International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 3, June 2016, pp. 925 ~ 935
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i3.9663

Automatic Extraction of Malay Compound Nouns Using A Hybrid of Statistical and Machine Learning Methods

Muneer A. S. Hazaa[1], Nazlia Omar[2], Fadl Mutaher Ba-Alwi[3], Mohammed Albared[3]
[1] Faculty of Computer Science and Information Technology, Thamar University, Yemen
[2] University Kebangsaan Malaysia, Faculty of Information Science and Technology
[3] Faculty of Computer and Information Technology, Sana'a University, Yemen

Article Info

Article history:

Received Oct 12, 2015
Revised Dec 7, 2015
Accepted Dec 21, 2015

Keyword:

Association Measures
Classification Algorithms
Compound Nouns
Malay Language

ABSTRACT

Identifying of compound nouns is important for a wide spectrum of applications in the field of natural language processing such as machine translation and information retrieval. Extraction of compound nouns requires deep or shallow syntactic preprocessing tools and large corpora. This paper investigates several methods for extracting Noun compounds from Malay text corpora. First, we present the empirical results of sixteen statistical association measures of Malay <N+N> compound nouns extraction. Second, we introduce the possibility of integrating multiple association measures. Third, this work also provides a standard dataset intended to provide a common platform for evaluating research on the identification compound Nouns in Malay language. The standard data set contains 7,235 unique N-N candidates, 2,970 of them are N-N compound nouns collocations. The extraction algorithms are evaluated against this reference data set. The experimental results demonstrate that a group of association measures (T-test , Piatersky-Shapiro (PS) , C_value, FGM and rank combination method) are the best association measure and outperforms the other association measures for <N+N> collocations in the Malay corpus. Finally, we describe several classification methods for combining association measures scores of the basic measures, followed by their evaluation. Evaluation results show that classification algorithms significantly outperform individual association measures. Experimental results obtained are quite satisfactory in terms of the Precision, Recall and F-score.

Corresponding Author:

Muneer A.S. Hazaa,
Faculty of Computer and Information Technology,
Dhamar University, Yemen.
Email: muneer_hazaa@yahoo.com

1. INTRODUCTION

Compound nouns are a commonly occurring construction in natural languages. Compound nouns are made up of two or more nouns which together function syntactically as single noun such as 'golf club 'or 'computer science'. The compound noun syntax and semantics are discussed in details in Levi [1]. Compound nouns which consist of two words are analyzed syntactically by means of the rule $N \rightarrow N\ N$ or the rule $N \rightarrow N\ N^{-}$ applied recursively. Compounds of more than two nouns are ambiguous in syntactic structure.

Noun-Noun compounds, as a subset of compound nouns, characteristically occur with high frequency and high lexical and semantic variability [2]. Noun compounds (or NCs) have received a significant deal of attention in recent years in computational linguistic literature. Identification of compound noun Multiword Expression (MWE) and understanding their syntax and semantics is difficult but important for many Natural Language Processing (NLP) applications, particularly parsing, and dictionary-based

applications like machine translation [3] and question answering [4],[5]. Extracting Malay compound nouns is challenging task in terms of obtaining accurate results. Hence, this study attempts to improve the effectiveness of Malay noun compound extraction by proposing a hybrid of statistical and machine learning methods. The main activity in our research work is to observe and find an acceptable technique to extract a pair of compound nouns in Malay.

Compounds have thus been a recurrent focus of attention within theoretical, cognitive, and in the last decade also within computational linguistics. Considerable research has been proposed on automatic identification of multiword units and noun-noun compounds and on to classify semantic relationships between Compounds components. Most of these studies on noun-noun compounds only deal with English and some other languages but not much research have been carried at this level for Malay. Various Lexical association measures have been suggested in literature for identification of MWEs. These association measures are mathematical formulas that compute an association score between two or more words based on their occurrences and co-occurrences in a text corpus. The scores indicate the potential for a candidate to be a collocation. They can be used for ranking (candidates with high scores at the top), or for classification (by setting a threshold and discarding all bigrams below this threshold). An overview of the most widely used techniques is given in [6]-[10].

Compound nouns in Malay have been classified into three major types based on their syntactic structure, as discussed in [11].The syntactic structure of the first and the second categories is noun followed noun, for example “gunung-ganang ” (mountain) and “kapal layar” (sailing ship). For the third category, the syntactic structure is a noun followed by a noun word. The POS of the non-noun word can be a determiner, verb, adjective, adverb, preposition phrase or ordinal. In this study, our work is focused on the automatic extraction of the N-N Malay compound nouns multiword expression.

In this paper, first, several statistical association measures [7],[8],[12]-[15] have been investigated for the identification noun–noun compounds in Malay corpus. After that, we present an automatic noun–noun compounds extraction based on weighted combination of multiple lexical association measures lists. Finally, we describe several classification methods which uses association measures scores as their feature sets. Experiments presented in this paper were performed on Malay data and our attention was restricted to the first and second categories of Malay noun compounds.

This paper is organized as follows: In Section 2, we give a summary of related. Section 3 describes our Malay Noun compounds extraction methods. Section 4 presents the evaluation methods, the experimental results and discussion on the results. Finally, Section 5 concludes the study and gives some future work.

2. RELATED WORK

Several approaches have been proposed have been carried out regarding MWE in various languages like English, German and some other languages Generally speaking, these approaches can be divided into four mainstream methodologies: statistical approaches [9],[16],[17], linguistic methods [18],[19] and Hybrid Methods [16],[20],[21], and machine leaning methods [7],[8].

In statistical methods for MWE extraction, Church and Hanks [22] presented the concept of association measures firstly, and then proposed Mutual Information (MI) as an objective measure for estimating word association. Pecina (2005) present empirical evaluation of a comprehensive list of automatic collocation extraction methods (84 kinds of association measures for bigram collocation extraction) and concluded that in Czech data, MI has the best performance. Yoshida et al. [23] propose a new method (Enhanced Mutual Information and Collocation Optimization) to extract MWE from text. The results show that the new method significantly improves the performance of multiword expression extraction in comparison with a classic MI extraction method. Chakraborty [24] and Dandapat, Mitra et al. [25] have used statistical measurements to extract Noun-Noun (N-N) and Noun-Verb (N-V) collocations as MWE in Bengali Corpus respectively. Kunchukuttan and Damani [26] developed a system for Hindi compound noun MWE extraction from a Hindi corpus. Their extraction methods are based on statistical co-occurrence measures.

The linguistic methods for MWE extraction is based on words’ POS tags that form the grammatical and syntactical requirement for a word sequence to be a MWE. Bourigault [27] propose grammatical analysis method for the extraction of terminological noun phrases. Argamon, Dagan et al. [28] proposed a memory-based approach to learn language patterns from corpora. Their method relies on local POS information of a word sequence instead of full parsing a sentence. The hybrid approach combines both statistical and linguistic information of word sequences. Dias [20] proposed a hybrid system which uses mutual expectation to score both the association of words and the association of POS patterns in the tagged corpora. Su, Wu et al. [29] designed an automatic compound retrieval to extract compounds within a text. They use n-gram mutual information, relative frequency count and POS as the features for compound extraction. In

machine leaning methods, Pecina [6],[7] used machine learning approach for MWE extraction. Their method uses 55 kinds of association measures, such as joint probability, MI and t-score, to score each compound noun candidate. After that, a machine learning method (linear logistic regression, linear discriminant analysis and neural net) is used to classify new coming collocation candidates using the association measures' scores as features and to determine whether or not they are MWEs. The machine learning methods significantly improved ranking of collocation candidates on all of their data sets than the best association measure. Duan, Lu et al. [30] developed a bio-inspired approach for multi-word expression Extraction.

3. RESEARCH METHOD

We have developed a system that extracts bigram compound nouns MWEs from a text corpus. The compound nouns extractor creates a ranked list of Malay compound nouns. Several approaches which mainly rely mainly on the statistical co-occurrence information of the compound nouns and POS patterns have been implemented. Basic system architecture is shown in Figure 1. The following subsections will discussed in detail the extraction methods used.

3.1. Corpus Acquisition

Corpora have been extensively employed in several NLP tasks as the basis for automatically learning models for language analysis and generation. In this step, we crawl and collect Malay news articles which are written in Malay language from Malaysian National News Agency (BERNAMA) news source [http://www.bernama.com/bernama/v6/index.php]. The size of the corpus is 49661 news article and 13,346,381 token.

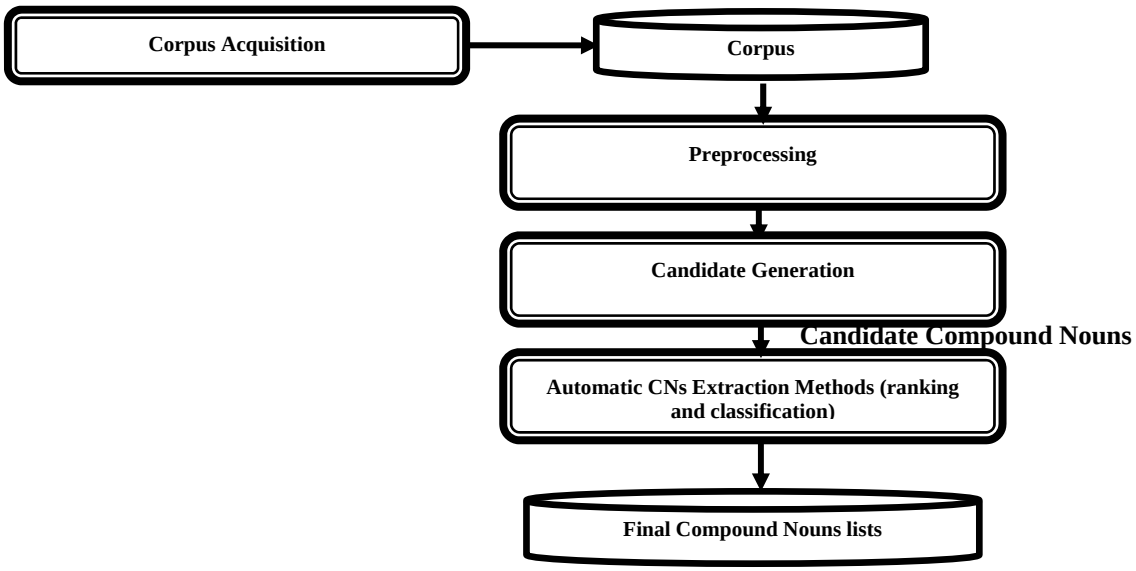

Figure 1. Extraction and Filtration of Compound Nouns Multiword Units

3.2. Preprocessing

In this phase, all crawled web pages are preprocessed by removing all HTML tags, identifying main content, automatic noise removal and breaking the content down to a sequence of individual tokens. After that, all-uppercase, capitalized and mixed case words were lowercased. Punctuations, special symbols and numbers are removed. Table 1 shows the n-gram statistic of our corpus.

Table 1. Statistics of the Malay corpus

Number of types	54742
Number of tokens	13,346,381
Number of unique bi-grams	705,680
Number of bi-grams	13,296,724
Number of unique tri-grams	1,730,916
Number of tri-grams	13,247,067

3.3. Candidate Generation

In this phase, we have tagged all nouns in the text corpus given a list of Malay noun list obtained from a manually annotated small tagged corpus and Malay lexicon which contain Malay words with their possible POS tags. This phase gives all possible N-N collocations that occur in a corpus. From the tagged corpus, if two consecutive words tagged as Noun and Noun respectively is extracted as a candidate N-N collocation. These compound nouns candidates are then passed to the next phase for automatic compound nouns extraction method. Compound nouns candidates which occur with very low frequency are discarded. Only candidate compound nouns collocations whose frequency in the corpus are greater than or equal to three are considered.

3.4. Automatic Extraction

Once we have extracted the candidate N -N compounds in the compound nouns candidate generation phase, we have ranked or classified each compound noun MWE candidate extracted from a corpus. In our task, several statistical co-occurrence measures and sequence type concerned model are calculated on each of the extracted candidates, and the candidate collocations are ranked or classified by these measures. They can be used for ranking (candidates with high scores at the top), or for classification (by setting a threshold and discarding all bigrams below this threshold).

3.4.1. Statistical co-occurrence association model

The major statistical measures used and evaluated in N -N compounds recognition in our study are presented.

Pointwise Mutual Information (PMI), Z-socre, T-test: These methods try to compare the observed frequencies of collocation candidates with the expected frequencies based on the assumption of independence in the target pairs (w1,w2). Krenn [31] did a thorough evaluation among t-score, z-socre and MI measures and showed that t-score over performed the other association measures for <PP+Verb> collocations in a German corpus. However, the statistical measures t-score, z-score, and MI are formulated below:

$$t-score = \frac{O-E}{\sqrt{E}};\quad z-score = \frac{O-E}{\sqrt{O}}$$

$$PMI - log(\frac{O}{E});\quad E = \frac{f_{w_1} \cdot f_{w_2}}{N}$$

where N: of the total instances of NNCs ; O: of the total instances of pair (w_1;w_2).
f_{w1}: of the total instances of w_1; f_{w2} : of the total instances of w_2.

Chi-square test (χ^2-test) : Pearson's χ^2 test of independence can be used to test if the words in the collocation are independent of each other. The χ^2-test is a classical method that is widely used for this type of analysis. The χ^2-test is formulated below:

$$\chi^2 = \frac{N(f_{w1}.O_{\overline{w1}\overline{w2}} - O_{\overline{w1}w2}O_{w1\overline{w2}})}{f_{w1}f_{w2}(O_{\overline{w1}w2} + O_{\overline{w1}\overline{w2}})(O_{w1\overline{w2}} + O_{\overline{w1}\overline{w2}})}$$

where N: of the total instances of NNCs ; O: of the total instances of pair (w_1;w_2)
f_{w1}: of the total instances of w_1; f_{w2} : of the total instances of w_2
$O_{\overline{w1}\overline{w2}}$: of pairs do not contain w_1 and w_2 simultaneously
$O_{\overline{w1}w2}$: of pairs contain w2 but not w1; $O_{w1\overline{w2}}$: of pairs contain w_1 but not w_2

Phi coefficient: In statistics, the Phi coefficient Φ is a measure of association for two binary variables. The Phi coefficient is adopted in several works for compounds extraction [8],[24],[32] The Phi coefficient is formulated below:

$$Phi = \frac{p(w_1w_2) - p(w_1*)p(*w_2)}{\sqrt{p(w_1*)p(*w_2)(1-p(w_1*))(1-p(*w_2))}}$$

Log Likelihood Ratio (LLR) : The likelihood-ratio test is a more general test of significance compared to the χ^2test and makes no assumptions of approximation to the normal distribution. The LLR has

proved to give better results [33]. 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, 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, then

$$LLR(w1,w2) = 2(a\ln a + b\ln b + c\ln c + d\ln d + (a+b+c+d)\ln(a+b+c+d) - (a+b)\ln(a+b) - (a+c)\ln(a+c) - (b+d)\ln(b+d) - (c+d)\ln(c+d))$$

Other methods: in addition to the methods described above, other statistical association measures such as dice coefficient, odds ratio and Jaccard (J), Normalized Expectation (NE), Mutual Dependency (MD), and Mutual Expectation (ME) are also used. These methods are widely used in the collocation extraction [6]-[9],[17],[24],[25],[32],[34]. These methods are formulated below:

$$MD = \log\left(\frac{p(w_1w_2)^2}{p(w_1*)p(*w_2)}\right); \quad NE = \frac{2f(w_1w_2)}{f(w_1*)+f(*w_2)}$$

$$ME = \frac{2f(w_1w_2)}{f(w_1*)+f(*w_2)} \times p(w_1w_2); \; Dice(w1,w2) = \frac{2*a}{2*a+b+c}$$

$$Odds\,ratio = \frac{ad}{bc}; \; Jaccard = \frac{a}{a+b+c}$$

3.4.2. The statistics of compound nouns and their components concerned methods

The C-value Approach: The C-value method is an efficient domain-independent multi-word term recognition method [35], which combines linguistic and statistical information [13],[14],[36]. C-value is sensitive to the nested compounding by its enhanced statistical measure of frequency of occurrence. C-value is defined as:

$$C-value(CN) = Log_2|CN|(f(CN) - \frac{1}{P(T_{CN})}\sum_{b\in T_{CN}} f(b))$$

where CN is a candidate compound noun, $|CN|$ is the number of simple nouns that consist of CN, $f(.)$ is its frequency of occurrence in the corpus, T_{CN} is the set of extracted candidate terms that contain CN, $P(T_{CN})$ is the number of these candidate terms. c(CN) is the number of those term candidates.

Combining frequency and geometric mean of nouns (FGM) : the main advantage of this method is that it manages to take into account both statistics of compound noun space and actual use in a corpus within one scoring function [20],[37],[38].

$$GMF(N) = GM(CN) * f(CN)$$

where $GM(CN) = \prod_{i=1}^{k}((LN(N_i)+1)(RN(N_i)+1))^{\frac{1}{2k}}$

$$LN(N) = \sum_{i=1}^{\#LDN(N)} \#L_i \text{ and } RN(N) = \sum_{j=1}^{\#RDN(N)} \#R_j$$

where f(CN) is the number of independent occurrences of noun CN, # LN(N) and # RN(N) are the number of distinct simple words which directly precede or succeed N and LN(N) and RN(N) are the frequencies of nouns that directly precede or succeed N.

3.4.3. Rank combination

Each of the above association measures methods gives a ranked list. We tried the following approach to combine these ranked lists:

Rank Aggregation (RA): The aim is to combine ranked lists produced by several association measures using information of the ordinal ranks of the elements in each list. The weighted combination method has proved to give better results their individuals [24]-[26]. Given multiple ordered lists $\mathbf{L_1}$, $\mathbf{L_2}$...$\mathbf{L_k}$ of CNs, the rank aggregation problem is to combine these lists into a single ranked list. We use the following rank aggregation heuristic which is called Borda's positional ranking:

Given lists $\mathbf{L_1}$, $\mathbf{L_2...L_m}$, where m ≤k for each candidate c ∈ NNCs and list $\mathbf{L}_i$, the score $B_{L_i}(c)$ is the number of candidates ranked below c in $\mathbf{L}_i$. The total Borda score is $B(c) = \sum_{i=1}^{m} B_{L_i}(c)$. The candidates are then sorted by descending Borda scores.

3.4.4. Statistical Classification

The main idea is to feed statistical and linguistic information about two adjacent Malay nouns to a machine learning classification framework. As shown in the previous sections, the statistical association measures are only measure *the association strength* of pairs of words. After that, their scores are usually ranked. Then, thresholds or evaluation points are set by users to evaluate them given a standard test. However, their scores even after ranking cannot indicate explicitly whether pairs of words scored are compound nouns or not. For example, "kad kredit" "credit card" word pair is scored " 61.65 " by t_test and ranked ninth in t_test list, but all these information cannot tell clearly weather the "kad kredit" is a Malay CN or not.

However, compound nouns extraction problem can be formulated as a binary classification problem [7] in which each candidate is assigned one class: $f(w_1, w_2) \rightarrow \{CN, not\ CN\}$. Each compound nouns candidate x is described by the feature or attribute vector$x = (x_1, \dots x_k)$, x_i is the statistical score given by one of the above association measures. We have several association scores given by several association measures methods for each candidate and want to combine them together to achieve better performance. In other words, the classification algorithms integrate all the association measures described above, and use their scores as attributes or features to classify N-N candidates. We evaluated several classification methods for compound nouns extraction.

Linear Logistic Regression: Logistic regression predicts the probability of an outcome that can only have binary response Logistic regression can handle several predictors (numerical and categorical). The multiple logistic regression model has the form :

$$logit(y\ is\ CN) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

The model defines the predicted probability as:

$$f(x) = P(x\ is\ CN) = \frac{exp^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + exp^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

where the coefficients β_i controls the effect of the of the predictor . The farther a β_i falls from 0, the stronger the effect of the predictor x_i.

Linear Discriminant Analysis: Linear Discriminant Analysis (LDA) is a popular tool for multiclass discriminative dimensionality reduction. The basic idea of LDA is to find a one-dimensional projection defined by a vector v that maximizes class separation. This method maximizes the ratio of between-class variance S_B to the within-class variance S_W in any particular data set thereby guaranteeing maximal separability.

$$\max_{v} \frac{v^t S_B v}{v^t S_W v},$$

Support Vector Machines: SVM proposed to solve two-class problems by finding the optimal separating hyper-plane between two classes of data. Suppose that X is set of labeled training points (feature vector) $(x_1, y_1),..., (x_n, y_n)$, where each training point $x_i \in R^N$ is given a label $y_i \in \{-1, +1\}$,where i = 1,. . .,n. The goal in SVM is to estimate a function $f(x) = w.x_i + b$ and to find a classifier $y(x) = sign(f(x))$ which can be solved through the following convex optimization:

$$\min_{w,b} \sum_{i=1}^{n} [1 - y_i(w.x_i + b)] + \frac{\lambda}{2}\|w\|$$

with λ as a regularization parameter.

4. EXPERIMENTS AND DISCUSSION

4.1. Data set and experiment setup

To create an evaluation gold standard, manual identification of compound nouns MWEs was done on a Malay corpus. All N-N compound nouns collocations are manually annotated by a native speaker. The entire reference data set containing 16535 N-N candidates (7235 unique N-N candidates), 2970 of the 7235 are N-N compound nouns collocations. We evaluate the extraction algorithms against the reference set of compound nouns collocations manually extracted from the 8200 files.

As described above, the collocation statistics were collected from a larger corpus of 49661 Malay news documents (13,346,381 words) from Malaysian National News Agency (BERNAMA) news source [http://www.bernama.com/bernama/v6/index.php]. Using a larger corpus provided more evidence for the statistical measures we used.

Since we manually annotated the entire reference data set, we have used standard metrics Precision and Recall for evaluating automatic compound nouns extraction method. These metrics are computed at different ranks, called Evaluation Points (EP) in the following way [6],[7],[24]-[26]:

Precision at evaluation point k is defined as:

$$P_K = \frac{\#\ \text{correctly extracted compund ncuns}}{k}$$

Recall at evaluation point k is defined as:

$$R_K = \frac{\#\ \text{correctly extracted compund nouns}}{\#\ \text{total compund nouns in the gold standard list}}$$

F-1 score at evaluation point k is defined as:

$$F_{1K} = \frac{2 \times P_K \times R_K}{P_K \times R_K}$$

4.2. Experimental results and analysis

In our experiment, we incrementally examined the n-highest ranked candidate lists returned by each method. The precision values are calculated for the first 100, 200, 500, 1000 and 2000 top ranked candidates. The precision metrics for different methods are shown in Figure 2. The x-axis represents the Evaluation Points, while the y-axis represents the precision values (the percentage of true N-N Compound nouns) achieved at these Evaluation Points. The performance metrics (Precision, Recall and F-score) for all methods are also shown in Table 2.

A first analysis of the precision curves and other metrics in Table 2 reveals distinction in two curve classes. Some of the methods start with very high precision and then decreases quite substantially. On the contrary, other methods start with low Precision and then slightly increase. The precision curve of each measure is important in this purpose because the monotonously decreasing graph indicates the more number of N-N compound nouns collocations in upper ranks rather than in lower ranks. Although all methods approximately have the same precision at 3000 top ranked list, finding a bigger proportion of the true N-N compound nouns at an early stage is simply more economical.

It is quite prominent from the results of Table 2 and Figure 2 that T-test , PS, C_value and FGM prove to be good measures for automatic extraction of Malay N-N compound nouns collocation as MWEs, since their Precision scores are higher at almost all evaluation points, while the worst measure appears to be CS method. As example, 99, 99, 98 and 98 of the top 100 ranked N-N by T-test, PS, C_value and FGM, respectively, are N-N compound nouns collocation. The top five candidates for each method and their corresponding tags are shown in Table 4. In fact, these methods show an interesting behavior compared to their behavior in other languages. The results obtained using these algorithms on Malay corpus are better than their results reported by other evaluation studies for other languages [6],[7],[9],[24]-[26].

It is important to note from Table 2 and Table 3 that some methods which are not mathematically equivalent (i.e., assigning identical scores to input candidates) such as T-test and PS achieve the same average precision and produce the same lists of ranked candidates. The ability to identify such groups of association measures may help in simplifying their formulas [39].

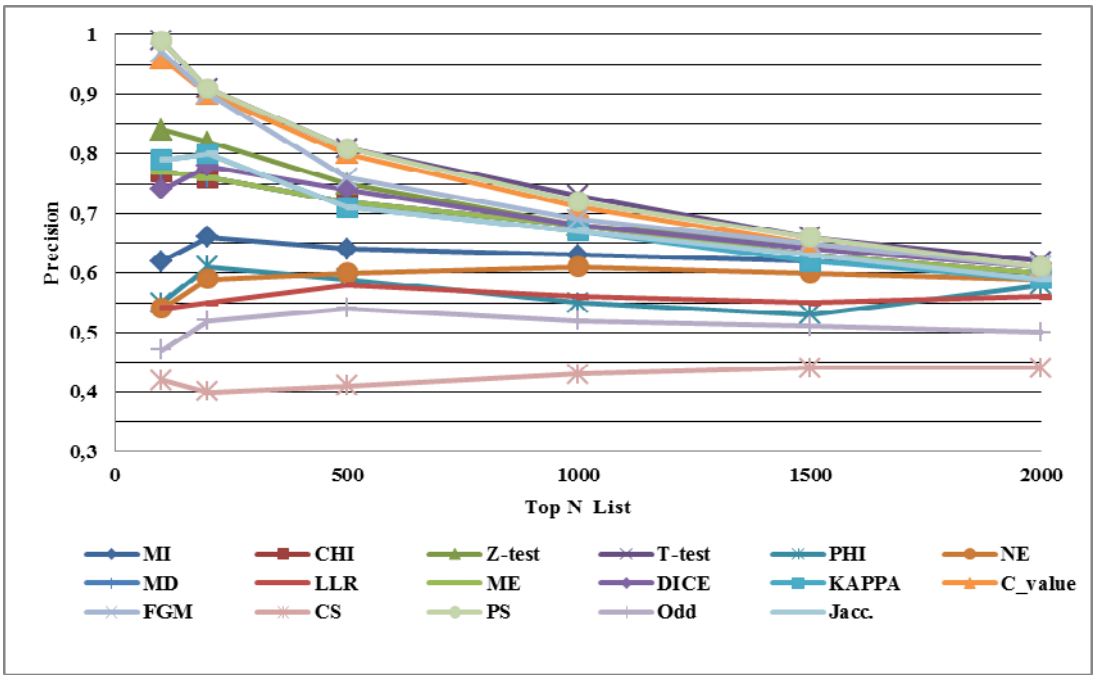

Figure 2. Overall Precision of different measures at different evaluation points

For the rank combination experiments, we combined the best four methods (T-test, PS, C_value and FGM). Table 4 shows Borda's positional ranking method's performance (Precision, Recall and F-score) and the top five candidates. Borda's positional ranking that does an approximate aggregation of the ranked list has been used as standard ranking function in previous studies [26],[40]. However, in our case, the Borda's positional ranking behaves in the same way as its individuals.

Table 2. The performance metrics (Precision, Recall and F-score) for all methods at different evaluation point

Evaluation	MI			CHI			T-test			PHI		
Point	P	R	F	P	R	F	P	R	F	P	R	F
100	0.62	0.02	0.04	0.77	0.03	0.05	0.99	0.03	0.06	0.55	0.02	0.04
200	0.66	0.04	0.08	0.76	0.05	0.09	0.91	0.06	0.11	0.61	0.04	0.08
500	0.64	0.11	0.18	0.72	0.12	0.21	0.81	0.14	0.23	0.59	0.1	0.17
1000	0.63	0.21	0.32	0.68	0.23	0.34	0.73	0.24	0.36	0.55	0.19	0.28
1500	0.62	0.31	0.41	0.63	0.32	0.42	0.66	0.33	0.44	0.53	0.27	0.36
2000	0.6	0.4	0.48	0.6	0.41	0.48	0.62	0.41	0.5	0.58	0.39	0.47
Evaluation	**LLR**			**MD**			**NE**			**ME**		
Point	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
100	0.54	0.02	0.04	0.77	0.03	0.05	0.54	0.02	0.04	0.77	0.03	0.05
200	0.55	0.04	0.07	0.76	0.05	0.09	0.59	0.04	0.07	0.76	0.05	0.09
500	0.58	0.1	0.17	0.72	0.12	0.21	0.6	0.1	0.17	0.72	0.12	0.21
1000	0.56	0.19	0.28	0.68	0.23	0.34	0.61	0.21	0.31	0.68	0.23	0.34
1500	0.55	0.28	0.37	0.63	0.32	0.42	0.6	0.3	0.4	0.63	0.32	0.42
2000	0.56	0.38	0.45	0.6	0.41	0.49	0.59	0.39	0.47	0.6	0.41	0.49
Evaluation	**DICE**			**KAPPA**			**CV**			**FGM**		
Point	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
100	0.74	0.02	0.05	0.79	0.03	0.05	0.96	0.03	0.06	0.97	0.03	0.06
200	0.78	0.05	0.1	0.8	0.05	0.1	0.9	0.06	0.11	0.9	0.06	0.11
500	0.74	0.12	0.21	0.71	0.12	0.21	0.8	0.13	0.23	0.76	0.13	0.22
1000	0.68	0.23	0.34	0.67	0.23	0.34	0.71	0.24	0.35	0.69	0.23	0.35
1500	0.64	0.32	0.43	0.62	0.32	0.42	0.65	0.33	0.44	0.65	0.33	0.43
2000	0.61	0.41	0.49	0.59	0.4	0.48	0.61	0.41	0.49	0.61	0.41	0.49
Evaluation	**CS**			**PS**			**Odd**			**Jacc.**		
Point	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
100	0.42	0.01	0.03	0.99	0.03	0.06	0.47	0.02	0.03	0.79	0.03	0.05
200	0.4	0.03	0.05	0.91	0.06	0.12	0.52	0.03	0.07	0.8	0.05	0.1
500	0.41	0.07	0.12	0.81	0.14	0.23	0.54	0.09	0.16	0.71	0.12	0.21
1000	0.43	0.15	0.22	0.72	0.25	0.37	0.52	0.18	0.26	0.67	0.23	0.34
1500	0.44	0.22	0.3	0.66	0.33	0.44	0.51	0.26	0.34	0.63	0.32	0.42
2000	0.44	0.3	0.35	0.61	0.42	0.49	0.5	0.34	0.41	0.59	0.4	0.48

To avoid incommensurability of association measures in our experiments, we used a common pre-processing technique for scores standardization: all association measure values are centered towards zero and scaled them to unit variance.

To evaluate machine learning methods Precision, recall and F1-measure of all classification methods were obtained by vertical averaging in ten-fold cross validation on the same reference data as in the earlier experiments. In each cross--validation step, nine folds were used for training and one fold for testing.

All classification methods performed very well. Detailed results (Precision, recall and F1-measure) of all classification methods are given in Table 5. The best result was achieved by a support vector machines. SVM achieves precision, recall and F1-measure of 75.44%, 87.78% and 81.14 % respectively.

Experiments show that classification algorithms which combine association scores given by several association measures methods lead to a significant performance improvement in comparison with individual basic methods. In fact, Experimental results obtained are quite satisfactory, especially when being compared to results obtained in other works [6],[7]. In [6],[7] a hybrid method of linguistic and statistical approaches has been proposed in terms of identifying compound nouns. Its clear that the hyprid method which combine both statistical and machine learning is outperformed the hybrid method of linguistic approach and statistical methods.

Table 3. Top 10 Malay N-N candidates extracted by different methods

MI		**CHI**		**T-test**		**PHI**	
panggung wayang	CN	sahabat handai	CN	kenaikan harga	CN	pengarah syarikat	CN
sahabat handai	CN	lubuk yu	CN	ehwal pengguna	CN	makanan ternakan	CN
karenah birokrasi	CN	jem madu	CN	harga minyak	CN	pakej umrah	CN
makhluk perosak	CN	pendingin hawa	CN	kementerian perdagangan	CN	perlindungan harta	NCN
kanun keseksaan	CN	laman web	CN	bahan api	CN	produk buatan	CN
jejari kentang	CN	harta intelek	CN	kerajaan negeri	CN	pegawai jabatan	NCN
adat resam	CN	kanun keseksaan	CN	ketua pegawai	CN	bot pukat	CN
akar umbi	CN	hukum syarak	CN	harga barang	CN	bulan april	CN
wakaf mempelam	CN	panggung wayang	CN	kad kredit	CN	permohonan lesen	CN
karbon dioksida	NCN	khabar angin	CN	musim perayaan	CN	muka surat	CN
LLR		**MD**		**NE**		**ME**	
sahabat handai	CN	jem madu	CN	lubuk yu	CN	jem madu	CN
panggung wayang	CN	sahabat handai	CN	sahabat handai	CN	sahabat handai	CN
karenah birokrasi	CN	lubuk yu	CN	jem madu	CN	lubuk yu	CN
barah otak	CN	pendingin hawa	CN	panggung wayang	CN	pendingin hawa	CN
paya pahlawan	NCN	laman web	CN	nira nipah	NCN	laman web	CN
hukum syarak	CN	harta intelek	CN	karenah birokrasi	CN	harta intelek	CN
makhluk perosak	CN	kanun keseksaan	CN	barah otak	CN	kanun keseksaan	CN
ais krim	NCN	hukum syarak	CN	era globalisasi	CN	hukum syarak	CN
milo ais	CN	panggung wayang	CN	ais krim	NCN	panggung wayang	CN
tumbuhan ubatan	NCN	khabar angin	CN	kondominium pangsapuri	NCN	khabar angin	CN
DICE		**KAPPA**		**CV**		**FGM**	
jem madu	CN	jaksa pendamai	NCN	kenaikan harga	CN	kenaikan harga	CN
sahabat handai	CN	laman web	CN	ehwal pengguna	CN	harga minyak	CN
lubuk yu	CN	kanun keseksaan	CN	harga minyak	CN	kementerian perdagangan	CN
pendingin hawa	CN	pendingin hawa	CN	kementerian perdagangan	CN	ehwal pengguna	CN
laman web	CN	harta intelek	CN	kerajaan negeri	CN	kerajaan negeri	CN
harta intelek	CN	musim perayaan	CN	bahan api	CN	harga barang	CN
kanun keseksaan	CN	penghilang dahaga	CN	ketua pegawai	CN	bahan api	CN
hukum syarak	CN	tali pinggang	CN	harga barang	CN	ketua pegawai	CN
panggung wayang	CN	topi keledar	CN	kad kredit	CN	harga bahan	CN
khabar angin	CN	akar umbi	CN	musim perayaan	CN	stesen minyak	CN
CS		**PS**		**Odd**		**Jacc.**	
sanak sudara	CN	kenaikan harga	CN	roti canai	CN	jaksa pendamai	NCN
pustaka sufi	NCN	ehwal pengguna	CN	mahkamah sesyen	CN	laman web	CN
angin sakal	NCN	harga minyak	CN	kanun keseksaan	CN	kanun keseksaan	CN
tuanku maharajalela	NCN	kementerian perdagangan	CN	sungai nyiur	NCN	pendingin hawa	CN
pokok mempisang	NCN	bahan api	CN	penyaman udara	CN	harta intelek	CN
online kegilaan	NCN	kerajaan negeri	CN	musim tengkujuh	CN	musim perayaan	CN
emas kerajang	NCN	ketua pegawai	CN	kacang buncis	CN	penghilang dahaga	CN
penubuhan platun	CN	harga barang	CN	muka sauk	CN	tali pinggang	CN
mesyuarat informal	NCN	kad kredit	CN	setebal muka	NCN	topi keledar	CN
bukit tekoh	CN	musim perayaan	CN	panggung wayang	CN	akar umbi	CN

Table 4. Results for rank combination Method

Evaluation Point	Precision	Recall	F-score	Top 5 ranked candidates	
100	0.98	0.03	0.06	kenaikan harga	CN
200	0.92	0.06	0.12	ehwal pengguna	CN
500	0.79	0.13	0.23	harga minyak	CN
1000	0.72	0.24	0.36	kementerian perdagangan	CN
1500	0.66	0.33	0.44	bahan api	CN

Table 5. Performance of Classification Methods Combining All Association Measures

Method	Precision	Recall	F1
SVM	75.44	87.78	81.14
LDA	72.21	81.09	76.39
GLM	69.94	83.48	76.11

5. CONCLUSIONS

In the present work, we have developed a compound noun MWE extraction system which ranks collocations using statistical methods. We developed and manually annotated a reference data set containing 5,610 Malay N-N bigrams, 1,854 of them were agreed to be a N-N compound noun. We implemented several lexical association measures, employed them for N-N compound noun extraction and evaluated them against the reference data set. The results obtained using these algorithms on Malay corpus are better than their results reported by other evaluation studies for other languages. The results also show that T-test, SP, C_value, FLR and RC are good measures for automatic extraction of Malay N-N compound nouns collocation. Finally, we employ three classification models (linear logistic regression, linear discriminant analysis and support vector machines) to combine association scores of the individual measures. Evaluation results show that these models significantly outperform individual association measures. SVM achieves precision, recall and F1-measure of 75.44%, 87.78% and 81.14 %, respectively.

In the future, we will implement, and evaluate other available methods suitable for this task. In addition, we will focus especially on automatically interpreting compound nouns relations and improving quality of the training and testing data. Finally, we will attempt to demonstrate contribution of collocations in selected application areas, such as machine translation or information retrieval.

REFERENCES

[1] A. Rahman, *et al.*, "Construction of compound nouns (CNs) for noun phrase in Malay sentence," *Presented at Information Retrieval & Knowledge Management (CAMP)*, 2012 International Conference on.
[2] Ahn, K., *et al.*, "Question Answering with QED at TREC-2005," *Presented at Proceedings of TREC,* 2005.
[3] Alias, N. A. R., *et al.*, "Application of semantic technology in digital library."
[4] Argamon, S., *et al.*, "A memory-based approach to learning shallow natural language patterns," *Presented at Proceedings of the 17th international conference on Computational linguistics, volume 1*.
[5] Baldwin, T. and Tanaka, T., "Translation by machine of complex nominals: getting it right," *Presented at Proceedings of the Workshop on Multiword Expressions: Integrating Processing*.
[6] Bourigault, D., "An endogeneous corpus-based method for structural noun phrase disambiguation," *Presented at Proc.*
[7] Church, K. W. Hanks, P., "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol/issue: 16(1), pp. 22-29, 1990.
[8] Ckakraborty, T., "Identification of Noun-Noun (NN) Collocations as Multi-Word Expressions in Bengali Corpus," *Presented at Student Session, International Conference of Natural Language Processing (ICON)*.
[9] Dandapat, S., *et al.*, "Statistical investigation of Bengali nounverb (NV) collocations as multi-wordexpressions," *Proceedings of Modeling and Shallow Parsing of Indian Languages (MSPIL)*, 2006, pp. 230-233.
[10] Dias, G., "Multiword unit hybrid extraction," *Presented at Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment, volume 18,* 2003.
[11] Duan, J., *et al.*, "A bio-inspired approach for multi-word expression extraction," *Presented at Proceedings of the COLING/ACL on Main conference poster sessions.*
[12] Frantzi, K., *et al.*, "Automatic recognition of multi-word terms:. the C-value/NC-value method," *International Journal on Digital Libraries*, vol/issue: 3(2), pp. 115-130, 2000.
[13] Gurrutxaga, A. and Alegria, I., "Measuring the compositionality of NV expressions in Basque by means of distributional similarity techniques," 2012.
[14] Hoang, H. H., *et al.*, "A re-examination of lexical association measures," *Presented at Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications.*
[15] Kit, C. and Liu, X., "Measuring mono-word termhood by rank difference via corpus comparison," *Terminology*, vol/issue: 14(2), pp. 204-229, 2008.

[16] Korkontzelos, I., *et al.*, "Reviewing and evaluating automatic term recognition techniques," *Advances in Natural Language Processing*, 2008, pp. 248-259.
[17] Krenn, B., "Empirical implications on lexical association measures," *Presented at Proceedings of The Ninth EURALEX International Congress.*
[18] Kunchukuttan, A. and Damani, O. P., "A System for Compound Noun Multiword Expression Extraction for Hindi," *Presented at 6th International. Conference on Natural Language Processing.*
[19] Levi, J. N., "The syntax and semantics of complex nominals: Academic Press," 1978.
[20] Liu, Y., and Tie, Z. "Automatic extraction and filtration of multiword units1," *Presented at Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on.*
[21] Manning, C. D. and Schütze, H., "Foundations of statistical natural language processing," MIT press.
[22] Merkel, M. and Andersson, M. "Knowledge-lite extraction of multi-word units with language filters and entropy thresholds," *Presented at Proc. 2000 Conf. User-Oriented Content-Based Text and Image Handling (RIAO'00).*
[23] Mima, H. and Ananiadou, S., "An application and e aluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese," *Terminology*, vol/issue: 6(2), pp. 175-194, 2001.
[24] Nakagawa, H., "Automatic term recognition based on statistics of compound nouns," *Terminology*, vol/issue: 6(2), pp. 195-210, 2001a.
[25] Nakagawa, H., "Experimental evaluation of ranking and selection methods in term extraction," Bourigault D, L'Homme M.-C., Jacquemin C.(éd.), *Recent advances in computational terminology*, John Benjamins Publishing Company, Amsterdam, pp. 303-326, 2001b.
[26] Nakagawa, H. and Mori, T., "A simple but powerful automatic term extraction method," *Presented at COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14.*
[27] Nakagawa, H. and Mori, T., "Automatic term recognition based on statistics of compound nouns and their components," *Terminology*, vol/issue: 9(2), pp. 201-219, 2003.
[28] Pearce, D., "A comparative evaluation of collocation extraction techniques," *Presented at Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002).*
[29] Pecina, P., "An extensive empirical study of collocation extraction methods," *Presented at Proceedings of the ACL Student Research Workshop.*
[30] Pecina, P., "Lexical association measures and collocation extraction," *Language resources and evaluation*, vol/issue: 44(1), pp. 137-158, 2010.
[31] Pecina, P. and Schlesinger, P., "Combining association measures for collocation extraction," *Presented at Proceedings of the COLING/ACL on Main conference poster sessions.*
[32] Piao, S. S., *et al.*, "Comparing and combining a semantic tagger and a statistical tool for MWE extraction," *Computer Speech & Language*, vol/issue: 19(4), pp. 378-397, 2005.
[33] Ramisch, C., *et al.*, "An evaluation of methods for the extraction of multiword expressions," *Presented at Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions (MWE 2008).*
[34] Saif, A. M. and Aziz, M. J. A., "An automatic noun compound extraction from Arabic corpus," *Presented at Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on.*
[35] Seretan, V., "Syntax-based collocation extraction," Springer, 2010.
[36] Seretan, V. and Wehrli, E., "Accurate collocation extraction using a multilingual parser," *Presented at Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.*
[37] Su, K. Y., *et al.*, "A corpus-based approach to automatic compound extraction," *Presented at Proceedings of the 32nd annual meeting on Association for Computational Linguistics.*
[38] Tanaka, T. and Baldwin, T., "Noun-noun compound machine translation: a feasibility study on shallow processing," *Presented at Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18.*
[39] Wu, C. C. and Chang, J. S., "Bilingual collocation extraction based on syntactic and statistical analyses," *Computational Linguistics and Chinese Language Processing*, vol/issue: 9(1), pp. 1-20, 2004.
[40] Zhang, W., *et al.*, "Improving effectiveness of mutual information for substantival multiword expression extraction," *Expert Systems with Applications*, vol/issue: 36(8), pp. 10919-10930, 2009.