言語に依存しない形態素解析ツールキットの開発

山下 達雄，松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{tatuo-y,matsu}@is.aist-nara.ac.jp

形態素解析処理において、日本語などのわかち書きされない言語と英語などのわかち書きされる言語では、形態素辞書検索のタイミングや辞書検索単位が異なる。これらを同じ枠組で扱うことにより、辞書検索部の多言語化を行った。また、これに関連し、形態素解析処理のモジュール分割を行い、多言語形態素解析ツールキットとして実装した。実験として日本語、英語、中国語、韓国語での実装を行った。

[キーワード] 形態素解析，多言語処理，辞書検索

Language Independent Morphological Analysis Tool Kit

Yamasita Tatuo, Matsumoto Yuji
Graduate School of Information Science, Nara Institute of Science and Technology

To develop a multi-lingual morphological analyzer, we need to recognize crucial differences between segmented languages, like English, and non-segmented languages, like Japanese. One difference is the way the dictionary is looked up. We developed a framework of dictionary look-up to achieve a uniform treatment of both types of languages. Processing undefined words and inflection is another difference. To minimize these language dependencies, we divided whole system into some modules. We implemented a multi-lingual morphological analyzer, in which morphological analysis of Japanese, English and Chinese are experimentally implemented.

[keyword] multi-lingual processing, morphological analysis, part of speech tagging

1 はじめに

インターネット上で様々な言語のテキストが行き交う現代において、特定の言語に依存しない、多種多様な言語を視野に入れた自然言語処理が必要とされている。その中でも形態素解析処理は、高度な自然言語処理の前処理であるだけでなく、情報検索の分野でも幅広く利用され、ある程度実用化された技術であると言える。形態素解析処理の多言語化はクロスリンガル情報検索等に非常に有用であると考えられる。

本研究では、様々な言語で共通に利用できる形態素解析の枠組の提案と、それに関連して形態素解析システムのモジュール分割を行った。

形態素辞書や形態素 (品詞) 間接続表といった言語依存なデータのみを取り換えることで、特定の言語に依存しない形態素解析処理の枠組の構築を目指す。利点として、

- 新たな言語の実装が容易
- 複数言語からなる文書の形態素解析が可能
- 自然言語以外のデータの自動タグ付けも可能

といった点が挙げられる。

形態素解析処理の言語依存部分として、「辞書検索方針」「活用処理」「未定義語処理」などが挙げられる。辞書検索方針とは、解析対象文をどういう単位で解釈するかという方針である。各言語における検索のタイミングや辞書エントリを構成する最小単位の定義などが関係する。これについては形態素片という概念を導入することによって、言語に依存しない統一的な処理を可能にした (第 2章)。活用処理、未定義語処理は、言語依存性が高く、処理方法も様々なバリエーションがあるため一般化するのは難しい[1]。

第 3章では、多言語化作業に関連した、形態素解析システムのモジュール分割について述べる。

第 4章では、本研究に基づいて作成した形態素解析システム MOZ での、日本語、英語、中国語、韓国語の実装について述べる。

2 形態素片の導入による辞書検索部の統一

わかち書きされる言語とそうでない言語[2] の形態素解析処理における差異の一つに辞書検索のタイミングが挙げられる。これは、解析対象文のどの部分が形態素[3]となりうるかという問題である。この章では、辞書検索アルゴリズムや辞書検索単位を統一し、この問題に対処する枠組を提案する。

2.1節では、それぞれの言語で一般的に用いられる辞書検索アルゴリズムである exact match と common prefix search について簡単に解説する。2.2節では、両タイプの言語を common prefix search で統一的に扱うために「形態素片」という概念を導入する。2.3節では、「形態素片」による common prefix search について具体例を用いて解説する。

2.1 辞書検索のアルゴリズム

形態素辞書検索の実装の際には、わかち書きされる言語では単純な exact match、わかち書きされない言語では common prefix search が一般的に用いられる。

例えば英語のようなわかち書きのされている言語では、“This is my pen.” のように単語がブランクや記号文字で明確に区切られているので、“This”, “my” といった文字列ごとに、その文字列が辞書に存在するか否かという基準で形態素辞書検索を行う (exact match) のが単純で効率的な実装である。

これに対し、日本語などのわかち書きされていない言語では、区切り認識も同時に行う必要がある。そのため、解析対象文の先頭から一文字ずつずらしながら、その文字位置から始まる文字列と一致する全ての単語を辞書から検索 (common prefix search) するのが現実的である。

common prefix search について簡単に解説する。例えば、 辞書に “a”, “abc”, “ac”, “abcb”, “ba” というエントリがあるとする。“abcbac” という文字列に対し common prefix search を行うと、結果は “a”, “abc”, “abcb” となる。common prefix search のためのデータ構造として一般的にトライが用いられる。トライは全てのエントリの共通接頭辞 (common prefix) を併合して作られる木構造である。図 1にトライの例を示す。図 1のトライで、「有効っぽい」という文字列に対して common prefix search を行うと、「有 (名詞)」「有効 (名詞)」「有効っぽい (形容詞)」が結果となる。詳細は文献 [1] を参照されたい。

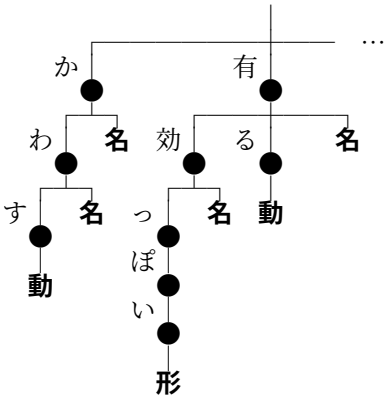

図 1: 日本語形態素辞書のトライ構造

明らかに common prefix search は exact macth を内包しているので、本研究では、両タイプの言語の形態素辞書検索を common prefix search に統一することを考える。

[1]第 3章で詳しく述べる。

[2]わかち書きされるか否かは、正書法という社会習慣的な問題であって、言語の本質的な問題ではないが、コンピュータ処理の際には重要な問題である [1]。

[3]本研究では、形態素とは「システムが扱う辞書にエントリされている文字列」と定義する。もちろん厳密な言語学的な用法とは異なっている。辞書エントリには、品詞を表すタグなどの情報も付与されている。

2.2 形態素片の認識

わかち書きされる言語の扱いをわかち書きされない言語と同じ枠組 (common prefix seach) で扱うことにより辞書検索処理の統一を行う。そのためには辞書検索単位を統一する必要がある。例えば、英語の単語と日本語の文字を同じ単位として扱えば、辞書検索処理を統一できる。この単位を「形態素片」と呼ぶことにする。

「形態素片」を以下のように定義する。

- **形態素片とは形態素として認識される可能性のある最小単位の文字列である**。わかち書きされていない言語では、その言語体系での文字一文字であり、わかち書きされている言語では、ブランク等で区切られた文字列である。

  例: You are wrong. → [You] [are] [wrong][.]

  例： ございます。 → [ご][ざ][い][ま][す][。]

  [ ] は形態素片を表す。

- **形態素片が一つ以上連続して形態素を構成する**。

  例： [ご] + [ざ] → ござ/名詞

- **形態素片を構成する文字列の途中から辞書検索を始めること、途中で辞書検索を終わることはできない**。つまり、形態素片が認識できれば、解析対象文の中の、辞書を引き始めることが可能な位置と形態素が終わることが可能な位置が分かる。

  例えば、形態素片 [wrong] が認識されたとしたら、たとえ辞書にあったとしても ong や wr は検索結果として認められない。

ここで、英語を例にわかち書きされる言語から形態素片を認識する方法を考える。まず単純に、ブランクで区切られた単位を形態素片とすることが考えられる。

例： You are wrong. → [You] [are] [wrong.]

しかし、wrong. を [wrong] [.]、John's を [John] ['s] と認識したいといった例のように、ブランク以外にある種の記号文字列も形態素片認識には必要となる。

ブランクや記号といったものは、言語 (文字体系) に依存するものであるので、形態素片の認識には、解析対象言語ごとに以下の情報が必要になることが分かる。

その言語でのブランク文字 形態素片の区切りを示すが、この文字自体は辞書検索の際、無視される文字。形態素の先頭と末尾には絶対現れない。

その言語での記号文字 形態素片の区切りを示す文字。この文字自体も形態素片になる。

その言語での文字と形態素片の関係 その言語で一文字を一形態素片とするか否かの情報。わかち書きされていない言語か否かの情報でもある。

これらの情報を用いた解析対象文からの形態素片認識の例を示す。

例 1 英語 文字コードは ASCII、ブランク文字は 0x20(スペース) と 0x09(タブ)、記号文字は [.][,][']['][:][!][?][-] で、「わかち書きされる言語」として扱うとする。解析対象文「`I'm in New York.`」の形態素片認識を行うと、以下のようになる。

[I][']['][m]␣[in]␣[New]␣[York][.]

ブランクは ␣ で表す[4]。

例 2 日本語 文字コードは日本語 EUC、ブランク文字、記号文字はなし、「わかち書きされない言語」として扱うとする。解析対象文「今日もしないとね。」の形態素片認識を行うと、以下のようになる。

[今][日][も][し][な][い][と][ね][。]

2.3 形態素片による common prefix search

前節の例 2 のように、わかち書きされない言語では全ての文字が形態素片であるとみなされるので、形態素片による common prefix search は 2.1節で解説したものと同じである。わかち書きされる言語に関して、前節の例 1 の「`I'm in New York.`」の形態素片認識結果 [I][']['][m]␣[in]␣[New]␣[York][.] を用いて説明する。

英語等のわかち書きされる言語では、形態素辞書は図 2のような、形態素片ごとにノードが構成されるトライになる。

[4] ブランクは存在したという情報だけが重要であり、複数連続で出現しても "␣" 一文字に置き換える。ブランクは検索の際、形態素片に挟まれた場合にのみ考慮される。2.3節の例を参照。

common prefix search を、二つ目の形態素片が始まる位置、つまり ['][m]_[in]... という形態素片列で行うと、「括弧類 '」「be 動詞 'm 」が検索結果となる。五つ目の形態素片が始まる位置 ([New]_[York][.]) だと、検索結果は「形容詞 New」「固有名詞 New_York 」となる。

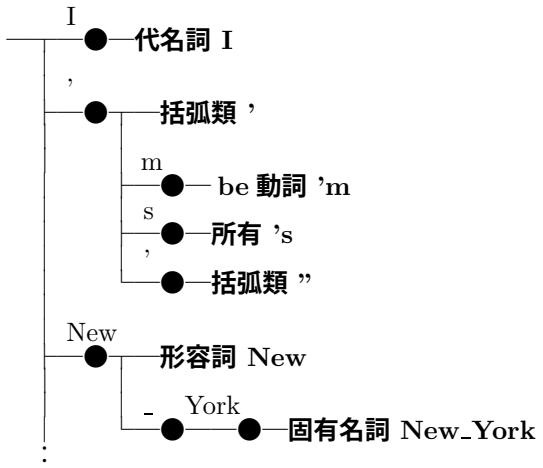

図 2: 英語形態素辞書のトライ構造

3 形態素解析処理のモジュール分割

コスト最小法による形態素解析システムである茶筌 [2] を参考にして、形態素解析における様々な処理のモジュール化を行なう。モジュール化の利点として、保守の容易性、言語依存部分の縮小、他のプログラムへの再利用といった点が挙げられる。

まず、形態素解析システム内の処理の中から言語非依存な処理のみを選び、それらをモジュール化し、形態素解析ツールキット "LimaTK"[5] としてプログラム言語 C++ で実装した (図 3)。辞書検索部には高速文字列検索ライブラリ SUFARY [3] を使用した。

各モジュールについて解説する。

形態素片認識

各言語ごとに定義された形態素片 (辞書検索最小単位) の認識を行う。与えられた文字列 (解析対象文) に対する形態素片認識結果を返す。

[5] **L**anguage **I**ndependent **M**orphological **A**nalysis **T**ool **K**it

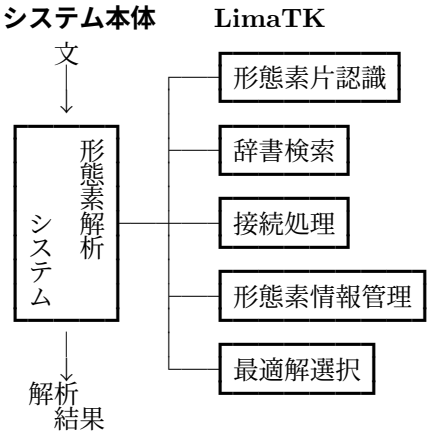

図 3: 形態素解析処理

辞書検索

形態素片を検索単位とした common prefix search によって、形態素辞書検索を行う。与えられた文字列に対する、辞書に存在する全ての common prefix を返す。

接続処理

接続処理は基本的に状態遷移モデルに基づいている。コード化された現状態と次の単語 (品詞) の二つ組に対し、その接続コストと遷移先状態を返す。品詞 bigram モデルとして扱うならば、現在の単語 (品詞) と次の単語 (品詞) の二つ組に対しての接続コストを返すと解釈すれば良い (遷移先状態は不要)。このような単純な枠組なので、文脈によって前件の長さ (N-gram の N) を変化できる可変長 gram モデル [4] の実装も容易である。

形態素情報管理

辞書検索の結果の個々の形態素の情報を管理する。

最適解選択

ラティス構造で管理されている形態素の中から、解として最適なものを計算する。

未定義語をどう扱うかといった方針は言語や用途に依存する。さらに様々なバリエーションがあり統一的に扱うのは難しい。現時点では、未定義語処理は図 3でいうところのシステム本体で行うべき処理と考えている。

活用処理に関しても、統一的な処理は難しい。第4章での日本語形態素解析の実装の際には、活用処理をプログラム側では行わず、辞書エントリに全ての活用形を展開することで他言語と同じ枠組で扱うようにしている。また、語幹と活用語尾を別エントリとする方法でも同じ枠組で扱える[6]。もちろん、日本語の活用語尾処理を辞書検索モジュールの内部に組み込めば、言語依存性は他の処理に波及しないように実装できる。しかし、各形態素の各活用形ごとの出現確率などを考慮して解析することを考えると、特にシステム内で活用処理を行う利点を見出せない。以上の理由から、今回の枠組では活用処理は考慮されていない。

4 様々な言語の実装

LimaTK を用いて、簡単な形態素解析システム "MOZ"[7]を作成し、日本語、英語、中国語、韓国語など様々な言語を実装した。図 4に解析結果の例を示す。

日本語 RWCP の品詞タグ付きコーパス [6] から品詞 bigram モデルで学習を行い、さらに茶筌 [2] の形態素辞書エントリを追加した。

英語 Penn Treebank[7] の品詞タグ付きコーパスから品詞 bigram モデルで学習を行い、Oxford Advanced Learner's Dictionary の電子化テキスト版 [8] の辞書エントリも追加した。語幹 (stem) 情報も同じく Oxford Advanced Learner's Dictionary から補完した。

中国語 台湾の中央研究院の品詞タグ付きコーパス [9] から品詞 bigram モデルで学習を行った。

韓国語 平野 [10] による韓国語形態素解析のデータを変換して利用した。

5 おわりに

本研究に関する情報は以下の URL にて公開している。プログラムや解析用データも入手できる。

`http://cl.aist-nara.ac.jp/~tatuo-y/ma/`

[6]日本語の活用語尾の扱いについては久光らの研究 [5] を参照されたい。

[7]**Mo**rphological Analy**z**er

謝辞

英語形態素解析の実装にあたっては奈良先端科学技術大学院大学自然言語処理学講座の玉野健一氏、中国語形態素解析の実装にあたっては同講座の Md Maruf Hasan 氏、韓国語形態素解析の実装にあたっては同講座 OB の平野善隆氏に御協力頂いた。ここに感謝の意を表する。

参考文献

[1] 永田昌明. "形態素解析", 岩波講座 言語の科学 3 "単語と辞書", 岩波書店, pp.53-92, 1997.

[2] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明. "日本語形態素解析システム『茶筌』version1.0 使用説明書", NAIST Technical Report, NAIST-IS-TR97007, February 1997.

[3] 奈良先端科学技術大学院大学 自然言語処理学講座. "高速文字列検索ライブラリ SUFARY Version 2.0", June 1998.
http://cl.aist-nara.ac.jp/lab/nlt/ss/

[4] 北内啓, 山下達雄, 松本裕治. "日本語形態素解析システムへの可変長連接規則の実装", 言語処理学会第 3 回年次大会, pp.437-440, March 1997.

[5] 久光徹, 新田義彦. "日本語形態素解析における効率的な動詞活用処理", 情報処理学会研究会報告, 94-NL-103, pp.1-7, September 1994.

[6] "RWC テキストデータベース報告書", 新情報処理開発機構, 1996.

[7] Beatrice Santorini. "Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd printing)", June 1990.

[8] Roger Mitton. "A Description of A Computer-Usable Dictionary File Based on The Oxford Advanced Learner's Dictionary of Current English", June 1992.
ftp://ota.ox.ac.uk/pub/ota/public/dicts/710/

[9] "中央研究院平衡語料庫的内容興説明", Technical Report no.95-02, 中文詞知識庫小組, 1995.

[10] 平野善隆. "用言の活用を考慮した韓国語品詞体系の提案とそれを用いた韓国語形態素解析", 奈良先端科学技術大学院大学修士論文, NAIST-IS-MT9551092, March 1997.

日本語　フォーマット: 形態素文字列 形態素コスト 品詞コード 1 品詞コード 2 [その他の情報]

なんだかやってられないって感じ。

```
なんだか 1051 149 149 [Y:ナンダカ POS:副詞-一般 Pr:6/3097]
やっ 783 115 115 [Y:ヤッ STEM:やる POS:動詞-自立/五段・ラ行/連用タ接続 Pr:91/9561]
て 66 65 65 [Y:テ POS:助詞-接続助詞 Pr:14172/21074]
られ 118 126 126 [Y:ラレ STEM:られる POS:動詞-接尾/一段/未然形 Pr:168/340]
ない 399 74 74 [Y:ナイ POS:助動詞/特殊・ナイ/基本形 Pr:2843/30431]
って 778 62 62 [Y:ッテ POS:助詞-格助詞-連語 Pr:50/5092]
感じ 1392 152 152 [Y:カンジ POS:名詞-一般 Pr:37/144546]
。 0 15 15 [Y:。 POS:記号-句点 Pr:27418/27452]
EOS
```

英語　フォーマット: 文中での出現位置 形態素文字列 形態素コスト 品詞コード 1 品詞コード 2 [その他の情報]

```
An essay's body paragraphs may be arranged in many ways that are familiar to readers.

0 An 5275 43 43 [DT]
3 essay 9091 20 20 [NN]
8 's 62 25 25 [POS]
11 body 7061 20 20 [NN]
16 paragraphs 9429 27 27 [NNS STEM:paragraph]
27 may 2080 32 32 [MD]
31 be 1621 36 36 [VB]
34 arranged 6533 42 42 [VBN STEM:arrange]
43 in 1554 28 28 [IN]
46 many 3677 19 19 [JJ]
51 ways 5442 27 27 [NNS STEM:way]
56 that 634 40 40 [WDT]
61 are 1018 34 34 [VBP STEM:be]
65 familiar 5516 19 19 [JJ]
74 to 6 65 65 [TO]
77 readers 6380 27 27 [NNS STEM:reader]
84 . 9 16 16 [.]
EOS
```

中国語　フォーマット: 形態素文字列 形態素コスト 品詞コード 1 品詞コード 2 [その他の情報]

我真正開始聽我的問題。

```
我 1144 85 85 [Nh Pr:12297/53410]
真正 5013 78 78 [D Pr:272/169553]
開始 2066 135 135 [VL Pr:730/10355]
聽 3240 82 82 [VE Pr:620/39727]
我 1144 85 85 [Nh Pr:12297/53410]
的 75 27 27 [DE Pr:100462/110636]
問題 4033 28 28 [Na Pr:2118/375262]
。 87 31 31 [PERIODCATEGORY Pr:71507/79982]
EOS
```

我真正开始听我的问题。

```
我 1144 85 85 [Nh Pr:12297/53410]
真正 5013 78 78 [D Pr:272/169553]
开始 2066 135 135 [VL Pr:730/10355]
听 3240 82 82 [VE Pr:620/39727]
我 1144 85 85 [Nh Pr:12297/53410]
的 75 27 27 [DE Pr:100462/110636]
问题 4033 28 28 [Na Pr:2118/375262]
。 87 31 31 [PERIODCATEGORY Pr:71507/79982]
EOS
```

図 4: 形態素解析結果