基般

3.1 形態素解析

森 信介（京都大学学術情報メディアセンター）　鍛治 伸裕（ヤフー株式会社）
村脇 有吾（九州大学大学院システム情報科学研究院）
斉藤 いつみ（NTT メディアインテリジェンス研究所）

タスクの定義

日本語は，英語等と異なり，単語を空白等で区切って表記する習慣を持たない．そのため，日本語テキストに対して情報抽出，評判分析，テキスト検索，機械翻訳などのアプリケーションを実現するためには，まずテキストを形態素に分割し，各形態素に品詞を割り当てる処理を行うということが必要不可欠となる（**表 -1**）．こうした一連の処理は，形態素解析と呼ばれ，日本語テキスト処理の重要な研究分野となっている．

主な手法

形態素解析の手法は，**図 -1** のように，形態素単位の手法[1),2)]と文字単位の手法[3)]に大別できる．形態素単位の手法は，入力文のすべての部分文字列を辞書引きし，ノードを作成し，文頭から文末までを過不足なく被覆し，ある評価関数値が最大となる形態素列を出力する．辞書引きの結果，各形態素の品詞も決定される．評価関数値としては，人手によるコスト（JUMAN☆1），隠れマルコフモデルや条件付き確率場による確率（MeCab☆2）が用いられる．機械学習による方法では，パラメータを学習コーパスを用いて決定しておく．

文字単位の手法は，各文字間に対して単語境界か否かを示すタグ（Y または N）を分類器を用いて推定して単語分割を行い，その結果得られる各単語に対して品詞を別の分類器を用いて推定する（KyTea☆3）．

形態素	品詞
形態	普通名詞
素	普通名詞
解析	普通名詞
器	普通名詞
を	格助詞
使って	動詞
、	読点

形態素	品詞
日本	地名
語	普通名詞
処理	サ変名詞
を	格助詞
行い	動詞
ます	動詞性接尾辞
。	句点

表 -1　JUMAN☆1 による形態素解析結果の例

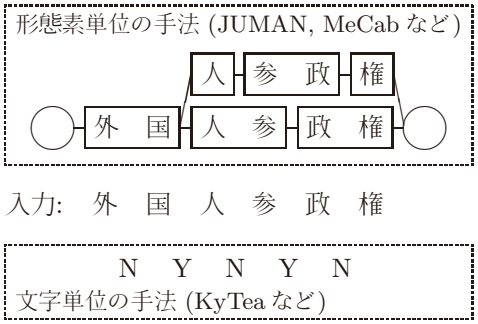

図 -1　形態素解析の手法（品詞は省略）

エラーの分類と取り組み

形態素解析の誤りにはある程度傾向がある．以下では，誤りの分類とその対策を紹介する．

未知語問題と語彙資源

形態素解析器は，知っている（あらかじめ辞書に記述された）言葉であれば高精度に解析できるが，未知語の解析を誤りやすい．たとえば，未知語を含むラジオ番組名「でじこさん」は「で / じこ / さん」といった風に誤分割され，この結果を用いて日英翻訳を行うと“In the accident Mr.”（Google 翻訳）が出力される．この問題の解決策の 1 つは辞書を拡張して未知語を既知語に変えることであり，計算機

☆1 http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
☆2 http://taku910.github.io/mecab/
☆3 http://www.phontron.com/kytea/index-ja.html

自身による辞書の（半）自動拡張が研究されてきた．

応用上重要と思われる未知語は商品名やサービス名等だが，それらは膨大な数に上り，解析の専門家による辞書整備は現実的ではない．既存の大規模語彙知識源として，Wikipedia やはてなキーワード等があるが，それらの外部資源は，分割基準の不一致から，単語辞書として容易には利用できない．2015 年には，分割基準を無視して外部資源を取り込んだ，実用優先の辞書 NEologd☆4 が登場し話題となった．

Twitter の派生的未知語と対策

今回のエラー分析では，Twitter データにおいて解析誤りの原因となった未知語の分布調査を行った．Twitter データは，ランダムにサンプリングし前処理とアノテーションを行った 2,976 文を用いた．Twitter の未知語を分類した結果，新語・低頻度語（23.0%），表記揺れ（21.9%），固有名詞（20.3%），顔文字・アスキーアート（12.8%），長音記号・小書き文字・母音字・促音文字の挿入（11.7%）の順に出現頻度が高かった．

表記揺れや長音記号・小書き文字・母音字・促音文字の挿入等に関しては，既知の辞書語を静的・動的に展開し解析する手法が提案されている．たとえば，既知の辞書語の表記揺れであるひらがな表記（テスト→てすと）は辞書の読み情報を用いてあらかじめ辞書に展開できる．また，長音の挿入（おいしい→おいしーい）などは，動的に長音記号「ー」を削除するルールを適用しながら辞書引きを行うことで，既知の辞書語を動的に拡張することができる[4), 5)]．

☆4 https://github.com/neologd

展望

あらゆるタイプの未知語に対応するために，人の新語生成の過程をモデル化し，新語を自動獲得することが研究課題であろう．加えて，人が単語と判定した確実な情報を蓄積していくことも重要である．各所での判断結果を文脈とともに共有する枠組みが，日本語の処理にとって重要である．

参考文献

1) 永田昌明：統計的言語モデルと N-best 探索を用いた日本語形態素解析法，情報処理学会論文誌，Vol.40, No.9, pp.3420-3431 (Sep. 1999).
2) 工藤　拓，山本　薫，松本裕治：Conditional Random Fields を用いた日本語形態素解析，情報処理学会研究報告，Vol. NL161 (2004).
3) 森　信介，中田陽介，Graham, N.，河原達也：点予測による形態素解析，自然言語処理，Vol.18, No.4, pp.367-381 (2011).
4) Sasano, R., Kurohashi, S. and Okumura, M. : A Simple Approach to Unknown Word Processing in Japanese Morphological Analysis, Proc. of IJCNLP2013, pp.162-170 (2013).
5) Kaji, N. and Kitsuregawa, M. : Accurate Word Segmentation and POS Tagging for Japanese Microblogs : Corpus Annotation and Joint Modeling with Lexical Normalization, Proc. of EMNLP2014, pp.99-109.

（2015 年 10 月 29 日受付）

森 信介（正会員）forest@i.kyoto-u.ac.jp
1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了．博士（工学）．同年日本アイ・ビー・エム（株）入社．2007 年より京都大学学術情報メディアセンター准教授，現在に至る．

鍜治 伸裕（正会員）nkaji@yahoo-corp.jp
2005 年東京大学大学院情報理工学系研究科博士課程修了．情報理工学博士．東京大学生産技術研究所特任准教授，情報通信研究機構主任研究員などを経て，2015 年よりヤフー株式会社 Yahoo! JAPAN 研究所上席研究員．

村脇 有吾（正会員）murawaki@ait.kyushu-u.ac.jp
2011 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．同年同大学術情報メディアセンター特定助教．2013 年九州大学大学院システム情報科学研究院助教，現在に至る．

斉藤 いつみ（正会員）saito.itsumi@lab.ntt.co.jp
2012 年東京大学大学院工学系研究科都市工学専攻修士課程修了．修士（工学）．同年 NTT 入社，現在に至る．