Journal of Theoretical and Applied Information Technology
30th June 2015. Vol.76. No.3

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

EXTRACTION OF ARABIC NESTED NOUN COMPOUNDS BASED ON A HYBRID METHOD OF LINGUISTIC APPROACH AND STATISTICAL METHODS

[1] MARYAM AL-MASHHADANI, [2] NAZLIA OMAR

Center for Artificial Intelligence Technology (CAIT)

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi Selangor, MALAYSIA

E-mail: [1]m1-2100@yahoo.com, [2]nazlia@ukm.edu.my

ABSTRACT

Arabic noun compound are phrases that consist of two or more nouns. The key characteristic behind noun compounds lies on its frequent occurrences within the text. Therefore, extracting these noun compounds is essential for several domains of research such as Information Retrieval, Sentiment Analysis and Question Answering. Many research efforts have been proposed in terms of extracting Arabic noun compounds using linguistic approaches, statistical measures or a combination of both. Most of the existing methods have concentrated on the extraction of bi-gram or tri-gram noun compound. However, extracting 4-gram or 5-gram noun compound or nested noun compound is challenging due to the difficulty of selecting an appropriate method with effective results. Multiple features have a significant impact on the effectiveness of extracting noun compound such as contextual information, unit-hood and term-hood. Thus, there is still room for improvement in terms of enhancing the effectiveness of nested noun compound extraction. Therefore, this study proposed a combination of linguistic approach and statistical measures in order to enhance the extraction of nested noun compound. Several preprocessing steps are involved including transformation, normalization, tokenization, and stemming. The linguistic approach that have been used in this study is Part-of-Speech tagging. In addition, a new linguistic pattern for named entities has been utilized using a list of Arabic named entities in order to enhance the linguistic approach in terms of noun compound recognition. The proposed statistical measures consists of NC-value, NTC-value and NLC-value. The experimental results have demonstrated that NLC-value has outperformed NTC-value and NC-value regarding to nested noun compound extraction by achieving 83%, 76%, 72% and 65% for bigram, trigram, 4-gram and 5-gram respectively.

Keywords: *Multi-word Expressions, Noun Compound, Nested Noun Compound, Statistical method, POS tagging*

1. INTRODUCTION

Noun compound (NC) is a sequence of two or more nouns which are associated with space such as “bus station”, without space such as “football” or with hyphenated such as “water-bottle” [1]. Arabic language is one of the most popular languages that contain several kinds of NCs. One of these kinds is Nested Noun Compound (NNC) which consists of multiple NCs whether it is associated such as “enterprise resource planning” which could be divided into “enterprise resource” and “resource planning”, or it may formed as non-associated such as “United States Secretary of Transportation” which could be divided into “United States” and “Secretary of Transportation” [2]. This scenario is occurring frequently in Arabic Language where nested noun compound is consisting of 3 words, 4 words or 5 words. Extracting Arabic NNC is challenging task in terms of obtaining accurate results. Hence, this study attempts to improve the effectiveness of Arabic nested noun compound extraction by proposing a hybrid of linguistic and statistical methods.

Many methods have been proposed in order to identify Arabic Noun Compounds. These methods have used linguistic approaches, statistical approaches and hybrid approaches in order to extract bi-gram and tri-gram. Nonetheless, Arabic contains a type of NCs which is called nested noun compound (NNC). Such type contains multiple NCs which make the extraction process more complicated. The process occur because of the need

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

of identifying more than bi-gram noun compounds such as 'الطاقة الذرية' which means *atomic energy* to 5-gram noun compounds such as' أعضاء الوكالة الدولية للطاقة الذرية' which means *International atomic energy agency members*. El-Balushi et al. [3] have proposed a hybrid method of linguistic and statistical approaches for extracting Arabic NNC including bi-gram, tri-gram, 4-gram and 5-gram candidates. Nonetheless, this method has a low performance due to some limitations such as the following:-

The first aspect is the linguistic patterns where it consist a number of limited a number of linguistic patterns (*e.g. noun-adjective*, *noun-noun*) are used. Second, is the statistical aspect where the statistical measures (NC-value , LLR , PMI) which have been proposed by El-Balushi et al. [3] such as the extraction of the bi-gram and tri-gram NCs which are more effective than the extraction of the 4-gram and 5-gram NCs. Basically, such measures have an important element or features contain important separate features separately such as contextual information, term-hood and unit-hood. Hence, having the combination of these features in a single statistical measure can be a useful method which fits the case of NNCs. Therefore, this study attempts to propose a hybrid method of linguistic and statistical approaches that can overcome these limitations. The proposed linguistic approach contains a new linguistic pattern based on named entity list and the proposed statistical methods are extensions of NC-value which have been introduced by Al-Mahdaouy et al. [4].

2. RELATED WORK

Many method have been presented in order to identify Arabic Multi-word Expressions. Most of them have used statistical method, linguistic method or a combination. For example Attia [5] has developed a semi-automatic approach for identifying Arabic multi-word. In fact, the author proposed a customized two-side regular expression transducer in terms of tracing the lexical-side and then generating the morphological lexicon. This approach has the ability to identify fixed and semi-fixed multi-words but it is not able to identify the flexible expressions.

Boulaknader et al. [6] have developed a prototype for identifying Arabic multi-words by proposing a hybrid method of linguistic and statistical. The authors have used the POS tagger that have introduced by Diab et al. [7] to facilitate the process of filtering candidates of multi-words based on patterns including Nouns, Adjectives and Preposition. Moreover, the authors have taken into the consideration the inflectional variations such as plural and singular and the syntactic variations such as synonyms with the internal alteration. Finally, they have applied Log-Likelihood Ratio, Mutual Information and T-score statistical measures in terms of ranking the candidates.

Bounhas & Slimani [8] have presented a hybrid approach of linguistic and statistical methods for identifying Arabic multi-words using an Arabic specialized corpora. In fact, the authors have used a tool called AraMorph which has been introduced by Hajic et al. [9] in order to form syntactic rules from the morphological features. Then the authors have use a POS tagger of Diab [7] in terms of filtering the candidates based linguistic filter. Hence, a rule based approach have been applied based on the POS tagger and AraMorph in order to identify multi-words. Finally, they have applied a statistical LLR method in terms of ranking the candidates.

Attia et al. [10] have presented a three complementary methods in order to identify and validate multi-words using an Arabic corpora automatically. The authors have applied a cross-lingual correspondences asymmetry which extracts MWEs from Arabic Wikipedia [11] in order to generate MWEs candidates using a multi-lingual lexicon for named entities. Then, they have translated MWEs in English that have been extracted from Princeton WordNet (PWN) to Arabic in terms of validating the candidates. Finally, a hybrid method of POS tagging and Point-wise Mutual Information has been used in terms of generating MWEs candidates in unigram, bi-gram and tri-gram.

Saif & Aziz [12] presented a hybrid method of linguistic and statistical methods for identifying Arabic collocations from newspaper corpus. The authors have used lemmatization and POS tagging as a linguistic method for filtering and generating unigram and bigram candidates. Then, they have used LLR, PMI, chi-square and Enhanced Mutual Information as statistical methods in terms of ranking the candidates based on co-occurrence. They have demonstrated that LLR is outperforming the other statistical measures.

El Mahdaouy et al. [4] have proposed a combination of linguistic and statistical methods for identifying Arabic multi-words terms. The linguistic part consists of POS tagger of Diab et al. [7] in order to provide tags for each word which is the core of filtering candidates based on the linguistic patterns. Then the authors have used three

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

statistical methods which are NC-value, NTC-value (combination of T-score and NC-value) and NLC (combination of LLR and NC-value). They reported that NLC-value has outperformed the other methods.

Al –Balushi et al. [3] have proposed a hybrid method of linguistic and statistical methods in order to identify nested noun compound in Arabic. The authors have used lemmatization and POS tagging in order to facilitate the process of filtering candidates based on the linguistic patterns. Then they have applied three statistical measures which are LLR, PMI and NC-value in terms of ranking the candidates. They have reported that NC-value has outperformed the other statistical measures.

3. PROPOSED METHOD

The proposed method as shown in Fig. 1 contains several phases including; (i) the corpus that has been used, (ii) transformation tasks that aims to turn the data into an internal representation, (iii) pre-processing tasks which aims to eliminate the noisy data, (iv) candidate extraction which contains POS tagging, (v) statistical methods including NC-value, NTC-value and NLC-value, and finally (vi) the evaluation metric that has been used to evaluate the proposed method.

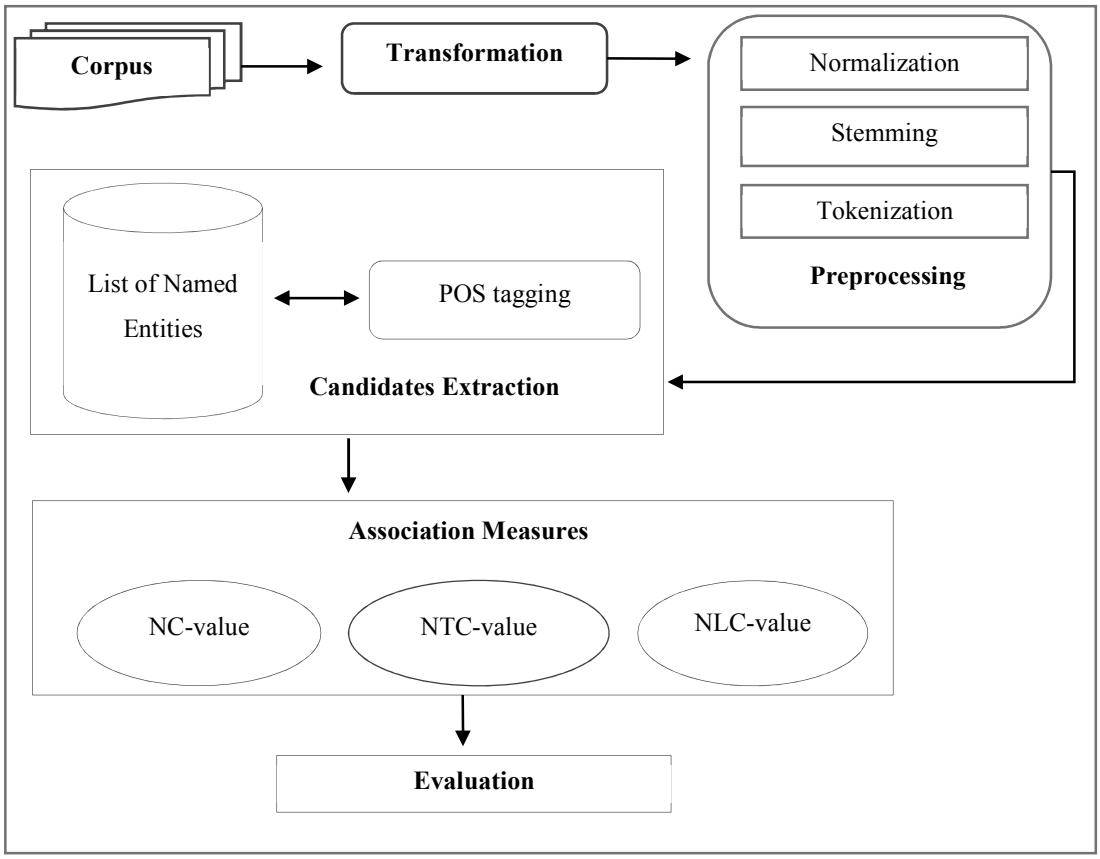

Figure1 Research Design of the Proposed Method

3.1. Corpus

The corpus that used in this study is a collection of an online Arabic newspaper which introduced by Saif & Aziz [12]. In fact, the corpus has been collected from two online news agencies which are Almotamar.net and Al-jazeera.net.

3.2. Transformation

This task aims to turn the data into an internal scheme that can enable processing. The key characteristic behind this task is to turn the files into UTF-8 encoding in order to recognize Arabic letters.

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

3.3. Preprocessing

This task aims to format the data in order to facilitate the process of identifying noun compounds. This has been performed within three sub-tasks; Normalization which aims to eliminate the unwanted and noisy data such as stop-words, punctuation and numerical values. Stemming which aims to retrieve the stem (root) of each word for instance, the word "*reading-*قراءة" would be stemmed into "*read-*قرأ". In this study, Buckwalter stemmer which has been introduced by Buckwalter [13] is used in order to stem the words. The algorithm works by splitting a certain input into prefix, stem and suffix and then match these segments with the corresponding lexicons [14]. Finally , Tokenization which aims to split words from text into different sets of consecutive morphemes, one of which typically corresponds to the word stem. For instance, "*United States of America*" after performing the tokenization it would be "*United_States_of_America*", or in Arabic the phrase '*مهرجان دبي السنوي*' which means *Dubai's annual festival* after tokenization it would be '*مهرجان_دبي_السنوي*'

3.4. Candidates Extraction

This phase relies on two main approaches which are Part-of-speech (POS) tagger and Named Entity list. These two approaches aim to generate n-gram lists of noun compound which will be filtered based on the linguistic patterns. Such approaches are illustrated as follows:

3.4.1 POS tagging

POS tagging is one of word sense disambiguation methods that aims to provide a tag such as noun, adverb, adjective or verb for each word in a given text [15]. There are many possibilities of tags for each words so that, the key characteristic lies behind POS tagger is the ability to assign each word with the exact tag. The POS tagger that has been used in the study is the one introduced by AlGahtani et al. [16] which is a transformation and learning-based algorithm.

In this study, POS tagger works within multiple tasks. Firstly, based on the linguistic pattern such as (Noun + Noun) a unigram list has been created. Every word from these list will be assigned with another word from the corpus that seems to be potential corresponding to form a noun compound. These possibilities will be stored with its linguistic tag and frequent occurrence in a list called 5-gram list. Sequentially, based on the POS tagger, the 5-gram list will be descent into several 4-gram noun compound possibilities. These possibilities will be stored with their POS tag and frequent occurrence. In the same manner, this 4-gram list will be descent into several tri-gram noun compound possibilities and stored with their tag and frequent occurrence in a list called tri-gram list. Finally, the POS tagging will descent this tri-gram list into several bi-gram noun compound possibilities and stored with their POS tag and frequent occurrence in a bi-gram list.

It aims to generate n-gram lists including bi-gram list, tri-gram list, 4-gram list and 5-gram list of noun compound candidates, then those lists have to be filtered according to the structural patterns. Firstly, it brought the words from the unigram list that have been acquired from the pre-processing phase. Each word will be assigned with a word from the corpus that seems to be possible combination with each other. These combinations will be stored using POS tagger with their linguistic category and frequent occurrences in a list called 5-gram. From the 5-gram list POS tagging will select a 4-gram combination that seems to be a candidate based on the linguistic structural patterns and store it in a list called 4-gram list with their linguistic category and frequent occurrences. Then, POS tagging will select from the 4-gram list a 3-gram combination that seems to be a candidate based on the structural patterns 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.

3.4.2 Proposed Named Entity (NE) pattern

This study proposed a new linguistic pattern called Named Entity pattern in order to enhance the process of identifying nested noun compounds in Arabic. Since a large portion of noun compounds are named entities such as "United Nation" thus, to facilitate the process of extracting noun compounds, a domain-specific list for named entities has been built which contains many types of names such as persons' names, locations' names and organizations' names. This approach has the ability to extract an exact noun compounds by checking the existence of these noun compounds from the proposed list. Note that, this list contains various forms of n-gram named entities including bi-gram, tri-gram. Table.1 shows a sample of the list contents.

Journal of Theoretical and Applied Information Technology
30th June 2015. Vol.76. No.3

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

Table .1 Sample of Named Entities

Word	Translation	Class	N-gram
نيو يورك	New York	Location	Bi-gram
مجلس التعاون الخليجي	Gulf Countries Council	Organization	Tri-gram
جبران خليل جبران	Gibran Khalil Gibran	Person	Tri-gram

Basically, the list approach will summarized the linguistic pattern used by the POS tagging. For instance, the noun compound of "Barack Obama president of United States" has a linguistic pattern of (N+ N + ADJ + PRE + N + N). Since, "Barack Obama" is a person's name and it is located in the list thus, it will be replaced with one tag which is Named Entity NE. This makes the linguistic pattern to become (NE + ADJ + PRE + N + N). Furthermore, since "United States" is a location's name which means it is located in the list therefore, it will be replaced with one tag which is NE. This makes the linguistic pattern to become (NE + ADJ + PRE + NE). This approach could significantly enhance the process of identification nested noun compound due to its ability to include several noun compounds in the meantime.

Since these named entities usually contain two or more words so the process of matching named entities would summarize and reduce the pattern. This makes the identification of nested noun compounds better. However, there are some named entities which enable swapping between the words for instance, the named entity "Google Company" enables the swapping of "company of Google" or in Arabic the named entity 'محمد النبي' which means Muhammad the prophet enables swapping of 'النبي محمد' which means the prophet Muhammad. This issue is challenging regarding to the accurate matching from the proposed list. This can be handled as the following:

Giving a swapping named entity S_{NE} that contains two words W_{S1} and W_{S2}, there is a list L where all the words of S_{NE} are stored separately. So that, for each word $W_i \in C$ where C is the corpus, a matching process will be took a place as follow:

$$when\ W_i \cong W_{Si} \in L \begin{cases} if\ W_{i-1} \cong W_{Si} \rightarrow (W_{i-1}, W_i)\ is\ S_{NE} \\ if\ W_{i+1} \cong W_{Si} \rightarrow (W_i, W_{i+1})\ is\ S_{NE} \end{cases}$$

3.5. Candidates Ranking

The process of ranking candidates of noun compound aims to select the best results which relying on frequency of occurrence and co-occurrence in the corpus. Since, the candidate extraction phase collects both syntactic information and information about their occurrence in the corpus. This phase aims to calculate the association measures for the extracted candidates in the n-gram lists where it assigns to each candidate a score of association strength. Each pair of words are extracted from the corpus to have a single real value which gives an indication on the amount of statistical association between the two words. In this study, three statistical measures have been used namely the NC-value, NTC-value and NLC-value which are illustrated as follows:

3.5.1 NC-value

NC-value is a statistical method which has been proposed Frantzi et al. [17]. It contains two major phases; first is identifying the C-value associated with measuring the term-hood for noun compound candidate as follow:

$$C-value(a) = \begin{cases} \log_2(|a|).f(a) & if\ a\ is\ not\ nested \\ \log_2(|a|).\left(f(a) - g(a)\right) & otherwise \end{cases}$$

Where $|a|$ refers to the length in words of candidate term a, while f(a) is the number of occurrences of a and:

$$g(a) = \frac{1}{|T_a|}\sum_{b \varepsilon T} f(b)$$

Where T(a) refers to the set of longer candidate terms into which a appears ($|T_a|$ is the cardinality of this set).

The second phase is combining the C-value with the contextual information (calculated by N-value)

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

which refers to the measure of the terminological status of a context for given candidate term. It is calculated as follows:

$$NC-value(w) = \alpha c-value(w) + (1-\alpha)\sum_{b\varepsilon C_w} f_w(b)weight(b)$$

Where Cw is the set of distinct context words of w, f (b) is the frequency of b as a MWE context word of w, weight (b) is the weight of b as a MWLU context word. Also, α is the weight assigned to the two factors of NC-value, and C-value.

3.5.2 NTC-value

This method has been introduced by Vu et al. [18] which combines the unit-hood feature based on T-score with NC-value in terms of enhancing the performance. It is defined as:

$$Ts\left(w_i, w_j\right) = \frac{P\left(w_i, w_j\right) - P(w_i).P(w_j)}{\sqrt{\frac{P(w_i, w_j)}{N}}}$$

Where $P\left(w_i, w_j\right)$ indicates the probability of w_i, w_j in the corpus, while $P(w_i)$ is the probability of word w_i.

Hence, the integrated with NC-value can be obtained by:

$$NTC\ (a) = 0.8\ .TC(a) + 0.2\ .Nvalue(a)$$

3.5.3 NLC-value

This method is a combination of Log-likelihood Ratio (LLR) which introduced by [19] with NC-value. This could provide more accurate unit-hood regarding to the ability of LLR to recognize the actual co-occurrence (whether it happens by chance or not). It is defined by:

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

Hence, the integrated NLC can be defined as:

$$NLC\ (a) = 0.8\ .LC(a) + 0.2\ .Nvalue(a)$$

3.6 Evaluation

The evaluation which has been used in this study is n-best method where it has been introduced by Evert [20] in terms of evaluating the proposed method. Basically, this method consists of three phases. The first phase is the n-best selection which collects the greatest value of association for the candidate ranking. Second step is the annotation which held by 10 experts in Arabic language where the correct noun compound is manually annotated with one and the incorrect noun compound is annotated with zero. Finally, the calculation of the precision for the annotated noun compounds based on the following equation:

$$Precision = \frac{TP}{TEC}$$

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

4. RESULTS

In this section, the results of the three association measures which are NC-value, NTC-value and NLC-value are identified based on two aspects. First is the results of such association measures without using the proposed NE pattern and the second is with using the proposed pattern. Table. 2 shows the two aspects' results.

Table . 2 Results of The Association Measures

Without NE pattern				
Association	**2-gram**	**3-gram**	**4-gram**	**5-gram**
NC-value	0.81	0.59	0.29	0.18
NTC-value	0.82	0.70	0.65	0.55
NLC-value	0.83	0.72	0.68	0.60
With NE pattern				
Association	**2-gram**	**3-gram**	**4-gram**	**5-gram**
NC-value	0.81	0.65	0.38	0.30
NTC-value	0.82	0.74	0.69	0.59
NLC-value	0.83	0.76	0.72	0.65

As shown in Table. 2, the greatest value of precision has been obtained by bi-gram candidates for both of using NE pattern and without using which reported 81%, 82% and 83% for NC-value, NTC-value and NLC-value respectively. Moreover, NLC-value has outperformed the other association measures in terms of bi-gram, tri-gram, 4-gram, 5-gram, with and without using NE pattern.

5. DISSCUSION

To sum up, NLC-value has outperformed NC-value and NTC-value in terms of extracting bi-gram and tri-gram candidates which is expected from the

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

study of El Mahdaouy et al.[4]. However, this study has demonstrated a similar outperformance for NLC-value compared with NC-value and NTC-value in terms of identifying nested noun compounds including Bi-gram, Tri-gram, 4-gram and 5-gram. This is due to NLC-value is a combination of LLR and NC-value which in other word means a combination of multiple features which are contextual information, unit-hood and term-hood features. Contextual information provides a measure of the terminological status of the context of a given candidate term. Whereas, unit-hood feature provides the degree of strength for combinations or

collocations [21]. Finally, the term-hood treats the terms as a linguistic unit [18]. These features have the ability to enhance the process of identifying nested noun compound in Arabic language.

On other hand, using the proposed linguistic pattern of Named Entity has the ability to enhance the process of extracting nested noun compounds for NC-value, NTC-value and NLC-value by facilitating the task of recognizing named entities which usually occurred as noun compounds.

In order to clarify the enhancement, a comparison with a baseline has to be taken place. The baseline of this research is the study of Al –Balushi et al. [3] where a hybrid method of linguistic and statistical approaches has been proposed in terms of identifying nested noun compound. Table. 3 shows the experiment results for the baseline and for the proposed method of this study.

Table .3 Comparison With A Baseline

Baseline results				
Association	**Bi-gram**	**Tri-gram**	**4-gram**	**5-gram**
NC-value	0.81	0.59	0.29	0.18
LLR	0.82	0.13	0.08	0.06
PMI	0.80	0.14	0.09	0.05
The proposed method				
Association	**Bi-gram**	**Tri-gram**	**4-gram**	**5-gram**
NC-value	0.81	0.65	0.38	0.30
NTC-value	0.82	0.74	0.69	0.59
NLC-value	0.83	0.76	0.72	0.65

As shown in Table 3, the baseline has obtained the greatest results with NC-value compared to LLR and PMI. In general, This study proposed statistical method which consists of NC-value, NTC-value and NLC-value without the consideration of applying the NE pattern has demonstrated that NTC-value and NLC-value have outperformed NC-value in terms of extracting noun compound which is expected from the study of El Mahdaouy et al. [4]. However, applying the NE pattern has reported an enhancement for NC-value, NTC-value and NLC-value in terms of identifying nested noun compounds including tri-gram, 4-gram and 5-gram by achieving 65%, 38% and 30% for NC-value, as well as 74%, 69% and 59% for NTC-value and 76%, 72% and 65% for NLC-value. Hence, the objectives of this study have been accomplished.

6. CONCLUSION

This study aims to address the problem of Arabic nested noun compound extraction by proposing a hybrid of linguistic approach and statistical methods. The linguistic approach consists of POS tagger and a proposed Named Entity pattern. Whereas, the statistical consists of three association measures which are NC-value, NTC-value and NLC-value. The experiment results have demonstrated an improvement with using NE pattern as well as, reported that NLC-value outperforms the other association measures. Note that, the Named Entity list that has been used in this study is restricted to a specific domain which is news. Hence, extending such list to cover various domains, may contribute toward enhancing the process of extraction for the nested noun compound.

7. REFERENCES

[1] Geoffrey K Pullum, "Syntactic Theory and English Syntax, Semester 1, 2007–2008," 2008Retrieved from.

[2] Katerina T Frantzi and Sophia Ananiadou, "Extracting nested collocations," in *Proceedings of the 16th conference on Computational linguistics-Volume 1*, 1996, pp. 41-46.doi:Retrieved from.

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

[3] Hamed Al-Balushi, Mohd Juziaddin Ab Aziz, k vidyavathi, rs sabeenian, k selvavinayaki, dre karthikeyan, s bhuvaneswari, ts subashini, p essaki muthu, and rmo gemson, "a hybrid method of linguistic approach and statistical method for nested noun compound extraction," *Journal of Theoretical and Applied Information Technology,* vol. 67, 2014Retrieved from.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Ibrahim Bounhas and Yahya 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:Retrieved from.

[9] 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.

[10] 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.

[11] Ar.Wikipedia. (2014). *Arabic Wikipedia*. Available: http://ar.wikipedia.org/wiki/

[12] 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.

[13] Tim Buckwalter, "Buckwalter {Arabic} Morphological Analyzer Version 1.0," 2002Retrieved from.

[14] 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.

[15] Roberto Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys (CSUR),* vol. 41, p. 10, 2009Retrieved from.

[16] 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.

[17] 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.

[18] Thuy Vu, Ai Ti Aw, and Min Zhang, "Term extraction through unithood and termhood unification," in *In Proc. of Int'l Joint Conf on Natural Language Proc*, 2008Retrieved from.

[19] 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.

[20] Stefan Evert, "The statistics of word cooccurrences," Dissertation, Stuttgart University, 2005Retrieved from.

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

[21] 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.