複数素性の順次適用による文節まとめあげ

白木 伸征　梅村 祥之　原田 義久

株式会社 豊田中央研究所

1 はじめに

近年の高度情報化の流れにより，種々の情報機器が自動車にも搭載されるようになり，さまざまな情報通信サービスが広がりつつある．その中には，交通情報，観光情報，電子メール，一般情報（例えば新聞記事）などが含まれるが，このような情報はディスプレイ上に文字で表示するよりも，音声により提供する方が望ましいとされている．

文字情報を音声に変換する技術の研究開発は進んでいるが，その合成音声の韻律は不自然であるという問題がある．その原因として大きな割合を占めるものはポーズ位置の誤りであり，その改善が重要となっている．ポーズ位置を制御するために係り受け解析を利用する手法が研究され，その有効性が報告されている[1]．本研究では，この係り受け解析において重要な位置を占めている文節まとめあげに注目した．そして複数素性[1]の順次適用による文節まとめあげという新しい手法を考案し，それにより99.38%という高い精度を得られた．

2 従来の研究

文節まとめあげに関する従来の研究は，人手により文節まとめあげの規則を書き下す方法と，大規模コーパスから機械学習により得た統計情報を利用する方法の2種類に大きく分けられる．

人手により作成した文節まとめあげの規則を利用する最も良く知られているツールに，knp2.0b4[2]がある．knpは文節に関する規則を人手で網羅することにより，非常に高精度な文節まとめあげを実現している．knpの文節まとめあげの規則は906行のファイルに148種類の規則が記述されている．文節まとめあげにはこのように多数の規則が必要であるため，人手で作成するためには修正・追加を繰り返さなければならないという問題がある．また，knpへの入力は形態素解析ツールJumanの出力を前提にしているのに対し，音声合成システムに含まれる形態素解析のデータ形式はそれぞれ異なる．そのため，knpを車載機器に実装するのが難しいという問題もある．

このような問題に対処でき，最近最も盛んに研究されているのが，大規模コーパスからの機械学習により得た統計情報を利用して文節まとめあげを行う手法である[3] [4]．これまで行われた研究で最も精度の高い結果を得ているのが，村田らによる研究である[4]．村田らの手法は，学習コーパス中での出現確率が100%である規則を排反な規則と呼び，その規則を最優先で利用する方法である．排反でない規則を利用するということは，あらかじめ誤る可能性のあるものを利用するということであるため，高い精度を望むことができない．そこで，排反な規則を重要視することが必要である，と主張している．この手法による文節まとめあげは，最高で99.17%という高い精度を得ている．ただしこの手法は，文節まとめあげの排反な規則を152種類も利用するため，処理が複雑かつデータが膨大になるという問題があり，車載情報機器への実装は困難である．

3章では，これらの問題を解決するための新しい手法について述べる．

3 文節まとめあげの手法

3.1 複数の形態素n-gramの順次適用による文節まとめあげ

本研究では，従来手法の問題点を解決するために，次のような目標を立てた．

- 学習が容易で，素性の数が少ないこと

[1] 文節まとめあげに用いる規則や特徴量．

- 素性を利用して文節をまとめる方法が従来手法より簡明であること
- 精度が従来手法と同程度かそれ以上となること

これらを実現するために，**複数素性の順次適用による文節まとめあげ**という新しい手法を提案する．

機械学習を用いる従来の文節まとめあげ手法は，大規模コーパスから多くの形態素の n-gram，主に 2-gram から 4-gram 程度を利用して，形態素間が文節の境界となるかどうかを判定している．村田らの方法では，前後 4 形態素を組み合わせて 152 種類の素性を考慮している．しかし，1 つの形態素の隙間に対して 152 種類の素性を考慮すると，1 文あたり 152×20= 約 3000 回もの計算をしなければならない[2]．

本研究で提案する手法は，1 つの形態素の隙間に対して 6 種類の素性だけを考慮する．具体的には，品詞，単語表記，品詞細分類，品詞＋単語表記の 4 種類の形態素 2-gram と，品詞，単語表記の 2 種類の形態素 3-gram を利用する．形態素の隙間ごとにこれらの素性を調べて，それにより文節に区切るか区切らないかを決定する．

例えば，ある文中の X，Y，Z という連続する 3 つの形態素の Y と Z の間の文節区切りを決定する処理は次の通りである．

1. X，Y，Z の形態素情報を得る．
2. 6 種類の素性 (X，Y，Z から得られる 3-gram と Y，Z から得られる 2-gram) を順番に調べる．調べる順番は 4章の実験により決定する．
   (a) 調べている素性が学習結果のデータベース中にあり，文節に区切る事象の方が接続する事象よりも多い場合には区切りを入れ，少ない場合には接続する．ここで Y と Z の文節区切りを確定し，処理を終了する．
   (b) 調べている素性がデータベース中にない場合，または文節に区切る事象と接続する事象が等しい場合には，次の素性を調べる．
3. 6 種類すべての素性を調べた結果，文節に区切るか区切らないか確定しない場合，文節に区切るものとする．

[2]京大コーパス [5] 中の毎日新聞 1995 年 1 月 1 日の記事中では，1 文平均 20 の形態素の隙間がある．

本研究の手法の最大の特徴は，**6 種類の素性を順番に調べるだけで文節まとめあげを行う**，という非常に簡明な点である．

3.2 形態素 n-gram の取得方法

文節まとめあげに用いる形態素 n-gram は，大規模コーパスから機械学習によって得る．本研究では，京大コーパス [5] を利用した．京大コーパスにはあらかじめ詳細な形態素の情報と文節区切りの情報が付与されているので，形態素の隙間ごとに文節区切りがあるかどうかを数えて，それを尤度比[3]の高い順に並べて保持する．

4 実験

本研究の文節まとめあげ手法の性能を評価するため，評価システムを作成して以下の 3 種類の実験を行った．

1. 形態素 2-gram，3-gram の各素性を適用する順番や数を変化させる実験
2. 学習コーパスの量を変化させる実験
3. 学習結果の一部を削除する実験

1.,3. の実験の学習コーパスには，京大コーパスの最初の 10000 文を利用し，2. の実験には，京大コーパスの最初から 1000 文づつ 10000 文まで変化させて利用した．また，すべての実験のテストコーパスは京大コーパスの 10001 文目から 5000 文を利用した．

また本研究での文節まとめあげの評価基準は，村田らと同じ F 値を用いた[4]．

4.1 実験結果

評価システムで行った 3 種類の実験の結果を以下に報告する．

1. 形態素 2-gram，3-gram の各素性の数や適用する順序を変化させる実験

   本研究の手法は最大 6 種類の素性を用いるので，素性の数や適用する順序による精度の変化を調べた．6 種類の素性のすべての組み合わせに関して調べたところ，図 1のような結果を得た．図中の

[3]本研究では尤度比とは対数尤度比のことをさす．

[4]F 値は適合率と再現率の調和平均である．

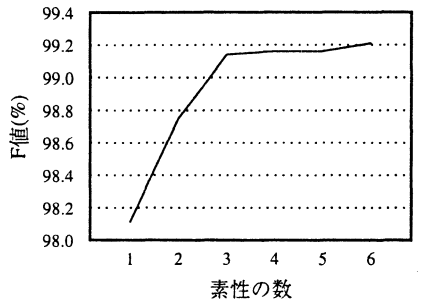

図 1: 素性の数や順序による精度の変化

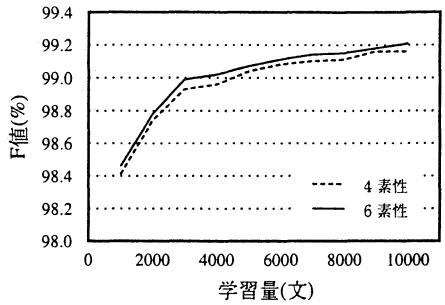

図 2: 学習コーパスの量による精度の変化

F 値は，最も精度の良い素性の組み合わせの結果である．

この結果から，素性の数を増やすと精度が上昇することがわかった．

2. 学習コーパスの量を変化させる実験

学習コーパスの量を 1000 文から 10000 文まで変化させた時の精度の変化を調べた．その結果を図 2に示す．図中の 4 素性と 6 素性は，実験 1 でのそれぞれの素性の組み合わせを用いた結果である．

この結果から，学習量を増加させると精度も上昇することがわかった．また 10000 文よりさらに学習量を増やすと，より精度が上昇すると思われる．

3. 学習結果の一部を削除する実験

機械学習の結果は尤度順に並べるが，そのうち尤度が高いものだけを利用する時に精度がどのように変化するか調べた．この実験は，車載情報機器への実装時にはデータ容量が小さいことが必要であるため，データ容量による精度変化を調べるためのものである．学習結果の利用割合を 10% から 100% まで変化させて実験を行ったところ，図 3の結果を得た．

この結果から，およそ 60% の学習データを用いれば 100% 用いる時とほぼ同じ精度を得られることがわかった．

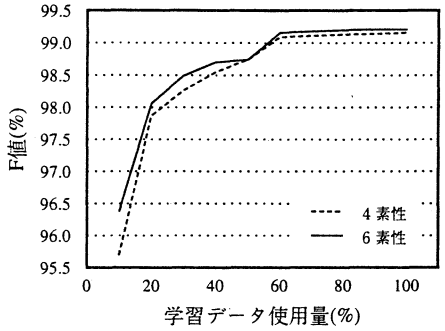

図 3: 学習結果の一部削除による精度の変化

4.2 素性や条件の追加実験

本手法の最大の特徴は，非常に簡明な方法で充分な精度を得られることである．非常に簡明であるので，従来手法の長所だけを組み合わせることも容易である．そのことを示すため，学習コーパスを 10000 文に固定して以下のような 2 種類の追加実験を行った．

- 1-gram を利用する方法

  2-gram や 3-gram だけでなく，1-gram が非常に有効となる場合も考えられる．そこで，6 種類の素性に加えて 1-gram を用いて実験を行ったところ，F 値が 99.29% に上昇した．
- 排反な規則を用いる方法

  排反な規則を最優先で利用すると高い精度を得られることが，村田らにより報告されている．6 種類の素性の中の排反な規則を最優先で利用するようにしたところ，F 値が 99.24% に上昇した．

6 種類の素性に加えて，上記 2 種類の手法をすべて組み合わせて実験を行ったところ，99.38% の F 値を得ることができた．

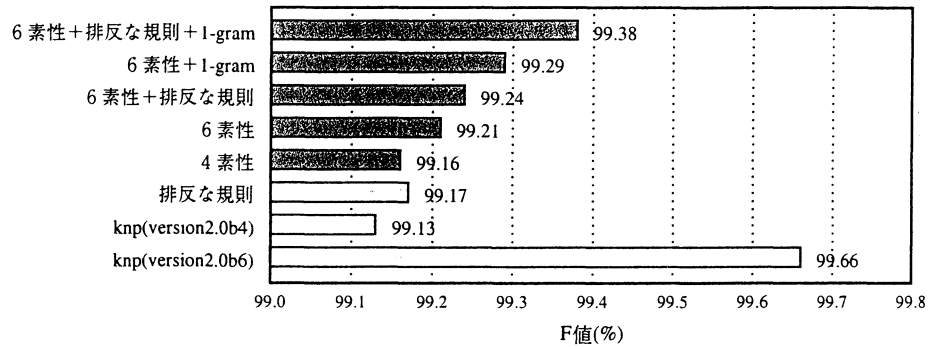

図 4: すべての実験の結果

4.3 実験のまとめ

以上の実験の結果を図4のグラフにまとめた．比較のため，knp2.0b4 の精度と，knp2.0b4 と機械学習を組み合わせた knp2.0b6[6] の精度も示した．

村田らによる手法と本研究の手法は，学習コーパスとテストコーパスが異なる[5]という点では，図4の結果を単純に比較するのは難しい．しかし本研究の手法が非常に簡明であること，車載情報機器への実装を最大の目標としていることを考慮すると，99.38% は非常に高い精度であるといえる．knp2.0b6 の精度が非常に高いのは，京大コーパスが knp2.0b4 の出力を人手で修正して作成され，その結果を学習したためである．

5 おわりに

本研究で提案した文節まとめあげの手法は，複数の素性を順次適用して文節の区切りを行うだけであり，非常に簡明である．しかしその精度は従来の手法と比較しても遜色はないことが示された．また，本手法は非常に簡明であるため，他の手法の長所のみを導入することが容易であり，そのことを 1-gram や排反な規則を組み合わせることにより示した．また 3.1節で立てた3つの目標をすべて達成することができた．

本研究で得られた 99.38% という精度は，あらゆる素性の組み合わせを調べた結果であり，本手法の枠組みにおいては最高の精度であると考えられる．今後は，本手法を係り受け解析の技術と融合させ，品質の高い音声合成技術に応用したいと考えている．

[5]村田らの報告では，学習コーパスもテストコーパスも 1000 文強のデータ量である．

参考文献

[1] 清水 司, 梅村 祥之, 原田 義久: 係り受け構造を用いたポーズ位置およびポーズ長の決定, 言語処理学会第 5 回年次大会発表論文集, pp.414-pp.417, 1999.

[2] 黒橋 禎夫: 日本語構文解析システム KNP 使用説明書 version2.0b4, 京都大学大学院情報学研究科, 1997.

[3] Zhang Y. and Ozeki K.: The Application of Classification Trees to Bunsetsu Segmentation of Japanese Sentences, *Journal of Natural Language Processing*, Vol.5, No.4, pp.17-33, 1998.

[4] 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－, 情報処理学会自然言語処理研究会 NL128-4, pp.23-30, 1998.

[5] 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 回年次大会発表論文集, pp.115-118 1997.

[6] 黒橋 禎夫: 日本語構文解析システム KNP 使用説明書 version2.0b6, 京都大学大学院情報学研究科, 1998.