文節まとめあげと形態素解析の融合

浅原 正幸　　松本 裕治

奈良先端科学技術大学院大学 情報科学研究科
{masayu-a,matsu}@is.aist-nara.ac.jp

Abstract

文節まとめあげは一般に形態素解析と構文解析の間の処理として行われる。既存の文節まとめあげ法は、形態素解析済みデータを前提としているものが多い。本稿では、形態素解析時に文節まとめあげを同時に行う方法を提案する。一般の形態素解析の品詞情報に、その形態素の直前の文節区切り情報を付与することにより実現した。日本語と英語について実際に解析を行い、その精度を検証した。形態素解析と文節まとめあげを逐次的に行う方法よりも、両者を同時進行的に行う方が、高い精度を達成することがわかった。また、べた書きかな文の文節まとめあげはかな漢字変換の前処理として行われるが、上記と同様な方法を用いることによりべた書きかな文の文節まとめあげも行った。今回提案する方法で、短時間で高精度な文節まとめあげができるようになった。

1 はじめに

文節まとめあげは係り受け解析の前処理として行われる。係り受け解析時に二要素間の距離が重要な情報となる。形態素単位ではなく、文節単位により、二要素間の距離を用いることにより、係り受け解析の精度を上げることが可能となる。べた書きかな混じり文の文節まとめあげはかな漢字変換の精度をあげる前処理として行われる。

既存の文節まとめあげのための手法は、形態素解析結果を前提とし、係り受け解析的な手法を用いているものが多い。具体的な手法として、決定木学習 [8]、最大エントロピー法、用例ベースによる (類似度を利用する) 手法、排反な規則を用いる手法、排反な規則と類似度を共に用いる手法 [11] がある。

形態素解析を必要のない文節まとめあげとして、点訳のための文節区切りを目的とした、表層情報からの解析する手法 [7] がある。また、べた書きかな文から文字 n-gram による解析 [3][4] や、文節単位の形態素解析 [10] などの手法がある。

本研究では、形態素解析と文節まとめあげを同時に行う方法を提案する。また関連して、品詞情報と文節区切り情報を利用したべた書きかな混じり文の文節まとめあげについても述べる。

2 形態素解析への文節区切り情報付与

形態素の品詞情報に文節区切りの情報を組み込んだマルコフモデルを作成することにより、文節境界の推定を行った。英語、日本語、べた書きかな文について解析を行った。英語については、最小の名詞句のまとめあげという問題を取り扱った。

2.1 共通した方法

形態素解析は文中の単語の語幹・接辞・語形変化を同定する処理のことをいう。形態素解析の統計的な手法として、単語と品詞の間の隠れマルコフモデルを用いて解析する方法がよく知られている。この隠れマルコフモデルは外部で観測される単語の列に対して、その内部状態として一意には決定できない品詞の遷移があると考えるモデルである。品詞の遷移確率と単語の出力確率をコーパスから学習し、学習データから最も大きい確率を生成する入力文に対する品詞付けを選択することによって解析が行われる。

今回提案する方法は、この内部状態に対して文節区切り情報を付与した隠れマルコフモデルを作成し、形態素解析と文節まとめあげの同時に行う手法である。品詞の遷移確率として、2つの形態素間の遷移において文節が区切られるか区切られないかの情報を付与する。また、単語の出力確率についてもその単語の前において

表 1: BaseNP の文節区切り情報

タグ	タグの意味
S	BaseNP のはじまりである GAP
E	BaseNP のおわりである GAP
B	BaseNP のはじまりであっておわりでもある GAP
C	BaseNP の内部であり、はじまりもおわりでもない GAP
N	BaseNP の外部であり、はじまりもおわりでもない GAP

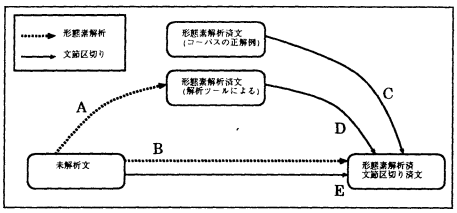

図 1: 5 つのモデル

文節が区切られるか区切られないかの情報を付与する。この文節区切り情報と品詞情報の 2 つ組を新たに品詞とみなす。このように作成した隠れマルコフモデルによって、文節区切り位置が同定することができる。付与する文節区切り情報については、後の節で詳しく述べる。

2.2 英語の解析

文節を BaseNP という再帰的に定義されない最小の名詞句という単位で定義する。[1] で紹介されている手法を用いて文節の区切り情報を定義する。具体的には単語間の空白に表 1に示すタグをつける。単語の前の空白についたこのタグと品詞情報を両方用いて、品詞 bigram マルコフモデルを構成した。

学習用のコーパスとして Penn Treebank [2] の Wall Street Journal の tagged file [1]を使用した。

2.3 日本語の解析

形態素の各品詞情報に、その形態素の前で文節が区切られるかどうかの二値 (区切られる = 1 , 区切られない = 0) をつけ、品詞 bigram マルコフモデルを構成した。

学習用のコーパスとして京都大学テキストコーパス [6] を使用した。

2.4 べた書きかな文の解析

べた書きかな文の文節まとめあげはかな漢字変換の前処理として行われる。

形態素の読みをエントリとし、品詞としてその形態素とその形態素の前で文節が区切られるかの情報を持つ辞書を作成し解析を行った。

[1]元コーパスにいくつかの修正を行って使用した。一つは `doesn't` 等 `does/VBZ n't/RB` として登録されているものを`doesn't/VBZ+RB` と変更した. もう一つは二重引用符``,''を全て"に統一した。

これにより、形態素単位に区切られると同時に、その形態素間で文節が区切られるかどうかの情報が付与される。

学習用のコーパスとして京都大学テキストコーパス [6] を使用した。

3 解析精度と解析結果の検証

文節まとめあげは 3 つのモデル、形態素解析は 2 つのモデルにより解析を行い、精度を比較した。全てのモデルはコーパスから統計学習することにより作成した。5 つのモデルを図 1に示す。

形態素解析は以下に示す 2 つのモデルについて解析を行い比較を行った。モデル (A) は通常の形態素解析である。モデル (B) は文節情報を付与し、文節まとめあげを同時に行うときの形態素解析である。

文節まとめあげは以下に示す 3 つのモデルについて解析を行い比較を行った。モデル (C) はコーパスから抽出した形態素解析済データから文節まとめあげを行うモデルである。モデル (D) はモデル (A) による形態素解析済データから文節まとめあげを行うモデルである。この形態素解析結果には誤りが含まれている。モデル (E) は未解析文から同時に形態素解析と文節まとめあげを行ったときの文節まとめあげである。なお、モデル (B) とモデル (E) は取り出す対象が異なるだけで、同じモデルである。

表 2〜表 5 の中の「訓練セット」は学習に使用したデータを解析したときの精度である。「評価セット」は学習に使用していないデータを解析したときの精度である。「再現率」はシステム出力中の正解の数を正解データの数で割ったもの「適合率」はシステム出力中の正解の数をシステムの出力の数で割ったものである。

形態素解析の精度は、形態素単位の品詞の一

表 2: 形態素解析の精度 (英語)

	訓練セット		評価セット	
	再現率	適合率	再現率	適合率
モデル A	97.14%	97.34%	96.75%	96.95%
モデル B	97.48%	97.69%	96.85%	97.05%

表 3: 文節まとめあげの解析精度 (英語)

	訓練セット		評価セット	
	再現率	適合率	再現率	適合率
モデル C	97.48%	97.69%	96.82%	97.03%
モデル D	97.68%	97.86%	97.31%	97.49%
モデル E	97.96%	98.17%	97.20%	97.40%

表 4: 形態素解析の精度 (日本語)

	訓練セット		評価セット	
	再現率	適合率	再現率	適合率
モデル A	97.81%	97.91%	96.58%	96.44%
モデル B	97.96%	98.05%	96.72%	96.57%

表 5: 文節まとめあげの解析精度 (日本語)

	訓練セット	評価セット
モデル C	99.76%	99.42%
モデル D	99.67%	99.30%
モデル E	99.78%	99.46%
べた書きかな文	99.72%	99.40%

致率による。英語の文節まとめあげの精度は、形態素解析結果が出力したわかち書き位置へのタグ付けを評価に対象にしている。英語の文節まとめあげ結果のわかち書き位置は、形態素解析時に、複合語などを分割する処理が行われるため、再現率と適合率は異なる。日本語の文節まとめあげについては、文字間へのタグ付けを評価対象にしている。このため、日本語の解析結果の再現率と適合率は一致する。

3.1 英語の解析

英語の解析は、コーパスを 5 つに分け、4 つを訓練セット、1 つを評価セットとした。そのような学習セットを 5 組作成し、その結果の平均をとった。

表 2に英語の形態素解析の解析精度を示す。

英語の形態素解析品詞タグ付けについては、文節区切り情報を付与しないモデル (A) より付与したモデル (B) の方が精度が高かった。

表 3に英語の文節まとめあげの解析精度を示す。

訓練セットについては、形態素解析と文節まとめあげを融合したモデル (E) が最も精度が良かった。評価セットについては、逐次処理するモデル (D) が最も精度が良かった。

3.2 日本語の解析

日本語の解析は、コーパスを 9 つに分け、8 つを訓練セット、1 つを評価セットとした。そのような学習セットを 9 組作成し、その結果の平均をとった。

表 4に日本語の形態素解析の解析精度を示す。

日本語の形態素解析品詞タグ付けについても、文節区切り情報を付与しないモデル (A) より、付与したモデル (B) の方が精度が高かった。

表 5に日本語の文節まとめあげの解析精度を示す。

日本語文節まとめあげについては、形態素解析と文節まとめあげを逐次処理するモデル (D) より、融合モデル (E) の方が精度が良かった。また、正しい形態素解析結果から文節まとめあげを行うモデル (C) よりも、融合モデル (E) の方が高い精度を示した。

3.3 べた書きかな文の解析

べた書きかな文の解析は日本語 (かな漢字混じり文) と同様に 9 つのセットを作成し平均をとった。なお、べた書きかな文の文節まとめあげの正解率は文字間単位で計算を行った。

表 5の最下行に日本語の文節まとめあげの解析精度を示す。

3.4 解析結果の検証

形態素解析については、英語、日本語ともに文節区切り情報を付与することにより、解析精度をあげることができた。

文節まとめあげについては、英語と日本語とで異なった結果が見られた。

英語については、形態素解析を行ってから文節まとめあげを行うモデル (D) が最も結果が良かった。英語の誤り箇所を見ると、一つの誤り箇所の後に複数の誤り箇所が続く現象が見られた。これは、英語の文節を名詞句の範囲を示す括弧で定義されていて、一つ間違えると括弧の

整合性にひきずられて誤りが連鎖するからだと考えられる。

日本語については、形態素解析と文節まとめあげを分離するモデルより、融合モデルのが結果が良かった。

計算時間は、日本語の新聞記事1000文を解析を行うのに、融合モデル (E) は 15.75 sec 要した。また、形態素解析を行ってから文節まとめあげを行う場合、形態素解析 (A) に 12.39 sec 文節まとめあげ (D) に 11.06 sec 要した[2]。計算時間についても、融合モデル (E) の方が優れていると言える。

べた書きかな文の解析も、かな漢字混じり文と同等な精度で解析ができた。この手法では、形態素解析を同時に行なっているため、辞書のエントリーに変換される漢字のリストを保持することができる。この漢字のリストを用いて、文節区切りと同時に内部に漢字の変換候補を持つ束構造を持つことができる。前後に共起する単語情報により漢字候補の絞り込みを行うことによって、統計的手法を用いたかな漢字変換ができるだろう。

4 おわりに

本研究では、形態素解析と文節まとめあげを融合する方法の提案を行った。融合することにより、形態素解析の精度をあげることができた。また、今回提案する方法により、簡単に文節まとめあげを行うことができるようになった。

現在は品詞情報だけに基づく統計モデルを用いたが、今後の方向性として、現在の解析で誤りの原因となっている単語について、単語の情報まで見ることによる解析精度の向上を考えている。また、文節の主辞となっている単語の情報を付与することにより、主辞の特定を同時に行えるようなシステムの開発を行いたいと思っている。

謝辞

今回の解析には、形態素解析ツールキットプロジェクト [12] による形態素解析器 MOZ [3] を利用した。今回の研究にあたって MOZ の開発者である奈良先端科学技術大学院大学自然言語処理学講座の山下達雄氏に、様々な助言を頂きました。ここに感謝の意を表します。

参考文献

[1] Michael John Collins A New Statistical Parser Based on Bigram Lexical Dependencies, ACL-96, pp.184-191, June ,1996.

[2] M. Marcus, B. Santorini and M. Marcinkiewicz. Building a Large Annotated Corpus of English: Penn Treebank. *Computational Linguistics,* 19(2):313-330

[3] 荒木哲郎, 池原悟, 土橋潤也. 2重マルコフ連鎖モデルを用いたべた書きかな文の文節先頭位置推定法の評価. 情報処理学会研究報告, 93-NL-94, March 1993.

[4] 荒木哲郎, 池原悟, 土橋潤也, 笹島伸一. 3重マルコフモデルによるべた書きかな文の仮文節推定法. 情報処理学会研究報告, 94-NL-102, July 1994.

[5] 北研二, 中村哲, 永田昌明. 音声言語処理-コーパスに基づくアプローチ- 森北出版, 1996.

[6] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会 第3回年次大会, pp.115-118 ,1997.

[7] 鈴木恵美子, 小野智司, 平岡大樹, 狩野均, 西原清一. 知識ベースに基づく点字翻訳のための日本語文節区切り手法. 情報処理学会研究報告, 97-NL-120. July 1997.

[8] 張玉潔, 尾関和彦. 分類木を用いた日本語文の自動文節分割. 情報処理学会研究報告, 97-NL-121, Sep 1997.

[9] 長尾真 編. 自然言語処理 pp.117-137, 岩波出版, 1996.

[10] 兵藤安昭, 池田尚志. 文節単位のコスト最小法による日本語形態素解析. 信学技報 NL98-2, 電子情報通信学会, May 1998.

[11] 村田真樹, 内元清貴, 馬青, 井佐原均. 学習による文節まとめあげ -決定木学習, 最大エントロピー法, 用例ベースによる手法と排反な規則を用いる新手法の比較-. 情報処理学会研究報告, 98-NL-128, Nov 1998.

[12] 山下達雄, 松本裕治. 言語に依存しない形態素解析ツールキットの開発情報処理学会研究報告, 98-NL-128. Nov 1998.

[2] 解析に要した時間は実時間である。システム CPU 時間は、(E) 0.51sec (A)0.31 sec (D) 0.63 sec である。(UltraSPARC-II 296MHz, Memory 917.5MB による)

[3] MOZ のパッケージは以下の URL で公開されている。
http://cl.aist-nara.ac.jp/student/tatuo-y/ma/