Vol.2011-SLP-87 No.4
2011/7/21

講演に対する読点の複数アノテーションに基づく自動挿入

秋田祐哉[†1]　河原達也[†1]

音声認識結果の可読性と有用性を高めるためには，句読点を自動的に挿入することが不可欠である．本稿では，単語・係り受け・ポーズの情報を素性とする条件付き確率場（Conditional Random Fields, CRF）に基づく読点の自動挿入について述べる．読点の挿入箇所は人により大きく異なるため，我々は複数のアノテータによる句読点ラベルを利用して，各アノテータの挿入傾向をモデル化した．そして，これらを投票と補間の枠組みにより組み合わせる．日本語話し言葉コーパス（CSJ）の講演を用いた評価実験では，個々の句読点モデルを組み合わせることで，それぞれのアノテータの読点と，全てのアノテータに共通する読点について高い挿入精度が得られることが示された．

Automatic Comma Insertion in Lecture Transcripts Based on Multiple Annotations

YUYA AKITA [†1] and TATSUYA KAWAHARA [†1]

To enhance readability and usability of speech recognition results, automatic punctuation is an essential process. In this paper, we address automatic comma prediction based on conditional random fields (CRF) using lexical, syntactic and pause information. Since there is large disagreement in comma insertion between humans, we model individual tendencies of punctuation using annotations given by multiple annotators, and combine these models by voting and interpolation frameworks. Experimental evaluations using lectures of the CSJ demonstrated that the combination of individual punctuation models achieves higher prediction accuracy for commas agreed by all annotators and those given by individual annotators.

†1 京都大学 学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University

1. はじめに

音声認識の研究対象は，講義[1] や講演・演説[2]，議会[3] など，さまざまな話し言葉音声に拡大してきている．このような話し言葉音声認識は音声翻訳や字幕付与，また音声の文書化への貢献が期待されているが，音声認識システムの出力には句読点が含まれないことが一つの問題となっている．長時間に及ぶ話し言葉音声では，読みやすい字幕や文書を効率的に実現するためには，音声認識結果を適切な単位に自動的に区切って句読点を付与する必要がある．また，音声認識の後段に行われる機械翻訳などの自然言語処理においても，句読点の付与されたテキストを入力として想定しているため，これらの単位は不可欠である．句読点の自動挿入は，人間の利用・機械の利用のいずれの場合でも重要な課題となっている．

音声の書き起こしに対する句読点の自動挿入については，主に句点（すなわち文境界推定）を対象として，放送ニュースや電話会話などのタスクで多くの研究が行われてきた．最大エントロピー法やサポートベクターマシン（SVM），条件付き確率場（CRF）などの機械学習の枠組みに基づき，韻律やポーズ・言語的情報を用いて挿入を行う手法が一般的である[4]．我々も，『日本語話し言葉コーパス』（CSJ）の講演を対象に，ポーズと言語的情報を素性とする SVM により推定を行う手法を提案している[5]．これに対して，読点やコンマの推定に関するこれまでの研究は限られている[6)–8)]．日本語の読点については，書き言葉の新聞記事が対象ではあるが，村田ら[8] により読点の用法の分類と分析，および最大エントロピー法に基づく自動推定が提案されており，評価実験で 0.76 の F 値を得ている．

これまでの研究では，正解として単一のアノテータにより付与された句読点を利用している．しかし，読点の挿入は句点よりも高頻度でかつ主観的であることから人により読点の挿入箇所は大きく異なり，単一の読点ラベルでは必ずしも信頼できないといえる．そこで，本研究では複数のアノテータにより付与された，異なる句読点ラベルを利用する．

本稿では，複数の句読点ラベルを用いた，講演音声の書き起こしへの句読点の自動挿入について述べる．まず，各アノテータによる講演の書き起こし中の句読点を比較し，アノテータ間の相違について分析を行う．これに基づき，CRF の枠組みに基づく自動挿入を検討する．具体的には，各アノテータの句読点挿入傾向を個別に CRF でモデル化し，これらを組み合わせることでより信頼できる挿入を目指す．本研究ではアノテータに共通の読点と個別の読点の挿入モデルをそれぞれ構築し，CSJ の講演音声において評価を行う．

Vol.2011-SLP-87 No.4
2011/7/21

2. コーパスとアノテーション

本研究では，『日本語話し言葉コーパス』（CSJ）[9] の講演音声の書き起こしに対して人手により句読点の挿入を行い，その傾向を分析した．対象としたのは CSJ で「コア」と呼ばれる 177 講演（学会講演 70・模擬講演 107，総単語数 365,305）である．CSJ には音声・書き起こしに加えてポーズや非流暢現象などのアノテーションが含まれているが，句読点は与えられていない．そのため，プロフェッショナルの速記者 3 名をアノテータとして，それぞれ独立に句読点の付与を行った．なお，CSJ では発話が忠実に書き起こされているが，これに対してフィラーや口語表現の整形をあらかじめ別の作業者により行った整形テキストに句読点の付与作業を行っている．また，句読点の付与に際してアノテータは音声を聴取せず，書き起こしのテキストのみ参照している．句読点の分析に先立って，書き起こしを自動解析して単語・文節への分割を行った．

日本語の文における典型的な読点の用法には，(1) 節の終端を明示，(2) “A, B, C” のように複数の要素を列挙，(3) 文内の係り受け構造（どの文節がどの文節に係っているか）の明確化，(4) 単語列が読みやすくなるよう分割，の 4 つが考えられる．このうち (1) 〜 (3) は英語のような他の言語でも共通する用法であるのに対して (4) は日本語に特有の用法である．(3) と (4) の用法は主観的であり様々な読点の入り方がみられるが，多くの読点を挿入するのは好まれず，したがって直後の文節を修飾する文節には読点が置かれない傾向がある．

3. 句読点の分析

本節では複数のアノテータにより付与された句読点の差異について調査する．また，言語的情報やポーズがどの程度読点と関連しているのかについても調べる．

3.1 アノテータ間の句読点の比較

まず，書き起こしに付与された句点と読点の数，およびアノテータ間の重複について比較する．表 1 に 3 名のアノテータ（A・B・C）ごとの読点と句点の総数を示す．句点の総数は 3 名のアノテータともほぼ同等であるのに対して，読点の数はアノテータによって顕著に異なる．最も少ないアノテータ C は，最も多いアノテータ A の 3 分の 2 程度である．図 1 は各アノテータにより付与された読点の重複の度合いを示しているが，3 名とも一致した読点は 15,027 箇所であり，これは A・B・C の各アノテータが付与した読点のそれぞれ 51%・64%，76%である．一方，A の読点の 20%（6,015 個），B の読点の 8%（1,855 個），C の読点の 7%（1,333 個）がそれぞれ単一のアノテータのみにより付与されている．多くの読点がアノテータにより異なる場所に付与されているが，これはたとえプロフェッショナルの速記者であっても，読点の数や位置は主観的影響を受けることを示している．なお，句点では 16,462 箇所（各アノテータが付与した句点の 97%）で 3 名の一致がみられており，句点にはアノテータによる揺れが少ないことが確認された．

表 1 各アノテータによる句読点の数

アノテータ	読点	句点
A	29,393	16,958
B	23,371	16,972
C	19,854	16,969

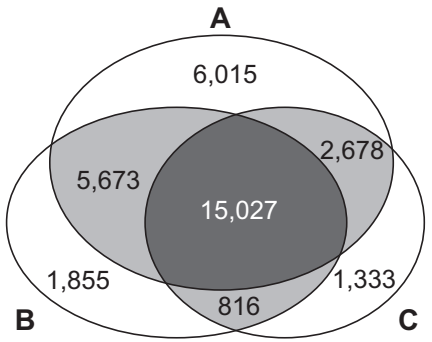

図 1 3 名のアノテータによる読点の重複の度合い

3.2 代表的な言語表現

文章を記述する際には，句読点（特に読点）を含む，人それぞれの表現法があると考えられる．読点の挿入における個人的な傾向を検証するため，ここでは読点とともに出現する言語的表現について調べる．

具体的にどのような箇所でアノテータ間に違いが生じるか調べるため，アノテータ 1 名のみにより付与された読点（すなわち図 1 における白地の部分 3 カ所）について，前後の単語・文節の比較を行った．この結果，特徴的な傾向としてアノテータ A では格助詞（「は」「が」等）の直後，たとえば「〜ということは」といった文節の直後への挿入が多くみられた．アノテータ B・C は接続詞の後に読点を挿入する傾向があったが，具体的な単語は異なり，

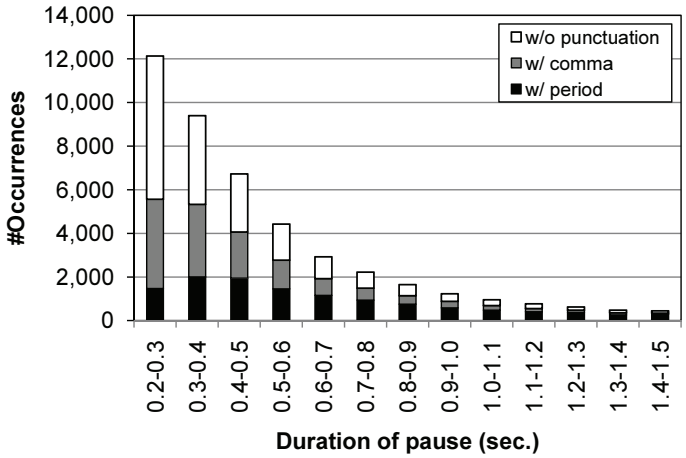

図 2 ポーズと句読点の相関

たとえばアノテータ B では「そして・つまり・それで・すなわち」など，アノテータ C では「あるいは，それから」や副詞「まず」などの後において多数の挿入が観測された．

3.3 ポーズとの相関

音声認識結果の入力を想定して句読点挿入（文境界推定）を行う際に，ポーズは一般的に用いられる特徴である．本研究のアノテータは句読点の付与の際に音声を聴取していないためポーズを手がかりとしては使用していないが，参考のためにポーズと句読点との対応を調査した．

図 2 はポーズ長を 0.1 秒ごとに区切って作成したポーズ長のヒストグラムで，ポーズ箇所のうち句点または読点が付与された内訳をあわせて示している．本研究では CSJ で人手により付与された転記基本単位の時刻情報をもとにポーズを抽出して利用したが，0.2 秒未満のポーズは CSJ ではアノテーションされていないため図 2 には含まれていない．なお，どのアノテータの句読点でも同様の分布を示したため，ここではアノテータ A の場合の統計を示している．

図 2 から，ポーズが長くなるほど句点に対応する割合が大きくなることがわかるが，読点の頻度は非常に小さい．一方，読点と関連したポーズの大半は長さが 0.5 秒以下である．この場合でも句点や読点に対応するポーズは 45%～53%にすぎないが，それでもこの比率（ポーズの存在する箇所に句読点が挿入される比率）は，ポーズなしの場合の比率と比べて明らかに大きいと考えられる．またこの比率はポーズの長さに関係なく同程度であった．つまり，ポーズの出現は読点の予測の手がかりとなりうるが，長さの情報はそうではないといえる．

4. 自動挿入手法

4.1 CRF によるモデル化

これらの分析を踏まえて，本研究では句読点挿入のための CRF に基づく識別器を構成した．CRF の実装には CRF++[⋆1] を利用する．識別に利用する特徴は単語（出現形），品詞（大分類），文節境界，直後の文節への係り受け情報[10] およびポーズである．ここで直後の文節への係り受け情報は，係り受け解析結果のうち隣接する文節に係るもののみを抽出したものである．このような係り受けがある場合は文節の結びつきが強く，読点が置かれにくいと考えられる．また，長い係り受け構造の推定は容易ではないが，このような直後の文節への係り受けは頑健に推定できると期待される．なお，形態素解析は ChaSen+IPADIC，文節境界の推定と直後の文節への係り受け推定は解析器 Cabocha によって自動的に行われている．ポーズの素性としては 0.2 秒以上のポーズの有無を抽出して利用する．図 2 で示したようにポーズ長と読点との間の相関は強くないため，ポーズ長は素性として使用しない．これらの特徴はそれぞれ前後 3 単語分まで識別器に入力される．

この挿入手法について，2 節で述べた CSJ の 177 講演における 10-fold の交差検定により評価を行った．以降で示す評価指標の値は，特に記述のない限り，この交差検定における平均値である．

4.2 素性の比較

まず，それぞれの素性の効果を測るために，素性のさまざまな組み合わせについてモデルを学習し評価を行った．ここでは，アノテータ 3 名中 2 名以上（多数決）により句点または読点と判断された箇所を正解とした．表 2 に，CRF の学習に用いた素性の組み合わせと，それぞれの場合の句点・読点の再現率・適合率・F 値を示す．表 2 より，句点は単語のみを素性としても高い精度が得られることがわかる．これは，人手により編集された書き起こしにて評価を行っているため，典型的な文末表現が容易に検出できることが理由と考えられる．一方，読点の予測ではどの素性も決定的ではなく，それぞれの素性が読点挿入における異なる要因を表現しているため，全ての素性が相乗的に性能を改善していくことがわかる．

⋆1 http://crfpp.sourceforge.net

表 2 句読点の自動挿入における素性の組み合わせの比較

使用した特徴	句点			読点		
	再現率	適合率	F 値	再現率	適合率	F 値
単語	0.972	0.969	0.971	0.611	0.729	0.665
単語+文節境界	0.975	0.974	0.975	0.647	0.764	0.700
単語+文節境界+直後係り受け	0.978	0.983	0.981	0.698	0.768	0.731
単語+品詞	0.974	0.973	0.974	0.624	0.764	0.687
単語+品詞+文節境界	0.976	0.973	0.975	0.679	0.768	0.721
単語+品詞+文節境界+直後係り受け	0.979	0.983	0.981	0.713	0.774	0.742
単語+品詞+文節境界+直後係り受け+ポーズ	0.975	0.984	0.980	0.734	0.784	0.758

4.3 アノテータ共通の読点の挿入

次に，CRF に基づく読点挿入手法を異なる種類の句読点ラベルに対して適用し，評価を行った．なお，以降の実験では読点のみを評価の対象とし，前節における全ての素性を使用する．

本研究では句読点ラベルとして 6 種類を用意した．まず，「アノテータ共通」の句読点ラベルとして，3 名のアノテータ間の共通性に基づき “3”・“2+”・“1+” の 3 種類のラベルを定めた．“3” ラベルは，3 名のアノテータが一致して付与した句読点のみをラベルとして用いるものである．同様に，“2+” ラベルは少なくとも 2 名により付与されたもの，“1+” ラベルは任意の 1 名以上により付与されたものである．これらのラベルは複数の人間の判断により選択されたものであるので，一般的であると考えられる．これに対して，それぞれのアノテータにより付与された句読点ラベルを「アノテータ個別」のラベル “A”・“B”・“C” として用いる．

共通の読点の挿入に際しては，“3”・“2+”・“1+” のアノテータ共通句読点ラベル 3 種類でそれぞれ CRF のモデルを直接的に学習し，それぞれの挿入を行う．さらに，アノテータ個別句読点ラベル “A”・“B”・“C” を用いて対応する個別モデルを学習し，これらの 3 つのモデルの挿入結果を基に投票を行う手法も導入する．投票の方法としては，どれか 1 つ以上のモデルが投票した場合に句読点を挿入する “Any”，2 つ以上のモデルの投票による “Majority”，全てのモデルの投票による “Consensus” の 3 種類を行う．これらは “1+”・“2+”・“3” のラベルから直接学習したモデルとそれぞれ比較可能である．換言すれば，直接モデルを学習した場合は投票がラベル作成の時点で行われているのに対して，個別モデルによる投票は挿入の段階で行われるといえる．

表 3 にモデルの直接学習と個別モデルの投票による読点挿入の結果を示す．評価ラベルが “3” の場合，すなわち全てのアノテータに共通の読点の場合，“3” モデルによる F 値は 0.620 であった．一方評価ラベルが “1+” の場合，すなわち読点の置かれうる全ての点を予測すべき場合は，“1+” モデルによる F 値は 0.822 であった．これらの結果は，読点の置かれうる点の予測が，共通の（必ず置かねばならない）読点の予測に対して比較的容易であることを示している．これらの結果と投票の結果を比較すると，“Consensus” 投票では “3” モデルよりも高い F 値が得られたのに対して，“Majority” 投票は “2+” モデルの場合とほぼ同等であり，“Any” では “1+” モデルに対して性能の改善が得られなかった．これらの結果から，異なるラベルを用いて独立に学習された複数のモデルの組み合わせは，ある基準に基づいて選択的に付与されている読点の挿入には有効であり，直接的な学習は任意に置かれうる読点をよくモデル化しているといえる．

表 3 アノテータ共通の句読点ラベルに対する挿入結果

（1）モデル直接学習

評価ラベル	1+	2+	3
学習ラベル	1+	2+	3
再現率	0.814	0.734	0.559
適合率	0.830	0.784	0.695
F 値	0.822	0.758	0.620

（2）A,B,C モデルによる投票

評価ラベル	1+	2+	3
学習ラベル 投票の種類	A,B,C Any	A,B,C Majority	A,B,C Consensus
再現率	0.774	0.729	0.633
適合率	0.849	0.786	0.652
F 値	0.810	0.756	0.642

1+: 1 名以上のアノテータにより付与された句読点
2+: 2 名以上のアノテータにより付与された句読点
3: 全てのアノテータにより付与された句読点
A/B/C: それぞれのアノテータにより付与された句読点

4.4 アノテータ個別の読点の挿入

次にアノテータ個別の読点のモデル化について検討する．個別の句読点ラベル “A”・“B”・“C” に対して，それぞれで学習された個別モデルにより挿入を行い評価を行った．これに加えて，表 3 で任意に置かれうる読点に対して高い再現率・適合率を実現した “1+” モデルについても評価を行った．さらに個別モデルと “1+” モデルの補間手法も導入する．CRF の枠組みでは全ての出力候補に対して確率が計算され，この確率に基づいて識別が行われ

表 4 アノテータ個別の読点ラベルに対する挿入結果

評価ラベル		A	B	C
個別モデル（A/B/C）のみ	再現率	0.772	0.712	0.617
	適合率	0.799	0.776	0.711
	F 値	0.785	0.743	0.661
1+モデルのみ	再現率	0.832	0.877	0.859
	適合率	0.758	0.635	0.529
	F 値	0.793	0.737	0.655
重み付き補間（A/B/C & 1+）	再現率	0.803	0.793	0.741
	適合率	0.786	0.725	0.644
	F 値	0.795	0.758	0.689

A・B・C および 1+については表 3 を参照のこと．

る（最大の確率を得た候補が結果として選択・出力される）．ここで，素性ベクトル X が入力された場合に，個別モデルと “1+” モデルが出力候補 C に対して与える確率をそれぞれ $P_{\mathrm{personal}}(C|X)$ および $P_{1+}(C|X)$ とすると，次式のようにこれらの確率を補間して最終的な識別に利用する．

$$P(C|X) = \lambda P_{\mathrm{personal}}(C|X) + (1-\lambda)P_{1+}(C|X). \quad (1)$$

本実験では，補間重み λ は事後的に 0.6 に設定した．

表 4 に，A・B・C の 3 名の個別の読点ラベルについて，対応する個別モデル，“1+” モデル，および補間手法を用いて挿入した結果を示す．これらのモデルの中で，補間手法が最も高い性能を実現した．個別モデルを強化する上で，他のアノテータの情報を組み合わせることが有用であるといえる．

4.5 音声認識結果における評価

最後に，これらの挿入モデルを講演音声の認識結果に適用し評価した．この実験では 177 講演中のうち，CSJ の音声認識テストセットに含まれている 8 講演のみを使用する．テストセットの総単語数は 17,925 で，単語誤り率は 17.1%であった．

表 5 に音声認識結果における挿入結果と，対応する人手の書き起こしにおける挿入結果を示す．使用したラベルは，学習・評価ともに “1+”・“2+”・“3” の 3 種類である．書き起こしにおける結果と比較して，音声認識結果では性能の大きな低下がみられた．この理由としてはまず音声認識誤りがあるが，2 節で述べた，書き起こし・認識結果に対して適用する整形処理も要因として挙げられる．ここでは音声認識結果における発話末の表現に対して単純な規則的変換を行って整形しているが，句読点の挿入のためにはこれでは不十分であるといえる．より高い句読点挿入精度を得るためには，この整形処理の改善[11] も必要である．

表 5 音声認識結果における読点挿入の結果

評価ラベル		1+	2+	3
人手による書き起こし	再現率	0.821	0.735	0.525
	適合率	0.827	0.775	0.715
	F 値	0.824	0.754	0.605
音声認識結果	再現率	0.601	0.493	0.315
	適合率	0.494	0.435	0.354
	F 値	0.542	0.462	0.334

1+・2+・3 については表 3 を参照のこと．

5. おわりに

本稿では，単語・ポーズ・係り受け情報を素性とする CRF に基づく，講演音声の書き起こしへの句読点自動挿入手法について述べた．まず，プロフェッショナルの速記者であっても読点の挿入傾向が人により異なることを確認した．そして，個別の読点挿入のモデルを作成して，これらを組み合わせる手法を検討した．複数のアノテータにより付与された異なる句読点ラベルを用いることで，それぞれのアノテータのための句読点挿入モデルが学習される．これらの個別モデルを組み合わせることで，アノテータに共通の基準，およびアノテータ個別の基準に基づく読点に対して挿入性能の改善を得ることができた．

謝辞：本研究は JST CREST 及び科学研究費補助金によって行われた．

参考文献

1) Glass, J., Hazen, T., Cyphers, S., Malioutov, I., Huynh, D. and Barzilay, R.: Recent Progress in the MIT Spoken Lecture Processing Project, *Proc. Interspeech*, pp.2553–2556 (2007).
2) Alberti, C., Bacchiani, M., Bezman, A., Chelba, C., Drofa, A., Liao, H., Moreno, P., Power, T., Sahuguet, A., Shugrina, M. and Siohan, O.: An Audio Indexing System for Election Video Material, *Proc. ICASSP*, pp.4873–4876 (2009).
3) Akita, Y., Mimura, M., Neubig, G. and T.Kawahara: Semi-automated Update of Automatic Transcription System for the Japanese National Congress, *Proc. Interspeech*, pp. 338–341 (2010).
4) Liu, Y., Shriberg, E., Stolcke, A., Peskin, B., Ang, J., Hillard, D., Ostendorf, M., Tomalin, M., Woodland, P. and Harper, M.: Structural Metadata Research in the EARS Program, *Proc. ICASSP*, Vol.5, pp.957–960 (2005).
5) Akita, Y., Saikou, M., Nanjo, H. and Kawahara, T.: Sentence Boundary Detection of Spontaneous Japanese Using Statistical Language Model and Support Vector Machines, *Proc.*

Vol.2011-SLP-87 No.4
2011/7/21

Interspeech, pp.1033–1036 (2006).
6) Batista, F., Caseiro, D., Mamede, N. and Trancoso, I.: Recovering Punctuation Marks for Automatic Speech Recognition, *Proc. Interspeech*, pp.2153–2156 (2007).
7) Favre, B., Hakkani-Tur, D. and Shriberg, E.: Syntactically-informed Models for Comma Prediction, *Proc. ICASSP*, pp.4697–4700 (2009).
8) 村田匡輝，大野誠寛，松原茂樹：読点の用法的分類に基づく自動読点挿入，情報処理学会研究報告，2010-SLP-81-8 (2010).
9) Furui, S., Maekawa, K. and Isahara, H.: Toward the Realization of Spontaneous Speech Recognition —Introduction of a Japanese Priority Program and Preliminary Results—, *Proc. ICSLP*, pp.518–521 (2000).
10) 西光雅弘，秋田祐哉，高梨克也，尾嶋憲治，河原達也：局所的な係り受けの情報を用いた話し言葉の節・文境界の推定，情報処理学会論文誌， Vol.50, No.2, pp.544–552 (2009).
11) Neubig, G., Akita, Y., Mori, S. and Kawahara, T.: Improved Statistical Models for SMT-based Speaking Style Transformation, *Proc. ICASSP*, pp.5206–5209 (2010).