形態素解析器を利用した講演書き起こしの文境界検出について

田島 幸恵† 難波 英嗣‡ 奥村 学††
† 東京工業大学大学院 総合理工学研究科 ‡ 広島市立大学 情報科学部 †† 東京工業大学 精密工学研究所

1 はじめに

音声情報と文字情報を比較すると、音声情報が一時的であるのに対して文字情報が半恒久的であることや、文字情報の方が二次利用がしやすいことが文字情報の利点として挙げられる。このことから、音声情報の文字化が求められていると考えられる。現在、音声情報を文字化するには、人手による書き起こしと音声認識器の利用の二つの方法がある。

音声情報は一般に話し言葉である。しかし、話し言葉には書き言葉とは異なる特徴があり二次利用が難しい。例えば、書き言葉には「文」という言語単位が存在しているが、話し言葉には「文」という単位が明確には存在しないという特徴がある。そのため、話し言葉をそのまま文字化しても文境界は明示されない。

自然言語処理の第一段階である形態素解析は、文が解析単位であり、文境界を持たない話し言葉に対して正確な結果を出すことができない。そこで、本論文ではこの問題を解決するため、文境界を自動検出する手法を提案する。

先行研究 [下岡 02] では、話者がとるポーズの長さと境界前後に現れる単語情報を用いた文境界の検出が行われていて、再現率 78.4%、精度 85.1%、F 値 0.816 の結果が得られている。本手法では、[国研 01][国研 02] を対象コーパスとして、ポーズ情報が得られない状況を考えテキスト情報のみを用いた検出を行う。[国研 01][国研 02] は年齢と性別をバランスさせた話者による講演の書き起こしテキストである。

2 提案手法

本節では文境界検出のための提案手法について述べる。提案手法は以下の二段階で構成されている。なお、本研究では文境界箇所は句点が挿入できる箇所であると定義している。

(1) 句点挿入候補箇所の特定

(2) 不適切な候補箇所の削除

2.1 句点挿入候補箇所の特定

句点挿入候補箇所の特定は以下の手順に従う。

1. 人手で句点を挿入した訓練コーパスの持つ、句点前後に現れる形態素が限られているという特徴を用いる。

   人手で句点を挿入したコーパスから句点の前後に現れる形態素を抽出し、その形態素の出現の前後が句点挿入候補箇所であるとする。これにより文末に現れる形態素「です」の後や文頭に現れる形態素「で」の前などが挿入候補箇所となる。

2. コスト最小法による形態素解析器 [松本 02] の持つ、正しい形態素の並びに対して低いコストを出力するという特徴を用いる。

   前段階で得られた挿入候補箇所に、句点を挿入した場合の出力コストと挿入しない場合の出力コストの比較を行う。挿入した場合の方がコストが下がるなら、その箇所へ句点を挿入することで形態素解析がより正しく行われたことになる。そこで挿入後の方が出力コストが低下する場合は、挿入箇所が文境界候補であるとする。

2.2 不適切な候補箇所の削除

2.2.1 文境界に現れやすい形態素を用いた削除

2.1 節で得られた候補箇所前後に現れる各形態素に対して訓練コーパスに対する式 (1) の値を求める。この値が大きい形態素は文境界に現れやすく、小さい形態素は文境界に現れにくい。値が小さい形態素の前後を選択している場合、誤った箇所である可能性が高いと考えられる。

$$\frac{\text{2.1 節で挿入した候補箇所でありかつ人手で挿入した箇所の数}}{\text{2.1 節で挿入した候補箇所数}} \tag{1}$$

そこで 2.1 節で句点を挿入した候補箇所について、句点前の形態素と後の形態素に対しての式 (1) の値の和を求め、この値が閾値より小さくなる箇所は不適切な候補箇所に当たるとして削除する。閾値は経験的に 0.75 としている。

2.2.2 文境界に現れにくい形態素を用いた削除

「文境界に現れやすい形態素を用いた削除」では、削除できない不適切な候補箇所がある。例えば「ですが」は文末に現れやすい形態素「です」と文頭に現れやすい形態素「が」の連続である。したがって、「です。が」のように間が文境界候補箇所と推定された場合、2.2.1 節の方法ではこの箇所の句点は削除されない。

このような候補箇所を削除するために以下の方法を用いる。

文末に現れやすい形態素と文頭に現れやすい形態素の組み合わせを作る。その組み合わせの中で人手で句点を挿入したコーパスで間に句点を持たず閾値回以上出現する組み合わせを、間に句点を持たない組み合わせとして選ぶ。前節までの手法で、その組み合わせの間が候補箇所になっている場合にその箇所を削除する。閾値の回数は経験的に 10 回とする。

3 実験

3.1 評価尺度

評価尺度として精度と再現率と F 値を以下のように定義する。

$$\text{精度} = \frac{\text{人手で挿入した句点と本手法で挿入した句点の一致箇所数}}{\text{本手法で挿入した句点数}}$$

Detecting sentence boudaries in speech transcription using a morphological analyzer

† Sachie TAJIMA (tajima@lr.pi.titech.ac.jp)
‡ Hidetsugu NANBA (nanba@its.hiroshima-cu.ac.jp)
†† Manabu OKUMURA (oku@pi.titech.ac.jp)
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology(†)
School of Information Sciences, Hiroshima City University(‡)
Precision and Intelligence Laboratory, Tokyo Institute of Technology(††)

$$\text{再現率} = \frac{\text{人手で挿入した句点と本手法で挿入した句点の一致箇所数}}{\text{人手で挿入した句点数}}$$

$$\text{F 値} = \frac{2*\text{精度}*\text{再現率}}{\text{精度}+\text{再現率}}$$

コーパスに人手で句点を挿入して得られた 3499 文を 5 分割し、5 分割交差検定により実験を行った。

3.2 候補箇所特定の方向性

2.1 節で述べた手法はテキストの先頭と最後のどちらからでも行える。そこで、両方でまず実験を行った。その結果、どちらの F 値も 0.8215 となったが、異なる候補箇所が存在した。2 通りの手法で得られた候補箇所の和集合と積集合をとると、積の F 値が 0.8227、和の F 値が 0.8218 になった。

このことから、先頭からと最後からの 2 通りの方向から句点の挿入を行って得られる 2 つの結果の積集合を用いるのが適当であると考える。以降この方法を用いる。

3.3 各手法の有効性評価

2 節で述べた個々の手法の有効性について確かめるため実験を行った。行った実験は以下の 4 つである。

(1) 特定手法のみ利用

(2) 特定手法と文境界に現れやすい形態素を用いた削除手法を利用

(3) 特定手法と文境界に現れにくい形態素を用いた削除を利用

(4) 全ての手法を利用

結果を表 1 に示す。

表 1: 実験結果

	(1)	(2)	(3)	(4)
再現率	0.8184	0.7813	0.8182	0.7724
精度	0.4602	0.8571	0.4719	0.8800
F 値	0.5889	0.8174	0.5982	0.8227

表 1 から (1) の特定手法の段階で再現率が 81%であることがわかる。削除手法を用いることで再現率は低下してしまうため、より高い再現率が望まれると言える。

(1) と (2) の結果から、(2) の削除手法によって再現率を 4%下げる代わりに精度を 40%上げることができた。(3) と (4) の結果からも同様のことがわかる。

(1) と (3) の結果から、(3) の削除手法によって再現率を 0.02%下げる代わりに精度を 1.2%上げることができた。(2) と (4) の結果からも同様のことがわかる。

各削除手法共、精度の上昇と比較すると再現率の低下は僅かであり、F 値は上昇している。このことから、2 手法は効果のある手法であると言える。

3.4 考察

本手法での検出の失敗例には「いたしましたでえーネットの」や「見えるんです。けれどもこれは」のようなものがある。

「いたしましたでえーネットの」では「た」と「で」の間の文境界が検出されていない。このように言い間違いや言い淀み、フィラーを含む場合に文境界が誤って検出されるか文境界が検出されないことがある。これは、これらの情報が形態素解析を行う際にノイズとして働いてしまい、誤った形態素解析がなされることが原因である。

「見えるんです。けれどもこれは」では「です」と「けれども」の間を文境界として検出してしまっている。これは、主に間に文境界を持たない形態素の組み合わせが希に間に文境界を持つ場合に、2.2.2 節の方法では文境界候補を削除できないことが原因である。

4 書き言葉コーパスへの適用

話し言葉に対して提案している本手法を書き言葉コーパスへ適用することは興味深いと考え、日経新聞 (2000 年版より 1000 文を抽出) と天声人語 (1985 年版より 5000 文を抽出) の、本来テキスト中に存在する句点を削除したテキストに 3 節で用いた方法と同様の 5 分割交差検定を行った。テキスト中に存在した句点を正解とした。その結果、日経新聞に対して再現率 14.8%、精度 87.0%、F 値 0.254、天声人語に対して再現率 44.1%、精度 87.0%、F 値 0.585 が得られた。日経新聞に対する結果と比較すると天声人語の結果の方が優れているが、共に効果的な結果は得られていない。

本手法は、文境界に頻出する形態素を用いて挿入箇所を選択している。書き言葉では、話し言葉で「で」「ます」などが頻出するように一定の形態素が文境界に偏って出現することは少なく、挿入箇所が選択できず、そのため効果的な結果が得られないと考えられる。このことから、書き言葉に対しては本手法は有効ではないといえる。

5 おわりに

本手法により再現率 77.2%、精度 88.0%、F 値 0.8227 で、文境界が与えられていない講演音声の書き起こしテキストを文に区切ることができた。コーパスが異なるため正確な比較はできないが、これは同様の実験をポーズ情報を用いて行っている [下岡 02] の結果である再現率 78.4%、精度 85.1%、F 値 0.816 を上回っており、本手法は有効な手法だと言える。

本手法を用いることで、用いない場合と比較すると講演音声を正確に形態素解析できるようになった。

本手法で得られる再現率は 77.2%であり 88.0%である精度と比較すると低い値になっている。そこで今後の課題は再現率の向上であると考える。

参考文献

[下岡 02] 下岡和也, 河原達也, 奥乃博. “講演の書き起こしに対する統計的手法を用いた文体の整形”. 情報処理学会 音声言語情報処理研究会, No.041-003. 2002.

[松本 02] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. “日本語形態素解析システム『茶筌』 version2.2.9 使用説明書”. 2002.

[国研 01] “『日本語話し言葉コーパス』モニター公開のご案内”. http://www.kokken.go.jp/public/monitor_kokai001.html. 2001.

[国研 02] “『日本語話し言葉コーパス』(モニター版 2002) 公開のご案内”. http://www.kokken.go.jp/public/monitor_kokai002.html. 2002.