Identifying and classifying unknown words in Malay texts

Bali Ranaivo-Malançon[1] Chong Chai Chua[2] Pek Kuan Ng[3]

[1,2,3] School of Computer Sciences, Universiti Sains Malaysia,
11800 USM, Penang, Malaysia
Tel. +60-12-5702934, Fax.: +60-4-6563244
e-mail : ranaivo@cs.usm.my, chongchai@gmail.com, wave_ng@yahoo.com

Abstract

In this paper, we propose a method based on a chain of filters to handle the problem of identifying and classifiying unknown words in Malay texts. A word is identified as unknown when it is not listed in the lexicon. The system presented in this paper classifies unknown words into four types: proper names, abbreviations, loanwords, and affixed words. One of our objectives is to reduce step by step the initial set of unknown words through a chain of filters: lookup wordlists, proper name identification, abbreviation identification, loanword identifier, and affixed word analyser. The experimental results reveal a good performance of our proposed method. Our two other objectives are to determine the types of words that remain unknown at the end of the whole process, and to make use of these information to specify the weaknesses of our identifiers so as to improve their accuracy.

1 Introduction and related works

When a text analysis module of any Natural Language Processing (NLP) application has to process a word that is not listed in its lexicon, it can either just tag it as “unknown” or try to classify it. A robust text analyser must be able to process all words contained in any kind of input texts. This means that one of the objectives when building a text analyser is to make it robust and thus finding a technique for processing unknown words is the key for robustness.

One solution that avoids the problem of unknown words is to list in a lexicon all possible word forms. This is illusory and does not take into account the dynamism of natural languages. At any time, new words can be created or borrowed. From the definition given here for unknown words, their number is tightly related to the size of the lexicon. But whatever the number of unknown words (small or large), these words preclude the achievement of most of NLP applications.

We can roughly divide the methods of processing unknown words into three groups: (1) the main objective is not to classify unknown words. However, their identification is required and it is included during the process (e.g. spelling correction, part of speech (POS) tagging, named-entity recognition, lexical knowledge acquisition, text segmentation, etc.); (2) the main objective is to distinguish unknown words from known words. Further classification of unknown words is not required; (3) the main objective is to identify unknown words and then classify them into different types. The work presented in this paper belongs to this group. In 2000, Toole mentioned the small number of works focusing on the identification and classification of unknown words (ICUW). This situation has not really changed seven years later. Toole (2000) used decision trees to classify unknown words. Her unknown word categoriser achieved 86.6% precision on the task of misspellings and names identification. The features used to train the decision tree for name recognition were POS and specific POS. Mikheev (2002) applied a document-centered approach to handle proper names and abbreviations. The disambiguation is based on information distributed across the entire document. Mikheev's system best achieved 95.12%-97.17% precision on proper name disambiguation and 98.8%-99.2% precision on abbreviation recognition. Goh et al. (2005) used a hierarchical model with multi-classifiers for the detection of numbers, time nouns, and person names. Each type of unknown words is processed by a specific support vector machine classifier. They reported higher precision (88.91%) compare to the method of using only one classifier for all types of unknown words (86%).

In this paper, we present a chain of filters for the ICUW in Malay texts using Latin alphabet[1]. Malay is understood here as the official language of Malaysia. Unknown words will be classified as "proper name", "abbreviation", "loanword", or "affixed word". We have three objectives: reducing the number of unknown words, determining the classes of words that remain unknown at the end of the whole process, and finally using these results to determine clearly the type of improvement needed for all our identifiers.

2 Types of unknown words

The common types of unknown words are misspellings, proper names, abbreviations, derived words, compounds, loanwords, foreign words, and neologisms. Other classes of unknown words have been proposed. Thai unknown words are classified by Kawtrakul et al. (1997) as explicit unknown words (they are not listed in the lexicon) and hidden unknown words (some substrings are known words). For Chinese, Chen and Bai (1998) proposed two groups: unknown words with syllabic morphemes and unknown words composed with multi-syllabic words only.

In this work, we try to identify four types of unknown words: proper names, abbreviations, loanwords, and affixed words. We do not include purposely in our ICUK the problem of spelling errors. The only Malay spelling checker available during our research is an interactive spelling checker. It uses exactly the same list of words as our Malay wordlist and contains the same affixed word analyser as we use in this work.

2.1 Proper names

Proper names (names of persons, locations, and organisations) correspond to open-class words. The simplest but very common method to recognise proper names is based on capitalisation. Other methods can be found in the area of information extraction where one of the subtasks is named-entity recognition.

2.2 Abbreviations

Abbreviations are perpetually created. They represent the shortened form of a word or a sequence of words. One possible approach is to maintain a list of known abbreviations and apply some guessing heuristics which examine the surface form of candidate abbreviations.

2.3 Affixed words

Malay can use different processes to derive complex words. It adds affixes to a base, duplicates a base by inserting a hyphen between the two elements (e.g. *penemuan-penemuan* 'discoveries'), or combines two bases (e.g. *memutarbelitkan* 'to twist' from *putar* 'turn' and *belit* 'around'). Affixation is a productive process in Malay, and therefore it is not possible to get an extensive list of affixed words. A complete morphological analyser should be able to recognise all morphologically complex words.

2.4 Borrowings: foreign words and loanwords

Any language needs to create or borrow words in order to express new concepts which often arise from new technologies. Foreign words are borrowed words that are used in the receiving language without any changes in their form and meaning. A language identifier that can guess the correct language of short words can help to identify foreign words. Loanwords are lexical units borrowed from another language but with their surface form adapted to the graphotactic and phonetic rules of the receiving language. In Malay, most of loanwords do not show the same graphotactic and morphological patterns as native words. A word is classified as loanword if (at least) one of these patterns is found in its structure (Ranaivo, 1996).

3 Classifying unknown words

[1] Also known as *Rumi*. Malay using Arabic alphabet is called *Jawi*.

The ICUW are performed through successive filters. After each filter, only words that are labelled "unknown" are retained to be the input of the next filter.

3.1 Lookup wordlist

3.1.1 Lookup Malay list of word forms

The first step in our proposed method is to get from a test corpus the list of words that is not in our list of 60,082 Malay word forms. This list contains roots, affixed words, compound words written without space, reduplicated words, and some loanwords.

3.1.2 Lookup list of proper names

After looking up to the Malay wordlist, the rest of unknown words are scanned for proper names. We use a list of 1,369 Malaysian names of person.

3.1.3 Lookup list of abbreviations

The remaining list of unknown words from the previous lookup is compared to a list of 293 abbreviations.

3.2 Abbreviation identifier

3.2.1 Identification by parentheses

If a sequence of letters is within two parentheses, and if the initial character of the previous words correspond to each of this sequence of letters, then the sequence of letters is retained as an abbreviation. For example, by applying this rule in the following text, *KPPK* and *GCR* are identified as abbreviations.

> *Kesatuan Perkhidmatan Perguruan Kebangsaan (KPPK) hari ini mencadangkan agar faedah "Pemberian Wang Tunai Gantian Cuti Rehat" (GCR) diperluaskan kepada semua guru biasa di negara ini.*

3.2.2 Identification by common formats

We have chosen some reliable rules that represent the majority of abbreviation formats.

- Any sequence of letters, each separated by a full-stop;
- Any sequence of capital letters with two, three, or four letters;
- Any sequence of consonants in upper case;
- Any sequence of vowels in upper case.

3.3 Proper name recogniser

3.3.1 By the definition of abbreviations

In step 3.2.1, we have identified some abbreviations preceded by their definitions. We capture all these definitions, and use each element of these definitions as proper names. Each element in *Kesatuan Perkhidmatan Perguruan Kebangsaan* 'Union of national education service' is recognised as a proper name when it appears in other place in the text – not at the beginning of a sentence – with identical spelling, that is, starting with a capital letter. Only the elements of the sequence *Gantian Cuti Rehat* will be considered as proper names as they correspond to the abbreviation *GCR*.

3.3.2 By specific titles

The use of a person title before the name is a sign of respect in Malaysia. We make use of title as a good marker of the beginning of a sequence of names of persons. There are many Malaysian titles so we reduce our list to "Tan Sri", "Tan Seri", "Toh Puan", "Datuk Seri", "Datuk", "Dato", "Datin", "Prof", and "Dr". All sequences of words that begin in capital case after these titles are considered as proper names.

3.4 Loanwords identifier

The loanwords identifier searches specific patterns (a letter or a sequence of letters). The tool discards loanwords from Malay native words.

3.4.1 Specific subset of letters

Among the 26 letters of the Latin alphabet, five of them, that is 'f', 'q', 'v', 'x', and 'z', appear only in loanwords.

3.4.2 Position of a letter or a sequence of letters

By studying the structure of Malay native words and "reversing" the Malay orthographic rules proposed by Mabbim (1992) in adapting loanwords, we have established a list of letters and sequence of letters that appear only in loanwords.

- Initial: ae, kh, gh, sy, abs, eks, auto, heks, hipo, homo, hiper, inter, intro, proto, super, hetero, C_1C_1 (the consonant must be the same);
- Medium: ae, sh, th;
- Final: e, o, c, j, w, y, ks, ans, oid, asma, isme, logi, grafi;
- Anywhere: ee, oo, uu, ie, bb, cc, dd, hh, jj, ll, mm, pp, qq, rr, ss, tt, vv, ww, xx, yy, zz, ph, sequence of three consonants (not necessarily the same).

3.4.3 Specific morphographemic rules

In Malay, the adjunction of one of these three affixes, meN-, peN-, and peN-an to a base must take into account different properties of the base: the number of syllables, the type of the initial letter, and the origin (native vs. borrowed). The rules are the same for the three affixes. To illustrate our purpose, only the rules for the prefix meN- are given as examples.

- Rules for monosyllabic bases
  - if the base is monosyllabic, then N → nge (e.g. meN-+cat > mengecat 'to paint');
- Rules for bases that are not monosyllabic
  - if the base starts with 'k'
    - if the base is native, then N+k → ng (e.g. meN-+kipas > mengipas 'to fan'),
    - otherwise, N+k → ngk (e.g. meN-+kritik > mengkritik 'to criticise').
  - if the base starts with 'p'
    - if the base is native, then N+p → m (e.g. meN-+pacu > memacu 'to spur'),
    - otherwise, N+p → mp (e.g. meN-+proses > memproses 'to process').
  - if the base starts with 's'
    - if the base is native, then N+s → ny (e.g. meN-+seduh > menyeduh 'to infuse'),
    - otherwise, N+s → ns (e.g. meN-+sabotaj > mensabotaj 'to sabotage').
  - if the base starts with 't'
    - if the base is native, then N+t → n (e.g. meN-+timbang > menimbang 'to measure'),
    - otherwise, N+t → nt (e.g. meN-+tradisi > mentradisi 'to make sthg a tradition').

An informal summary of these rules could be: nasal assimilation is for native words, and nasal insertion for loanwords. For example, for loanwords starting with one of the letter listed in 3.4.1, we have the following rules.

- If the loanword begins with 'f', then N+f → mf (e.g. meN-+fotostat > memfotostat 'to photostat').
- If the loanword begins with 'v', then N+v → mv (e.g. meN-+veto > memveto 'to veto').
- If the loanword begins with 'q', then N+q → nq (e.g. meN-+qada > menqada 'to perform a religious obligation').
- If the loanword begins with 'z', then N+z → nz (e.g. meN-+zeroks > menzeroks 'to xerox').
- If the loanword begins with 'x', then N+x → ngx (e.g. meN-+x-ray > mengx-ray 'to take an x-ray of').

3.4.4 Consonant-Vowel structures

The basic structure of a Malay syllable is *[C]V[C]* where *C* stands for consonant, *V* for vowel, and the square brackets for "optional". Malay has six vowels ('a', 'e', 'i', 'o' and 'u'), three diphthongs that we consider as *V* in our description ('ai', 'au', and 'oi'), and 23 consonants. The sequences 'ng', 'ny', and 'sy' are considered as three consonants. We have determined the different Malay CV structures of mono-, di-, and trisyllabic roots (Table 1). The dot indicates a syllable boundary.

Table 1 : CV-structures of Malay roots

Syllables	
1	CV,VC,CVC
2	V.V, V.VC, V.CV, V.CVC, VC.CV, VC.CVC, CV.V, CV.CV, CVC.CV, CVC.CVC
3	CV.CV.CV

3.5 Affixed word analyser

Our rule-based Malay affixed word analyser (Ranaivo-Malançon, 2004) extracts the root of a given affixed word. The program uses a list of Malay roots and some infixed words (infixation is no longer productive in Malay).

The analyser is an interactive tool. It displays all possible segmentations of a given word. One property that makes this affixed word analyser very powerful is that it always displays among the list of possible segmentations the correct one. If the analyser cannot determine it, it means that the root is not listed in its database yet. In this case, the user has to add the new root to the database, and in the next use, all words derived from the same root will be analysed correctly. When the case of missing root appears, we do not insert it manually into the database. The idea behind this is that we want the whole process of ICUK to be fully automatic. This means that unknown affixed words will remain unknown at the end of the whole process.

4 Experiment and results

In our experiment, a word is considered as unknown if it is not listed in our Malay wordlist. The corpus test corresponds to the compilation of Malay journalistic texts containing 105,069 tokens corresponding to 12,159 types (the tokenisation is case sensitive). We have eliminated from this list all numbers, alphanumerals, one letter, and url. We started our experiment with 12,022 word types. After looking up in the 60,082 Malay wordlist, 3,680 word types have been found "unknown" (about 30%). Table 2 shows the results of our experiments.

Table 2: Identification of unknown words

After ...	Unknown	Identified	Errors
Lookup Malay wordlist	3,680	8,342	-
Lookup proper names	3,418	262	-
Lookup abbreviations	3,351	67	-
Applying abbreviation rules (see 3.2.1)	3,287	64	-
Applying abbreviation rules (see 3.2.2)	3,014	273	20
Applying proper name rules (see 3.3.1)	2,987	27	0
Applying proper name rules (see 3.3.2)	2,811	176	2
Identifying loanwords	1,713	1,098	954 (or 804?)
Affixed word analysis	184	1,529	-

The column "Errors" correspond to the errors done among the "Identified" class of words. For example, during the application of abbreviation rules, 273 abbreviations have been identified. 20 of them are not abbreviations, and therefore classified as "errors".

The number of unknown words dropped abruptly after the affixed word analyser. This indicates that many of those unknown words (1,713) are new affixed words (1,529).

The set of words that remains unknown contains 83 proper names, 50 morphologically complex words, 32 misspelled words, 11 reduplicated words, 6 loanwords, 1 neologism, and 1 abbreviation.

4.1 Evaluation of the errors

The results given in Table 2 show that our method works well in reducing the number of unknown words: from 3,680 to 184. It is not evident to give an overall evaluation of the whole process as the errors could be done at any level of identification, and thus increasing the number of remaining unknown words.

Some reduplicated words remain unknown (e.g. *ekonomi-ekonomi* 'economies', *isteri-isteri* 'women') as they do not contain any affix. Our affixed word analyser extracts only the root if the given word is affixed.

Among 273 abbreviations identified by common format rules, 20 are found wrongly tagged. 19 of these words have length four (the maximum value used in one of the rules). They have been identified as abbreviations because they are all in capital case. It means that applying this simple rule in any abbreviation identifier will automatically create some errors.

The two errors in the identification of proper names by rules are *Ir* (the abbreviation of 'engineer', a title used mainly in Indonesia) and *M.Kayveas*. The tokeniser did not separate the sequence and since *Kayveas* as been identified as a new proper name in the previous identification (by the definition of abbreviations), all sequences with *Kayveas* are tagged proper names.

The last set of errors – done during loanword identification – needs some clarifications. This set contains 747 proper names, 150 foreign words, 29 abbreviations, and 28 spelling errors. We mention two values for the total number of errors: 954 and 804 (without the 150 foreign words). The reason is that, many rules used to identify loanwords are also valid for foreign words. Malay often borrows words without any transliteration making the separation of loanwords and foreign words not very clear.

5 Conclusion and future works

We have proposed in this paper a chain of filters for the ICUW in Malay texts. Through our experiment, we have reached one of our objectives. The number of unknown words has dropped spectacularly. In the same time, we have found that this small amount of unknown words is not the actual value. Additional unknown words may come from the errors done during each step of the process. If we add all errors, the total of real unknown words is 1,160 (= 20 + 2 + 954 + 184). This means that one third of the total number of the initial set unknown words have not been identified and classified correctly. But it also means that two third of the initial set of unknown words have been identified and classified correctly.

Our second objective is to determine the classes of words that remain unknown at the end of the whole process, and in the same time provide good indication in the improvement of all our identifiers. The problem of an automatic identification of proper names appears at any level of our method. This means that in order to improve the result of our method, we need to increase the number of proper names in our initial list (only 1,369), and find an accurate method for the proper name identification. The lack of full morphological analysis has left over the complete analysis of complex words and reduplicated words. In our future work, we plan to complete the Malay affixed word analyser with the analysis of reduplicated and compound words.

The classification of unknown words into only four types is not our final objective. As we have mentioned in section 2, other types of unknown words exist. In our future works, we plan to integrate other identifiers (e.g. neologism identifier, compound word identifier) that can classify unknown words into more specific classes.

The scope of this study is the identification and classification of unknown words. However, all tools and rules used in this study can be also applied to the classification of known words.

Acknowledgement

We are grateful to the anonymous reviewers who provided us valuable comments.

References

K.-J. Chen, M.-H. Bai. 1998. Unknown word dectection for Chinese by a corpus-based learning method. *Computational Linguistics and Chinese Language Processing*, 3(1): 27-44.

C.-L. Goh, M. Asahara and Y. Matsumoto. 2005. Training multi-classifiers for Chinese Unknown word detection. *Journal of Chinese Language and Computing*, 15(1): 1-12.

A. Kawtrakul, C. Thumkanon, Y. Poovorawan, P. Varasrai and M. Suktarachan. 1997. Automatic Thai unknown word recognition. In *Proc. of the Natural Language Processing Pacific Rim Symposium*, Phuket, Thailand, pp. 341-348.

A. Mikheev. 2002. Periods, Capitalized Words, etc. *Computational Linguistics* 28(3): 289-318.

Mabbim (Majlis Bahasa Brunei Darussalam-Indonesia-Malaysia). 1992. *General guidelines for the formation of terms in Malay.* DBP, Malaysia.

B. Ranaivo. 1996. *Automatic identification of foreign words in scientific and technical Malay texts.* D.E.A. Dissertation, INALCO, France.

B. Ranaivo-Malançon. 2004. Computational Analysis of Affixed Words in Malay Language. *ISMIL8*, Penang, Malaysia.

J. Toole. 2000. Categorizing unknown words: using decision trees to identify names and misspellings. In *Proc. of the 6th Conference on Applied Natural Language Processing,* Seattle, Washington, pp. 173-179.