現代モンゴル語の接辞処理と索引語抽出への応用

Khaltar Badam- Osor　　藤井 敦
筑波大学大学院図書館情報メディア研究科
E-mail: {khab23, fujii}@slis.tsukuba.ac.jp

1. はじめに

現代モンゴル語（以下、単に「モンゴル語」）はキリル文字を使用し、文は文節の単位で分かち書きされる。文節は自立語に付属語が接続して構成される。接続の際に、自立語と付属語の語形が変化することがある。モンゴル語の自立語と付属語を分割し、更に原形を特定することは、自然言語処理や種々の応用において重要である。

情報検索では、文書の内容を表す索引語を抽出するために、付属語を分割して、自立語を使う。この処理を「接辞処理」という。現在、Google や Yahoo!などの検索エンジンでは、モンゴル語の接辞処理が行われていない。例えば、「Үндсэн хууль（憲法）」を検索質問として検索した場合、「Үндсэн хууль（憲法）」という原形を含むページは検索される。しかし、原形を含まずに、「Үндсэн хуулийн（憲法の）」や「Үндсэн хуулиас（憲法から）」のような語形変化だけを含むページは検索することができない。

本研究は、モンゴル語の接辞処理手法を提案し、情報検索の索引語抽出に応用する。

2. モンゴル語における自立語と付属語の接続

モンゴル語では、名詞に接続する付属語は格助詞である。形容詞には、名詞に接続する格助詞のうち複数形を表す助詞以外の格助詞を接続することができる。動詞に接続する付属語は動詞の活用形を表す。本研究では、名詞と形容詞の格助詞と動詞の活用語尾を「接辞」と総称する。

モンゴル語では、同じ意味を表す複数の接辞がある。例えば、属格（「の」）を表す接辞には「**ын**」、「**ийн**」、「**ы**」、「**ий**」、「**н**」がある。これらは自立語に接続するとき、自立語に含まれる母音の性と自立語の末尾が重要である。母音の性には、男性（**а, о, у**）、女性（**ө, ү, э**）、中性（**и**）がある。例えば、男性自立語に接続する属格の付属語は「**ын**」である。しかし、自立語の末尾が「**ж, ч, ш, г, ь, и**」のいずれかであれば「**ийн**」が接続する。「**ийн**」は女性自立語にも接続する。以下の例では、「**ах**（兄）」が男性自立語であり、末尾が「**ж, ч, ш, г, ь, и**」ではないので、「**ын**」が接続する。「**ээж**（母）」は女性自立語であるため、「**ийн**」が接続する。

ах + ын → ахын
兄　の　兄の
ээж + ийн → ээжийн
母　の　母の

モンゴル語における自立語と接辞の接続パターンについて図 1 を用いて説明する。

(ア)では、「**ном**（本）」に「**ын**（の）」が語形変化せずに接続している。それに対して(イ)～(オ)では、語形変化が生じている。(イ)では、「**яв**（行く）」に「**х**（未来形）」が接続する際、「**а**」が挿入されている。(ウ)では、「**байшин**（建物）」に「**аас**（から）」が接続する際、「**г**」が挿入されている。(エ)では、「**харь**（帰る）」に「**в**（終止形）」が接続する際、「**харь**（帰る）」の最後の「**ь**」が削除され、接辞の先頭文字が「**и**」に変化している。(オ)では、「**ажил**（仕事）」に「**ын**（の）」が接続する際、「**и**」が削除されている。

(ア) 語形変化せずに接続する	**ном + ын → номын** 本　の　本の
(イ) 母音の挿入	**яв + х → явах** 行く　未来形　行く
(ウ) 子音の挿入	**байшин + аас → байшингаас** 建物　から　建物から
(エ) 記号文字「ь」が削除され、接辞の先頭文字が「и」に変化	**харь + аад → хариад** 帰る　副動詞形　帰って
(オ) 母音の削除	**ажил + ын → ажлын** 仕事　の　仕事の

図 1　自立語と接辞の接続パターン

自立語が外来語である場合は、モンゴル語固有の接続パターンに従わないことがある。例えば、以下の例では、「**станц**（ステーション）」という外来語は男性自立語であり、末尾が「**ж, ч, ш, г, ь, и**」のいずれでもない。そこで、本来ならば「**ын**」が接続する。しかし、実際は「**ийн**」が接続する。

станц　+　ийн　→　станцийн
ステーション　の　ステーションの

また、図 1 の(オ)は自立語が外来語の場合は生じない。そのため、外来語には外来語特有の接続規則が必要である。

外来語には名詞だけではなく、形容詞や動詞もある。モンゴル語において外来語動詞は外来語名詞にモンゴル語の動詞形成接尾辞が接続することで生成される。例えば、「**систем**（システム）」という外来語名詞に「**чил**（～化）」という動詞形成接尾辞が接続して「**системчил**（システム化）」という動詞に派生する。外来語名詞に動詞形成接尾辞が接続するために、外来語名詞の接続規則が必要である。生成された動詞は、モンゴル語固有の動詞と同じように活用する。ただし、本研究では、動詞に派生した単語は名詞に復元しない。

3. 先行研究の検討と本研究の位置付け

Sanduijav ら[1]は名詞と動詞の自立語に付属語が接続する際の音韻的・形態論的制約を人手で作成して、その制約に基づいて自立語と付属語の活用形を自動的に作成した。Sanduijav らはその結果を語形変化テーブルに登録して、語形変化テーブルを参照することで接辞処理を行った。ウェブ上のモンゴル語新聞 1.5 年分から無作為に抽出した 680 語を対象に実験を行った結果、680 語のうち 587 語が辞書に載っており、これらは全て自立語と付属語に正しく分割された。

江原ら[2]は日本語の形態素解析システム茶筌を処理系としてモンゴル語文の形態素解析を行った。江原らは自立語と付属語の語形変化を人手で作成した。

Sanduijav らと江原らの手法では、名詞辞書を利用して

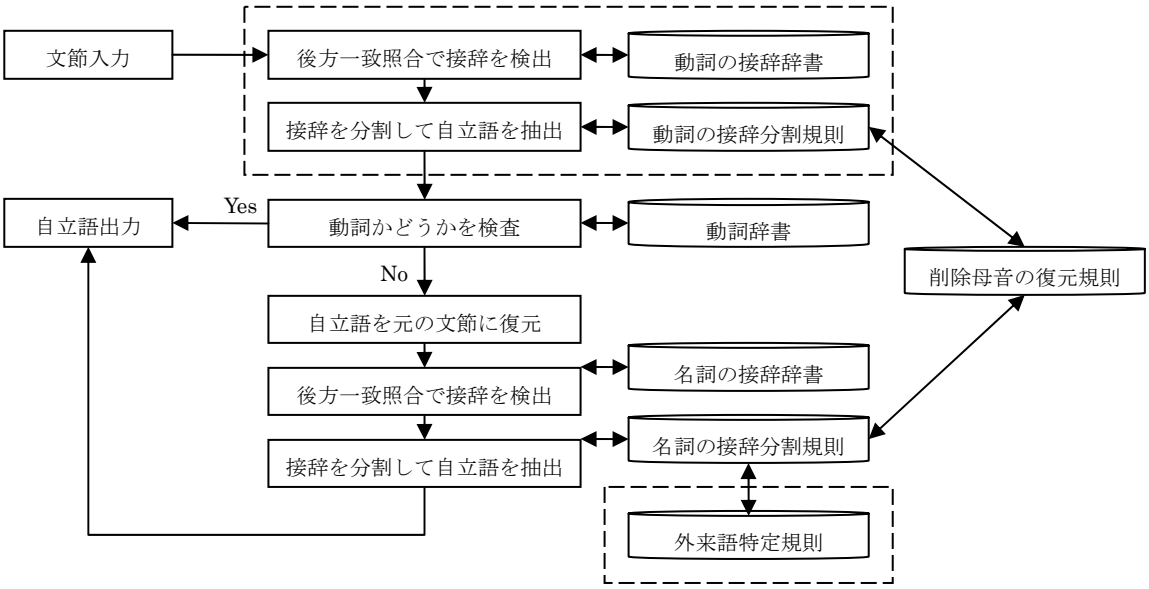

図２　モンゴル語を対象とした接辞処理システムの構成

いるため、辞書に登録されていない名詞は接辞処理を行うことができないという問題がある。

Khaltar ら[3]は名詞辞書を利用しない接辞処理手法を提案した。しかし、接辞処理において、外来語特有の接続規則を考慮していない。本研究では、外来語の接続規則を考慮し、更に動詞も対象にすることで Khaltar らの手法を拡張する。

4. 本研究で提案する接辞処理の手法

索引語は名詞、動詞、形容詞などの自立語に限定されることが多いため、本研究は、名詞、動詞、形容詞の接辞処理手法を提案する。

専門用語、固有名詞などの新語は名詞であることが多いため、本手法は名詞辞書を利用しない。しかし、動詞は名詞に比べると新語の出現が少ないため、動詞辞書を利用する。形容詞は名詞と同様に活用するため、名詞と同一の接辞処理手法を提案する。以降、名詞と形容詞の接辞処理をあわせて「名詞の接辞処理」と呼ぶ。接辞処理システムの構成を図 2 に示す。図 2 において、破線で囲んだ部分は今回拡張した部分であり、それ以外は Khaltar らの手法と共通である。

図 2 では、まず、入力された文節を「動詞の接辞辞書」と後方一致照合し、接辞を検出する。次に「動詞の接辞分割規則」を利用して接辞を分割して、原形に復元して、動詞を抽出する。

名詞と動詞に接続する接辞には同じ形の接辞がある。そこで、名詞の接辞が動詞の接辞として誤って処理される問題が生じる。この問題を解決するために、本研究では動詞辞書を利用して、抽出した自立語が動詞辞書に登録されているかを検査する。

対象の単語が「動詞辞書」に存在すれば自立語として出力する。しかし、「動詞辞書」に存在しない単語は最初に入力された文節の形で名詞の接辞処理に渡す。入力された文節を「名詞の接辞辞書」と後方一致照合し、接辞を検出する。次に、「名詞の接辞分割規則」を利用して接辞を分割し、さらに名詞を原形に復元して抽出する。

図 1 の(ア)~(エ)は「接辞分割規則」で扱われており、名詞と動詞によって分割規則が異なる。しかし、図 1 の(オ)は「削除母音の復元規則」で扱われており、「削除母音の復元規則」は名詞と動詞に共通である。

しかし、外来語名詞の場合は母音の削除が生じないため、母音を復元してはいけない。そのため、名詞が外来語であるかを特定する必要がある。そこで、今回新たに「外来語特定規則」を追加した。外来語名詞に接辞が接続するときに、モンゴル語固有の規則と異なることがあるため、接辞を分割する規則も異なる。そのため、外来語と特定された自立語には、外来語特有の接辞分割規則を適用する必要がある。

自立語に複数の接辞が連続して接続することがあるため、自立語の末尾が接辞辞書中の項目と一致しなくなるまで処理を再帰的に繰り返す。

入力された文節が動詞と名詞の接辞辞書のどちらにも後方一致しない場合は、接辞が接続されていないと見なし、そのまま出力する。

以下、「動詞と名詞の接辞辞書」、「動詞辞書」、「動詞と名詞の接辞分割規則」、「削除母音の復元規則」、「外来語特定規則」について説明する。

動詞と名詞の接辞辞書

「動詞の接辞辞書」には動詞に接続する接辞が 126 件登録されている。接辞が接続する時に、図 1(エ)のように接辞が語形変化する場合があるため、語形変化後の形態も登録されている。「名詞の接辞辞書」には名詞に接続する接辞と語形変化後の接辞が 38 件登録されている。

動詞辞書

Sanduijav[1]らが作成した動詞辞書を利用する。この辞書には 1254 の動詞が登録されている。

動詞と名詞の接辞分割規則

接辞分割規則は、名詞と動詞に接辞が接続する時の語形

変化を考慮して作成した。同一の接辞であっても自立語が名詞か動詞によって文節を分割する境界の位置が異なる。名詞と動詞の接辞分割規則数は 196 あり、その中に外来語特有の接辞分割規則が 23 ある。動詞の接辞分割規則数は 179 である。

図 3 に接辞分割規則の例を示す。属格の「**ийн**（の）」には①、②、③のいずれも後方一致する。その場合は、接辞の前にある文字と自立語の性、または自立語が外来語であるかによって、分割する時に使用する規則が異なる。文節の下線で示した部分が分割される。

①の場合、一致した「**ийн**」の前にある部分に女性母音が含まれている。「**ийн**」は女性自立語に接続する接辞なので、一致した接辞を分割して、残りを自立語として抽出する。②の場合、「長母音で終わる男性自立語に **ийн** を接続する時に子音 **г** を挿入する」という接続パターンを考慮する。一致した「**ийн**」の前にある部分に男性母音が含まれており、末尾が「**г**」である。さらに、「**г**」の前にある文字が「**аа**」長母音である。そこで、「**г**」は接続の際に挿入された子音と判定して、接辞と一緒に分割する。③の場合、外来語特定規則によって外来語であることが分かる。そのため、経験則によって最後の「**йн**」を分割する。

語尾種類	文節	語幹
ийн（の）	① **ээжийн**　（母の）	**ээж**（母）
	② **Хараагийн** （ハラー（川の名）の）	**Хараа** （ハラー）
	③ **геологийн** （地質学の）	**геологи** （地質学）

図 3　接辞分割規則の例

削除母音の復元規則

語形変化で図 1(エ)のように母音の削除が起こった場合は、自立語を抽出する際に削除母音を復元しなければならない。母音の削除があったかどうかは、接辞を分割した後で自立語の末尾にある 2 文字を調べることで分かる。

抽出された自立語の末尾にある 2 文字が子音の連続であった場合は、その子音の間にある母音が削除されたと判定する。しかし、元の自立語が子音の連続で終わる場合もある。そこで、モンゴル語の文法教科書[4]を参考にして、どのような子音連続の時に削除母音を復元するかを 6 通りの規則で表現した。例えば、抽出された自立語の最後にある 2 文字が子音「**м**」、「**г**」、「**л**」、「**б**」、「**в**」、「**р**」のいずれか 2 つであればそれらの間に母音を復元する。しかし、これらいずれかの後ろに子音「**ц**」、「**ж**」、「**з**」、「**с**」、「**д**」、「**т**」、「**ш**」、「**ч**」、「**х**」のいずれかが連続していれば、その間には母音を復元しない。

外来語特定規則

外来語特定規則として、Khaltar ら[3]が提案した(a)~(f)の外来語抽出規則を使用する。これらの規則は名詞、動詞、形容詞のいずれにも適用することができる。Khaltar らは以下の条件に合致する単語を外来語であると特定した。

- 外来語に特有な 4 つの子音を含む単語
- 母音調和規則に違反する単語
- 語頭が子音の連続である単語
- 語尾に特定の子音が連続する単語
- 「**в**」で始まる単語
- 「**р**」で始まる単語

Khaltar らはこれらの規則を用いて、モンゴル語コーパスから外来語を抽出した。しかし、本研究は接辞処理において、使用する規則を変更するために外来語特定規則を利用する。

5. 評価実験

本手法の有効性を評価するために、モンゴル語のウェブサイト（http://www.itpark.mn）から収集したモンゴル語の研究抄録 1102 件を用いて、実験を行った。研究抄録の延べ文節数は 178498、異なり文節数は 23369 である。提案した接辞処理の手法を「接辞処理の正解率」と「情報検索における有効性」という観点で評価した。5.1 節と 5.2 節でそれぞれの評価について説明する。

5.1 接辞処理

モンゴル語の抄録 1102 件に対して接辞処理を行った。接辞処理の正解は、モンゴル人の大学院生 2 名が個別に判定した。名詞、動詞、形容詞ごとに接辞処理の正解率を評価するために、判定者は品詞も特定した。品詞の種類として、「外来語名詞」、「外来語動詞」、「外来語形容詞」も含めた。

本研究では、評価の客観性を高めるために、判定者の判断がどのぐらい一致しているかを調べた。判定者間の一致率を式（1）の Kappa 統計量（K）を用いて計算した。

$$K = \frac{(\text{判断の一致率}) - (\text{偶然の一致率})}{1 - (\text{偶然の一致率})} \qquad (1)$$

判定者の判定が完全に一致している場合は K=1 になる。判定者 2 名の接辞処理と品詞特定の一致率はそれぞれ 0.94 と 0.79 であった。接辞処理の正解判定はほぼ一致しているものの、品詞特定の一致率は低い。多品詞語のために、品詞特定には名詞か形容詞かの判断に不一致があった。例えば、「**өргөн**」は名詞として「広さ」という意味を表し、形容詞として「広い」の意味を表す。

本研究は、判定者 2 名の判定が一致した接辞処理と品詞だけを正解として使用した。

本研究の接辞処理を Sanduijav ら[1]の手法と比較した。Sanduijav らの出力は彼らが作成した「名詞辞書と動詞辞書に登録されている単語」と「活用していない単語」である。この名詞辞書と動詞辞書には、それぞれ 1926 語と、1254 語が登録されている。また、外来語特定規則の有無による正解率の変化も評価した。接辞処理の正解率を比較した結果を表 1 に示す。

表 1 では、Sanduijav らの手法を「先行研究」、「外来語特定規則なし」を「なし」、「外来語特定規則あり」を「あり」と表記している。

表 1 より、本手法の「なし」と「あり」の両方も Sanduijav らの手法より正解率が高かった。ただし、本手法で使用する動詞辞書は Sanduijav らの動詞辞書と同じであるため、動詞の正解率は原理的に同じになる。

表 1 より、「外来語特定規則なし」と「外来語特定規則あり」の結果を比較すると、外来語特定規則を使用したときの外来語名詞と名詞の正解率がそれぞれ 4.3 と 0.7 ポイント向上した。また、形容詞の正解率が 1.1 ポイント向上し、全体の正解率が 1.2 ポイント向上した。

本手法は、外来語名詞の 172 文節に対して接辞処理を

誤って処理した。外来語特定規則で特定できなかった外来語に母音を誤って復元した失敗が 58 件、接辞を誤って分割した失敗が 114 件あった。外来語動詞の 57 文節は、動詞が対象の動詞辞書に登録されていなかったため失敗した。また、外来語形容詞の 1 文節の末尾を接辞として誤って分割した。

外来語以外では、名詞の 441 文節を誤って分割した。動詞の 2208 文節が動詞辞書に登録されていなかったため、誤って分割した。動詞 158 件と形容詞 33 件が誤って分割された。

表 1　接辞処理の正解率（%）

品詞	文節数	先行研究	本研究	
			なし	あり
外来語名詞	3510	49.4	90.8	95.1
外来語動詞	58	1.7	1.7	1.7
外来語形容詞	6	50.0	83.3	83.3
名詞	8612	71.9	94.2	94.9
動詞	4048	41.6	41.6	41.6
形容詞	613	63.3	93.5	94.6
全体	16847	60.8	80.4	81.7

5.2　接辞処理の情報検索における有効性

接辞処理が情報検索において有効であるかどうかを評価した。実験に用いた研究抄録の例を図 4 に示す。

タイトル： **Хангай-Хэнтийн атираат тогтолцооны тектоник, магматизм, алтны хүдэржилт**
指導者の名字：**Гомбосүрэн**
名前:**Бадарч**
終了した年：2003
キーワード: **террейн, геодинамик, бүслүүр, гранитоид…**
要約： **Хангай-Хэнтийн атираат бүслүүрийн тектоникийн тогтоцыг террейний үзэл баримтлалаар…**
結果： **Хангай-Хэнтийн атираат бүслүүрийн талаарх шинэлэг материалыг нэгтгэн дүгнэж, түүний үүсэл…**

図 4　モンゴル語抄録の例

抄録にはそれぞれキーワードが付いている。図 4 の例では、「**террейн**（地域）, **геодинамик**（ジオダイナミック）, **бүслүүр**（地帯）, **гранитоид**（御影石）」などのキーワードが付いている。

抄録に付いているキーワードを検索質問として利用した。検索質問を「一つのキーワードを一つの検索質問とした場合」と「一つの抄録についているキーワード集合全てを一つの検索質問とした場合」の 2 通り作成した。一つの質問あたりのキーワードは平均で 6.1 ある。検索質問に対する適合文書は、そのキーワードが付いていた抄録である。検索モデルには Okapi BM25 を使用した。

抄録を検索するときに抄録から索引語を抽出して、索引付けを行った。索引語を抽出する際、以下の手法を利用して、MAP（Mean Average Precision）を比較した。

(a) 接辞処理なし
(b) Sanduijav ら[1]の手法
(c) 本研究の手法（外来語特定規則を利用しない）
(d) 本研究の手法（外来語特定規則を利用する）
(e) 正しい接辞処理結果（人手判定）

実験結果を表 2 と表 3 に示す。全ての手法で MAP が 0 になる質問が多かったため、それらは評価対象から事前に削除した。対象となった質問の件数は「一つのキーワードを一つの検索質問とした場合」では 686 であり、「一つの抄録についているキーワード集合全てを一つの検索質問とした場合」では 273 であった。

表 2 と表 3 より以下の点について考察する。まず、接辞処理を行うことが検索において有効であるかを確認するために(a)と(e)の結果を比較する。検索質問の種別によらずに接辞処理をした場合の MAP が高かった。

次に、(b)、(c)、(d)の結果を比較する。本研究の手法は Sanduijav ら[1]の手法より MAP が高かった。

外来語特定規則が検索精度に及ぼす影響を評価するために、(c)と(d)の結果を比較する。外来語特定規則の適用によって、一つのキーワードを一つの検索質問とした場合に MAP が若干低下し、一つの抄録についているキーワード集合全てを一つの検索質問とした場合の MAP が向上した。

表 2　「一つのキーワードを一つの検索質問とした場合」の実験結果

手法	MAP
(a)	0.2312
(b)	0.2882
(c)	0.3058
(d)	0.3052
(e)	0.3268

表 3　「一つの抄録についているキーワード集合全てを一つの検索質問とした場合」の実験結果

手法	MAP
(a)	0.2766
(b)	0.2834
(c)	0.3039
(d)	0.3052
(e)	0.3187

6. おわりに

本研究では、モンゴル語の名詞、動詞、形容詞の接辞処理手法を提案した。名詞と形容詞の接辞処理は辞書を利用しない手法を提案した。また、外来語特定規則を利用することで接辞処理の正解率が向上した。接辞処理をモンゴル語の情報検索に応用した結果、検索精度が向上した。

参考文献

[1] Sanduijav Enkhbayar, 宇津呂武仁, 佐藤理史. 音韻論的・形態論的制約を用いたモンゴル語句生成・形態素解析. 自然言語処理, Vol.12, No.5, pp. 185－205. 2005.
[2] 江原輝将, 早田清冷, 木村展幸. 茶筌を用いたモンゴル語の形態素解析. 言語処理学会第 10 回年大会発表論文集, pp. 709－712, 2004.
[3] Badam-Osor Khaltar, Atsushi Fujii, and Tetsuya Ishikawa. Extracting loanwords from Mongolian corpora and producing a Japanese-Mongolian bilingual dictionary. Proc. of COLING/ACL. pp.65－664, 2006.
[4] Ц.Баярмаа. Монгол хэл I－IV анги. 2002. (和訳：1 年生か 4 年生のモンゴル語文法)