the appropriate substring and discard the unwanted ones. For instance, "New York Stock Exchange" could be divide into "New York", "York Stock" and "Stock Exchange", the first and third ones should be considered as noun compounds while the second one should be discarded. Therefore, extracting nested noun compound seems to be a challenging issue.

There are several approaches that have been proposed in terms of extracting Arabic noun compounds. Some of them have used linguistic techniques, other have used statistical methods and the rest have used a hybrid between them. Linguistic-based technique aims to use analysis tools such as, POS tagging, lemmatization and parser. While statistical methods aims to rank the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

candidates of noun compounds. The ranking of candidates in the hybrid methods is very crucial, it usually performed using the standard association measures. However, the performance of these measures is roughly similar in terms of extracting bi-gram noun compounds. Nonetheless, Arabic has tremendous classifications of noun compounds. There some classifications are based on nested of two or more words such as, tri-gram, 4-gram and 5-gram. For instance, "وزير الخارجية العمانية" this noun compound is consist of multiple bi-gram " الخارجية العمانية" and "وزير الخارجية". The problem becomes more difficult when dealing with those classifications due to the overlap that occurred between two or more noun compounds. Therefore, there is still a significant need for improving the extraction of Arabic nested noun compounds in terms of accuracy.

This research aims to extract Arabic nested noun compounds including bi-gram, tri-gram, 4-gram and 5-gram using a hybrid method of linguistic approach and statistical method. The linguistic approach that used in this research in consisting of stemming and POS tagging, while the statistical method is NC-value. The dataset is the same that have been used by Saif & Aziz [6] which contain an online Arabic newspaper archive from Aljazeara.net and Almotamar.net. Several pre-processing have been performed in this study in order to normalize, tokenize and stem the data. In addition, candidate identification step has been performed using the linguistic approach (stemming and POS tagger). Meanwhile, n-gram model has been used in order to generate bi-gram, tri-gram, 4-gram and 5-gram. In other hand, three statistical methods have been used which are NC-value, Point-wise Mutual Information (PMI) and Log-Likelihood (LLR) in order to rank the candidates. Eventually, n-best evaluation method has been used with a human annotation (expertise) in order to evaluate the proposed method.

2. RELATED WORK

Several approaches have been proposed in terms of extracting Arabic MWEs. Some of them have used statistical methods, other have used linguistic methods and the rest have used a combination between them. For instance, Attia [7] have proposed a semi-automatic method for extracting MWE using a manual lexicon. Basically, he create a specialized two-side transducer based on regular expression in order to analyze the lexical-side and generate the morphological lexicon. This method can extract fixed and semi-fixed MWEs however, it cannot treat the syntactically-flexible expressions. Apart from that, Boulaknader et al. [8] have proposed a MWE extraction program for Arabic language using a hybrid method of linguistic and statistical. They firstly used the part-of-speech tagging that have presented by Diab et al. [9] in order to filter the candidates of multiword terminology. Hence, they have used the linguistic filter to determine the multi-words patterns including N ADJ, NN and N PREP N. furthermore, their method considers the variations of MWE such as, a graphical variations (e.g. "Ha'a" and "Ta'a marbuta"), inflectional variations (e.g. plural, singular and define articles "ال"), morph-syntactic variations (e.g. synonyms) and internal modification. On other hand, they have used four statistical measures which are Log-Likelihood Ratio (LLR) [10], FLR [10]. Mutual Information (MI) and t-scores [11] for ordering the MWE candidates.

Bounhas & Slimani [12] have proposed a combination method to extract MWEs of Arabic terminology using an Arabic specialized corpora. Basically, they have used the Arabic morphological analyzer (AraMorph) that has been proposed by Hajic et al. [13]. The aim of using AraMorph is to compute the morphological features for the syntactic rules. In the same manner, they have used the POS tagger that have proposed by Diab [9] as a linguistic filter. Therefore, they integrate the POS tagger with the AraMorph in order to develop the Morph-POS matcher. Morph-POS matcher aims to reduce the morphological ambiguity by analyzing the context of each word. The sequence identifier in this method used for determining the MWEs from sequences and count the frequency of each entry that contains the POS and morpho-syntactic features. Therefore, a rule-based syntactic tool based on the POS and the morphological features have been used in order to detect the compound nouns. On other hand, they use an association measure which is LRR in order to identify the best solution. Although this method has outperformed the method that has been presented in Boulaknadel et al. [8], but it did not extract MWEs with lexical units such as, verb-particle constriction, prepositional phrases and Arabic collocations.

Attia et al. [14] have proposed an automatic extraction in terms of extracting and validating MWEs from Arabic corpora using three complementary approaches. The first one is cross-lingual correspondences asymmetry which extracts MWEs from Arabic Wikipedia (Ar.Wikipedia). Using a multi-lingual lexicon of named entities, this

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

approach aims to generate all candidates of MWE titles in Ar. Wikipedia. The second approach is a translation-based that aims to extract the English MWEs from Princeton WordNet (PWN) and then translate it to Arabic using Google translator. This method has been evaluated automatically using the gold standard PWN. Eventually, the third approach which is the corpus-based that aims to extracts MWEs using the statistical measures and POS-annotation filtering. It combines the linguistic and statistical methods in order to generate the unigram, bigram and trigram candidates. The association measures that have been used are the Point-wise Mutual Information (PMI).

Saif & Aziz [6] proposed a hybrid method of linguistic-based approach and statistical approach for extracting Arabic collocations from collected Arabic newspaper corpus. Firstly, a normalization step has been taken a place including define article elimination, non-Arabic words elimination and generating unigram list. Secondly, a linguistic approach have been used including lemmatization and POS tagging. Basically, the lemmatization aims to enhance the unigram list and generating a bi-gram candidates, and then the POS tagger used for disambiguate the words that have more than one linguistic categories. Note that, the POS tagger used in this phase contains the joint tagging and segmenting algorithm that have been presented by AlGahtani et al. [15]. As a result of linguistic phase, the bi-gram of candidates will be generated with their frequency. On the other hand, four statistical measures which are Log-Likelihood Ratio (LLR), chi-square, Point-wise Mutual Information (PMI) and Enhanced Mutual Information (EMI) have been applied in order to rank the candidates based on the co-occurrence. They reported that LLR has outperformed others association measures.

El Mahdaouy et al. [16] have used a hybrid method of linguistic including POS tagging with sequence identifier and a novel statistical method called NLC-value for extracting Multi-word Terms MWTs. Firstly, they have used AMIRA toolkit [9] which is a POS tagger that aims to assign each word in the corpus with its polarity, then the sequence identifier have been used in order to tokenize tag files in the corpus. Therefore, the candidate of MWTs will be filtered in order to fit the linguistic structural patterns. After that, several statistical methods have been used including C-value, NC-value, NTC-value and NLC-value in order to rank the candidates. Eventually, they have demonstrated that NLC-value outperforms the other association measures.

3. PROPOSED METHOD

The proposed method contains the details of the dataset that have been used. Moreover, it describes the pre-processing tasks that performed on the dataset including normalization, stemming and POS tagging. Normalization task aims to remove the noisy data such as stop-words, while stemming task aims to turn the inflected words into its own stem or root, whereas POS tagging task aims to identify words' categories such as verb, noun or adjective. On other hand, the proposed method includes the process of identifying the candidates of the noun compounds using n-gram model to generate bi-gram, tri-gram, 4-gram and 5-gram. Furthermore, the proposed method contains the process of ranking such candidates using the association measures including Log-Likelihood Ratio (LLR), Point-Wise Mutual Information (PMI) and NC-value. Eventually, the evaluation phase which aims to evaluate the extracted noun compounds. Fig. 1 shows the research design.

Journal of Theoretical and Applied Information Technology
30th September 2014. Vol. 67 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

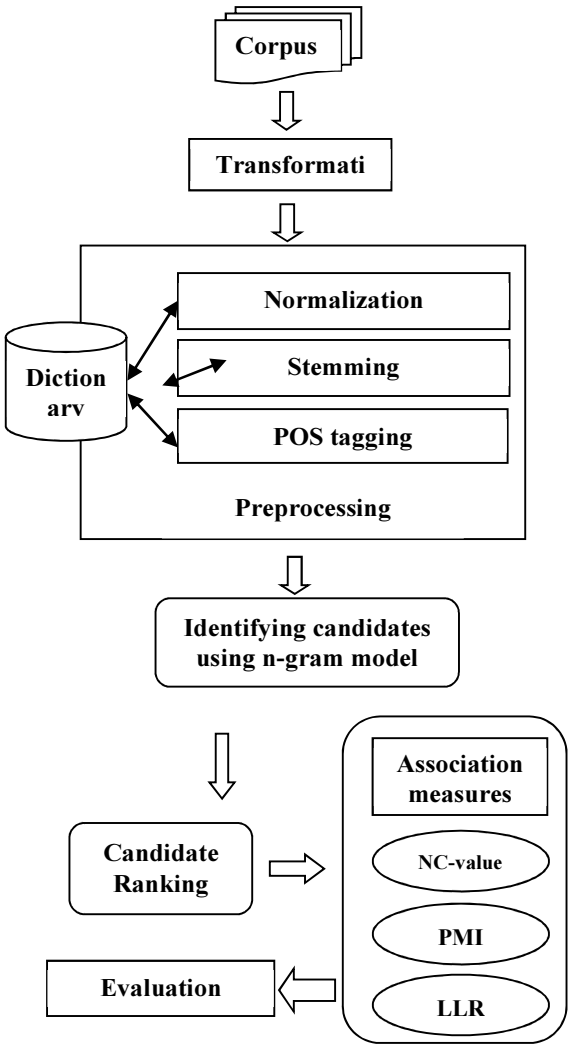

Figure 1.The research design

3.1. Transformation

This phase aims to turn the data into an internal representation that enables applying pre-processing steps. In fact, the data consists of text files of collected Arabic newspaper. Hence, there are several steps of transformation which illustrated as follows:

1. Turning the files into UTF-8 encoding

UTF-8 is an encoding that can be used to simulate any character in the Unicode character. It has been used in order to represent complicated languages such as Arabic language.

2. Arabic letters

Firstly, all the Arabic letters have to be transformed in order to recognize any Arabic word.

3. Arabic diacritics

Arabic language has numerous diacritics (tashkil تَشْكِيْل), it is vowel marks that indicate how to pronounce letters (e.g. short or long vowel).

3.2. Pre-processing

This phase aims to perform several steps including normalization, stemming and POS tagging in order to turn the data into a format that enables applying statistical processes. Note that, this phase is very crucial in terms of enhancing the results of extracting noun candidates of noun compounds. The steps of this phase are determined as follows:

3.2.1 Normalization

This step aims to clean the data by removing noisy or unwanted data such as, digits, stop-words and special characters. Such data has been explained as follows:

3.2.1.1. Remove special characters

Special characters such as "_+-:""?" have to be removed in order to clean the data.

3.2.1.2. Remove non-Arabic letters

Every character tend to be non-Arabic letter has to be removed.

3.2.1.3. Remove digits

Digits from 0 to 9 have to be removed.

3.2.1.4. Remove diacritics

As mention earlier, Arabic language has several forms of diacritics thus; those diacritics have to be removed.

3.2.1.5. Remove definite articles

It is a set of letters that could be prefix of suffix which indicates to the kind of reference that made by the noun

3.2.1.6. Tokenization

Tokenization is the task of dividing words from text into clusters of consecutive morphemes, one of which typically corresponds to the word stem. It shows how tokenizing the texts lead to the sequence of tokens as: الاتحاد العالمي لكرة القدم, after tokenizing become; (الاتحاد_العالمي_لكرة_القدم).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2.1.7. Remove stop-words

Stop words are words that connect the sentences which seem to be irrelevant and have to be removed.

3.2.2 Stemming

The stemmer that have been used in this research is based on the Buckwalter system which includes three distinct lexicon files (dictionary) listing all prefixes, suffixes, and stems. The algorithm works by dividing a given input word into prefix, stem and suffix and then match those segments with the corresponding lexicons [17].

3.2.3 POS tagging

Part-of-speech POS tagging is one of word sense disambiguation methods that aims to assign each word in certain text with a fixed set of parts of speech such as, noun, verb, adjective or adverb [18]. There are tremendous words that have several potential tags thus, POS have been came up in order to disambiguate these words. Therefore, the main role of POS is to deter-mine the exact tagging for each words in the corpus. Note that, the POS tagging that have been used in this research is the same that used by AlGahtani et al. [15] which is the joint tagging and segmenting algorithm.

3.3 Candidate identification

Basically, this phase is depending on linguistic analysis tools such as, stemming and POS tagging. It aims to select the n-gram lists including bi-gram, tri-gram, 4-gram and 5-gram lists of noun compound candidates according to the structural patterns. Firstly, it brought the words from the unigram list that have been acquired from the pre-processing phase. This process relies on several sub-steps which are; (i) detecting candidates that start with Noun, (ii) the followed components have to be either noun or adjective, (iii) the followed components have to be two or more words. Each combination will be stored based on its degree for instance, the combination with 3 words will be stored in a Tri-gram list and the combination with 4 words will be stored in a 4-gram list. Secondly, from the 5-gram list a 4-gram combination that seems to be a candidate based on the linguistic structural patterns will be selected and store it in a list called 4-gram list with their linguistic category and frequent occurrences. Then, from the 4-gram list a 3-gram combination that seems to be a candidate based on the structural patterns will be selected and store it in a list called tri-gram list with their linguistic category and frequent occurrences. In the same manner, bi-gram list will be built from the 3-gram list.

3.4 Candidate ranking

The candidate ranking depends on frequency information for the word occurrence and co-occurrence in the corpus. The candidate identification phase collects both syntactic information and information about their occurrence in the corpus. In this phase, the association measures are computed to the identified candidates in the n-gram lists that assigns to each candidate a score of association strength. For each pair of words extracted from a corpus, association score is a single real value that indicates the amount of statistical association between the two words. In this research, three statistical measures have been used including NC-value, Point-wise mutual information PMI and log-likelihood LLR, which are explained as follows:

3.4.1 C-value/NC-value

This method was introduced by Frantzi et al. [19]. This method consists of two steps: 1) the C-value is to enhance the extraction of nested multi-word terms, and 2) the NC-value that incorporates context information to the C-value method, in order to enhance multi-word term extraction in general. The C-value approach is a domain-independent method to extract MWEs which includes the processing the nested terms by using statistical measure of frequency of occurrence. The C-value statistical measure assigns the term-hood to a candidate of MWE w, ranking it in the output list of candidates. The measure of term-hood, called C-value is defined as

$$c\text{-}value(w)\begin{cases} log_2|w|.f(w) & w\ is\ not\ a\ nested \\ log_2|w|.\left(f(w)-\frac{1}{p(T_w)}\sum_{b\varepsilon T}f(b)\right) & otherwise \end{cases} \quad (1)$$

, w is the candidate string, f is its frequency of occurrence in the corpus, Tw is the set of candidate MWLUs that contain w, and P (Tw) is the number of these candidate MWLUs.

3.4.2 Point-wise Mutual Information (PMI)

This association measure was proposed as the association ratio to compute the word connectedness based on information theory concept by Church & Hanks [20]. Regarding to Zhang et

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

al. [21], in multi-word detection, the mutual information can be defined as "the amount of information provided by the occurrence of the word represented by Y about the occurrence of the word represented by X". The MI between the bigram candidate of x and y was defined as the following:

$$MI(x,y) = log_2 \frac{p(x,y)}{p(x)p(y)} \quad (2)$$

, where $P(x)$ is the occurrence probability of word x and $P(y)$ is the occurrence probability of word y in the corpus.

3.4.3 Log-Likelihood Ratio (LLR)

This measure has been introduced by [10]. It had been determined as a stable indicator of terminologically related collocations. In addition, it usually works with special purpose corpora due to its ability to extract rare collocations which is suitable for scattered data [22]. The log-likelihood is calculated with a formula adjusted for co-occurrence contingency table as follows. For a given pair of words W1 and W2 and a search window W, let a be the number of windows in which W1 and W2 co-occur, let b be the number of windows in which only W1 occurs, let c be the number of windows in which only W2 occurs and let d be the number of windows in which none of them occurs. The LLR is defined as the following:

$$\begin{aligned} LLR = 2((&alna + blnb + clnc + dlnd \\ &+ (a+b+c \\ &+ d)\ln(a+b+c+d)) \\ &- \big((a+b)\ln(a+b) \\ &+ (a+c)\ln(a+c) \\ &+ (b+d)\ln(b+d) \\ &+ (c+d)\ln(c+d)\big)) \end{aligned} \quad (3)$$

3.4 Evaluation

In this research, n-best evaluation method [23] have been used in order to evaluate the proposed method. It uses association scored to rank the candidates of noun compound. In fact, this method contains three steps; the first one is the selection of n-best list which aims to choose the highest association scores of candidate ranking. The second step is the human annotation (group of expertise-Arabic teachers) which aims to manually select the true noun compounds from the n-best list with three tags; the first tag is 1 which indicated the true noun compounds, the second tag is 0 which indicates the false noun compounds and the third tag is Err which refers to an error. The third step is evaluating the manual annotation of candidates using precision which computed as the following equation.

$$Precision = \frac{TP}{TEC} \quad (4)$$

Where: TP is the number of correct extracted noun compounds. While TEC is the total number of extracted noun compounds (the n-value for n-best list).

4.

5. RESULTS AND DISCUSSION

Basically, the three association measures which are NC-value, PMI and LLR have been carried out in order to rank the candidates of nested noun compound Table 1 shows the results with bi-gram, tri-gram, 4-gram and 5-gram candidates.

Table 1 Results precision for the three association measures

Association	2-gram	3-gram	4-gram	5-gram
NC-value	0.81	0.59	0.29	0.18
LLR	0.82	0.137	0.08	0.068
PMI	0.80	0.146	0.094	0.056

As shown in Table 1, LLR has outperformed MI and NC-value for the bi-gram candidates. As expected form other studies such as Boulaknadel et al. [8], Evert [24] and Saif & Aziz [6] PMI has lower results than LLR when extracting bi-gram candidates. In other hand, regarding to El Mahdaouy et al. [16], LLR is outperforming NC-value in terms of extracting bi-gram candidates due to the unit-hood feature which lies in LLR. Unit-hood feature provides the degree of strength for combinations or collocations [25], which fits the case of bi-gram candidates. However, NC-value has outperformed LLR and PMI in terms of extracting tri-gram, 4-gram and 5-gram candidates. This is due to the feature of term-hood that lies on NC-value which makes it outperforming LLR in terms of extracting 3-gram, 4-gram or 5-gram [16]. Regarding to Vu et al. (2008) the term-hood treats the terms as a linguistic unit, which fits the case of nested candidates (tri-gram, 4-gram and 5-gram). Therefore, the ability of NC-value to extract nested noun compounds is so remarkable.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. CONCLUSION

This research proposed a hybrid of linguistic approach and statistical method in order to extract Arabic noun compound. The dataset has been collected from online Arabic newspaper archive from Aljazeara.net and Almotamar.net. Several pre-processing steps have been carried out on the data including transformation, normalization, stemming and POS tagging. After that, the n-gram model has generated bi-gram, tri-gram, 4-gram, and 5-gram candidates of noun compound. Then three association measures which are NC-value, PMI and LLR have been used in order to rank the candidates. The evaluation has been performed using n-best method with a human annotation (manual selection by expertise). NC-value has outperformed PMI and LLR in terms of extracting tri-gram, 4-gram and 5-gram. In general, this study is focusing on the nested noun compound with a maximum combination of 5-gram and a specific pattern of POS tagging which are adjectives and nouns. Furthermore, this study have used a relatively small dataset. Those issues could be the limitation of this study.

The areas suggested for future research work are described as follows. Combining two statistical measures can contribute toward enhancing the effectiveness. By using more linguistic approaches such as parser can enhance the results of extracting Arabic noun compound. Finally, the evaluation of this research can be improved by replacing the manual annotation with an automatic annotation. This can be performed by using a huge dictionary that contains a tremendous amount of noun compounds.

REFERENCES

[1] Geoffrey K Pullum, "The evolution of model-theoretic frameworks in linguistics," in *Model-theoretic syntax (proceedings of the MTS@10 workshop, August 13-17, organized as part of ESSLLI 2007, the European Summer School on Logic, Language and Information)*, 2007, pp. 1-10.doi:10.1.1.132.6972Retrieved from.

[2] Mark Lauer, "Corpus statistics meet the noun compound: some empirical results," presented at the Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Cambridge, Massachusetts, 1995. doi:10.3115/981658.981665Retrieved from.

[3] Alan M. Buckeridge and Richard F. E. Sutcliffe, "Disambiguating noun compounds with latent semantic indexing," presented at the COLING-02 on COMPUTERM 2002: second international workshop on computational terminology - Volume 14, 2002. doi:10.3115/1118771.1118772Retrieved from.

[4] Preslav Nakov and Marti Hearst, "Using the web as an implicit training set: application to structural ambiguity resolution," presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, 2005. doi:10.3115/1220575.1220680Retrieved from.

[5] Katerina T. Frantzi and Sophia Ananiadou, "Extracting nested collocations," presented at the Proceedings of the 16th conference on Computational linguistics - Volume 1, Copenhagen, Denmark, 1996. doi:10.3115/992628.992639Retrieved from.

[6] Abdulgabbar M Saif and Mohd JA Aziz, "An automatic collocation extraction from Arabic corpus," *Journal of Computer Science,* vol. 7, p. 6, 2010.doi:10.3844/jcssp.2011.6.11Retrieved from.

[7] Mohammed Attia, "An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks," in *Challenges of Arabic for NLP/MT Conference, The British Computer Society, London, UK*, 200610.1.1.72.1482Retrieved from.

[8] Siham Boulaknadel, Beatrice Daille, and Driss Aboutajdine, "A Multi-Word Term Extraction Program for Arabic Language," in *International Conference on Language Resources and Evaluation*, 2008Retrieved from http://dblp.uni-trier.de/db/conf/lrec/lrec2008.html#BoulaknadelDA08.

[9] Mona Diab, Kadri Hacioglu, and Daniel Jurafsky, "Automatic tagging of Arabic text: From raw text to base phrase chunks," in *Proceedings of HLT-NAACL 2004: Short Papers*, 2004, pp. 149-152.doi:Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.2147&rep=rep1&type=pdf.

[10] Ted Dunning, "Accurate methods for the statistics of surprise and coincidence," *Comput. Linguist.,* vol. 19, pp. 61-74, 1993Retrieved from http://dl.acm.org/ft_gateway.cfm?id=972454&ftid=250605&dwn=1&CFID=391011153&CFTOKEN=66514564.

[11] Kenneth W. Church and Robert L. Mercer, "Introduction to the special issue on computational linguistics using large corpora," *Comput. Linguist.,* vol. 19, pp. 1-24, 1993Retrieved from http://dl.acm.org/ft_gateway.cfm?id=972452&ft

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

id=250603&dwn=1&CFID=391011153&CFTOKEN=66514564.

[12] I. Bounhas and Y. Slimani, "A hybrid approach for Arabic multi-word term extraction," in *Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on*, 2009, pp. 1-8.doi:10.1109/NLPKE.2009.5313728Retrieved from.

[13] Jan Hajic, Otakar Smrz, Tim Buckwalter, and Hubert Jin, "Feature-based tagger of approximations of functional Arabic morphology," in *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, 2005, pp. 53-64.doi:Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.880&rep=rep1&type=pdf.

[14] Mohammed Attia, Lamia Tounsi, Pavel Pecina, Josef van Genabith, and Antonio Toral, "Automatic extraction of Arabic multiword expressions," in *In Proceedings of the 7th Conference on Language Resources and Evaluation, LREC-2010*, 2010Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.187.575&rep=rep1&type=pdf.

[15] Shabib AlGahtani, William Black, and John McNaught, "Arabic part-of-speech tagging using transformation-based learning," in *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, Cairo*, 2009, pp. 66-70.doi:Retrieved from http://www.elda.org/medar-conference/pdf/43.pdf.

[16] Abdelkader El Mahdaouy, Saïd El Alaoui Ouatik, and Eric Gaussier, "A Study of Association Measures and their Combination for Arabic MWT Extraction," in *Proceedings 10th International Conference on Terminology and Artificial Intelligence*, 2013, pp. 45-52.doi:Retrieved from http://hal.archives-ouvertes.fr/docs/00/88/11/75/PDF/MWT_Latex_7_pages.pdf.

[17] Alexander Fraser, Jinxi Xu, and Ralph M Weischedel, "TREC 2002 Cross-lingual Retrieval at BBN," in *TREC*, 2002Retrieved from http://trec.nist.gov/pubs/trec11/papers/bbn.xu.cross.pdf.

[18] Roberto Navigli, "Word sense disambiguation: A survey," *ACM Comput. Surv.,* vol. 41, pp. 1-69, 2009.doi:10.1145/1459352.1459355Retrieved from.

[19] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima, "Automatic recognition of multi-word terms:. the C-value/NC-value method," *International Journal on Digital Libraries,* vol. 3, pp. 115-130, 2000/08/01 2000.doi:10.1007/s007999900023Retrieved from http://dx.doi.org/10.1007/s007999900023.

[20] Kenneth Ward Church and Patrick Hanks, "Word association norms, mutual information, and lexicography," *Comput. Linguist.,* vol. 16, pp. 22-29, 1990Retrieved from http://dl.acm.org/ft_gateway.cfm?id=89095&ftid=248552&dwn=1&CFID=391011153&CFTOKEN=66514564.

[21] Wen Zhang, Taketoshi Yoshida, Xijin Tang, and Tu-Bao Ho, "Improving effectiveness of mutual information for substantival multiword expression extraction," *Expert Syst. Appl.,* vol. 36, pp. 10919-10930, 2009.doi:10.1016/j.eswa.2009.02.026Retrieved from.

[22] Špela Vintar, "Comparative Evaluation of C-value in the Treatment of Nested Terms," in *Workshop Description*, 2004Retrieved from http://lrec.elra.info/proceedings/lrec2004/ws/ws6.pdf#page=54.

[23] Stefan Evert, "Title," unpublishedRetrieved from http://elib.uni-stuttgart.de/opus/volltexte/2005/2371.

[24] Stefan Evert, "Corpora and collocations," in *Corpus Linguistics. An International Handbook*, 200810.1.1.159.6220Retrieved from.

[25] I Fahmi, "C Value Method for Multi-word Term Extraction. Seminar in Statistics and Methodology. ," in *Alfa-informatica, RuG, May 23 (2005)*, 2005Retrieved from http://odur.let.rug.nl/~nerbonne/teach/rema-stats-meth-seminar/presentations/fahmi-statistics-c-value.pdf.