2007－SLP－67 (3)
2007／7／20

局所的な係り受けと韻律の素性を用いた話し言葉の節・文境界推定

尾嶋 憲治[†] 秋田 祐哉[†‡] 河原 達也[†‡]

[†]京都大学 情報学研究科
[‡]京都大学 学術情報メディアセンター
〒606-8501 京都市左京区吉田二本松町

あらまし 我々はこれまでにSVMを用いて節境界・文境界を自動的に推定する手法を提案しているが、本稿では、直後の文節への係り受け情報および一般的な韻律情報の導入による拡張を検討する。『日本語話し言葉コーパス』(CSJ)の講演音声を用いて評価実験を行った結果、隣接文節間の係り受け情報が文境界推定に対して有効であること、および韻律情報が音声認識結果における節・文境界推定に有効であることがわかった。

Clause and Sentence Boundary Detection Using Local Syntactic Dependency and Prosodic Features

Kenji OJIMA[†] Yuya AKITA[†‡] Tatsuya KAWAHARA[†‡]

[†]School of Informatics, Kyoto University,
[‡]Academic Center for Computing and Media Studies, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract We have proposed an approach utilizing support vector machines (SVM) for clause and sentence boundary detection of spontaneous Japanese. In this report, we present an extension of this approach by using local structure of syantactic dependency and prosodic features. We evaluated these methods on manual and automatic transcription of spontaneous lectures and speeches. As a result, the local syntactic dependency is effective in sentence boundary detection, and the prosodic features are effective for boundary detection in automatic transcription.

1 はじめに

近年、講演や会議などの音声認識技術の進展にともない、話し言葉音声を対象とした筆記録の作成や整形・要約などの研究が進められている [1]。しかし、多くの自然言語処理技術は書き言葉の「文」や「節」を処理単位としているのに対して、音声では句読点がなく同様の処理単位が自明ではないことが問題となっている。ポーズの検出により処理単位を認定することも一般的に行われているが、このような単位は言語的に均質なまとまりになっているとは限らず、文や節などの処理単位と必ずしも一致するとはいえない。

これに対して、入力音声を文の単位に自動分割する文境界推定の研究が行われている。欧米では、英語の放送ニュース音声タスク [2, 3, 4] および電話会話音声タスク [4, 5, 6] が主な研究対象である。最も一般的な手法は韻律と言語的情報をもとに決定木などを用いて判定する手法であり [3, 4]、これらのタスクで高い性能を示している。このほか韻律のみに基づいた手法も提案されている [2, 6]。一方、日本語を対象とした文境界推定としては [7] などが行われている。英語とは異なり、日本語では明らかな文末表現が存在するため、言語的手がかりを用いた手法が多く提案されている。我々も統計的言語モデルおよびサポートベクターマシン（SVM）を用いた手法を提案し、高い精度を達成している [8]。

本稿では、我々がこれまで提案してきた言語的素性によるSVM手法の拡張を提案する。新たな言語的素性として、隣接文節間に限定した局所的な係り受け関係 [9] を導入する。また，多くの先行研究で用いられている基本周波数（F0）などの韻律素性の導入も検討する。これらの拡張について、『日本語話し言葉コーパス』（Corpus of Spontaneous Japanese: CSJ）を対象として評価を行う。

2 CSJにおける節・文の単位

CSJは、学会講演や模擬講演などのモノローグを主な対象として収集・構築されたコーパスである。CSJに収録されている講演のうち、コアと呼ばれる一部の講演に対しては、書き起こしテキストのほかに形態素・係り受け・節単位などの言語的情報や韻律ラベルなどの音響的情報が付与されている。本研究では文の単位としてCSJにおける節単位の定義を採用している。

表 1: 節境界ラベルの例

境界の種類	節境界ラベル
絶対境界	文末・文末候補・と文末など
強境界	並列節「ケド」「ガ」など
弱境界	理由節「カラ」・タリ節・条件節「ナラ」「レバ」など

CSJにおける「節単位」とは、話し言葉の文に相当する統語的・意味的な妥当性を備えた単位であり、音声言語処理に有用な単位であると考えられる。この節単位は、より詳細な単位である節境界をまず自動推定し、それらの中から人手による修正を施した上で認定されている [10]。節境界は、その切れ目の大きさによって、絶対境界・強境界・弱境界の3種類に分けられている。節境界ラベルの例を表1に示す。このうち絶対境界・強境界は基本的に文境界（節単位）となり、弱境界については機能的に分割できると判断された箇所のみが文境界となる。節境界ラベルの推定にはプログラムCBAP-CSJ [11] が用いられているが、このプログラムはルールに基づいて判定しており、誤りを含む音声認識結果に対しては精度が著しく低下することが知られている。

3 SVMに基づく節・文境界推定

我々は、音声認識結果における文境界推定の手法として、統計的言語モデルを用いたものおよびSVMを用いたものの2つの手法を提案している [8]。統計的言語モデルを用いた手法では、文境界が付与された学習テキストからN-gram言語モデルを学習し、入力音声で検出された節境界の表現において、文境界の有無による言語尤度の違いをもとに文境界を決定する。CSJの

講演音声を用いた評価実験では、書き起こしで0.82、音声認識結果で0.71のF値を得ている。一方、SVMによる手法では、文境界推定をテキストチャンキングの問題として扱い、SVMベースのテキストチャンカを学習している。素性には前後3単語の単語情報（表層表現、読み、品詞）・ポーズ情報などを用いている。CSJの講演音声を用いた評価実験では、書き起こしで0.85、音声認識結果で0.78のF値を得ている。

我々は、このほかSVMチャンカを用いた節境界推定も検討している。CSJで定義されている3レベルの節境界に加え、体言止めなどの人手での修正により認定された境界を節境界と定義し、4種の境界と境界以外の5クラスを、pairwise法を利用したSVMに基づく多値分類器によりそれらを推定している[9]。CSJの講演音声を用いた評価実験では、絶対・強・弱境界に対して、書き起こしで0.95以上の、音声認識結果で0.63〜0.75のF値を得ている。

2つの手法を比較すると、SVMを用いた手法の方が高い精度を達成しており、本稿ではSVMを用いた手法の拡張を検討する。以下では、まず局所的係り受け情報の導入について検討し、次に韻律情報の導入について述べる。

4 局所的係り受け情報を利用した節・文境界推定

4.1 係り受け情報

係り受け情報は、ある文節が最も依存している他の文節を係り先という形で示したものであり、係り先は二文節の単語情報から、「主語と述語」「修飾語と被修飾語」といった二文節間の関係を考慮して決定される。日本語文においては、述語は基本的に節末に存在するという性質をもっていることから、係り受け情報は節境界・文境界を推定する手がかりとなりうる。

このとき定義されている係り受けは、一般的に書き言葉における文節間の係り受けである場合が多い。そのため、話し言葉における係り受け解析では、以下に示すような話し言葉特有の問題が生じる[12]。

1. 文境界が明示されていない
2. 係り先がない文節がある
3. 係り受け関係が交差する
4. 言い直しが多い
5. 倒置表現がある

以上の問題に加え、話し言葉では音声認識誤りや係り受け解析誤りの問題が避けられない。

これらの問題の多くは遠距離の係り受けで特に深刻であることから、係り受け情報を利用する場合には、局所的な係り受けに限定することで誤りの影響を抑えることができる。そこで、直後の文節への係り受けに着目する。我々は、字幕のような話し言葉の処理単位の生成のために、このような局所的係り受け情報を用いた段階的チャンキングを提案しており、その有用性を確認している[9]。ここではまず係り受け情報を用いて文節から構成要素と呼ばれる単位へのチャンキングを行い、次に生成された構成要素からの節・文境界推定を行っている。これに対し、本研究では、係り受け情報をSVMの素性として加えて、形態素列から直接節・文境界推定を行う。[9]では、構成要素のチャンキングの誤りが後段の処理に影響を及ぼすが、本手法は構成要素のような中間的かつ決定的な境界を定めないため、このような問題は発生しない。

一般的に、文節の係り受けは文内部で完結することが多い。CSJにおいても係り受け情報は文単位で付与されており、文間での係り受けが考慮されていないことから[13]、文境界直前の文節が後ろの文節に係る可能性は低いと考えられる。すなわち、直後の文節への係りが認められる文節は、文境界直前の文節にはなり得ないとみなすことができる。このことから、直後の文節に係る、係らない（係り先なしを含む）を素性とすることは妥当である。

4.2 評価実験

本実験では、CSJ公開版の音声認識テストセット（30講演）を評価に用い、これを除くコアのデータ（168講演）を学習セットとした。テキストチャンカにはYamChaを用いて、3次の多項式カーネルをカーネル関数としている。SVM

に与える素性として、前後3形態素の単語情報(表層表現、読み、品詞情報)、文節の境界、各講演の中で正規化したポーズ・フィラー情報および直後の文節への係りを用いた。さらに、文境界推定の実験においては、節境界の推定結果の情報も素性に加えている。なお、文節境界の推定精度(F値)は書き起こしで0.982、音声認識結果で0.793であり、直後の文節への係り受け精度(F値)は書き起こしで0.882、音声認識結果では0.665であった。また、ポーズは、CSJの転記基本単位の定義に従い、フィラーは、形態素単位にFタグが付与された感動詞、Dタグが付与された言いよどみとしている。

この他にも、一般には活用形なども素性として用いることが多いが、音声認識結果では終止形と連体形の混同が多く見られるので、本研究では用いていない。なお予備実験において、活用形を素性として用いない場合の方が、用いる場合と同等またはそれ以上の精度という結果を確認している。

節境界推定の実験結果を表2に、文境界推定の結果を表3に示す。文境界推定では、文節区切りと直後の文節への係りの素性を加えることにより、性能の向上が見られた。一方、節境界推定では、2つの素性を加えても、同等あるいはやや低い結果となった。これは、文境界とならない節境界においては、境界直前の文節であっても直後の文節に係る文節があったことが影響しているためと考えられる。

表2: 局所的係り受け情報を利用した節境界推定精度(F値)

対象	素性	絶対 (1794)	強 (1077)	弱 (4123)	全境界 (7215)
書き起こし	単語・ポーズ	0.958	0.967	0.969	0.961
	+文節	0.960	0.969	0.972	0.965
	+文節・直後への係り	0.953	0.970	0.962	0.960
音声認識結果	単語・ポーズ	0.759	0.747	0.649	0.715
	+文節	0.746	0.736	0.643	0.708
	+文節・直後への係り	0.726	0.738	0.641	0.703

表3: 局所的係り受け情報を利用した文境界推定精度

対象	素性	再現率	適合率	F値
書き起こし	単語・ポーズ・節	82.7%	88.0%	0.853
	+文節	83.2%	87.6%	0.854
	+文節・直後への係り	84.0%	92.0%	0.878
音声認識結果	単語・ポーズ・節	56.4%	66.0%	0.609
	+文節	56.9%	66.0%	0.612
	+文節・直後への係り	56.5%	67.5%	0.615

5 韻律素性を利用した節・文境界推定

5.1 韻律ラベルの情報

発話においては文全体あるいは句単位において、基本周波数(F0)が立ち上がった後に継時的に下降調になるというパターンがある。このような韻律的特徴を文境界推定の手がかりとして用いる試みがされている[14]。

CSJのコアを対象に付与されている韻律的特徴は、日本語(東京方言)の音韻的構造分析に基づくラベリング体系であるX-JToBIに準拠してラベリングされており[15]、これらのラベルは、F0のパターンや音韻の時間長変化によるリズムを考慮して定義されたものである。このX-JToBIによるラベル情報は、以下の6層から構成されている。

1. 単語層
2. 分節音層
3. トーン層
4. BI層
5. プロミネンス層
6. 注釈層

以上の6層のうち、本研究では韻律素性として、トーン層およびBI層を利用することを考える。トーン層は、イントネーションを構成するトーンの種類と時間軸上の位置を示しており、ラベルに対応するF0イベントの時刻に付与されている。ラベル例を表4に示す。また、BI層はすべての語境界に付与されているもので、基本的に韻律境界の深さに応じて1から3までの整数と、その韻律境界の深さが整数値の中間であると判断される場合の根拠を示すアルファベットから構成されている。

表 4: X-JToBI トーン層ラベルの例

ラベル	説明
A	語彙アクセント
%L	アクセント句頭境界
L%	下降調の句末境界
H%	単純な上昇調の句末境界
HL%	上昇下降調の句末境界

付与される BI ラベルのうち、もっとも深い韻律上の境界を表す「イントネーション句境界」は、そこでピッチレンジがリセットされる、アクセント句境界よりも深い境界と定義されており、整数値 3 が付与されている。学習セットとして用いた 168 講演を対象に、イントネーション句境界と節境界の関連を調査したところ、イントネーション句境界に一致する節境界の再現率は 66%であり、このときのイントネーション句境界の適合率は 38%である。

5.2 評価実験

CSJ では韻律情報はコアのみに含まれるため、本実験では音声認識テストセットのうちコアに含まれる 8 講演を評価に用い、前節と同様に 168 講演を学習セットとした。SVM に与える素性として、前後 3 形態素の単語情報（表層表現、読み、品詞情報）、各講演の中で正規化したポーズ・フィラー情報、X-JToBI におけるトーン層および BI 層のラベル、各講演で正規化した F0 を用いた。この F0 は、各形態素に対してトーン層のラベルが付与された時刻に記録された値のうち、最大の値をとっている。また前節と同様、文境界推定の実験においては、節境界の推定結果の情報も素性に加えている。SVM のその他のパラメータについても、前節の実験と同様である。

節境界推定の実験結果を表 5 に、文境界推定の結果を表 6 に示す。今回新たに加えた素性のうち、F0 値の導入により、適合率には向上が見られたが再現率は低下しており、結果として F 値はほとんど変化がなかった。本実験では F0 値

表 5: 韻律素性を利用した節境界推定精度（F 値）

対象	素性	絶対 (412)	強 (274)	弱 (1102)	全境界 (1828)
書き起こし	単語・ポーズ	0.951	0.965	0.967	0.967
	+トーン層	0.949	0.951	0.956	0.962
	+BI 層	0.957	0.966	0.957	0.970
	+F0	0.950	0.949	0.957	0.963
	+F0・トーン層	0.941	0.939	0.946	0.954
	+F0・トーン層・BI 層	0.943	0.943	0.945	0.957
音声認識結果	単語・ポーズ	0.767	0.770	0.676	0.734
	+トーン層	0.785	0.774	0.673	0.736
	+BI 層	0.808	0.795	0.683	0.753
	+F0	0.761	0.767	0.670	0.730
	+F0・トーン層	0.788	0.764	0.676	0.739
	+F0・トーン層・BI 層	0.816	0.778	0.679	0.750

表 6: 韻律素性を利用した文境界推定精度

対象	素性	再現率	適合率	F 値
書き起こし	単語・ポーズ	77.5%	86.9%	0.819
	+トーン層	75.5%	87.0%	0.809
	+BI 層	77.9%	86.8%	0.821
	+F0	76.6%	88.4%	0.821
	+F0・トーン層	75.1%	87.7%	0.809
	+F0・トーン層 ・BI 層	74.8%	87.5%	0.807
音声認識結果	単語・ポーズ	57.3%	79.4%	0.666
	+トーン層	57.6%	79.3%	0.667
	+BI 層	59.9%	82.6%	0.695
	+F0	56.5%	81.0%	0.666
	+F0・トーン層	56.5%	83.0%	0.672
	+F0・トーン層 ・BI 層	59.5%	85.9%	0.703

をトーン層ラベルの位置に基づいて定めているが、これが適切でなかった可能性がある。また、トーン層のラベルは、アクセント句の始端・終端をより明示的に表していることから、および BI 層のラベルは、上で述べたイントネーション句境界と節境界の相関があることから、精度の向上が予測されたが、実験により実際に確認された。BI 層のラベルを単独で加えた場合を除き、書き起こしの場合は韻律素性を加えることにより性能が低下しているが、音声認識結果に対する文境界推定では、F 値で 0.04 の向上が見られた。これは、認識誤りが不可避である音声認識結果において、提案した韻律素性が効果的であることを示している。ただし、今回利用している CSJ における X-JToBI のラベルは、人手で付与されており、音声からの自動推定が課題で

ある。

6 おわりに

本稿では、SVMに基づく節・文境界推定において、直後の文節への局所的な係り受け情報やCSJの韻律ラベル・基本周波数などの韻律情報の評価を行った。CSJの講演音声を用いた評価実験の結果、文境界推定に関して、書き起こしでは局所的な係り受け情報により0.02のF値の向上が見られた。また、韻律素性を導入することにより、文境界推定で0.04のF値の向上が見られるなど、音声認識結果に対して有効であることを確認した。今後は、今回用いた韻律情報と言語情報を複合させた場合の評価を行うほか、その他の韻律素性の検討を行う予定である。

参考文献

[1] 河原達也. 筆記録作成のための話し言葉処理技術. 電子情報通信学会技術研究報告, SP2006-120, NLC2006-64 (SLP-64-36), pp. 209–214, 2006.

[2] D. Wang, L. Lu, and H.-J. Zhang. Speech Segmentation without Speech Recognition. In *Proc. ICASSP*, 2003.

[3] A. Srivastava and F. Kubala. Sentence Boundary Detection in Arabic Speech. In *Proc. Eurospeech*, 2003.

[4] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper. Structural Metadata Research in the EARS Program. In *Proc. ICASSP*, 2005.

[5] J. Huang and G. Zweig. Maximum Entropy Model for Punctuation Annotation from Speech. In *Proc. ICSLP*, 2002.

[6] D. Wang and S. S. Narayanan. A Multi-pass Linear Fold Algorithm for Sentence Boundary Detection using Prosodic Cues. In *Proc. ICASSP*, 2004.

[7] 中嶋秀治, 山本博史. 音声認識過程での発話分割のための統計的言語モデル. 情報処理学会論文誌, Vol. 42, No. 11, pp. 2681–2688, 2001.

[8] H. Nanjo Y. Akita, M. Saikou and T. Kawahara. Sentence Boundary Detection of Spontaneous Japanese using Statistical Language Model and Support Vector Machines. In *Proc. ICSLP*, 2006.

[9] 西光雅弘, 河原達也, 高梨克也. 隣接文節間の係り受け情報に着目した話し言葉のチャンキングの評価. 情報処理学会研究報告, 2006-SLP-61-4, pp. 19–24, 2006.

[10] 高梨克也, 丸山岳彦, 内元清貴, 井佐原均. 話し言葉の文境界 -csj コーパスにおける文境界の定義と半自動認定-. 言語処理学会第9回年次大会, pp. 521–524, 2003.

[11] 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝. 日本語節境界プログラム cbap の開発と評価. 自然言語処理, Vol. 11, No. 3, pp. 39–68, 2004.

[12] 下岡和也, 内元清貴, 河原達也, 井佐原均. 日本語話し言葉の係り受け解析と文境界推定の相互作用による高精度化. 自然言語処理, Vol. 12, No. 3, pp. 3–17, 2005.

[13] 内元清貴, 丸山岳彦, 高梨克也, 井佐原均. 『日本語話し言葉コーパス』における係り受け構造付与 (version 1.0). 『日本語話し言葉コーパス』マニュアル, 2004.

[14] 野村和弘, 河原達也, 堂下修司. F0パターンに基づく講義音声の文単位へのセグメンテーション. 電子情報通信学会技術研究報告, SP99-13, 1999.

[15] 前川喜久雄, 五十嵐陽介, 菊池英明, 米山聖子. 『日本語話し言葉コーパス』のイントネーションラベリング version 1.0. 『日本語話し言葉コーパス』マニュアル, 2004.