Statistical-based Extraction for Malay Compound Nouns

Tuan Norhafizah Tuan Zakaria[1], Mohd Juzaiddin Ab Aziz[2], Mohd Rusmadi Mokhtar[3] & Saadiyah Darus[4]

[1,2,3]*Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, Selangor*
[4]*Fakulti Sains Sosial dan Komunikasi, Universiti Kebangsaan Malaysia, Selangor*

tn_hafizah@yahoo.com

Abstract. Collocation extraction is an important part in many natural language processing tasks such as machine translation, word sense disambiguation and information retrieval. This paper presents a statistical-based collocation extraction for Malay compound nouns. The Mutual Information was used to measure the association strength of compound nouns. Then, we used syntactic patterns and compared the precision using statistical -based and syntactic patterns to propose a hybrid approach (statistical-based and syntactical-based) to extract the compound nouns using collocation extraction methodology. The precision results show that the hybrid approach can be used for Malay collocation extraction for compound nouns.

Keywords: collocation; Malay; statistical-based; Mutual Information; syntax.

INTRODUCTION

Collocation is a word association method used to produce natural speech and writing [1]. For example, in English, it is inappropriate to use 'heavy wind' or 'strong rain'. The correct collocations for these phrases are 'strong wind' and 'heavy rain'. In Malay, the word '*tanak*' can be used with the word '*nasi*' and becomes inappropriate when we combine it with the word '*air*'. Generally, there are three main views for using collocation: views based on corpus research; views on discourse analysis; and views from linguistic field. Referring to views based on corpus research, [2] defined collocation as a combination of words that have been formed with perfect grammar. While in discourse analysis, [3] examines the phenomenon of psychological and language distribution. In linguistics, [4] defined collocation as "*A collocation is two or more words that tend to occur together. Collocations are frequent co- occurrences of lexical items or of particular construction."*

Collocations can be divided into two categories: grammatical collocations and lexical collocations [5]. Grammatical collocation contains a dominant word (usually a verb, noun or adjective) and a dependant word such as a preposition, or it contains certain pattern such as *dative-movement transformation, r^ar-clause*, atau *to + infinitival + gerund*. While lexical collocation contains two components such as verb + noun or adjective + noun.

Collocations play an important part in many natural language processing tasks such as machine translation, word sense disambiguation, information retrieval, natural language generation and lexicography.

Compounding refers to a process of forming a new word by combining two or more words. Collocation extraction can be used to extract compound words from a corpus. According to [6] and [7], compound nouns are a part of nominal phrases, consisting of two or more nouns with a space (strawberry juice) or without a space (blackboard). Normally, a compound noun has two parts: the head noun at the rightmost noun, and the modifier as the remainder component in a compound noun. The modifiers can be nouns, verbs, adjectives or adverbs. In Malay, there are three main categories for creating Malay compound nouns highlighted by [8] and [9]: (i) noun and noun; (ii) noun and non-noun modifier; and (iii) noun and noun modifier. Table (1) depicts examples of Malay compound nouns for each category.

TABLE 1. Examples of Malay Compound Nouns

Type of category	POS pattern	Compound noun (CN)
Noun and noun	*Gunung* (KN) + *ganang* (KN)	*Gunung-ganang* (mountains)
Noun and non-noun modifier	*Guru* (KN) + *besar* (KA)	*Guru besar* (headmaster)
Noun and noun modifier	*Ulat* (KN) + *buku* (KN)	*Ulat buku* (bookworm)

RELATED WORKS

Traditionally, collocations were identified manually and compiled from texts [10], [11]. However, these manual works were not ideal in terms of their coverage and consistency; it is also not suitable for computer processing of natural language application [12]. These manual works were also costly and time consuming.

The development of large-scale electronic texts has brought some improvement to the collocation extraction when automatic collocation extraction systems were developed. Most of the techniques followed the statistical-based (also called as window-based) approach [13], [14], [15], [16], [17]. This approach utilizes lexical statistics between a headword (also called a keyword) and its context words (collocate words) within a fixed-window span (called as distance) to estimate the relevance of the association between two or more words [12], [13]. This approach follows the definition of collocation by [18]: ―A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things.

This definition exhibits that a collocation is habitually dependent on its occurrence frequency. The co -occurrence of collocation was used as a basis to employ corpus-based association measures where the probability of differences between two or more samples were measured and compared. A variety of association measurements were involved in previous researches such as Dice coefficient [19], log-likelihood [20] and Mutual Information [21]. Generally, association statistical measurements can be applied in two ways: (i) as the primary step or (ii) as a baseline or assistant step integrated with linguistic knowledge such as syntactic patterns or semantic links (also called as hybrid approach)

With the improvement of accuracy and efficiency of parsing and tagging processes in natural language processing, the usage of syntactic-based collocation extraction approach has increased [22], [23], [24], [25]. Their collocation extraction approaches are based on the syntactic composition of collocations. Several measures are used for candidate ranking according to the collocational strength. The part-of-speech tag patterns [26] and dependency parser were applied to filter out pseudo collocations with high co-occurrence frequency. By restricting the candidate

search to syntactically dependent word combinations, syntax-based approach achieves good accuracy even for low-frequency collocations. However, this approach highly relies on the performance of the parser or tagger employed because the errors on parsing and tagging process will influence the performance of collocation extraction system. The precision and recall performance will decrease as well.

The collocation extraction studies have evolved over time. It is not only performed on the English language [25] but also extends to other languages such as Chinese [27], Arabic ([28], [29]), Thai [30] etc. [28] propose three approaches to extract Arabic multi-word expressions (MWEs). The first approach relies on the correspondence asymmetries between Arabic Wikipedia titles and titles in 21 different languages. For the second approach, he used Princeton WordNet 3.0 to collect English MWEs and translated the collection into Arabic using Google Translate, and utilized different search engines to validate the output. The third approach used lexical association measures to extract MWEs from a large unannotated corpus. [30] used statistical collocations and POS bigram probabilities without a POS tagger to extract Thai compounds. Probabilities of POS sequences estimated from compounds found in the dictionary were used to adjust the strength of collocation within a possible compound.

COLLOCATION VARIATION

Collocation depends on the nature of a language. Different language has its own variations. In order to improve the accuracy of collocation extraction, we need to determine the variety of the extracted candidates. In this paper, we take into account the morphological variations of Malay language: according to ([31], [32]), Malay language belongs to the family of agglutinative language and have different …… of morphology. There are some processes that are used by Malay morphology such as affixation, reduplication and compounding. However, some processes such as combination and separation in Malay morphology may change the nature of a particular word. Table (2) shows examples of combination and separation process in Malay language:

TABLE 2. Examples of Morphological Process in Malay

Type of word formation	Original word	POS	Affixation	New word	POS
Combination	semak	Verb	…an	semakan	Noun
Separation	kepantasan	Noun	ke…an	pantas	Adjective

STATISTICAL-BASED COMPOUND NOUNS EXTRACTION

Three stages are involved in this statistical-based compound nouns extraction.

A. Stage 1: Preparation of Training Corpus and Answer Set

A small training corpora needs to be constructed for this research. The raw texts were collected from several newspapers. All of the texts were segmented and tagged based on part-of-speech. An answer set will be constructed which contains a list of headwords and a list of true collocations based on the training corpus.

1) Step 1: Data Collection

The text was collected from several newspaper including Utusan Malaysia and Berita Harian. The total number of words were 7997.

2) Step 2*:* Word tagging

The process of tagging is important to enable more effective extraction of collocate words. The raw data in this corpus were manually annotated by a linguistic expert. The tagset used in this annotated corpus is shown in Table (3).

TABLE (3). The Tagset

Tagset	**POS**	**Example**
KN	Kata Nama (Noun)	*darjah* (degree), *pemain* (player)
KNK	Kata Nama Khas (Proper Nouns)	Kuala Lumpur, Noraina Abdul Samad
KK	Kata Kerja (Verb)	*mengurus* (manage), *membaiki* (repair)
KA	*Kata Adjektif* (Adjective)	*tinggi* (high), *cantik* (beautiful)
Bil	*Kata Bilangan* (Ordinal)	118

An example of the annotated text is shown in Figure 1:

KUALA LUMPUR<KNK>, 1 Feb<KN> (Bernama)<KN> -- Ketua Setiausaha Negara<KNK>, Tan Sri Mohd Sidek Hassan<KNK> hari<KN> ini<Ktgs> mengetuai<KK> senarai<KN> 256<bil> penerima<KN> darjah kebesaran<KN>, bintang<KN> dan<Ktgs> pingat<KN> Wilayah Persekutuan<KNK> sempena<KN> sambutan<KN> Hari Wilayah Persekutuan 2011<KNK>. Istiadat<KN> pengurniaan<KN> Darjah<KN> Seri Utama Mahkota Wilayah (SUMW)<KNK> kepada<Ktgs> beliau<KGN> disempurnakan<KK> Yang di-Pertuan Agong Tuanku Mizan Zainal Abidin<KNK> di<Ktgs> Balairong Seri Istana Melawati, Putrajaya<KNK>

FIGURE 1. Sample of Malay Tagged Corpus

B. Stage 2: Statistical-based Collocation Extraction

Statistical-based collocation extraction, also called as window-based technique is a foundation for all extraction systems. This technique will be the baseline system in this research. Three steps were involved in this stage.

1) Step 1: Generating the wordlist

According to [33], due to the multiple possibilities of word segmentation and part-of-speech tag, a wordlist needs to be constructed in the specified segmentation with the information of frequency and its POS tag. Table (4) shows an example of the wordlist.

TABLE 4. Wordlist of The Word with POS Information

Word	POS	Frequency
semak (check)	KK (verb)	20
semak (bush)	KN (noun)	5
cantik (pretty)	KA (adjective)	2

In the above example, semak (check) and semak (bush) are two different words because of their different POS tag.

2) Step 2: Determine the headwords

A list of headwords (W_h) were constructed based on the wordlist produced in Step 1. In this paper, twenty headwords were chosen.

3) Step 3: Association Measure

In order to measure the co-occurrence strength of the bi-gram candidates obtained in Step 2, Mutual Information (MI) measure was used to measure the association strength between the bi-gram.

This measure has been used to rank the candidates of collocation by [34] and was chosen because MI has the support from information theory and mathematical proof which makes it suitable for collocation ranking. The calculation was done as follows: For given two words x and y, P(x) is the occurrence probability of word x and P(y) is the occurrence probability of word y. [35] suggests that a MI score > 3 can be accepted as evidence of a valid collocation. Table (5) lists the MI score for some collocation candidates.

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

TABLE 5. MI Score

Collocation		
Keyword	**candidate**	**MI Score**
Jalan	jalan ladang	15.01690292
	jalan tersebut	13.20185184
	jalan di	13.04054737
	Jalan Kluang	12.616889
	Jalan Muar	11.78645134
	jalan kampung	11.54188347
	jalan raya	11.25593662
	jalan taman	11.20112514
	Jalan Taiping	11.20112514
	Jalan Angkasa	10.37086868

C. Stage 3: Performance Evaluation

Precision measures the percentage of correctly identified collocations as defined below:

$$\text{Precision} = \frac{\text{The number of correctly identified collocations}}{\text{Total number of extracted collocations}} \quad (2)$$

(2)

COMBINING SYNTACTIC PATTERNS FOR EXTRACTION

In this syntactical-based collocation extraction, the linguistic category (part-of-speech) of the word will be used to filter the collocation based on the defined collocation patterns. The candidate identification depends on the linguistic analysis (POS tag). The pseudo collocation (collocation candidates which do not fulfill the defined patterns) will be eliminated. In this paper, we adopted the compound noun patterns listed in [11] which consist of three categories: (i) noun and noun; (ii) noun and non-noun modifier; and (iii) noun and noun modifier. Table (6) shows examples of compound nouns.

TABLE 6. Compound Nouns

Word 1	Word 2	Compound Noun
darjah (KN)	*kebesaran* (KN)	*darjah kebesaran* (honors)
bilik (KN)	*operasi* (KN)	*bilik operasi* (operation room)
bilik (KN)	*tidur* (KK)	*bilik tidur* (bedroom)
hak (KN)	*cipta* (KK)	*hak cipta* (copyright)
lebuh (KN)	*raya* (KN)	*lebuh raya* (highway)

RESULT AND DISCUSSIONS

The correct compound nouns for this corpus were identified manually by a linguist. From Table VI, some collocation candidates were not correct although the MI score was high, for example '*jalan di*' and '*Jalan Kluang*'. It is due to the high frequency of co-occurrence but has no linguistic relations between the bi-grams. To evaluate whether adding syntactic patterns (POS sequences) could increase the precision value of extraction, two lists of outputs were constructed and compared. The first outputs were sorted by collocation scores only, while the other one was sorted by collocation scores and POS patterns. The precision value is shown in Table (7).

TABLE 7. Precision

Evaluation	MI	MI + POS
Percentage of Precision	25.6%	38.5%

The results show that the POS patterns could slightly increase the precision value of collocation extraction. However, the precision rates for both methods are not really high, perhaps due to the small corpus used.

CONCLUSIONS

In this paper, we present a statistical-based collocation extraction for Malay compound nouns. We used an association measure (MI) to rank the collocation candidates based on association strength of the bi-grams. Then, we added a POS patterns to compare the precision rate. The purpose of this comparison is to evaluate whether syntactic relations between the bi-

grams could increase the precision rate of collocation extraction. Some improvements need to be done to improve our research. For the future work, we plan to do the following:

i) To use a hybrid approach for collocation extraction which combine statistical-based (association measures) and syntactical-based approach (POS patterns).
ii) To extract other Malay multi words such as Verb Particle and Prepositional Phrase.
iii) To evaluate other association measures (AMs) such as log-likelihood, chi-square etc and choose the best AM for each type of Malay multi words.
iv) To build a large Malay corpus for this collocation extraction.
v) To construct more specific tagset for tagging process.

REFERENCES

1. Oxford Collocations Dictionary (2003). Oxford: Oxford University Press.
2. Kjellmer, G. (1994). A dictionary of English collocations. New York: Oxford University Press.
3. Nattinger, J. and DeCarrico, J. (1992). Lexical phrases and language teaching, Oxford: Oxford University Press.
4. Lewis, M. (2000). Teaching collocation: further development in the lexical approach, Hove: Language Teaching Publications.
5. Benson, M. (1990). Collocations and General Purpose Dictionaries, International Journal of Lexicography. vol. 3(1), pp. 23-35.
6. Huddleston, R. and Pullum, G. K. (2002). The Cambridge Grammar of the English Language. Cambridge University Press, Cambridge, UK.
7. Sag, I.A., Timothy, B., Francis, B., Ann, C. and Dan, F. (2002). Multiword expressions: A pain in the neck for NLP, in Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), Mexico City, Mexico. pp. 1–15.
8. Karim, N.S., Onn, F.M., Musa, H.H. & Mahmood, A.H. (2010). Tatabahasa Dewan, 3rd edition, Dewan Bahasa dan Pustaka (DBP).
9. Rahman, S.A., Omar, N. and Hassan, N.B.C. (2012). Construction of Compound Nouns (CNs) for Noun Phrase in Malay Sentence. pp. 22-25.
10. Sinclair, J. (1995). Collins COBUILD English Dictionary. Harper Collins.
11. Mei, J.J. (1999). Dictionary of Modern Chinese Collocations, Hanyu Dictionary Press.
12. Smadja, F. (1993). Retrieving collocations form text: X-tract, Computational Linguistics. vol. 19(1), pp. 143-177.
13. Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocation expressions in large textual database, in Proceedings of the RIAO Conference on User-Oriented Content-Based Text and Image Handling, Cambrigde, 1988, pp 21-24.
14. Church, K.W. and Mercer, R.L. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora,‖ Computational Linguistics. vol. 19, pp 1-24.
15. Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence, Computational Linguistics. vol. 19(1), pp 61-74.
16. Sun, M.S., Huang, C.N. and Fang, J. (1997). Preliminary Study on Quantitative Study on Chinese Collocations, ZhongGuoYuWen. vol. l, pp. 29-38.
17. Xu, R., Lu, Q. and Li, Y. (2003). An automatic Chinese Collocation Extraction Algorithm based on Lexical Statistics, in Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China. pp. 321-326.
18. Manning, C. and Sch'utze, H. (1999). Foundations of Statistical Natural Language Processing, M IT Press, Cambridge.

19. Smadja, F., Kathleen, R., Mckeown, and Hatzivassiloglou, V. (1999). Translation collocations for bilingual lexicons: a statistical approach, *Computational Linguistics*. vol. 22(l), pp. 3-38.
20. Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence, Computational Linguistics. vol. 19(1), pp 61-74.
21. Church, K. and Hanks, P. (1989). Word Association Norms, Mutual Information, and Lexicography, in Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics.
22. Lin, D. (1998). Extracting collocation from Text corpora, First Workshop on Computational Terminology. pp. 57-63.
23. Zhou, M. (2001). Improving translation selection with a new translation model trained by independent monolingual corpora, Computational Linguistics and Chinese Language Processing. vol. 6(1), pp. 1-26.
24. Krenn, B. and Evert, S. (2009). Can we do better than frequency? A case study on extracting pp-verb collocations, in Proceedings of the ACL Workshop on Collocations, France.
25. S. Violeta, ―Induction of Syntactic Collocation Patterns from Generic Syntactic Relations,‖ in Proceedings o f Nineteenth International Joint Conference on Artificial Intelligence (IJCAI2005), Edinburgh, Scotland, 2005, pp. 1698-1699.
26. Justeson, J.S. and Katz, S.M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text,‖ Natural Language Engineering. vol. 1, pp.
27. Duan, J., Zhang, M., Tong, L. and Guo, F. (2009). A Hybrid Approach to Improve Bilingual Multiword Expression Extraction, Proceeding of the 13th Pacific-Asia Conference on Knowledge Discovery and Data.
28. Attia, M., Toral, A., Tounsi, L., Pecina, P. and van Genabith, J. (2010). Automatic Extraction of Arabic Multiword Expressions‖, in Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010). pp. 18-26.
29. Saif, A.M., Aziz, M.J.A. (2011). An Automatic Collocation Extraction from Arabic Corpus, Journal of Computer Science. vol. 7(1), pp. 6-11.
30. Aroonmanakun, W. (2009). Extracting Thai Compounds Using Collocations and POS Bigram Probabilities without a POS Tagger.
31. Karim, N.S., Onn, F.M., Musa, H. and Mahmood, A.H. (2006). Tatabahasa Dewan. Dewan Bahasa dan Pustaka (DBP).
32. Omar, A. (2008). Ensiklopedia Bahasa Melayu, 2008, Dewan Bahasa dan Pustaka.
33. Yu, S.W. (1998). The Grammatical Knowledge-base of Contemporary Chinese: A Complete Specification, Tsinghua University Press, Beijing, China.
34. Zhang, W., Yoshida, T., Tang, X. and Ho, T.B (2009). Improving effectiveness of mutual information for substantival multiword expression extraction, Exp. Syst. Application. vol (36): 10919-10930. DOI: 10.1016/j.eswa.2009.02.026.
35. Hunston, S. Corpora in Applied Linguistics, Cambridge University Press, Cambridge.