Volume 18(2), May 2018 http://doi.org/10.17576/gema-2018-1802-07

Arabic Nested Noun Compound Extraction Based on Linguistic Features and Statistical Measures

Nazlia Omar
nazlia@ukm.edu.my
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia

Qasem Al-Tashi
Qasemacc22@gmail.com
Computer and Information Sciences,
Universiti Teknologi PETRONAS

ABSTRACT

The extraction of Arabic nested noun compound is significant for several research areas such as sentiment analysis, text summarization, word categorization, grammar checker, and machine translation. Much research has studied the extraction of Arabic noun compound using linguistic approaches, statistical methods, or a hybrid of both. A wide range of the existing approaches concentrate on the extraction of the bi-gram or tri-gram noun compound. Nonetheless, extracting a 4-gram or 5-gram nested noun compound is a challenging task due to the morphological, orthographic, syntactic and semantic variations. Many features have an important effect on the efficiency of extracting a noun compound such as unit-hood, contextual information, and term-hood. Hence, there is a need to improve the effectiveness of the Arabic nested noun compound extraction. Thus, this paper proposes a hybrid linguistic approach and a statistical method with a view to enhance the extraction of the Arabic nested noun compound. A number of pre-processing phases are presented, including transformation, tokenization, and normalisation. The linguistic approaches that have been used in this study consist of a part-of-speech tagging and the named entities pattern, whereas the proposed statistical methods that have been used in this study consist of the NC-value, NTC-value, NLC-value, and the combination of these association measures. The proposed methods have demonstrated that the combined association measures have outperformed the NLC-value, NTC-value, and NC-value in terms of nested noun compound extraction by achieving 90%, 88%, 87%, and 81% for bigram, trigram, 4-gram, and 5-gram, respectively.

Keywords: Arabic multi-word expressions; noun compound; nested noun compound; association measures; POS tagging

INTRODUCTION

Noun compound (NC) is a phrase that is made up of a combination of two or more nouns which are sometimes joined together: for example, the words ‘tooth’ and ‘paste’ are both nouns and if they are connected to each other they produce a new word “toothpaste”. Occasionally, compound nouns appear as two separate words such as “Christmas tree”; sometimes they are joined using a hyphen such as “father-in-law” (Albared et al., 2016). Currently, Arabic text has rapidly increased over the Internet whether in social media, news agencies or advertisements. Hence, extracting these noun compounds to meaningful information is an essential demand. Noun compounds (NCs) frequently appear in Arabic text, which make the extraction of these nouns an important role in the field of Information Extraction. Extraction of Arabic NCs is one of the challenging tasks in natural language

eISSN: 2550-2131
ISSN: 1675-8021

processing(NLP) where in Arabic the words do not have capital or small letters. Moreover, Arabic NCs may contain semantic ambiguity, for example, "باب اليمن" which means "the door of Yemen" and also a famous place in Yemen. This may lead to a misunderstanding when attempting to identify this noun compound. On the other side, most Arabic noun compounds rely on the occurrences of two or more words together such as "خبر عاجل" which means "breaking news". Such two words frequently occur together rather than with synonyms of them such as "خبر طارئ" which means "emergency news". Finally, there is a lack of available resources of Arabic noun compound lexicon. To overcome this limitation we need to extract those compound nouns to process it further.

The identification and extraction of noun compounds have been widely researched by a number of researchers. For example, the research by (Buckeridge & Sutcliffe, 2002) proposes that the modifier and the head should be nouns. One of the most popular languages is the Arabic language which also contains several kinds of NCs. Nested noun compound (NNC) is one of these kinds, which in turn consists of multiple NCs (Salehi, 2016). It may consist of two to five words. Moreover, the NNC is a word that is used very frequently where new NNC is created to describe the exact meaning of the language terms. It is difficult to identify those nested noun compounds manually, due to the high cost and time. The problem with the NNC is that it relies on its frequent occurrences within the text. Extracting these noun compounds are important for numerous domains of research such as Information Retrieval, Sentiment Analysis, and Question Answering, and seems to be a challenging issue (Korayem, Crandall & Abdul-Mageed, 2012).

There are some differences between Arabic and English noun phrases. While English has both definite and indefinite articles and both occur before the noun, in Arabic there is only the definite article *al* 'the' but no overt indefinite article. English and Arabic differ with regard to the position of ordinals in the noun phrase. Ordinals can only precede the headword of a noun phrase but in Arabic they can also follow the headword in the structure of a noun phrase. In English adjectives are not inflected for number and gender but in Arabic they are. In Arabic, the number and the gender of the possessor has an impact on the form of the headword (kitab-u-hu كتابه 'his book', kitab-u-ha كتابها 'her book', kitab-u-hum كتابهم 'their book') but in English the number and the gender has no impact on the form of the headword at all, e.g., his book, her book or their book. Therefore, this makes Arabic noun compound extraction more challenging compared to English.

Identifying Arabic noun compounds is an important issue in NLP. It is a necessity to automatically extract them before they are translated to other languages or used for various tasks or applications. Much research has been conducted on this issue and proposes many methods for identifying NCs (Hazaa, Omar, Ba-Alwi & Albared, 2016). Some research have used linguistic pattern methods, statistical methods, or a combination of both with a view to the extraction of bi-gram and tri-gram NCs. Nevertheless, the Arabic language comprises a multiple of noun compounds named the nested noun compound, which makes the process of extraction more difficult. The process occurs because of the need to extract more than two noun compounds such as "رئيس الوزراء", which means Prime Minister to five compounds such as "رئيس الوزراء أحمد بن دغر", which means Prime Minister Ahmed bin Daghr. A combination method of linguistic and statistical approaches has been proposed by (Al-Mashhadani & Omar, 2015) for extracting Arabic NNCs using one to five-gram candidates. However, this method has some limitations in terms of accuracy. The limitations can be stated as follow: First, there is a lack of extraction in terms of relying on just POS tagging alone where in such, the patterns are not accurate. Second, the statistical method where the association measures NC-value, NTC-value, and NLC are proposed by (Al-Mashhadani & Omar, 2015) lacks in terms of accuracy. Arabic language has numerous types of NCs which have been associated with complexities regarding to the morphological differences that lie in Arabic. Based on

Bounhas and Slimani (2009), Arabic NCs have five classes which are described as follows:

i. **Annexation Constituent (مركب إضافي):** this type contains two parts. The first part is called annexed "المضاف" and the second part is called annexer "المضاف اليه". An for this type of constituent is a person name "عبد العليم " which means " Abdu Al-Aleem".
ii. **Adjective Constituent (مركب وصفي):** this type consists of noun and adjective which are connected with each other such as 'قاسم سليم' which means'Qasem Saleem'.
iii. **Prepositional compound noun (أسماء مركبة مجرورة):** this type contains two nouns that are connected by a preposition. Some example includes 'أحمد في الجامعة' which means 'Ahmed at the University' and 'العصفور فوق الشجرة' which means 'The bird is above the tree'
iv. **Conjunctive compound noun (أسماء مركبة موصوله):** this type contains two nouns that are connected by a conjunction such as 'العلم و الإيمان' which means 'Science and faith' and 'الذكر و الانثى' which means 'male and female'.
v. **Compound nouns linked by composite relations (أسماء مركبة مجموعة العلاقات):** this type of NC contains two nouns that are connected by two or more prepositions, conjunctions or both of them. For example, the phrase 'شاب وفيه شيب' means ' A young man with white hair'.

Several approaches are introduced with a view to identify Arabic multi-word expressions. A wide range of them use the linguistic approach, statistical approach, or a combination of both. For example, Attia et al. 2010 have introduced three integral approaches with a view to automatically identify and evaluate multi-words for the Arabic dataset. In this study, a cross-lingual consistency asymmetry has been applied, aiming to extract multi-words from the Arabic Wikipedia (Ar.Wikipedia), with a view to generate multi-word candidates based on a multi-lingual lexicon for the named entities. Following this, the English multi-words that have been extracted from the Princeton WordNet have been translated to Arabic in order to validate the candidates. Lastly, a point-wise mutual information and POS tagging hybrid method has been used in order to generate multi-word candidates in unigram, bigram, and tri-gram.

Saif and Aziz (2011) propose a combination of linguistic and statistical methods in order to identify Arabic collocations from a newspaper dataset. Lemmatisation and POS tagging are used as a linguistic approach in order to generate and filter unigram and bigram candidates. Following this, the authors use statistical methods consisting of the following association measurers: PMI, chi-square, LLR and improved mutual information. The ranking process of the candidates used is based on co-occurrence. Lastly, the authors conclude that LLR is outperforming the other association measures.

Mahdaouy, Ouatik and Gaussier (2014) present a hybrid method of linguistic and statistical approaches in order to identify Arabic multi-words. First, the authors use a POS tagger as a linguistic pattern borrowed from Diab, Hacioglu and Jurafsky (2004) with a view to assign tags for every word which is essential for filtering candidates. Second, they use three statistical measures, namely NC-value, NTC-value, and NLC. Lastly, the authors illustrate that the NLC-value has outperformed the other measures.

Al-Balushi et al. (2014) present a combination of linguistic and statistical approaches with a view to detect the Arabic nested noun compound. First, lemmatisation and POS tagging are used as linguistic patterns to enable the process of filtering candidates. Second, in order to rank the candidates, the authors use three association measures: LLR, PMI, and NC-value. Lastly, the authors demonstrate that NC-value has outperformed the other association measures.

Al-Mashhadani and Omar (2015) suggest a combination of linguistic and statistical

approaches with a view to extract the Arabic nested noun compound. First, normalisation, stemming, and tokenisation are performed as pre-processing tasks which work to remove unwanted data already used. Second, the authors apply candidate extraction which contains POS tagging. Third, the authors use three statistical methods introduced by (Mahdaouy et al., 2014) which are NC-value, NTC-value, and NLC. Lastly, the authors report that the NLC-value has outperformed NTC-value and NC-value with regards to nested noun compound extraction by achieving 83%, 76%, 72% and 65% for bigram, trigram, 4-gram and 5-gram, respectively. [Table 1] shows a summary of the related work.

TABLE 1. Summary of related work

Author	Year	Approach	Dataset	Pattern	Accuracy
(Attia et al.)	2010	cross-lingual, translation-based, corpus-based and statistical (PMI)	Princeton WordNet & Ar. Wikipedia	Noun compound	71%
(Saif et al.)	2010	combination of linguistic (POS, lemmatization) and statistical (LLR, PMI, EMI, chi-square)	newspaper dataset in Arabic	Arabic collocations	83%
(El Mahdaouy et al.)	2013	combination of linguistic (POS tagging) and statistical NLC-value	environment dataset in Arabic	MWT	73%
(AL-Balushi et al. 2014)	2014	combination of linguistic approach and statistical approach	Newspaper dataset in Arabic	Nested Noun Compound	Bi-gram 81% Tri-gram 59% 4-gram 29% 5-gram 18%
(Al-Mashhadani et al.2015)		Hybrid of linguistic and statistical methods)	Arabic newspaper	Nested Noun Compound	Bi-gram 83% Tri-gram 76% 4-gram 72% 5-gram 65%

In short, according to Bounhas and Slimani (2009), identifying Arabic MWEs is a difficult task due to the complication that relies on semantic or syntactic of its morphological variations. Some example of the MWE variations includes graphical variants (the graphic alternations between the letters " ها" "ha'a" and "تاء مربوطه" "Ta'a marbutah"), inflectional variants (the number inflection of nouns, the number and gender inflections of adjectives and the definite article "ا ل" (al)), morphosyntactic variants (the synonymy relationship between two MWEs of different structures) and syntactic variants (the modifications of the internal structure of the base-term, without affecting the grammatical categories of the main item which remain identical). Moreover, the current methods have some limitations in terms of the extraction of nested noun compound. A wide range of the current approaches has been proposed to extract bi-gram and tri-gram candidates. However, there are two approaches that have been proposed in order to extract nested noun compound. The first approach is presented by Al-Balushi (2014). This method has some limitations which is described as follows; (i) the linguistic approach used is limited to simple linguistic patterns containing

Noun + Noun, Noun + Adjective and their extensions, (ii) the association measures that have been used are (LLR, PMI and NC-value) have a limitation regarding to their own features. The second approach has been proposed by Al-Mashhaddani et al. (2015). This method has some limitations in terms of its accuracy. The limitations can be stated as follows: (i) there is a lack of extraction in terms of POS tagging in which such patterns are not accurate. (ii) the statistical method where the following association measures: NC-value, NTC-value and NLC have been proposed by Al-Mashhadani et al. (2015) lacks in term of accuracy. Thus, this study aims to propose a combination method of POS tagging, named entity and the combination of the association measures that improve these limitations.

METHODOLOGY

Several phases are involved in the presented approach as shown in [Figure 1]. These phases include the following: (i) the dataset that is used in this study, (ii) the transformation phase that proposes to fit the data into an internal representation, (iii) the pre-processing phase which consists of two tasks: tokenisation, which intends to divide the words of the dataset into groups of consecutive morphemes; and normalisation, which works to remove the unwanted data, (iv) extraction of the candidate which also consists of two tasks: POS tagging, which aims to define the word categories such as verb, noun, or adjective; and named entity, which aims to improve the process of identifying Arabic nested noun compounds. Furthermore, the suggested method includes the procedure of detecting the noun compounds candidates using the n-gram model to produce bigram, tri-gram, 4-gram and 5-gram, (v) the association measures containing NC-value, NTC-value, and NLC-value, (vi) the combination mechanism of the three association measures, and lastly (vii) the evaluation of the presented method.

CORPUS

The dataset that is used in this study is presented by (Saif & Aziz, 2011), which is a collection of text files of two online Arabic newspapers, namely Al-jazeera.net and Almotamar.net. [Table 2] provides the numerical details about the Arabic corpus used. The distribution of the noun compounds across the corpus is also provided.

TABLE 2. Details of the corpus used

Name	Statistics Value
Size (MB)	12.3
Files	100
Words	2,325,152
Sentences	102
Bi-gram NCs	11151
Tri-gram NCs	4933
4-gram NCs	2352
5-gram NCs	239

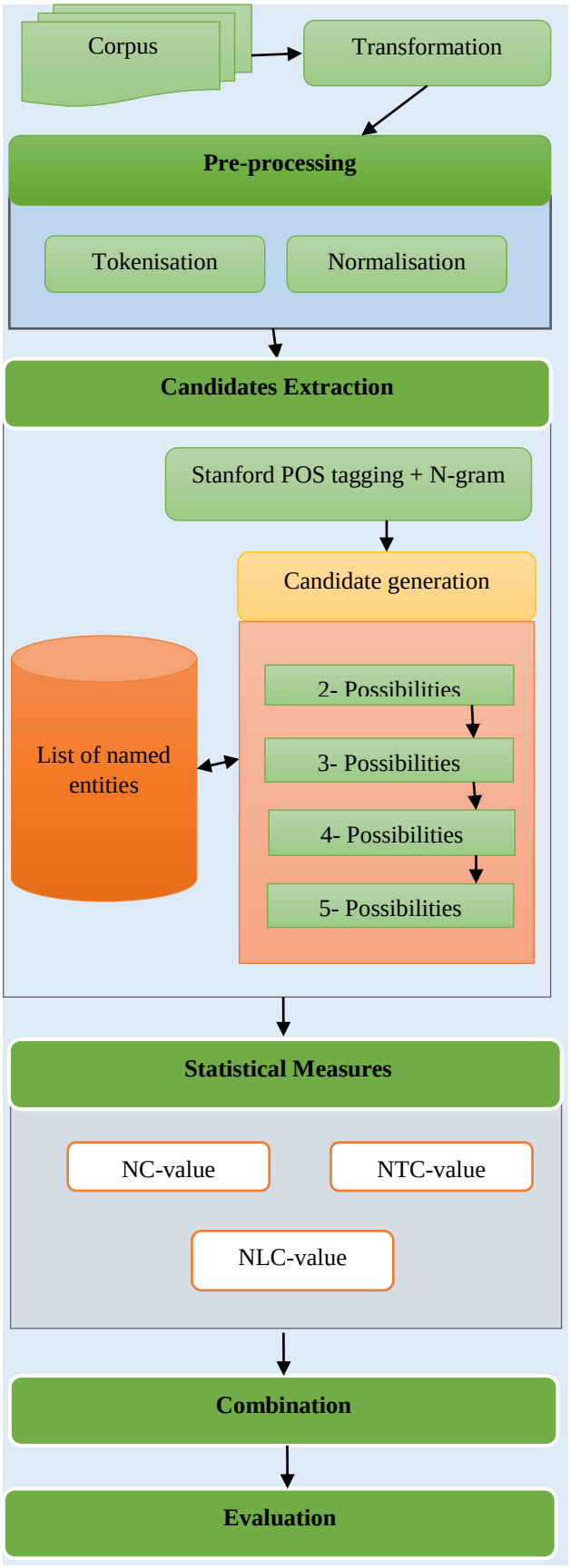

FIGURE. 1. Research design of the proposed method

TRANSFORMATION

The purpose of this phase is to convert the data into an internal illustration to obtain an accurate compiling that enables the application of the pre-processing steps. UTF-8 encoding has to be used in order to provide an illustration for the Arabic letters (Selamat & Ng, 2011) since the letters cannot be demonstrated by the ANSI code.

PRE-PROCESSING

The pre-processing phase performs numerous steps which aim to turn the data into a suitable format that allows the application of statistical measures by cleansing the dataset of unnecessary data. Therefore, two tasks of pre-processing phase are carried out, namely tokenisation and normalisation, and they are defined as follows:

TOKENIZATION

The task of tokenisation is to split words from text into sets of consecutive morphemes (Aliwy, 2012). For example, after applying the tokenisation process on *"United Arab Emirates"* it would become *"United_ Arab_ Emirates"*. Similarly, after applying the process of tokenisation on the Arabic phrase *'الثراث العربي الاسلامي'* which means Arab and Islamic heritage, it would become *'التراث_العربي_الاسلامي'* which means *"Arab_ Islamic_ Heritage"*.

NORMALIZATION

The aim of normalisation is to clean the data by excluding unwanted data such as digits or numbers, special characters, punctuation, and stop-words.

EXTRACTION OF CANDIDATES

Candidate extraction consists of two methods, namely Stanford POS tagger and a list of named entity. The goal of these two approaches is to produce a list of the n-gram of noun compounds in order to be clarified, depending on the linguistic patterns. These two methods are described as follows:

POS TAGGING

According to (Navigli, 2009), POS tagging is a method of word-sense disambiguation whose purpose is to assign tags such as adjective, noun, verb, or adverb for all the words in a text. There are a considerable amount of words that have numerous potential tags, and hence POS has been introduced with a view to disambiguate these words. Thus, the key characteristic of POS tagging lies in its ability to provide each word in the dataset with the exact tag. In this study, the Stanford POS tagging has been used. [Table 3] shows an example of Arabic POS tagging.

TABLE 3. Example of Arabic POS tagger

Word	Translation	POS
مدرسة	School	NOUN
يقراء	Read	VERB
صغير	Small	ADJ
سعيد	Happily	Adverb

Initially, a list of unigrams has been produced based on the linguistic pattern such as (Preposition + Adjective), which has been introduced by (Boujelben, Mesfar & Hamadou, 2010). Each word from this list will be allocated with another word from the dataset that

appears to be potentially consistent, to create a noun compound. These potentials will be kept with their linguistic tags and frequent occurrence in a list named the 5-gram list. Consecutively, relying on the POS tagger, the 5-gram list will be the ancestry leading into numerous 4-gram noun compound potentials. These potentials will be stored with their POS tags and frequent occurrence. Likewise, this 4-gram list will be the ancestry into numerous tri-gram noun compound potentials and stored with their tags and frequent occurrence in a list named the tri-gram list. Lastly, the POS tagging will descend this tri-gram list into numerous bi-gram noun compound potentials stored with their POS tags and frequent occurrence in a bi-gram list. Its purpose is to produce a list of n-gram holding bi-gram, tri-gram, 4-gram, and 5-gram lists of noun compounds. Subsequently, the lists have to be filtered based on the structural patterns. Initially, it fetches the words from the unigram list which was learnt during the pre-processing task. Every word will be allocated with a word from the dataset that it appears to have possible integration with. Using the POS tagger, these combinations will be stored with their linguistic classification and recurrent occurrences in a list named 5-gram. From the 5-gram list, POS tagging will choose a 4-gram combination that appears to be a candidate according to the linguistic structural patterns and store it in a list named the 4-gram list with its linguistic classification and recurrent occurrences. From the 4-gram list, POS tagging will previously choose a 3-gram combination that appears to be a candidate, relying on the structural patterns and store it in a list named the tri-gram list with the linguistic classification and recurrent occurrences. Likewise, the bi-gram list will be constructed from the 3-gram list.

NAMED ENTITY

The named entity pattern is proposed in order to improve the process of extracting Arabic nested noun compounds, which have been used by (Al-Mashhadani & Omar, 2015), since a great percentage of NCs are named entities – for example "Security Council". Therefore, to simplify the procedure of extracting noun compounds, a domain-specific named entities has been constructed which consists of various kinds of names (e.g. persons, locations and organisations). Thus, this method has the capability to extract accurate noun compounds by checking the availability of these compound nouns from the proposed list. Essentially, the list method will shorten the linguistic patterns that have been used by the Stanford POS tagging. For example, given the NC "Qaboos Said Sultan of Sultanate Oman", this compound noun has a linguistic pattern of (N+ N +ADJ + PRE + N + N). Meanwhile, "Qaboos Said" is located in the list and it is a person's name. Therefore, it will be swapped with one tag which is Named Entity (NE). Likewise, since "Sultanate Oman" is assigned in the list and it is a location, it will therefore be swapped with one tag which is NE. Thus, this method could sufficiently improve the process of extracting nested noun compounds because of its capability to contain numerous noun compounds. [Table 4] shows examples of patterns generated using Named Entity.

TABLE 4. Example of named entity patterns

Nested Noun Compounds	Translation	Pattern	NE pattern	N-gram
الرئيس باراك اوباما	President Barack Obama	N + N + N	N + NE	3-gram
الممثل الكوميدي ناصر القصبي	Comedian actor Naser Al- Qasabi	N + ADJ + N + N	N + ADJ + NE	4-gram
رئيس اتحاد الطلبة قاسم الطشي	Chancellor of Student Association Qasem Altashi	N + N + N + N + N	N + N + N + NE	5-gram

CANDIDATES RANKING

Candidate ranking phase aims to calculate the statistical measures for the candidates that have been extracted in the lists of n-gram which allocates each candidate a score of association strength (Ittoo & Bouma, 2013). The association measure that has been used, includes NC-value, NTC-value, NLC-value, and the combination of these three measures, where both term-hood and unit hood measures are considered. The definitions of these association measures are as follows:

NC-VALUE

NC-value has been proposed by (Frantzi, Ananiadou & Mima, 2000), whereas C-value is a statistical method that measures the term-hood of a candidate based on the following features: number of occurrences, term nesting, and term length. It is measured as:

$$C-value(a) = \begin{cases} \log_2(|a|).f(a) & if\ a\ is\ not\ nested \\ \log_2(|a|).\left(f(a)-g(a)\right) & otherwise \end{cases} \quad (1)$$

Where a indicates the length of the candidate term – for example, in the case of a non- nested noun compound such as 'الامعاء الغليظة', which means "large intestine". where a=2, it indicates that this phrase is a noun compound. But in case a is a nested noun compound, then we see an example such as "رئيس الوزراء اليمني عبد الكريم الارياني" which means "Yemeni Prime Minister Abdulkarim Alaryani", where a =6 while $|a|$=2 indicates " رئيس الوزراء", which means "Prime Minister". This part of the equation for noun compound takes other nouns to form a longer sentence.

|a| indicates the length of candidate term a in words, while $f(a)$ is the number of occurrences of a. For instance, if $f(a)$ = 12, it means the phrase "رئيس الوزراء اليمني عبد الكريم الارياني" has occurred 12 times in the dataset and:

$$g(a) = \frac{1}{|T_a|}\sum_{b\varepsilon T} f(b) \quad (2)$$

where T_a indicates the set of longer terms in which a appears ($|T_a|$ is the cardinality of this set).

For instance, the sentence "رئيس الوزراء اليمني عبد الكريم الارياني" has appeared 9 times and the sentence 'رئيس الوزراء اليمني' has appeared 17 times. Hence, a = "رئيس الوزراء اليمني", and T_a=9.

Furthermore, the NC-value combines the C-value together with contextual information which is calculated based on the N-value that indicates a measure of the terminological status of the context of a given candidate term. It is measured as:

$$Nvalue(a) = \sum_{b\in Ca} f_a(b).\ \frac{|T(b)|}{n} \quad (3)$$

Where Ca indicates the set of distinct context words of a, $f_a(b)$ indicates the number of times b occurs in the context of a and n is the total number of terms considered. This measure is then simply combined with the C-value to provide the overall NC-value measure:

$$NC-value(w) = \alpha c-value(w) + (1-\alpha)\sum_{b\varepsilon C_w} f_w(b)weight(b) \quad (4)$$

eISSN: 2550-2131
ISSN: 1675-8021

Where $f_w(b)$ is the frequency of b as a MWE context word of w, C_w is the collection of featured context words of w, weight (b) is the weight of (b) as an MWLU context word. Besides that, α is the weight assigned to the two factors of NC-value, and C-value.

NTC-VALUE

This method has been presented by (Vu, Aw & Zhang, 2008), which aims to combine the unit-hood feature based on the T-score with NC-value in order to improve the performance. It is measured as follows:

$$Ts\left(w_i, w_j\right) = \frac{P\left(w_i, w_j\right) - P(w_i).P(w_j)}{\sqrt{\frac{P(w_i, w_j)}{N}}} \quad (5)$$

Where $P(w_i, w_j)$ refers to the probability of the bigram (w_i, w_j) in the corpus, while $P(w_i)$ is the probability of word w_i.
For instance, if the two words "مدينة رداع" which means "Rada'a City" were applied using t-score, it would be:

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

Where $P\left(\text{رداع, مدينة}\right)$ is the probability of the two words, $P\left(\text{مدينة}\right)$ is the probability of the word "مدينة", $P(\text{رداع})$ is the probability of the word "رداع", and N is the total number of words in the dataset. After that, the T-score is combined in the NC measures through a re-weighting of the number of existences that privileges terms with a positive T-score:

$$F(a) = \begin{cases} f(a) & if\ \min\left(Ts(a)\right) \leq 0 \\ f(a)\ In(2 + \min\left(Ts(a)\right) & otherwise \end{cases} \quad (6)$$

Where $\min\left(Ts(a)\right)$ indicates the minimum T-score gained from all the word pairs in a. Replacing $F(a)$ to $f(a)$ in Eq. (1) produces the TC-value, which is then combined with the N-value as before, leading to the NTC-value:

$$NTC - value(a) = 0.8.TCvalue(a) + 0.2.Nvalue(a) \quad (7)$$

NLC-VALUE

NLC-value is a combination of NC-value which is proposed by (Frantzi et al., 2000), with LLR which is introduced by Dunning(Dunning, 1993). This could offer more exact unit-hood in terms of the capability of LLR to distinguish the actual co-occurrence. It is measured by:

$$\begin{aligned} LLR = 2((&alna + blnb + clnc + dlnd \\ &+ (a+b+c+d)\ln(a+b+c+d)) \\ &- \left((a+b)\ln(a+b) + (a+c)\ln(a+c)\right. \\ &\left. + (b+d)\ln(b+d) + (c+d)\ln(c+d)\right)) \end{aligned} \quad (8)$$

which leads to the NLC-value that integrates contextual information and both term-hood and unit-hood. The combined NLC can be illustrated as:

$$NLC\ (a) = 0.8\ .\ LC(a) + 0.2\ .\ Nvalue(a) \qquad (9)$$

COMBINATION

This method is a combination of the three association measures introduced above, where the NC-value, NTC-value and NLC-value are integrated together, which could lead to more accurate unit-hood in terms of the accuracy of the association measures used to provide the actual co-occurrence. The combination is introduced in two steps: first, the combination is made with 80% of the result of the NLC-value with 20% of the result of the NTC-value. It is defined as:

$$C(a) = 0.8\ .\ NLC(a) + 0.2\ .\ NTC(a) \qquad (10)$$

In the second step, the combination involves 80% of the result of the $C(a)$ value which indicates the combination of (the NLC-value with the NTC-value) as shown in Eq. (10) with 20% of the result of the NC-value. This leads to the combination-value that integrates contextual information and both term-hood and unit-hood:

$$Combination - value(a) = 0.8\ .\ C(a) + 0.2\ .\ NC(a) \qquad (11)$$

EVALUATION

The method of evaluation considered in this study is the n-best method which has been proposed by (Evert, 2005). Basically, three stages of this evaluation method have been used: The first is the n-best selection which gathers the highest value of association for the candidate ranking. The second stage is the annotation where the accurate noun compound is manually annotated with one and the incorrect noun compound is annotated with zero. Lastly, the precision calculation for the annotated noun compounds has been used according to the following equation:

$$Precision = \frac{TP}{TEC} \qquad (12)$$

Where TP is the number of correct noun compounds and TEC is the total number of extracted noun compounds.

RESULTS AND DISCUSSION

In this section, the results of the association measures, namely NC-value, NTC-value, NLC-value, and the combination for all of them, are identified. As shown in [Table 5] , it has been noticed that Bi-gram candidates have the greatest value of precision where the increasing of n-gram causes a decreasing of precision. This indicates the difficulty in extracting the accurate candidate when the n-gram is higher. On the other hand, the results when N = 100 are greater than the other values of N for all n-gram forms. This is because the possibilities of extracting incorrect candidates will increase where the process of identifying more than bi-gram candidates may result in the detection of invalid noun compounds. Obviously, Combination-value has outperformed NC-value, NTC-value and NLC-value due to the

combination between all three association measures. However, the greatest value of precision has been achieved when N = 100 with Bi-gram by obtaining 97%, while the lowest value of precision has been obtained when N = 500 with 5-gram by achieving 81%.

TABLE 5. Results of association measures

Association	Bi-Gram	Tri-Gram	4-Gram	5-Gram
NC-value	0.81	0.75	0.67	0.61
NTC-value	0.85	0.83	0.81	0.71
NLC-value	0.89	0.87	0.86	0.75
Combination- value	0.90	0.88	0.87	0.81

To summarise, the Combination-value has outperformed NLC-value, NTC-value, and NC-value in terms of the extraction of bi-gram and tri-gram. However, this study has demonstrated a similar performance for Combination-value compared with NLC-value, NTC-value, and NC-value in terms of the extraction of ANNCs involving Bi-gram, Tri-gram, 4-gram, and 5-gram. This is because Combination-value is a combination of NLC-value, NTC-value, and NC-value, which in other words means a combination of multiple features which are contextual information, unit-hood, and term-hood. Contextual information measures the terminological rank of a given candidate term. The unit-hood feature offers the degree of strength for combinations or collocations (Fahmi, 2005). Lastly, the term-hood treats the terms as a linguistic unit (Vu et al., 2008). These features have the ability to improve the extraction procedure of the nested noun compound in the Arabic language. Furthermore, using named entity as a linguistic pattern has the capability to improve the procedure of nested noun compounds extraction for Combination-value NLC-value, NTC-value, and NC-value, and in facilitating the task of recognising named entities which usually occur as noun compounds.

[Table 6] shows a sample result of bi-gram compound noun extraction based on the proposed method. The extracted candidate is the extracted compound noun. The combination value indicates the strong correlation between the two nouns extracted with the given pattern types. Similar sample results for tri-gram, 4-gram and 5-gram are shown in Table 7, 8 and 9 respectively.

TABLE 6. Sample results of Combination-value (Bi-gram)

Extracted Candidate	English Translation	Combination-value	Pattern
طوق الحمامة	The Ring of the Dove	15.79804188	NN ADJ
داود الأصفهاني	Dawood Al-Isfahani	12.78423524	NN ADJ
مكة المكرمة	Makkah Al Mukarramah	12.22951798	NN ADJ
موضوع الدراسة	Subject of study	11.17934141	NN ADJ

TABLE 7. Sample results of Combination-value (Tri-gram)

Extracted Candidate	English Translation	Combination-value	Pattern
ألفريد آدلر هناك	Alfred Adler there	19.49253276	NN NN NN
بروكلمان كراتشكوفسكي بتروف	Brockelman Kratzkowski Petrov	14.68157034	NN NN NN
ترجمة عادل نجيب	Translation Adel Najib	12.67826747	NN NN NN
داود الأصفهاني كتابه	Dawood Al-Isfahani Book	12.22951798	NN ADJ NN

eISSN: 2550-2131
ISSN: 1675-8021

TABLE 8. Sample results of Combination-value (4-gram)

Extracted Candidate	English Translation	Combination-value	Pattern
الدكتور خالد محمد القاضي	Doctor Khaled Mohd Al-Qadhi	19.49253276	NN NN NN ADJ
ترجمة عادل نجيب بشرى	Adel Najeeb Bushra Translation	15.79804188	NN NN NN NN
صفوت عبد الله الخطيب	Safwat Abdu Allah Al-Khatib	14.68157034	NN NN NN ADJ
فيلمز وزعته الأخيرة بالاشتراك	Films distribution last partnership	12.66987978	NN NN ADJ NN

TABLE 9. Sample results of Combination-value (5-gram)

Extracted Candidate	English Translation	Combination-value	Pattern
إسرائيل الولايات المتحدة المؤلفه جايمس	Israeli United States author James	19.49253276	NN ADJ ADJ NN NN
فاروق حسني الأمين العام للمجلس	Farouk Hosni Secretary General of the Council	15.79804188	NN NN ADJ ADJ NN
عبد الله بن الهداد العثماني	Abdu Allah bin Al-Haddad Al-Ottomani	13.18141536	NN NN NN ADJ ADJ
عالمتي الفلك القديرتين كانيس وأثيريا	The majestic astronomers Canis and Athria	12.4904649	NN ADJ ADJ NN NN
مخرج الفيلم التسجيلي جورج سكوت	Documentary film director George Scott	11.50963248	NN ADJ ADJ NN NN

With a view to clarify the improvement, a comparison with related work or baseline is performed. The related work to this research is the study of (Al-Mashhadani & Omar, 2015). The results shows that the proposed method clearly outperformed the work of (Al-Mashhadani & Omar, 2015). Here, a combination of Stanford POS tagging, named entity, and a combined association measures has been proposed in order to identify the nested noun compound. Table [10] shows the experiment results for the introduced method of this study and the related work.

TABLE 10. Comparison with Baseline

Baseline results (learning-based algorithm POS tagging)				
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
Proposed method (Stanford POS tagging)				
Association	**Bi-gram**	**Tri-gram**	**4-gram**	**5-gram**
NC-value	0.81	0.75	0.67	0.61
NTC-value	0.85	0.83	0.81	0.71
NLC-value	0.89	0.87	0.86	0.75
Combination- value	0.90	0.88	0.87	0.81

CONCLUSION

This study proposed a combination of linguistic and statistical methods for the extraction of Arabic nested noun compounds. The linguistic approach consists of POS tagging, which permits the process of selecting candidates depending on the word categories and linguistic patterns, and the named entity pattern which uses a list of Arabic named entities. The presented statistical approach consists of the following association measures: the

combination-value, NLC-value, NTC-value, and NC-value. The experimental results have been evaluated using the n-best method and have been compared with the related work. Essentially, Combination-value has outperformed the other three association measures in terms of identifying Arabic nested noun compounds. This research demonstrates that extraction of the nested noun compounds in Arabic especially the 4-gram such as, ' الممثل الكوميدي ناصر القصبي' which means 'Comedian actor Naser Al- Qasabi', and 5-gram such as 'مخرج الفيلم التسجيلي جورج سكوت' which means 'Documentary film director George Scott' can be improved using the proposed combination methods. The automatic extraction of the compound nouns may assist many language processing tasks such machine translation, named entity recognition and question answering. Moreover, extracting such noun compounds helps the field of Arabic language studies in term of discovering the different types of nested noun compound that exists in numerous linguistic patterns.

REFERENCES

Al-Balushi, H., Ab Aziz, M. J., Vidyavathi, K., Sabeenian, R. S., Selvavinayaki, K., Karthikeyan, D. R. E., … Others. (2014). A Hybrid Method Of Linguistic Approach And Statistical Method For Nested Noun Compound Extraction. *Journal of Theoretical and Applied Information Technology. Vol. 67*(3).

Al-Mashhadani, M. & Omar, N. (2015). Extraction Of Arabic Nested Noun Compounds Based On A Hybrid Method Of Linguistic Approach And Statistical Methods. *Journal of Theoretical and Applied Information Technology*. *Vol. 76*(3), 408-416.

Albared, M., Al-Moslmi, T., Omar, N., Al-Shabi, A. & Ba-Alwi, F. M. (2016). Probabilistic Arabic Part Of Speech Tagger With Unknown Words Handling. *Journal of Theoretical and Applied Information Technology. Vol. 90*(2), 236.

Aliwy, A. H. (2012). Tokenization as Preprocessing for Arabic Tagging System. *International Journal of Information and Education Technology. Vol. 2*(4), 348.

Attia, M., Tounsi, L., Pecina, P., van Genabith, J. & Toral, A. (2010). Automatic Extraction Of Arabic Multiword Expressions. In *Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010)*, pages 19–27,. Beijing, August 2010

Boujelben, I., Mesfar, S. & Hamadou, A. Ben. (2010). Arabic Compound Nouns Processing: Inflection And Tokenization. In *Proceedings of Nooj International Conference* (p. 40).

Buckeridge, A. M. & Sutcliffe, R. F. E. (2002). Disambiguating Noun Compounds With Latent Semantic Indexing. In *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14* (pp. 1-7).

Diab, M., Hacioglu, K., & Jurafsky, D. (2004). Automatic Tagging Of Arabic Text: From Raw Text To Base Phrase Chunks. In *Proceedings of HLT-NAACL 2004: Short papers* (pp. 149-152).

Dunning, T. (1993). Accurate Methods For The Statistics Of Surprise And Coincidence. *Computational Linguistics. Vol. 19*(1), 61-74.

Evert, S. (2005). The Statistics Of Word Cooccurrences: Word Pairs And Collocations. PhD thesis, University of Stuttgart.

Fahmi, I. (2005). C-Value Method For Multi-Word Term Extraction. In *Seminar In Statistics And Methodology*.

Frantzi, K., Ananiadou, S. & Mima, H. (2000). Automatic Recognition Of Multi-Word Terms:. The C-Value/Nc-Value Method. *International Journal on Digital Libraries. Vol. 3*(2), 115-130.

Hazaa, M. A. S., Omar, N., Ba-Alwi, F. M., & Albared, M. (2016). Automatic Extraction of Malay Compound Nouns using a Hybrid of Statistical and Machine Learning

Methods. *International Journal of Electrical and Computer Engineering. Vol. 6*(3), 925-926.

Ittoo, A. & Bouma, G. (2013). Term Extraction From Sparse, Ungrammatical Domain-Specific Documents. *Expert Systems with Applications. Vol. 40*(7), 2530-2540.

Korayem, M., Crandall, D. & Abdul-Mageed, M. (2012). Subjectivity and Sentiment Analysis of Arabic: A Survey. *Cs.indiana.edu*, 1–10.

Mahdaouy, A. El, Ouatik, S. E. L. & Gaussier, E. (2014). A Study of Association Measures and their Combination for Arabic MWT Extraction. *arXiv Preprint arXiv:1409.3005*.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR). Vol. 41*(2), 10.

Saif, A. M. & Aziz, M. J. A. (2011). An Automatic Collocation Extraction From Arabic Corpus. *Journal of Computer Science. Vol. 7*(1), 6-7.

Salehi, B. (2016). Flexible Language Independent Multiword Expression Analysis.

Selamat, A. & Ng, C.-C. (2011). Arabic Script Web Page Language Identifications Using Decision Tree Neural Networks. *Pattern Recognition. Vol. 44*(1), 133-144.

Vu, T., Aw, A. T. & Zhang, M. (2008). Term Extraction Through Unithood And Termhood Unification. In *In Proc. of International Joint Conference on Natural Language Processing*, 631-636.

ABOUT THE AUTHORS

Nazlia Omar is currently an Associate Professor at the Centre for AI Technology, Faculty of Information Science and Technology, Universiti Kebangsaan, Malaysia (UKM). She holds her PhD from the University of Ulster, UK. Her main research interest lies in the area of Natural Language Processing and Computational Linguistics.

Qasem Al-Tashi is currently doing his PhD on Information technology at the Universiti Teknologi PETRONAS. He holds his Master from Universiti Kebangsaan, Malaysia (UKM). He attained his bachelor's degree in computer science at University Technology Malaysia (UTM). His main research interest is in the area of Artificial Intelligence, Swarm Intelligence, Feature Selection, Natural Language Processing and Data Mining.