■2 群（画像・音・言語）- 10 編（自然言語処理）

2 章 形態素解析

【本章の構成】

本章では，形態素解析アルゴリズム（2–1 節），形態素解析の適応化（2–2 節）について述べる．

■2 群 - 10 編 - 2 章

2-1 形態素解析アルゴリズム

■2 群 - 10 編 - 2 章

2-2 形態素解析の適応化

（執筆者：浅原正幸）[2009 年 6 月 受領]

2-2-1 わかち書き体系と品詞体系

形態素解析器の多くは文内の候補語をトレリス状に展開し候補語の組み合わせの曖昧性を解消することにより構成されている．形態素解析器 JUMAN はこの曖昧性解消に人手による規則に基づく手法を用いている．この人手による規則は品詞タグ付きコーパスを作成しながら規則間の優先度を調整している．形態素解析器 ChaSen ではコーパスからの頻度を用いて語の生起確率を推定する手法を導入している．識別モデルの研究が進み，形態素解析器 MeCab では条件付き確率場というモデルを用い，実用的な形態素解析器を構成するに至った．これらの形態素解析器が出力するわかち書き体系や品詞体系は学習元となっているコーパスに深く依存する．新聞記事に基づく代表的なコーパスとして EDR コーパス，RWCP テキストコーパス，京都大学テキストコーパスの 3 つがある．EDR コーパスは，EDR 独自の品詞体系に基づきタグ付けされ，EDR 電子化辞書作成の典拠となっており新聞記事および雑誌記事から集められている．RWCP テキストコーパスは IPA 品詞体系と呼ばれる品詞体系に基づきタグ付けされている．京都大学テキストコーパスは益岡・田窪文法と呼ばれる品詞体系に基づきタグ付けされている．いずれの品詞体系も小中学校で習う学校文法とは違った品詞体系となっている．これらのテキストコーパスは主に書き言葉を対象にしており，これにより構成される形態素解析器が話し言葉や他分野の書き言葉を解析するには限界があった．2000 年代に入り，国立国語研究所を中心に，日本語話し言葉コーパス（CSJ コーパス）が構築され，新聞記事に表れない，話し言葉特有の活用や縮約形態に対するタグ付けが進んだ．この研究により学校文法に可換な品詞体系定義や単位の斉一化が進み，UniDic がリリースされた．2006 年以降，国立国語研究所では，大規模な書き言葉のバランスドコーパス「日本語コーパス」を構築し，UniDic の書き言葉の辞書項目の追加が進んでいる．UniDic では様々な用途に利用できるよう短単位・長単位の複数の単位を設計している．

2-2-2 分野間の齟齬と適応化

古くは新聞記事に基づいたコーパスしかなかったために，他の分野に対する形態素解析器の適応化についてはあまり研究されていなかったが，数少ない研究として，松本ら[12] のコーパスの混ぜ合わせがある．書き言葉と話し言葉間の齟齬を吸収する単純な手法として，大規模な新聞記事コーパスに少量の話し言葉コーパスを混ぜることにより，話し言葉に対する形態素解析器の性能の改善を報告している．その後，大規模な CSJ コーパスが整備され[6]などの話し言葉に対する形態素解析器が整備される．現状では言語資源や解析器の整備されている分野に偏りが見られるが，先に述べたバランスドコーパス「日本語コーパス」プロジェクトでは，大規模な生コーパスと小規模の形態素タグ付きデータの整備が進んでいる．これらを用いた半教師あり学習手法の発展が期待される．

半教師あり学習を用いた適応化の一手法として，坪井らの手法[11]がある．適応先の分野特有の単語のみについて語境界をタグ付けされたデータから条件付き確率場を用いた単語わかち書きモデルを学習する手法を提案している．この手法は単語わかち書きのみについての手

法であるが，形態素解析器のような品詞などの情報を含めると文字単位に付与する可能なラベル数が増えるために，学習の計算量や収束性などの点で検証が待たれる．機械学習の分野で他のタスクにおいて提案されている半教師あり学習手法[1, 4]などが日本語形態素解析器において利用可能であると考える．

2-2-3 未知語処理

日本語形態素解析手法の多くはトライなどの形式で構造化された辞書を用い，文中の可能な候補形態素すべてを辞書引きして枚挙し，その候補の中で最尤な形態素列を抽出することによって行われてきた．形態素解析における未知語の問題とは，辞書にない語をどのようにして同定するかである．中国語単語わかち書きでは，文字単位に形態素境界を付与する手法[7]が提案されている．ペン（Peng）らは，文字単位の条件付き確率場を適用する際に得られる周辺確率を用いて各未知語候補の尤度を推定する手法[3]を提案している．この手法を発展させたものとして，岡野原らの形態素の周辺確率を用いた確率的単語分割[9]がある．この研究では形態素の境界の曖昧性を保ったままコンパクトに境界情報を保持する方法を提案している．これらの手法では形態素境界は同定されても，品詞の情報までは同定されない．内元らは，文中の可能な 5 文字以下の部分文字列を辞書引きされた候補形態素とともにラティス状に枚挙し，最大エントロピー法を用いて最尤な品詞の情報を含む形態素境界を同定する手法[5]を提案している．この問題の定式化に対して，より性能の良い学習器である条件付き確率場を適用する手法[8] が提案されている．他の問題の定式化として未知語部分のみ文字単位で解析する手法[2]，構文解析などで見られる Shift-Reduce 操作による手法[10]が提案されている．

■参考文献

1) R. K. Ando and T. Zhang：“A High-Performance Semi-Supervised Learning Method for Text Chunking,” Proc. of ACL-2005, pp.1-9, 2005.
2) T. Nakagawa：“Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information,” Proc. of COLING-2004, pp.466-472, 2004.
3) F. Peng, F. Feng, and A. McCallum：“Chinese Segmentation and New Word Detection using Conditional Ramdom Fields,” Proc. of COLING-2004, pp.562-568, 2004.
4) J. Suzuki, A. Fujino, and H. Isozaki：“Semi-Supervised Structured Output Learning based on a Hybrid Generative and Discriminative Approach,” Proc. of EMNLP-CoNLL-2007, pp.791-800, 2007.
5) K. Uchimoto, S. Sekine, and H. Isahara：“The Unknown Word Problem: a Morphological Analysis of Japanese Using Maximum Entropy Aided by a Dictionary,” Proc. of EMNLP-2001, pp.91-99, 2001.
6) K. Uchimoto, C. Nobata, A. Yamada, S. Sekine, and H. Isahara：“Morphological Analysis of a Large Spontaneous Speech Corpus in Japanese,” Proc. of ACL-2003, pp.479-488, 2003.
7) N. Xue and S. P. Converse：“Combining classifiers for Chinese word segmentation,” Proc. SIGHAN-2002, 2002.
8) 東　藍, 浅原正幸, 松本裕治：“条件付確率場による日本語未知語処理,” 情処学研報, 2006-NL-173, pp.67-74, 2006.
9) 岡野原大輔, 工藤　拓, 森　信介：“形態素周辺確率を用いた確率的単語分割コーパスの構築とその応用,” NLP 若手の会第 1 回シンポジウム, 2006.
10) 岡野原大輔, 辻井潤一: “Shift-Reduce 操作に基づく未知語を考慮した形態素解析”, 言語処理学会第 14 回年次大会発表論文集, pp.400-403 (2008)

11) 坪井祐太, 鹿島久詞, 森　信介, 小田裕樹, 松本裕治：“部分的かつ曖昧なラベル付き構造データからのマルコフ条件付き確率場の学習,” 情処学研報, 2007-NL-182, pp.67-74, 2007.
12) 松本裕治, 伝　康晴：“話し言葉の形態素解析,” 情処学研報, 2001-NL-143, pp.49-54, 2001.