

Natural-Annotation-based Malay Multiword Expressions Extraction and Clustering

Wuying Liu¹ and Lin Wang²(✉)

¹ Laboratory of Language Engineering and Computing,
Guangdong University of Foreign Studies,
Guangzhou 510420, Guangdong, China
wylu@gdufs.edu.cn

² Xianda College of Economics and Humanities,
Shanghai International Studies University,
Shanghai 200083, China
lwang@xdsisu.edu.cn

Abstract. Multiword expression (MWE) is an optimal granularity of language reuse. However, no explicit boundaries between MWEs and other words causes a serious problem on automatic identification of MWEs for some non-common languages. This paper addresses the issue of Malay MWEs extraction and clustering, and proposes a novel unsupervised extraction and clustering algorithm based on natural annotations. In our algorithm, we firstly use a binary classification for each space character to solve length-varying Malay MWEs extraction, secondly transfer natural document-level category annotations to MWE-level ones for Malay MWEs clustering, and finally distill out a general MWEs resource and several domain resources. The experimental results in the Malay dataset of 272,783 text documents show that our algorithm can extract MWEs precisely and dispatch them into domain clusters efficiently.

Keywords: Multiword Expression, Natural Annotation, Extraction, Clustering, Malay

1. Introduction

Language granularity has always been a difficult puzzle for Machine Translation [1] and Natural Language Processing. A multiword expression (MWE) is a lexeme made up of a sequence of two or more lexemes and its meaning cannot be obtained from its parts. This MWE granularity, between words and sentences, will be a more suitable size for a language reuse efficiently [2].

The MWEs extraction for common languages [3][4] has been widely investigated since the early days of Natural Language Processing, and many rule-based or statistic-based algorithms have been proposed [5][6]. However, the investigation on the MWEs extraction for non-common languages is still rare now.

Modern Malay (Bahasa Melayu), a non-common language, is spoken by 290 million people in Brunei, Indonesia, Malaysia, and Singapore, whose alphabet is just the same to that of English. The fast-paced development of linguistic science and computing technology has led to the accumulation of large-scale Malay text documents on the In-

ternet in recent years. For instance, the Wikipedia contains more than 300 thousand Malay text documents today.

This explosion of language big data with natural annotations has brought a great opportunity to Non-common Language Processing, not only MWEs extraction but also MWEs clustering. This paper addresses the issue of Malay MWEs extraction and clustering, and tries to build a general Malay MWEs resource and several domain resources automatically.

2. Related Work

There have been many MWEs extraction algorithms proposed for common languages up to the present. The early supervised algorithm uses statistical context features to reach a relatively high accuracy on MWEs extraction for common languages [7]. Subsequently, the semi-supervised algorithm uses morphosyntactic features to create Arabic verbal MWEs resource [8]. The previous supervised and semi-supervised algorithms normally require manual annotations, and this procedure will be time-consuming and expensive, which may not keep up with the rapid expansions of big data.

Recent investigations have used an unsupervised algorithm to extract Vietnamese multisyllabic words [9], whose scientific problem is very similar to that faced by MWEs extraction. The unsupervised algorithm is a straightforward and efficient model for real incremental and online applications of big data, whose effectiveness is due to that simple models and a lot of data trump more elaborate models based on less data in contemporary big data era [10]. Unsupervised learning does not require any manual annotation [11], which will be more suitable for MWEs extraction from dynamic language big data.

Some researchers used mutual information and information entropy to extract Chinese MWEs from Internet news pages and obtained an excellent result [12]. Some investigations defined the tokenization issue of the non-common language as a binary classification algorithm for each space character to identify the token boundaries [13], which motivated us to implement MWEs extraction according to a binary classification algorithm. There have been also some studies to implement fine-grained instance-level annotating using multi-instance learning from coarse-grained document-level annotations [14], which has important reference significance for our MWEs clustering.

Based on the above motivations from related works, we design a novel unsupervised architecture, which takes full advantage of language big data and straightforward efficient model and uses a binary classification algorithm for each space character to implement MWEs extraction, and which makes full use of document-level natural annotations to implement MWEs clustering.

3. Architecture

Figure 1 shows our unsupervised extracting and clustering architecture for Malay MWEs extraction and clustering, which mainly includes two parts: **Extracting** and **Clustering**. The **Extracting** part receives large-scale Malay text documents and produces Malay MWEs, and the **Clustering** part receives the Malay MWEs generated by the **Extracting** part and outputs a general MWEs and several domain MWEs, which can be used as general and domain lexical resources for Natural Language Processing. From this global perspective, the architecture is a meta-level structure and various suitable extracting and clustering algorithms can be implemented within it.

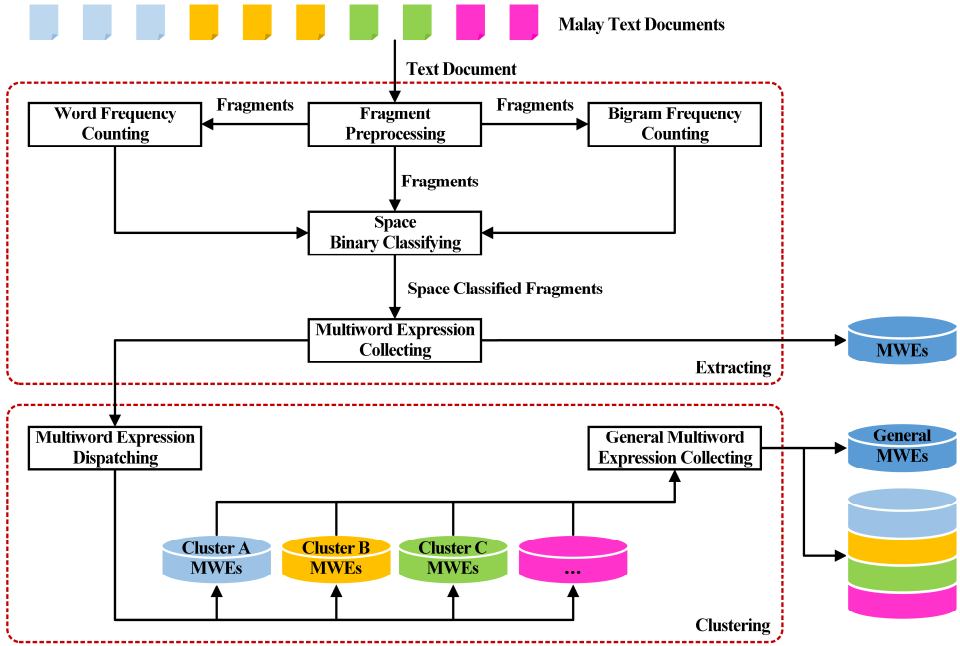


Figure 1. Unsupervised Extracting and Clustering Architecture.

In raw Malay texts, space character, similar to that in Vietnamese, can also be regarded as an overload character: a connector within a MWE or a separator between MWEs. Based on this understanding, we define the Malay MWEs extraction as a binary classification task for each space character in the **Extracting** part. There are total five processing units to complete the extracting work together. When a Malay text document arrives, the *Fragment Preprocessing* unit will be triggered firstly to split the text into several fragments, which are pure Malay word sequences without any punctuations, English words and other characters. Subsequently, the *Word Frequency Counting* unit and the word-level *Bigram Frequency Counting* unit receive the generated fragments respectively and count the corresponding frequency in parallel. The *Space Binary Classifying* unit receives the fragments and the unsupervised knowledges of word frequency and bigram frequency, and uses our proposed connectivity-based method to classify each space character: if a space character is a connector within a MWE, the method will output a dash character ('-') to replace it; if it is a separator between MWEs, the method will maintain it as a space character (' ') in the classified fragments. Finally, the *Multiword Expression Collecting* unit tokenizes all space classified fragments only by space characters and counts the frequency of each candidate MWE token, at least containing one dash character. According to the statistical results of frequency and a preset frequency threshold (it is 3 in this paper), the unit will collect a set of MWEs.

The effectiveness of above binary classifying method depends on the amount of Malay text documents. Fortunately, the explosion of language big data from Internet brings a lot of Malay news pages, which can be crawled automatically to form a large-scale dataset of raw text documents. Furthermore, most of news pages normally have manual category annotations. These natural document-level category annotations can be

transformed down to MWE-level cluster annotations in the **Clustering** part, where there are only two processing units. The **Multiword Expression Dispatching** unit incrementally creates MWE clusters named from category annotations of news text documents, and dispatches each extracted MWE into a related cluster according to the category annotation of the text document containing the MWE. The **General Multiword Expression Collecting** unit selects those MWEs with a high cluster frequency to form a set of general MWEs. At the same time, the unit removes the general MWEs from each cluster, and distills out multiple domain MWE sets.

4. Algorithm

Within the unsupervised extracting and clustering architecture, we design a detailed Malay MWEs extracting and clustering algorithm, which is shown in Figure 2.

```

1. // Malay MWEs Extracting and Clustering Algorithm
2. Input: Document[] mtds; // Malay Text Documents
3.     Float ct; // Connectivity Threshold
4. Output: String[] mwes; // Multiword Expressions
5.     String[] gmwes; // General Multiword Expressions
6.     Cluster[] dmwes; // Domain Multiword Expressions

```

```

7. Function String[]: extracting(Document[] mtds; Float ct)
8. String[] frags;
9. Map<String, Integer> bf;
10. Map<String, Integer> wf;
11. For Integer i ← 1 To mtds.size Do
12.     String[] frag ← fragmentpreprocessing(mtds[i].text);
13.     frags.merge(frag);
14.     bf.merge(bigramfrequencycounting(frag));
15.     wf.merge(wordfrequencycounting(frag));
16. End For
17. // Space Binary Classifying
18. For Integer j ← 1 To bf.keySet.size Do
19.     Integer bwf ← bf.get(bf.keySet[j]);
20.     Integer fwf ← wf.get(bf.keySet[j].firstword);
21.     Integer swf ← wf.get(bf.keySet[j].secondword);
22.     Float c ← (bwf/fgf+bwf/swf)/2;
23.     If (c>ct)
24.         Then frags.update(bf.keySet[j].firstword+'-'+bf.keySet[j].secondword);
25.     End If
26. End For
27. mwes ← multiwordexpressioncollecting(frags);
28. Return mwes.
29.
30. Function String[]: clustering.general(Document[] mtds; String[] mwes)
31. // Multiword Expression Dispatching
32. For Integer i ← 1 To mwes.size Do

```

```

33.   Document[] docs ← mtds.getdocuments(mwes[i]);
34.   For Integer j ← 1 To docs.size Do
35.       String category ← docs[j].category;
36.       If (dmwes.contain(category))
37.           Then dmwes.get(category).add(mwes[i]);
38.       Else dmwes.put(Cluster.new(category).add(mwes[i]));
39.       End If
40.   End For
41. End For
42. gmwes ← collectinggeneral(dmwes);
43. Return gmwes.
44.
45. Function Cluster[]: clustering.domain(String[] gmwes; Cluster[] dmwes)
46. For Integer i ← 1 To dmwes.size Do
47.     dmwes[i].remove(gmwes);
48. End For
49. Return dmwes.

```

Figure 2. Malay MWEs Extracting and Clustering Algorithm.

The *extracting* function is the pseudo-code implementation of the **Extracting** part of our above architecture. While the *clustering.general* function and the *clustering.domain* function together implement the **Clustering** part. The main contribution of the algorithm is the computing definition of the connectivity (line 22 in Figure 2). We regard the co-occurrence probability to individual occurrence probability of two neighboring words as a connectivity degree of the space character between the two words. This idea can avoid the non-fixed number of words problem in MWEs extraction. So, our algorithm can complete length-varying Malay MWEs extraction only by counting word frequency and bigram frequency. In our algorithm, we also make full use of document-level natural annotations to implement MWEs clustering straightforwardly. In the *clustering.general* function, we collect those MWEs occurring in more than 60% clusters to form the general MWEs.

During the total running process of the algorithm, the space overhead is mainly used to cache two indexes of word-frequency map and bigram-frequency map, which is negligible for the current 64bits memory capacity. Though the main time overhead is proportional to the size of the Malay text documents. The stream processing method, only one-pass scanning of each text document for two indexes construction, will make it time-efficient. The space-time complexity of the Malay MWEs extracting and clustering algorithm is acceptable in practical applications.

5. Experiment

In order to validate the effectiveness of our unsupervised extracting and clustering architecture and algorithm, we firstly prepare large-scale Malay text documents by crawling news pages on the Internet and gather total 272,783 Malay plain-text documents with natural category annotations. Furthermore, we already have a list with 85,539 Malay MWEs, which will be used as the golden standard to evaluate the experimental results. Secondly, we implement our unsupervised extracting and clustering algorithm, and run

extracting and clustering functions respectively in the above dataset. We also implement another unsupervised extracting algorithm according to mutual information and information entropy as the baseline [12]. Finally, we analyze the experimental results and suggest some discussions.

5.1. Result and Discussion about Extraction

In the unsupervised extracting part of experiments, we run our algorithm under different connectivity threshold from 0.1 to 0.9 and report the classical Precision (P), Recall (R) and F1-measure (F1) to evaluate the result of extraction. Figure 3 shows the detailed trends of the three measures. We can find that the P values and the R values have contrary trends with the increase of the connectivity threshold. When the connectivity threshold equals 0.3, the F1 value will climb the optimal peak, where the P, R and F1 values are 0.2965, 0.2311 and 0.2597 respectively. We also run the baseline in the same environment and gain its best P, R and F1 values (0.0693, 0.3013 and 0.1126). The experimental results prove that the performance of our straightforward unsupervised extracting is superior to that of mutual information and information entropy method.

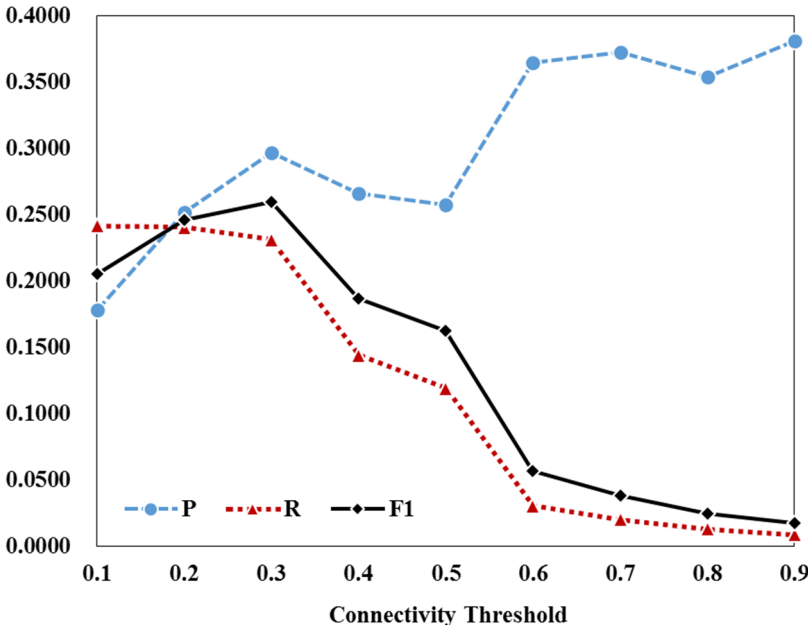


Figure 3. Evaluation Result of Extraction.

Though the experimental extracting precision is only about 30% under the situation of the best F1 value. The main reason is that the Malay text documents are independent on the golden standard. The actual extracting precision will high exceed the experimental result. When the F1 value equals the best 0.2597, we can obtain 66,625 Malay MWEs. Table 1 shows the partial examples of the extracted Malay MWEs, from which linguists estimate that the actual extracting precision will over 80%.

Table 1. Partial Examples of the Extracted Malay MWEs.

adat resam	anak watan	ayam pedaging	ARKADIUSZ Milik
adu domba	anak yatim	ayam piru	ASCOLI PICENO
aedes albopictus	anak yatim piatu	ayam tambatan	ATHLETIC BILBAO
ah long	angin ahmar	AARON Aziz	AUNG San Suu Kyi
ahlan wasahlan	angkasa lepas	ABDULLA Yameen	AYDA Jebat
ahli nujum	angkat sumpah	ADELIN Tsen	AYER KEROH
air kencing	angkatan tentera	AFRIKA SELATAN	AZ Alkmaar
air liur	anjakan paradigma	AHMAD NISFU	AZREL Ismail
air mani	anjing penghidu	AIDE Iskandar	Aamir Khan
air pancut	anugerah Grammy	ALEXANDRE Pato	Aaron Ago Dagang
air pasang	apam balik	ALI HAMSA	Aaron Cresswell
air zamzam	apit kanan	ALOR SETAR	Abby Fana
akad nikah	arang batu	AMELIA ALICIA ANSCALLY	Abdelaziz Bouteflika
akan datang	asam garam	AMERIKA SYARIKAT	Abdoulaye Faye
akar umbi	asid benzoik	AMPANG JAYA	Abdul Azeez Abdul Rahim
akil baligh	asid deoksiribonukleik	AMYRA Rosli	Abdul Gani Patail
alam barzakh	asid folik	ANDORRA LA VELLA	Abdul Ghani Minhat
alam sekitar	asid hidroklorik	ANGKAT BERAT	Abdul Halim
alam semesta	asid nitrik	ANGKATAN Bersenjata	Abdul Hamid
alat penimbang	asid sulfurik	ANNUAR Musa	Abdul Hamid Pawanteh
amar makruf nahi mungkar	asid urik	ANTHONY Kevin Morais	Abdul Khaliq Hamirin
anak bongsu	awan kumulonimbus	ANTOINE Griezmann	Abdul Muntaqim
anak didik	awan kumululus	ANUGERAH PLANET MUZIK	Abdul Rahman Dahlan
anak panah	awet muda	ANWAR ibrahim	Abdul Rahman Palil
anak pinak	ayam golek	ARITZ Aduriz	Abdul Taib Mahmud

5.2. Rusult and Discussion about Clustering

In the unsupervised clustering part of experiments, we run our algorithm to cluster the above 66,625 Malay MWEs according to their original category annotations, and finally we select out a general resource with 3,419 Malay MWEs and distill out total 9 domain resources. The detailed number of Malay MWEs in each domain cluster is shown in Table 2.

Table 2. Number of Malay MWEs.

General MWEs	Domain MWEs	
Number	Cluster Annotation	Number
3,419	dunia	8,284
	jenayah	5,733
	nasional	16,023
	semeasa	10,797
	sukan	13,531
	utusan borneo-berita iban	14,906
	utusan borneo-berita nasional	12,739
	utusan borneo-berita sarawak	20,530
	utusan borneo-sukan	7,628

The experimental results show that the unsupervised clustering function can transmit natural document-level category annotations to MWE-level ones efficiently, and support online incremental construction. If you want to obtain more fine-grained domain cluster, you can cluster the Malay text documents first and then use our unsupervised clustering function.

6. Conclusion

This paper proposes a natural-annotation-based MWEs extraction and clustering algorithm for Malay general and domain resources construction. The experimental results show that the effectiveness of our algorithm only depends on the context knowledge of co-occurrence frequency of words and natural document-level category annotations.

Further research will concern the influence of additional language knowledge such as word formation, stop word, syntax, and even semantics. We will also transfer above research productions to other suitable Austronesian languages like Indonesian, Filipino, and so on.

Acknowledgements. The research is supported by the Key Project of State Language Commission of China (No. ZDI135-26), the Featured Innovation Project of Guangdong Province (No. 2015KTSCX035), the Bidding Project of Guangdong Provincial Key Laboratory of Philosophy and Social Sciences (No. LEC2017WTKT002), and the Key Project of Guangzhou Key Research Base of Humanities and Social Sciences: Guangzhou Center for Innovative Communication in International Cities (No. 2017-IC-02).

References

1. Aiti Aw, Sharifah Mahani Aljunied, Haizhou Li. Malay Multi-word Expression Translation. Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation, 2009.
2. Helena Medeiros de Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, Aline Villavicencio. Alignment-based Extraction of Multiword Expressions. Language Resources and Evaluation, 44(1):59–77, 2010.
3. Campbell Hore, Masayuki Asahara, Yūji Matsumoto. Automatic Extraction of Fixed Multiword Expressions. Proceedings of the 2nd International Joint Conference on Natural Language Processing, 565–575, 2005.
4. Marion Weller, Ulrich Heid. Extraction of German Multiword Expressions from Parsed Corpora using Context Features. Proceedings of the 7th International Conference on Language Resources and Evaluation, 3195–3201, 2010.
5. Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, Josef Van Genabith. Automatic Extraction of Arabic Multiword Expressions. Proceedings of the Multiword Expressions: From Theory to Applications, 19–27, 2010.
6. Marie Dubremetz, Joakim Nivre. Extraction of Nominal Multiword Expressions in French. Proceedings of the 10th Workshop on Multiword Expressions, 72–76, 2014.
7. Meghdad Farahmand, Ronaldo Martins. A Supervised Model for Extraction of Multiword Expressions Based on Statistical Context Features. Proceedings of the 10th Workshop on Multiword Expressions, 10–16, 2014.
8. Mohamed Al-Badrashiny, Abdelati Hawwari, Mahmoud Ghoneim, Mona Diab. SAMER: A Semi-Automatically Created Lexical Resource for Arabic Verbal Multiword Expressions

- Tokens Paradigm and their Morphosyntactic Features. Proceedings of the 12th Workshop on Asian Language Resources, 113–122, 2016.
9. Wuying Liu, Lin Wang. Unsupervised Ensemble Learning for Vietnamese Multisyllabic Word Extraction. Proceedings of the 20th International Conference on Asian Language Processing, 353–357, 2016.
 10. Alon Halevy, Peter Norvig, Fernando Pereira. The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24(2):8–12, 2009.
 11. Andreas Vlachos. Evaluating Unsupervised Learning for Natural Language Processing Tasks. Proceedings of the 1st Workshop on Unsupervised Learning in NLP, 35–42, 2011.
 12. Jian Liu, Huifeng Tang, Wuying Liu. An Extraction Method for Chinese Terminology Based on Statistical Technology. China Terminology, 16(5):10–14, 2014.
 13. Wuying Liu. Supervised Ensemble Learning for Vietnamese Tokenization. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 25(2):285–299, 2017.
 14. Yi Zhang, Arun C. Surendran, John C. Platt, Mukund Narasimhan. Learning from Multi-topic Web Documents for Contextual Advertisement. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1051–1059, 2008.