社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－ 93 (6)
2009－ＮＬ－189 (6)
2009／1／22

マレー語のステマーの開発とその評価

リム・フイ・ティアン† 安川美智子†† 横尾 英俊††

† 群馬大学工学部 情報工学科
†† 群馬大学大学院工学研究科 情報工学専攻
376-8515 桐生市天神町 1–5–1

あらまし マレー語を対象とした接辞処理のアルゴリズムを提案する．マレー語には，接辞が文法上の重要な役割を果たすという特徴があり，接辞が語幹に付く仕組みが複雑である．英語などの他の言語では，末尾の語形変化や接尾辞を繰り返し除去する単純なステマーが用いられているが，そのようなステマーをマレー語に適用しても接辞処理の効果が得られない．マレー語の接辞の特徴を考慮した接辞処理のアルゴリズムが必要である．同一の語幹をもつ，意味の類似した派生形をまとめあげる接辞処理は，索引語の多様性を吸収する効果がある．しかし，意味の大きく異なる派生形を同一の語幹に関連づける過度な接辞処理を行うと，接辞処理がかえって索引処理に悪影響を及ぼす場合もある．このため，開発したステマーではマレー語の接辞ルールと二種類の単語辞書 (語幹辞書と派生語辞書) を用いた接辞処理により，派生形に対する過度な接辞処理を避け，効果的な接辞除去を行っている．WWW 上のマレー語のテキストデータを用いて行ったステマーの性能評価についても報告する．

キーワード マレー語，接辞処理，接辞ルール，索引処理，索引語，テキストマイニング

Development and Evaluation of a Malay Stemmer

LIM Hui Tian†, Michiko YASUKAWA††, and Hidetoshi YOKOO††

† Department of Computer Science, Gunma University
†† Graduate School of Engineering, Gunma University, Kiryu 376-8515, Japan

Abstract Stemming is the process for reducing various word forms to their root forms in order to improve the effectiveness of information systems. In the Malay language, affixes have important grammatical functions. Hence, the composition of Malay words is more complicated than that of English words. Although there are several types of stemming algorithms for English and some other languages, they have not overcome the difficulties in Malay words stemming. It is essential to avoid both over-stemming and under-stemming. We have developed a new Malay stemmer for removing inflectional and derivational affixes. Our stemmer uses a set of affix rules and two types of dictionaries. We conducted an evaluation experiment using text data from WWW to support our Malay stemming algorithm.

Key words Malay language, stemmer, stemming, affix rule, indexing, index term, text mining

1. はじめに

近年，インターネットやパソコンの普及に伴い，電子化された文書が劇的に増加している．Web，メール，掲示板などあらゆる文書データがディジタル化されている．このような状況を背景にして，テキストマイニング [1] の研究開発が急速に進み，実用化の面でも活用されるようになってきている．我々は，マレー語を対象としたテキストマイニング，特に，マレー語の文書群からの決定則の抽出に取り組んでいる．決定則は文書群に含まれる知識の要約となっていることから，人間の意志決定を支援する用途に用いることができ，また，計算機を用いた文書自動分類にも適用できることから，テキストマイニングの分野においては重要な技術の一つとなっている．

テキストマイニングは，文書群から有益な知識を発見・抽出するための技術である．データからの知識発見に用いられる技術としてデータマイニングの技術がある．データマイニングは，表 (spreadsheet) のような構造を持ったデー

タ (構造化データ) を対象とする．データマイニングのために開発された既存技術を，文書群に適用するためには，文書と単語の関係を表の形で表現することが必要となる．一般に，文書にはノイズとなる単語が多数含まれているため，文書と単語の関係を表現する際には，単語の中から，**文書の内容をよく表す単語** (以下,「**索引語**」と呼ぶ) を選定しなければならない．

文書から索引語を選定する際に，単語としてどのような単位を考えるかは，対象とする言語に依存する．英語のように単語と単語の間に空白を置く言語では，空白を手掛かりに単語を認定できる．しかし，日本語のように単語の間に区切り記号を置かない言語では，形態素解析 (morphological analysis) と呼ばれる処理によって，単語を認定することが必要となる [3].

マレー語は，表記には英語と同じアルファベットを使用し，区切り記号には空白を使用するため，字面は英語に似ている．しかし，マレー語には，**接辞** (affix) によって文法的機能を表すという特徴があり，接辞と**語幹** (stem) の間に区切り記号を置かない．日本語において自立語 (名詞や動詞など) に接語 (助詞や助動詞など) が付く (膠着する) ように，マレー語においては，文の意味を表す語幹に，文法的機能を表す接辞が付いて，文が構成される．このため，マレー語は言語としては日本語に似た側面を持ち，日本語のように，単語を認定するための特別な処理が必要となる．マレー語では，形態素解析ではなく，接辞を除去して語幹を得る**接辞処理** (stemming) と呼ばれる処理によって単語を認定する．

本稿では，接辞処理のアルゴリズムを提案し，提案法に基づくステマーの実装とその評価について述べる．提案するアルゴリズムは，マレー語の接辞のルールを定義したテンプレートファイルと語幹を検出するための語幹辞書，及び，過度な正規化を避けるための派生語辞書を用いることで，効果的な接辞除去が行える．

2. マレー語の接辞処理

マレー語はマレーシアの国語であり，7 世紀にスマトラ東海岸およびマレー半島で話されていたムラユ語 (Bahasa Melayu) が起源となっている [4]．現在は，マレーシア，シンガポール，ブルネイなど東南アジアの広範囲で話されている．また，マレー語とインドネシア語はもともと同じムラユ語から派生した言語であり，共通点が多い．

マレー語は表記に英語と同じアルファベットを使用するため，文字によって表記されたマレー語の文は英語の文に似ている．たとえば，日本語の文 (1) に対応する英語の文とマレー語の文はそれぞれ文 (2) と文 (3) のようになる．

文 (1)　彼は、本を読んだ。　(日本語)
文 (3)　He read a book.　(英語)
文 (3)　Dia membaca buku.　(マレー語)

表 1　日本語の文を形態素解析した結果

出現形	読み	原形	品詞
彼	カレ	彼	名詞-代名詞-一般
は	ハ	は	助詞-係助詞
、	、	、	記号-読点
本	ホン	本	名詞-一般
を	ヲ	を	助詞-格助詞-一般
読ん	ヨン	読む	動詞-自立
だ	ダ	だ	助動詞
。	。	。	記号-句点

表 2　マレー語の文を人手で接辞処理した結果

マレー語	英訳	日本語訳	説明
Dia	he/she	彼/彼女	代名詞
mem+		～を～する	接頭辞 (語幹 baca を動詞化)
baca	read	読むこと	語幹
buku	a book	一冊の本	語幹 (単独で名詞の単数形)
.	.	。	記号

英語の場合，空白を手掛かりとして，文の中に含まれる単語を容易に認定することができる．たとえば，上述の例から「He」「read」「a」「book」などの単語を認定することができる．しかし，日本語とマレー語には，単語間を区切るわかち書きの習慣がない．日本語の場合は「、」や「。」などの句読点を手掛かりに，また，マレー語の場合は空白を手掛かりに**文の成分となる文字列** (以下,「**連結成分**」と呼ぶ) を認定できる．たとえば，文 (1) からは「彼は」「本を読んだ」などの日本語の連結成分を，文 (3) からは「Dia」「membaca」「buku」などのマレー語の連結成分を容易に認定できる．しかし，日本語やマレー語の連結成分から索引語となる単語を認定することは，容易ではない．

一般的に，文書中には，表層的に異なる意味の同じ文字列が多数含まれている．たとえば，英語の文書中には read と reads など，日本語の文書中には「読む」と「読ん」など，マレー語の文書中には baca(読むこと) と membaca(～を～する+読むこと) などが含まれる．テキストマイニングの効果を高めるためには，このような表層的な異なりを取り除き，同じ意味の文字列をひとつの単語にまとめあげる接辞処理や形態素解析が必要になる．

英語の場合，加算名詞の複数を表す語尾+s (例: books) や，規則動詞の三人称単数現在形の活用語尾+s (例：reads) を接辞処理によって取り除き，語幹 (例: book や read など) を得る．日本語の場合，連結成分から助詞 (例: を) や助動詞 (例: だ) などを取り除き，名詞 (例：本) や動詞の原形 (例：読む) を得るために形態素解析を行う．

英語では，Porter [6] や Lovins [7] などの**接辞処理のアルゴリズム** (以下,「**ステマー**」と呼ぶ) が一般的に用いられている．また，日本語では，ChaSen [8] や MeCab [9] などの形態素解析器が一般的に用いられている．しかし，マレー語には，一般的に用いられているステマーは存在しない．

表 3　マレー語の接辞

	語幹	接辞	派生語	派生語の構造
接頭辞	baca（読む）	*mem+*	*mem*baca（～を読む）	接頭辞 + 語幹
接尾辞	baca（読む）	*+an*	baca*an*（読み物，読本）	語幹 + 接尾辞
接中辞	gigi（歯）	*+er+*	g*er*igi（鋸 (のこぎり)）	語幹前半 + 接中辞 + 語幹後半
接周辞	baca（読む）	*mem+kan*	*mem*baca*kan*（～を～に読んであげる）	接周辞前半 + 語幹 + 接周辞後半

表 4　語幹による接辞 *me+*, *pe+* の形態変化規則

形態	語幹の頭文字
me+, *pe+*	l, m, n, r, w
mem+, *pem+*	b, **f**, **p**, v
men+, *pen+*	c, d, j, **t**, z
meng+, *peng+*	a, **e**, g, h, i, **k**, o, u
meny+, *peny+*	**s**

前述の文 (1) を ChaSen で形態素解析した結果と，文 (3) の連結成分を人手で接辞処理した結果をそれぞれ表 1 と表 2 に示す．表 2 において，英訳と日本語訳は，マレー語の接辞処理の結果に対応する英語と日本語である．

英語の場合，時制によって動詞の語尾が変化する場合 (例: retrieve → retrieve*s*) や，加算名詞の語尾が複数形で変化する場合 (例: language → language*s*) があるが，マレー語には，このような語尾の変化はない．マレー語では，動詞に時制による変化がなく，名詞の複数形は語幹の繰り返し (重語) で表現することになっている．たとえば，動詞 membaca は現在，過去，未来で変化しない．時制は文脈または時を表す副詞 (besok; 明日) などで判断する．また，語幹 buku は単独で「一冊の本」を表し，語幹 buku を 2 つ反復してハイフンで結合した buku-buku は「複数の本」を表す．マレー語では，代名詞はごく少数に限られ，第三人称の男女の区別 (彼，彼女) はない．

日本語の場合 (表 1) は，形態素解析の結果から「名詞・一般」や「動詞・自立」の原形 (「本」や「読む」など) を取り出すことで，索引語となる単語を認定できる．マレー語の場合 (表 2) は，接辞処理の結果から，連結成分の語幹を取りだすことで，索引語となる単語を認定できる．マレー語の連結成分には，1 個以上の語幹に 0 個以上の接辞が含まれる．そこで，マレー語を対象としたテキストマイニングでは，表 2 のような接辞処理を，手動ではなく自動で行えることが必要になる．

接辞処理は，語幹に関する処理であることから，マレー語の語幹に対する接辞の付き方について，次に説明する．

マレー語の接辞は，接頭辞 (prefix)，接尾辞 (suffix)，接中辞 (infix)，接周辞 (confix) の大きく 4 つに分類される．表 3 にマレー語の接辞の例を示す．接中辞は語幹の間に挿入される特殊な接辞であり，接中辞がつく語幹はごく少数に限られている(注1)．接周辞は，接頭辞と接尾辞の組み合わせで構成される複合的な接辞(注2)であり，マレー語において広く用いられている．たとえば，接周辞 me+kan が語幹に付いて構成される連結成分は，「～を～にさせる」という使役の意味を持つ動詞や「～のために～をする」という便宜供与を表す動詞になる．

通常，マレー語の接辞はそのままの形態で語幹に付くが，接辞と語幹の先頭文字の組み合わせによっては，接辞の形態が変化することがある [5]．たとえば，接辞 *me+*，*pe+* の形態変化規則は表 4 のようになっている．接辞 *me+* は語幹の先頭文字によって，*mem+*, *men+*, *meng+*, *meny+* などの形態変化をした後に付く．表 3 の例では，接辞 *me+* が *mem+* の形態に変化して，b で始まる語幹 baca に付いている．

また，表 4 において太字で示した **f, p, t, k, s** を先頭文字に含む語幹に対して，接辞 *me+*，*pe+* が付く場合は，語幹の先頭文字が消失するという規則がある．たとえば，語幹 sedia(準備する) に，接周辞 *meny* + *kan* が付く場合，“*meny* sedia *kan*” ではなく，“*meny* edia *kan*” となる．

接辞処理は，基本的には，上記のようなルールで語幹についた接辞を除去して語幹に戻す処理となる．たとえば，連結成分 “*mem* baca *kan*” から接周辞 *mem* + *kan* を取り除いて語幹 baca を得ることができる．また，連結成分 “*meny* edia *kan*” から接周辞 *meny* + *kan* を取り除いて，消失した先頭文字 s を復活させると語幹 sedia を得ることができる．

マレー語の話者が接辞処理を行う際には，経験上，接辞や語幹になる文字列パターンや文字列の意味が分かるので，連結成分から接辞と思われる部分を適切に取り除くことができる．計算機に接辞処理を行わせる際には，過不足なく接辞処理を行って接辞処理の誤りを軽減しつつ，かつ，計算量が爆発しないように効率よく接辞処理を行うアルゴリズムが必要になる．

マレー語のステマーに関する先行研究は，情報検索システムにおける検索性能向上を目的としているものがほとんどであり ([10]～[13])，先行研究の評価実験 [13] で扱っている文書は，「コーラン (イスラム教の聖典)」のマレー語の翻訳 10 章分など，文書の数やトピックの範囲が限られている．これに対して，本研究では，多様な Web 文書を対象として，テキストマイニングにおいて効果的なステマーの開発を目的としている．

(注1)：日本語にはない．

(注2)：日本語では，敬語の接周辞「お + する」や「お + になる」などがある．例:「お伝えする」「お話になる」

3. マレー語のステマーの開発

我々のステマーの目的は，マレー語の文書群から，精度の良い決定則の抽出が行えるようにすることである．決定則は，適合文書 (ポジティブクラス) と不適合文書 (ネガティブクラス) のラベルが付与された文書群 (学習用データ) から抽出され，抽出された決定則を用いて，未知の文書群 (テスト用データ) をポジティブクラスとネガティブクラスに分類する．たとえば，トピック "baca"(読む) をポジティブクラスとして，マレー語の文書をポジティブクラス (READ) とネガティブクラス (¯READ) に二値分類する場合に，図 1 のような決定則を抽出すると，(1)〜(3) 以外の派生語 (たとえば，"*pem* baca" や "*pem* baca *an*") についての文書がポジティブクラスに分類されなくなる (参考: 表 3 の派生語)．文書群に対して接辞処理を行って，図 2 のような決定則を抽出すれば，語幹 baca を含む文書が漏れなくポジティブクラスに分類されるようになる (参考: 表 3 の語幹)．文書群からの決定則抽出の流れは以下のようになる．

（1） 連結成分の抽出: 空白を手がかりとして，マレー語の文書から連結成分を抽出する．

（2） 接辞処理: 抽出した連結成分に対して接辞処理のアルゴリズムを適用して，不要な接辞を除去する．

（3） 索引処理: 不要な接辞を除去した連結成分を索引語として，文書と索引語の対応関係を作成する．

（4） 決定則抽出: 学習用データ (ラベル付きの文書群) から，ポジティブクラスに含まれ，かつ，ネガティブクラスに含まれない索引語を抽出する．

接辞処理に過不足があると，決定則の分類精度・再現率が悪くなる．接辞処理が不足すると，たとえば，図 1 のように，適合文書に対して網羅的でない決定則が抽出され，文書分類の再現率が悪くなる．また，接辞処理が過度に行われ，意味の異なる語幹や派生語までが同一視されると，適合文書と不適合文書が区別できなくなり，文書分類の精度が悪くなる．ステマーが避けるべき接辞処理の過不足は，以下の 2 つのタイプに分類される．

Type1: 除去すべき接辞を除去しない．(under-stemming)
Type2: 除去すべきでない接辞を除去する．(over-stemming)

```
(1) membaca --> READ
(2) bacaan --> READ
(3) membacakan  --> READ
(4) [TRUE] --> ¯READ
```

図 1 決定則の例 (接辞処理・なし)

```
(1) baca --> READ
(2) [TRUE] --> ¯READ
```

図 2 決定則の例 (接辞処理・あり)

Input: 連結成分を要素とする配列 IN，接辞ルールを要素とする配列 RULE，派生語辞書 D，語幹辞書 S
Step-0 カウンタ変数 *i* を初期化する．(*i* = 0)
Step-1 配列 IN から *i* 番目の要素を読みだし，変数 *string* にコピーする．読みだす要素が無い場合は，処理を終了する．
Step-2 派生語辞書 D の中に，変数 *string* の値と一致する単語が存在する場合は，Step-9 へ．
Step-3 変数 *string* の値を，変数 *stem* にコピーする．
Step-4 カウンタ変数 *j* を初期化する．(*j* = 0)
Step-5 語幹辞書 S の中に，変数 *stem* の値と一致する単語が存在する場合は，Step-8 へ．
Step-6 接辞ルールの配列 RULE から *j* 番目の要素を取り出し，変数 *stem* に接辞ルールを適用（接辞を除去）する．読みだす要素が無い場合は，Step-9 へ．
Step-7 カウンタ変数 *j* をインクリメントして，Step-5 へ．
Step-8 変数 *stem* の値を配列 OUT の *i* 番目にコピーして，Step-10 へ．
Step-9 変数 *string* の値を配列 OUT の *i* 番目にコピーして，Step-10 へ．
Step-10 カウンタ変数 *i* をインクリメントして，Step-1 へ．
Output: 接辞処理済みの連結成分を要素とする配列 OUT

図 3 接辞処理のアルゴリズム

Type2 は，さらに 2 つのタイプに分けられる．

Type2-1: (語幹に対する over-stemming)
語幹の中に，接辞ではないが，接辞と一致する文字列が含まれており，接辞相当の文字列を語幹から除去することで，別の意味の語幹に変わってしまう．
Type2-2: (派生形に対する over-stemming)
連結成分の意味の中心となる語幹が，接辞との組み合わせで意味を持つ派生形になっており，接辞を除去することで，連結成分の意味が大きく変わってしまう．

我々の接辞処理のアルゴリズムでは，Type1 と Type2-1 の誤りを軽減する工夫をした上で，さらに，Type2-2 についても考慮に入れている点に特徴がある．具体的には以下の対策を行っている．

- マレー語の接辞を網羅的に定義した接辞ルールの集合を参照し，Type1 の誤りを軽減する．
- 接辞を除去する際に，語幹辞書 (マレー語の語幹のリスト) を参照し，Type2-1 の誤りを軽減する．
- 接辞を除去する際に，派生語辞書 (語幹に接辞がついた派生形のうち，マレー語の文書においてよく使用される派生語のリスト) を参照し，Type2-2 の誤りを軽減する．

提案する接辞処理のアルゴリズムを図 3 に示す．接辞ルールの適用と辞書の参照について，詳細を次に説明する．

3.1 接辞ルールの適用

開発したステマーは，Othman ら，および，Ahmad らが定義した 432 個のマレー語の接辞ルール(注3)から構成される接辞ルールの集合を参照する．接辞ルールとは，表 3 に示した接辞 mem+，+an などの接辞の文字列パターンである．Ahmad は，マレー語の新語に対応するための 129 個の接辞ルール(注4)についても定義したが，新語のための接辞ルールは，新語以外の単語に対して誤った接辞処理を行ってしまうため，我々のステマーでは用いないこととした．また，接辞ルールの適用のみでは接辞付加による形態変化に対応できないため [14]，形態変化有りの接辞を除去する際に，表 4 の規則に従い，語幹の補正を行うこととした．

連結成分に対して適用可能な接辞ルールは複数存在し得るため，接辞ルール適用の順序を工夫する必要がある．先行研究では，コーラン (イスラム教の聖典) や研究論文の概要 10 件などの特定の小規模なデータセットを対象とした実験を行い，データセットごとに適切なルール適用順序を採用することで，接辞処理のエラーが軽減できると報告している [10]．しかし，我々のステマーでは未知の多様な文書データを対象としているため，接辞ルールの適用順序を特定の小規模なデータセットに合わせて固定することは望ましくない．マレー語の文書中で接中辞が用いられることは，ほとんどなく，Type2-1 のエラーを回避するためには，接中辞の適用は，他のルールで語幹が得られない場合に限定するべきである．そこで，我々のステマーでは，マレー語の文書において使用頻度が高い，接周辞，接頭辞，接尾辞，接中辞の順で接辞ルールを適用することとした．さらに，連結成分によっては，除去の候補となる接周辞，接頭辞，接尾辞がそれぞれ複数存在し得るため，接周辞，接頭辞，接尾辞ごとの接辞ルール適用の順序についても考慮する必要がある．我々のステマーでは，接辞の文字列長が長い順で接辞ルールを適用し，語幹が得られなかった例外的な連結成分には，さらに，文字列長が短い順で接辞ルールを適用することとした．

3.2 語幹辞書の参照

語幹に対する接辞ルールの適用 (Type2-1 のエラー) を回避するため，Ahmad らは接辞処理の際に，マレー語の語幹辞書を参照し，語幹に対する接辞ルールの適用を回避することを提案している [10]．そこで，我々の開発したステマーでも，Ahmad らと同様に接辞ルール適用の際にマレー語の語幹辞書を参照することとした．Ahmad らは，マレー語の辞書 “Kamus Dewan” の見出し語を参照し，ステマーの評価を行った結果，辞書の見出し語 22,393 語には，語幹と接辞で構成される派生語も多数含まれていため，接辞処理の漏れ (Type1 のエラー) が生じたと報告している [10]．そこで，我々のステマーでは，マレー語の基本的な語幹のみが掲載されている辞書 “Kamus Dwibahasa” [17](以下，「語幹辞書 KD」と呼ぶ) の見出し語 5,774 語を参照することとした．

3.3 派生語辞書の参照

英語を対象とした先行研究では，接辞処理により，派生形の最小語幹による関連づけが行われると，検索精度が大幅に低下する場合があると報告されている [15]．

マレー語を対象とした先行研究では，この問題は議論されていないが，マレー語においても英語の場合と同様に，派生形に対する過度な接辞除去は避けるべきである．たとえば，“waris*an*”(文化遺産) と “*me*waris*i*”(受け継ぐ) から接辞を除去すると，語幹 “waris”(相続人) が得られる．このような派生形をすべて語幹 “waris” に変換すると，「文化遺産」の文書と「遺産相続」の文書が区別できなくなるなどの問題が生じる．そこで、このような過度な接辞処理による悪影響を避けるため，我々のステマーでは，一般的に用いられるマレー語の派生語は接辞処理の対象外とすることとした．具体的には，マレー語とインドネシア語のスペルチェッカー [18] の辞書に掲載されている単語を用いて，派生語辞書 (以下，「派生語辞書 MYID」と呼ぶ) を作成し，派生語辞書 MYID に含まれる単語には接辞処理を行わないこととした．派生語辞書 MYID にインドネシア語の単語を含めたのは，マレー語の Web 文書にはインドネシア語の表記が混在する場合があり，インドネシア語の単語に対しても過度な正規化を回避する必要があるためである．

4. 評価実験

提案アルゴリズムの有効性を評価するため，ステマーを PHP で実装し，実装したステマーと実際の WWW 上の多様な文書データを用いて，評価実験を行った．

4.1 実験用データセット

検索エンジン API [16] を用いて表 5 に示すトピックごとに Web 検索を行い，Web ページを収集した．具体的には，トピックを表すマレー語の単語 (表 5 の agama など) を検索語として用いて，各トピックごとに約 1000 件の URL を検索し，検索結果 URL に対応付けられている Web ページ (キャッシュページ) をダウンロードして，各 Web ページからテキストデータの抽出を行い，1 件の Web ページから 1 件の文書データを作成した．さらに，トピック間で重複している文書データは取り除き，トピック間で重複のない文書データを各トピックの適合文書とした．各トピックごとに，平均 382.5 件の適合文書データを収集した．トピックごとの検索結果 URL 数，Web ページ数 (トピック間での重複含む)，トピックに適合な文書データ数 (トピック間での重複なし) を図 4 に示す．

4.2 ステマーの単体評価

開発したステマーを，派生語辞書なしで，単体で動作させ，接辞除去 (図 3 の Step-3 から Step-9) の正確さを評価す

(注3)：文献 [10] の付録 A と付録 B.
(注4)：文献 [10] の付録 C.

表 5　実験用のトピック

agama(宗教), budaya(文化), ekonomi(経済), hiburan(娯楽), kesihatan(健康), komputer(コンピュータ), makanan(グルメ), pendidikan(教育), politik(政治), sukan(スポーツ)

表 6　トピックの関連単語

umat masjid sembahyang warisan seni pertumbuhan pembangunan berekreasi beriadah berlagu lagu lagumu lagunya muzik muzikal pemuzik rekreasi riadah senilagu hospital klinik vitamin sakit internet gourmet parti kerajaan sukma liga piala

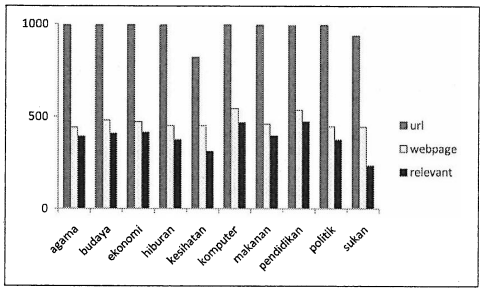

図 4　実験用データ (トピックごとのファイル数)

る．表 5 の単語，及び，各トピックと関連の深いキーワード (表 6) を，部分文字列として含む連結成分 943 個を実験用データから抽出し，開発したステマーを適用して接辞処理を行った．ステマーにより誤った接辞除去が行われた連結成分は 20 個，行うべき接辞除去が行われなかった連結成分は 13 個であった．評価用の連結成分に対するエラー率は約 3.5%である．誤った接辞処理が行われた連結成分を表 7 に示す．また，接辞処理が行われなかった連結成分を表 8 に示す．主なエラーの原因は，以下の通りである．

- 連結成分の語幹が語幹辞書に含まれていない．
- 連結成分の接辞がルール集合に含まれていない．
- 接辞ルールの適用順序が固定されているため，複数の接辞ルールの適用可能性や，複数の語幹候補がある場合に柔軟に対応できない．

マレー語の話者は，連結成分に未知の接辞や未知の語幹が含まれている場合に，前後の文脈と経験的な知識を用いて柔軟に対応できるが，開発したステマーは，未知の文字列に対応できない．接辞や語幹のどちらかが分かれば不要な接辞部分を推測して除去する工夫や，前後の文脈に応じて最適な語幹の選択を行えるようにする工夫が必要である．

4.3　ステマーの結合評価

決定則抽出におけるステマーの有効性を検証するため，実験用データと文書群からの決定則抽出を行うソフトウエア，RIKTEXT(Rule Induction Kit for Text) [2] を用いた評価実験を行った．

決定則の分類精度を調べるためには，各トピックごとの適合・不適合のラベルが付与された学習用データとテスト用データが必要である．そこで，実験用データから，各トピックごとに 10 個ずつ，トピックに適合なデータを手動で取り出し，合計で 100 個の文書データから成るテスト用データを作成した．また，テスト用データとして選ばなかった残りの文書データを学習用データとした．

提案アルゴリズムは，派生語辞書を参照するという点に先行研究とは異なる特徴がある．そこで，接辞処理なし (従来法 1)，及び，先行研究で行われている派生語辞書なしの接辞処理 (従来法 2) と，接辞ルール，語幹辞書，派生語辞書の 3 つを参照する接辞処理 (提案法) の比較を行った．トピックごとの決定則の精度と再現率を図 5 と図 6 に示す．

図 5 を見ると分かるように，提案法 (過度な正規化を避けた接辞処理) では，多くのトピックにおいて，従来法よりも精度が向上している．トピック「agama(宗教)」では，従来法 1 と比較して提案法は精度がやや悪化しているが，これは，トピック「agama(宗教)」は，トピック「budaya(文化)」と関連があるため，もともとトピック「agama(宗教)」とトピック「budaya(文化)」が分離できない内容であったことが原因と考えられる．具体的には，提案法では，トピック「agama(宗教)」の決定則に「buddhism（仏教)」などマレーシアの文化と宗教の両方に関連する単語が含まれ，これにより，決定則の精度が悪くなってしまった．

他のトピックでは，提案法は，従来法 2 が従来法 1 よりも精度が良い場合 (接辞処理を行うことが効果的である場合) は，従来法 2 と同程度の精度となり，また，従来法 1 の方が従来法 2 よりも精度が良い場合 (接辞処理を控えた方が効果的である場合) は，従来法 1 と同程度の精度となっている．つまり，提案法は，従来法 1 と従来法 2 の処理を選択的に行い，過度な正規化を避け，効果的に接辞を除去していると言える．

先行研究では，情報検索システムを対象としてステマーの評価を行い，ステマーによって，再現率は向上するが，精度は低下すると報告されている [13]．これに対して，我々の接辞処理のアルゴリズムでは，同じ意味を持つ派生形が適切にまとめられ，文書分類の精度が向上することが確認された．従来法 1，従来法 2，提案法の精度，再現率，F 値の平均値を表 9 に示す．F 値を見てみると，従来法 2 の性能が従来法 1，提案法と比べて，悪くなっていることから，従来法 2 のような過度な接辞処理は，文書検索や文書分類の性能向上には役立たないと言える．

提案法は，従来法 1 と比較して，F 値ではあまり差がないが，精度の平均値では，提案法は従来法を上回っている．このことから，提案法は，Web 文書など大量の文書データを対象として，再現率よりも，精度を重視する文書自動分類を行いたい場合に適していると言える．

表 7　誤った接辞処理が行われた単語

単語	語幹 (正解)	ステミングの結果	エラーの種類
*anti*poligami	poligami（一夫多妻）	gam（糊 (のり)）	over-stemming
*ber*maklumat	maklumat（情報）	maklum（知る）	over-stemming
*ber*pasukan	pasukan（チーム）	pasu（花瓶）	over-stemming
*ber*poligami	poligami（一夫多妻）	gam（糊 (のり)）	over-stemming
*ber se*pasukan	pasukan（チーム）	pasu（花瓶）	over-stemming
*ber*tema*kan*	tema（テーマ）	makan（食べる）	over-stemming + under-stemming
*di*poligami	poligami（一夫多妻）	gam（糊 (のり)）	over-stemming
gema*kan*	gema（反応）	makan（食べる）	over-stemming + under-stemming
*ke*sukan*an*	sukan（スポーツ）	kanan（右）	over-stemming + under-stemming
makan*mu*	makan（食べる）	mak（母）	over-stemming
*me*masu*kan*	masuk（入る）	pasu（花瓶）	over-stemming
*me*masuk*kannya*	masuk（入る）	pasu（花瓶）	over-stemming
pasukan*nya*	pasukan（チーム）	pasu（花瓶）	over-stemming
*pen*didik*an*	didik（教育する）	di（〜で (場所を表す前置詞)）	over-stemming
peny(s)emak*an*	semak（検査する）	makan（食べる）	over-stemming + under-stemming
poligami	poligami（一夫多妻）	gam（糊 (のり)）	over-stemming
politik*us*	politik（政治）	tikus（鼠 (ねずみ)）	over-stemming + under-stemming
politik*uslah*	politik（政治）	tikus（鼠 (ねずみ)）	over-stemming + under-stemming
ragam*an*	ragam（態度）	raga（誇示する）	over-stemming
*se*pasukan	pasukan（チーム）	pasu（花瓶）	over-stemming

表 8　接辞処理が行われなかった単語

単語	語幹 (正解)	ステミングの結果	エラーの種類
*ber*vitamin	vitamin（ビタミン）	bervitamin ([派生語] ビタミンを〜する)	under-stemming
*di*nasihat	nasihat（アドバイス）	dinasihat ([派生語] アドバイスする)	under-stemming
internet*ku*	internet（インターネット）	internetku ([派生語] 私のインターネット)	under-stemming
internet*nya*	internet（インターネット）	internetnya ([派生語] 彼らのインターネット)	under-stemming
klinikal*nya*	klinik（診療所）	klinikalnya ([派生語] 彼らの臨床)	under-stemming
makan*lah*	makan（食べる）	makanlah ([派生語] 食べなさい)	under-stemming
*me*maksima*kan*	maksima（最大）	memaksimakan ([派生語] 最大化する)	under-stemming
*me*minima*kan*	minima（最小）	meminimakan ([派生語] 最小化する)	under-stemming
*meng*optima*kan*	optima（最適）	mengoptimakan ([派生語] 最適化する)	under-stemming
muzik*al*	muzik（音楽）	muzikal ([派生語] 音楽の)	under-stemming
*pe*nasihat	nasihat（アドバイス）	penasihat ([派生語] 相談相手)	under-stemming
politik*al*	politik（政治）	politikal ([派生語] 政治の)	under-stemming
seni*nya*	seni（芸術）	seninya ([派生語] 彼らの芸術)	under-stemming

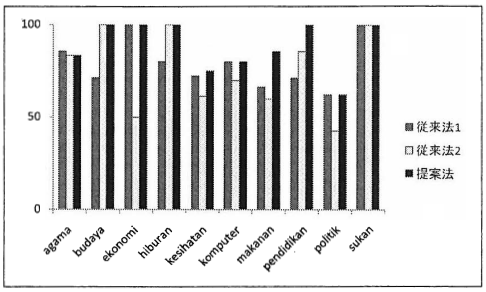

図 5　決定則の精度 (%)

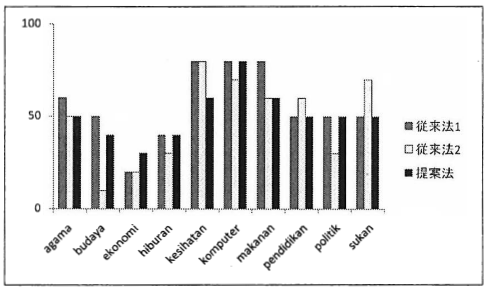

図 6　決定則の再現率 (%)

表 9　決定則の精度・再現率・F 値の平均

	従来法 1	従来法 2	提案法
精度	79.0	75.3	88.7
再現率	56.0	48.0	51.0
F 値	62.6	54.3	62.9

5. おわりに

本稿では，マレー語を対象とした接辞処理のアルゴリズムを提案した．また，提案法に基づくステマーを実装し，ステマーを適用することで文書自動分類の精度が向上することを実際のマレー語のテキストデータを用いた評価実験により確認した．

単語の多様性を正規化する接辞処理は，接辞が文法上重要な役割を果たすマレー語においては，特に重要なテキスト処理である．しかし，意味の異なる派生語を関連付ける，過度な接辞の除去を行うと，かえって，接辞処理が逆効果になる場合もある．この問題を解決するため，提案アルゴリズムでは，できるだけ派生形を同一語幹にまとめつつ，使用頻度の高い一般的な派生形に対しては接辞処理を控える，という選択的な接辞処理を行っている．

提案アルゴリズムは，接辞ルールの適用順序を固定しているが，接辞ルールの候補が複数あるときは，文脈に応じた接辞ルール適用を行うことが望ましい．今後の課題として，あらかじめ大量のコーパスで接辞ルールの適用順序を学習しておき，文脈に応じた接辞ルールの適用を行うことを考えている．

文　献

[1] Weiss, S.M., Indurkhya, N., Zhang, T., and Damerau, F., "Text Mining," 2005.

[2] Indurkhya, N., "RIKTEXT:Rule Induction Kit for Text," 2004.

[3] 徳永健伸, 情報検索と言語処理, 東京大学出版会, 1999.

[4] ファリダ・モハメッド, 近藤由美, "CD エクスプレス マレー語," 白水社, 2005.

[5] 正保 勇, 聴いて, 話すための-マレーシア語基本単語 2000, 語研, 1993.

[6] Porter, M.F., "An Algorithm for Suffix Stripping, Program," Vol. 14, No. 3, pp. 130–137, 1980.

[7] LOVINS, J.B., "Development of a Stemming Algorithm," Mechanical Translation and computation Linguistics, Vol. 11 No. 1, pp. 23–31, 1968.

[8] ChaSen, http://chasen.aist-nara.ac.jp/

[9] MeCab, http://mecab.sourceforge.net/

[10] Ahmad, F., Yusoff, M., Sembok, T.M., "Experiments with a Stemming Algorithm for Malay Words," JASIS, Vol. 47, No. 12, pp.909–918, 1996.

[11] Abdullah, M.T., Ahmad, F., Yusoff, Mahmod, R., M., Sembok, T.M., "A Stemming Algorithm for Malay Language," Proc. of International Conference on Information Technology in Asia, pp.181–186, 2005.

[12] Abdullah, M. T., Ahmad, F., Mahmod, R., and Sembok, T. M., "Improvement of Malay Information Retrieval Using Local Stop Words," Proc. of International Advanced Technology Congress, 2005.

[13] Tai, S.Y., Ong, C.S., Abdullah, N.A., "On Designing an Automated Malaysian Stemer for the Malay Language." Proc. of The Fifth International Workshop on Information Retrieval with Asian Languages, Hong Kong, September 2000.

[14] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M.M., Williams, H.E, Stemming Indonesian:A Confix-Stripping Approach, ACM Transactions on Asian Language Information Processiong, Vol. 6, No. 4, Article 13, 2007.

[15] 本間咲子, 全文検索における英語接辞処理の評価, 自然言語処理, pp.138–8, 2000.

[16] Yahoo!JAPAN DEVELOPER NETWORK, http://developer.yahoo.co.jp/

[17] Kamus dwibahasa: Inggeris-Malayu Malayu-Inggeris, pelanduk, 1993.

[18] Dicitionaries, OpenOffice.org, http://wiki.services.openoffice.org/wiki/Dictionaries