Sentence-Level Novelty Detection in English and Malay

Agus T. Kwee, Flora S. Tsai, and Wenyin Tang

Nanyang Technological University,
School of Electrical & Electronic Engineering,
Singapore
atkwee@ntu.edu.sg,fst1@columbia.edu,wenyintang@ntu.edu.sg

Abstract. Novelty detection (ND) is a process for identifying information from an incoming stream of documents. Although there are many studies of ND on English language documents, however, to the best of our knowledge, none has been reported on Malay documents. This issue is important because there are many documents with a mixture of both English and Malay languages. This paper examines multilingual sentence-level ND in English and Malay documents using TREC 2003 and TREC 2004 Novelty Track data. We describe the text processing for multilingual ND, which consists of language translation, stop words removal, automatic stemming, and novel sentence detection. We compare the results for sentence-level ND on English and Malay documents and find that the results are fairly similar. Therefore, after preprocessing is performed on Malay documents, our ND algorithm appears to be robust in detecting novel sentences, and can possibly be extended to other alphabet-based languages.

Keywords: Novelty detection, sentence novelty, multilingual, automatic stemming, stop words removal, Malay.

1 Introduction

Nowadays, it becomes more and more convenient to find relevant information regarding a certain topic. With the development of Internet and information technology, lots of information sources are now available such as news articles, scientific papers, blogs, forums, etc. The bloom of information brings us rich useful information as well as many redundant information. People who are interested in a certain topic may read similar or even duplicate information over and over again. To solve this problem, a novelty detection system can be designed to retrieve only novel information regarding a certain topic. Novel information refers to documents or sentences that contain fresh content, and novelty detection (ND) is the process of singling out novel information from a given set of text documents. An automatic ND system consists of three main modules: (i) preprocessing, (ii) categorization, and (iii) novelty detection. The first module preprocesses the text documents by removing stop words and performing automatic stemming. Categorization classifies each incoming document or sentence

T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 40–51, 2009.

into its relevant topic bin. Then, within each topic bin containing a group of relevant documents or sentences, the ND module searches through the time sequence of documents or sentences and retrieves only those with "novel" information.

The pioneer work for ND was contributed by [16]. This paper gave the definition of "novelty" which was the opposite to "redundancy". A document that was less similar to its history documents was regarded as more "novel". [16] introduced ND at the document level. To serve users better, it could be more helpful to further highlight novel information at the sentence level [2]. Later studies focused on sentence-level ND, such as those reported in TREC 2002-2004 Novelty Tracks [12][14][13], which compared various novelty metrics [2][17], and integrated various natural language techniques [15][9][6]. This paper also focuses on sentence-level ND, which is the Task 2 of the TREC 2003/2004 Novelty Track. In this task, one needs to identify all novel sentences from groups of relevant sentences in all documents.

Although previous ND studies have been conducted on the English language, to best of our knowledge, no ND studies have been reported for Malay. This issue is important because there are many documents with a mixture of both English and Malay languages. The Malay language is spoken by more than 20 million people residing in the Malay Peninsula, southern Thailand, Philippines, Singapore, and Indonesia [7]. If a ND algorithm is tuned for the English language, is it necessary to develop a different ND algorithm to process documents of other languages such as Malay? This paper attempts to answer this question by investigating multilingual ND on both English and Malay documents.

If the documents were written in English, preprocessing steps are needed before applying the ND algorithm on Malay documents. Preprocessing steps includes language translation to Malay, Malay stop words removal, and Malay word stemming. Each of these steps will be discussed in detail in this paper.

For this paper, a list of Malay stop words is proposed and an adapted Malay stemming algorithm is introduced. After employing the Malay stop words and integrating the Malay stemming algorithm with the existing system, our ND system, which was originally developed for English sentences, can work well in detecting novel sentences in Malay.

This paper is organized as follows. A novelty detection algorithm for the English language is described in Section 2. A Malay ND algorithm, including Malay stop words and Malay stemming algorithm is presented and explained in Section 3. In Section 4, experiments and results are presented and discussed. Section 5 concludes the paper.

2 Novelty Detection in the English Language

Novelty detection in English involves the following steps:

2.1 Given all relevant sentences in all documents, first all stop words are removed. Stop words are words that need to be filtered out before doing the ND process as it affects the novelty prediction. Words like 'a', 'an', 'the', 'before', etc are considered stop words.

2.2 Next, remaining words are stemmed using Porter stemming algorithm [10]. Stemming [4], by definition, is a process of reducing the inflected (or sometimes derived) words to their root forms. Words like 'gives' and 'given' become 'give' after undergone the stemming process.

2.3 Finally, the novelty of each incoming sentence is determined using cosine similarity. The cosine similarity novelty metric calculates the similarities between the current sentence s_t and each of its history sentences s_i ($1 \leq i \leq t-1$), which determines the novelty score (S) for s_t as shown in Eq. (1).

$$S_{cos}(s_t) = \min_{1 \leq i \leq t-1} S_{cos}(s_t, s_i) \tag{1}$$

$$S_{cos}(s_t, s_i) = -\frac{\sum_{k=1}^{n} w_k(s_t) \cdot w_k(s_i)}{\|s_t\| \cdot \|s_i\|}$$

where $S_{cos}(s)$ denotes the cosine novelty score of sentence s and $w_k(s)$ is the weight of k^{th} element in sentence weighted vector s. The weighting function used in our work is tf.isf (term frequency multiply inverse sentence frequency) weighting function as defined below.

$$w_k(s_i) = tf_{w_k,s_i} log\left(\frac{n+1}{sf_{w_k} + 0.5}\right) \tag{2}$$

tf_{w_k,s_i} is the frequency of the word w_k in sentence s_i; sf_{w_k} is the number of sentences, in which the word w_k appears in the collection; n is the number of sentences in the collection.

The summary of novelty detection process in English is shown in Fig. 1.

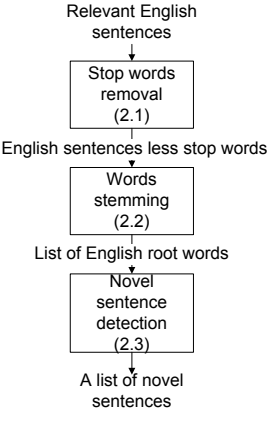

Fig. 1. English Novelty Detection Process

3 Novelty Detection in the Malay Language

"Is it necessary to develop a new ND algorithm for non-English documents?"

This paper addresses the question by adapting the existing ND algorithm on non-English documents, such as Malay. To tackle this issue, all sentences in English documents are first translated into Malay. Then, stop words are removed and the remaining words are stemmed. Lastly, novelty detection is carried out to retrieve all novel sentences. The overall process of ND in Malay can be seen in Fig. 2.

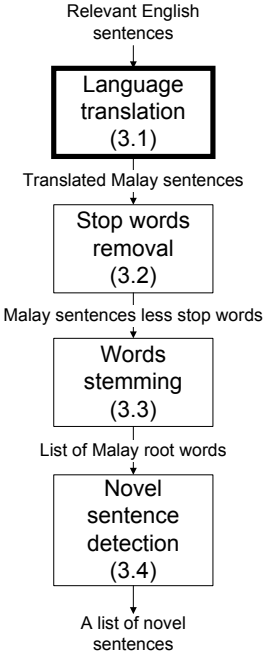

Fig. 2. Novelty Detection Process in Malay

3.1 Language Translation

In this first step, all English sentences are translated to Malay. During this process, we investigated issues in machine (automatic) vs. manual translation. Testing was conducted using TREC 2003 Novelty Track data on topics N1, N2, and N3. These topics were translated using two methods. First, we used only machine translation EBMT (Example-Based Machine Translation) [5]. Next, we manually corrected the machine translation. The F-score, precision and recall were calculated for each of 3 topics, and averaged across the topics. We compared the ND results of the two translations, which are shown in Table 1.

$$\text{Precision} = \frac{M}{S} \qquad \text{Recall} = \frac{M}{A} \tag{3}$$

Table 1. Comparison between machine translation and corrected translation

	Machine Translation	Corrected Translation	% difference
Average Precision	0.7599	0.7557	0.5527
Average Recall	0.9542	0.9564	0.2306
Average F-Score	0.8348	0.8330	0.2156

where S is the number of novel sentences selected by the novelty detection system; A is the number of novel sentences selected by the NIST assessor; and M is the number of matched novel sentences (number of novel sentences selected by both novelty detection system and NIST assessor). In TREC Novelty Track, the novel sentences marked by the NIST assessor is regarded as the truth.

$$F\text{-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where F-Score represents the harmonic mean of precision and recall.

As shown in Table 1, the average results differ only slightly (<1%) between machine and corrected translations. Since the results differed only slightly, we used machine translation for the remaining documents.

3.2 Stop Words Removal

Since stop words are language dependent, different languages have different set of stop words. For this paper, a list of Malay stop words was generated. Some stop words were directly translated from English stop words, such as 'each' becomes 'setiap', 'after' becomes 'sesudah' and 'therefore' becomes 'maka'. Some others were gathered from Malay documents, such as 'ayuh', 'alamak', and 'amboi'. A list of 216 Malay stop words used in this paper is shown in Table 2.

3.3 Malay Word Stemming

Similar to stop words removal, word stemming is also language dependent. Therefore, each language has its own rules for stemming. In this section, a Malay word stemming algorithm is introduced.

Algorithm

The Malay stemming algorithm used in this paper is adapted from Abdullah [1] and Mangalam [8]. After some modifications, the flowchart is shown in Fig. 3.

In this flowchart, the difference between backward and forward stemming is the affixes (either prefixes or suffixes) which are stemmed first. In the backward stemming, suffixes are stemmed first, while in the forward stemming prefixes are stemmed first.

Table 2. List of Malay stop words

kah	arkian	di	kemudian	nun	semua
lah	atau	dia	kenapa	oh	seperti
tah	au	ee	kendatipun	oleh	serba
acap	auh	eh	kepada	oleh itu	serta
acap kali	ayuh	ehem	kerana	pabila	sesungguhnya
adakala	ayuhai	enggan	ketika	pada	setelah
adakalanya	bagai	entah	kian	padahal	setiap
adalah	bagaimana	entahkan	kita	paling	sewaktu
adapun	bagaimanapun	gamaknya	lagi	para	siapa
adoi	bagi	ha	lagikan	pelbagai	sila
aduh	bahawa	haah	laksana	perlu	sini
aduhai	bahawasanya	hanya	laksana	pernah	situ
agak	bahkan	harapnya	lalu	pun	sudah
agaknya	banyak	harus	macam	sahaja	sungguh
agar	barangkali	hatta	maha	saja	sungguhpun
ah	beberapa	hei	mahu	sambil	supaya
ahai	belum	helo	mahupun	sampai	syabas
ahem	benar	hendak	maka	sangat	syahadan
ai	berapa	hingga	malah	saya	tatkala
aja	betul	ia	malahan	sebab	telah
akan	bila	ialah	mana	sebagai	tentang
alah	bilamana	ini	manakala	sebagaimana	terhadap
alahai	boleh	itu	manalagi	sebermula	terlalu
alamak	boleh jadi	jangan	masih	sedang	tetapi
alhasil	buat	jemput	masing-masing	sedikit	tiap-tiap
alkisah	bukan	jika	memang	segala	tidak
amat	celaka	jikalau	mengapa	sejak	tolong
amboi	cis	jua	meskipun	sekali	umpama
andai	dalam	juga	mesti	sekali peristiwa	umpama
andai kata	dan	kadang	minta	sekalian	untuk
aneka	dapat	kadangkala	misal	sekiranya	usah
antara	dari	kalakian	mungkin	selalu	wah
apa	darihal	kalau	nampaknya	seluruh	wahai
apabila	daripada	kami	namun	semasa	walau
apakala	demi	ke	nan	sementara	walaupun
apalagi	dengan	kecuali	nian	semoga	yang

Malay Affixes

Mangalam [8] points out the affixes in Malay that need to be stemmed:
Prefixes: me-, pe-, be-, ke-, se-, te-, di-
Suffixes: -wati, -wan, -man, -nya, -lah, -kah, -kan, -pun, -ita, -ku, -mu, -an, -i

Rules for Stemming Malay Words

Ranaivo [11] explains the rules for stemming Malay words.

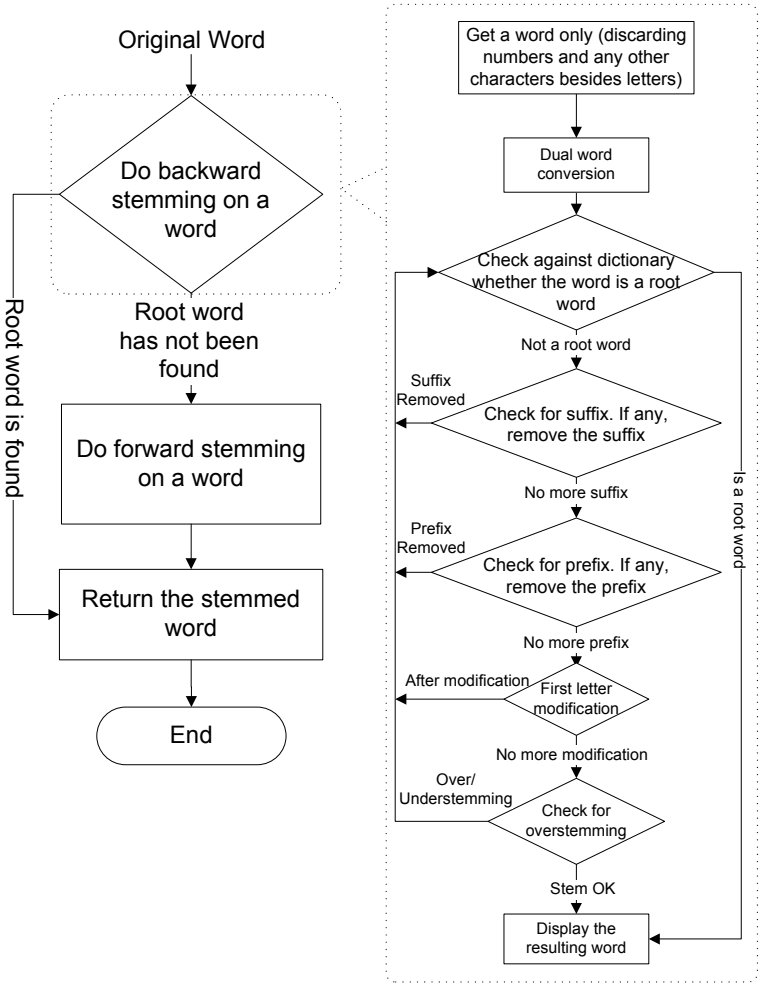

Fig. 3. Malay Word Stemming Flowchart with details on Backward Stemming Flowchart

Dual Word Conversion
In Malay, plural forms of words are represented by their dual word forms, such as 'anak - anak' (kids), 'murid - murid' (students), etc. In this step, all plural words are converted to their singular forms. For example, 'houses' is represented as 'rumah - rumah' in Malay. Having undergone this step, the word 'rumah - rumah' becomes 'rumah' (house), which is the singular form of 'rumah - rumah'.

Prefix
Rule 1: same prefix cannot occur more than once consecutively
If this happens; then, the second prefix must be part of the root word

Rule 2: after prefix te-, or pe-, no more prefixes are possible
Rule 3: prefix di- must appear first
Rule 4: prefix di- and me- can never coexist
Rule 5: prefix be- and te- can never coexist
Rule 6: prefix me- and be- can never coexist
Rule 7: prefix te- must appear after prefix me-
Rule 8: prefix pe- must appear after prefix me-
Rule 9: prefix be-, te-, me- and pe-
If the next two letters after truncating the above prefixes are
-ng, then remove 'ng'
-nC (C: consonant), then remove 'n'
-mC (C: consonant), then remove 'm'
-rC (C: consonant), then remove 'r'
-lV (V: vowel), then remove 'l'

Suffix
Rule 1: suffix -i and -kan can never coexist
Rule 2: suffix -wati, -wan, -man, and -ita can never coexist with any prefixes
Rule 3: suffix -kah and -kan can never coexist
Rule 4: suffix -lah and -kan can never coexist
Rule 5: suffix -an and -kan can never coexist
Rule 6: suffix -i and -an can never coexist

First Letter Modification
After taking out all prefixes and suffixes, each resulting word needs to be checked for its first letter.
Rule 1: if the first letter is 'm', then change it to 'p' or 'f'
Rule 2: if the first letter is 'n', then change it to 't'
Rule 3: if the first letter is 'y', then change it to 's'
Rule 4: if the first letter is vowel ('a', 'i', 'e', 'u', 'o'), then add 'k' at the beginning

Check for Overstemming
Rule 1: if the last prefix truncated is me- and first letter of the resulting word is 's', then check the word against dictionary by changing the first letter 's' with 'ny'. If the word is found in the dictionary, then overstemming has occurred.

Rule 2: if the last suffix truncated is -kan, then check the word against the dictionary by adding 'k' at the end. If by adding the letter 'k' results in a root word, then overstemming has occurred.

Drawback
One drawback of this algorithm is that the algorithm relies heavily on the dictionary to check for root words. Therefore, in order to improve the performance we need to compile a 'complete' list of Malay root words. Currently, the dictionary consists of 5300 Malay root words. Those words were taken from Bhanot's Malay - English dictionary [3].

At the end of stemming process, a list of root words for each sentence is generated.

3.4 Novel Sentence Detection

The last step of the ND process in Malay is novel sentence detection. This step is exactly the same as Section 2.3 of the ND process in the English language. After this step, a list of novel sentences is retrieved and the novelty detection process in Malay ends.

4 Experiments and Discussions

This paper used TREC 2003 and TREC 2004 Novelty Track data as the experimental data. TREC 2003 Novelty Track data consists of 50 topics (N1 - N50) and TREC 2004 Novelty Track data also consists of 50 topics (N51 - N100). The summary of these data is given in Table. 3.

Table 3. Data Statistics of TREC 2003 and TREC 2004 Novelty Track

Statistics	TREC 2003	TREC 2004
# topics	50 (N1-N50)	50 (N51-N100)
# relevant sentences	15557	8343
# novel sentences	10226	3454

In this experimental study, the focus was on novelty detection rather than relevant sentence categorization. Therefore, our experiments started with given all relevant sentences, from which the novel sentences were identified.

All sentences were originally written in English. We followed the novelty detection steps as described in Section 3 on the datasets to retrieve a list of Malay novel sentences. The cosine similarity novelty metric was employed and the threshold of novelty scores was predefined in a range from 0.05 to 0.95 with a step value of 0.10. The total of 40 runs were performed for both English and Malay on TREC 2003/2004 Novelty Track data. For each run (or each threshold), precision and recall (Eq. (3)) for each topic were calculated and the average precision and recall over 50 topics were plotted.

In information retrieval, precision-recall (PR) curves are always used to compare algorithms, where the algorithm with larger area under the curve is regarded as a better algorithm. Fig. 4 and Fig. 5 shows the PR curves of the English version vs. Malay version ND system on both TREC 2003 (Fig. 4) and TREC 2004 (Fig. 5) Novelty Track data. From these figures, we observe that the results of Malay ND is only slightly lower than English ND in the TREC 2003 Novelty Track data (with the average precision loss of 0.77% and the average recall loss of 0.96%), and is very similar in the TREC 2004 Novelty Track data (with the average precision loss of 0.09% and the average of recall loss of 0.16%). The

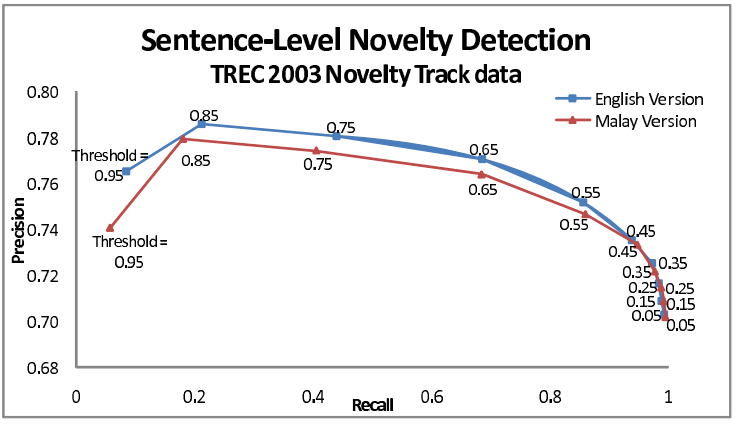

Fig. 4. Comparison between English and Malay version on TREC 2003 Novelty Track data

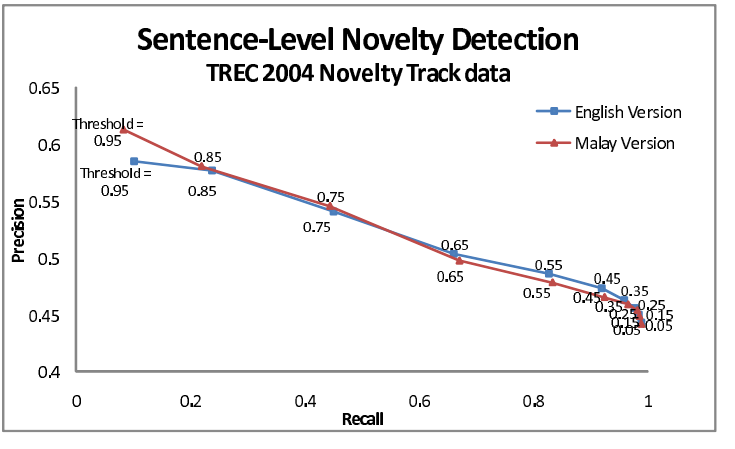

Fig. 5. Comparison between English and Malay version on TREC 2004 Novelty Track data

average precision and recall loss is the percentage difference of average precision and recall between English and Malay over all thresholds.

These experimental results indicate that, except for using different stop word lists and word stemming algorithms, the ND algorithm designed for the English language also performs well in detecting novel sentences in Malay.

5 Conclusion and Future Works

Sentence-level novelty detection (ND) is a process of retrieving novel sentences from a stream of incoming documents, based on a user's history and preferences.

In this paper, we focused on Task 2 of the TREC 2003/2004 Novelty Track to identify all novel sentences, given the relevant sentences in all documents. The document sets in the Novelty Track were all in the English language; thus, the corresponding ND algorithms were all designed for the English language. This paper addressed the issues involved in adapting existing ND algorithms for non-English languages such as Malay, which, to best of our knowledge, had not been previously studied.

Since the existing Novelty Track data is in English, in order to try the existing algorithms for documents in Malay, we first translated all relevant sentences using the EBMT (Example-Based Machine Translation) tool. Next, all Malay stop words from the translated sentences were removed, and the remaining words were converted into their basic forms using automatic Malay word stemming. After these preprocessing steps, the existing ND algorithm was performed on the resulting Malay sentences, and compared the results with original English sentences. The results show that by replacing English stop words with our proposed Malay stop words and English stemming algorithm with our adapted Malay stemming algorithm, existing ND algorithms that were originally developed for English can also be used for Malay sentences, without much loss in precision (<1%) and recall (<1%). Therefore, this paper shows that by replacing the language-specific processes (stop words and stemming), existing ND algorithms appear to be robust in detecting novel sentences in the Malay language, and can possibly be extended to other alphabet-based languages.

References

1. Abdullah, M.T., Ahmad, F., Mahmod, R., Sembok, T.M.T.: A stemming algorithm for Malay language. In: CITA, pp. 181–186 (2005)
2. Allan, J., Wade, C., Bolivar, A.: Retrieval and novelty detection at sentence level. In: SIGIR 2003, Toronto, Canada, pp. 314–321. ACM, New York (2003)
3. Bhanot, D.K.: The first online Malay — English dictionary (June 2008), http://dictionary.bhanot.net/
4. Datasegment.com online dictionary (August 2008), http://onlinedictionary.datasegment.com/word/stemming
5. U. R. Group. Example-Based Machine Translation (EBMT) prototype, Universiti Sains Malaysia (July 2008), http://utmk.cs.us.my:8080/ebmt-controller/index.jsp
6. Li, X., Croft, W.B.: An information-pattern-based approach to novelty detection. Information Processing and Management 44(3), 1159–1188 (2008)
7. Malay language. In: Wikipedia, The Free Encyclopedia (2008), http://en.wikipedia.org/wiki/Malay_language
8. Mangalam, S.S.V.: Malay-language stemmer. Sunway Academic Journal 3, 147–153 (2006)
9. Ng, K.W., Tsai, F.S., Goh, K.C., Chen, L.: Novelty detection for text documents using named entity recognition. In: 2007 6th International Conference on Information, Communications and Signal Processing, pp. 1–5 (2007)
10. Porter, M.F.: An algorithm for suffix stripping. In: Readings in information retrieval, pp. 313–316 (1997)

11. Ranaivo, B.M.: Computational analysis of affixed words in Malay language. In: 8th International Symposium on Malay/Indonesian Linguistics (ISMIL) (2004)
12. Robertson, S., Soboroff, I.: The TREC 2002 Filtering Track report. In: TREC 2002 - the 11th Text REtrieval Conference (2002)
13. Soboroff, I.: Overview of the TREC 2004 Novelty Track. In: TREC 2004 - the 13th Text REtrieval Conference (2004)
14. Soboroff, I., Harman, D.: Overview of the TREC 2003 Novelty Track. In: TREC 2003 - the 12th Text REtrieval Conference (2003)
15. Zhang, H.-P., Sun, J., Wang, B., Bai, S.: Computation on sentence semantic distance for novelty detection. Journal of Computer Science and Technology 20(3), 331–337 (2005)
16. Zhang, Y., Callan, J., Minka, T.: Novelty and redundancy detection in adaptive filter. In: SIGIR 2002, Tampere, Finland, pp. 81–88. ACM, New York (2002)
17. Zhao, L., Zheng, M., Ma, S.: The nature of novelty detection. Information Retrieval 9, 527–541 (2006)