4T-01 単語の共起知識を利用した講演文のテキストセグメンテーション

伊藤 山彦† 松本 賢司† 谷田 泰郎‡ 柏岡 秀紀† 浦谷 則好†

†(株)エイ・ティ・アール音声言語通信研究所 ‡(株)国際電気通信基礎研究所

1．はじめに

昨今の急速な情報通信技術の発達に伴い、大量の情報から速やかに要点を把握することの重要性が増している。テキストからの重要文の抽出は、従来から自動要約技術として研究されており、主に新聞記事や論説文など、書き言葉を対象とすることが多かった。これに対し、我々は講演文を対象とした自動要約技術の研究を進めている。

講演において、講演者はテーマに関する背景、経緯、具体的な事例等、複数の話題を組み合わせて主張したい結論を導き出す。要約文の生成には、話題の構成を反映させるのが望ましく、そのためには、話題の境界を判定する技術(テキストセグメンテーション)が必要である。本稿では、NHK番組「あすを読む」の書き起こし原稿を対象とし、テキスト中に出現する単語の分布を利用したテキストセグメンテーションの実験について述べる。また、同じ話題の中で同じ単語が繰り返し出現しない場合も、話題の境界の判定を可能とするために、新聞記事から抽出した単語の共起知識を利用する手法を提案する。

2．テキストセグメンテーションの課題

従来のテキストセグメンテーションの手法は、(1)接続詞や文末表現など、話題の転換となる手掛かり表現を利用する方法(文献[1])と、(2)テキスト中に出現する単語の分布から話題の範囲を判定する方法(文献[2])の2つに大別できる。(1)は、講演者の癖に依存する場合が多く、汎用性に欠けるため、本研究では(2)のアプローチを採用する。

文献[2]では、テキスト中の各位置(基準点)の前後に一定の大きさの窓を設け、2つ窓にどれくらい同じ単語が出現しているかにより、窓間の意味的関連性の強さ(結束度)を、単語の頻度ベクトルの余弦値によって測定する。テキストの先頭から末尾まで、基準点を一定の幅で移動し、結束度の値が極小となる位置を話題の境界と判定する(図1)。

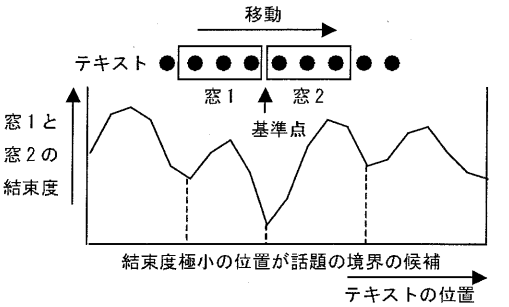

図1 単語の分布によるテキストセグメンテーション

「あすを読む」は10分間の番組であり、書き起こすと約3000文字程のテキストとなる。文献[2]は科学記事が実験の対象であるが、それに比べてテキストのサイズが小さく、話題の範囲も短い。そのため基準点の前後が同じ話題であっても、同一の単語が現れず、結束度が低いと判定される場合がある。

この問題に対し、予め共起性の高い単語の知識を持つことにより、同じ単語が現れる回数が少ない場合も、高い結束度の範囲を判定する手法を提案する。「あすを読む」は時事問題をテーマとすることが多く、新聞記事と単語の出現傾向が類似している。そのため、本研究では、日本経済新聞の記事から抽出した単語の相互情報量の値を用いる(図2)。

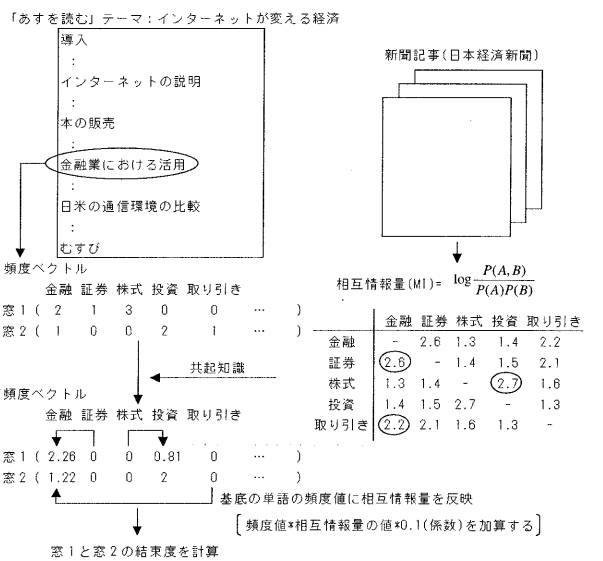

図2 共起知識の利用

3．共起知識を利用した結束度の補正

窓1中の単語の頻度値をf(i)、窓2中の単語の頻度値をg(j)とし、窓1と窓2の頻度ベクトルV(窓1)、V(窓2)を、図3のように表す。

Text Segmentation of Lecture Sentences Using the Knowledge of Word Cooccurrence

Takahiro ITO†, Kenji MATSUMOTO†, Yasuo TANIDA‡, Hideki KASHIOKA†, Noriyoshi URATANI†

†ATR Spoken Language Translation Research Laboratories.

‡ Advanced Telecommunications Research Institute International.

2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, JAPAN.

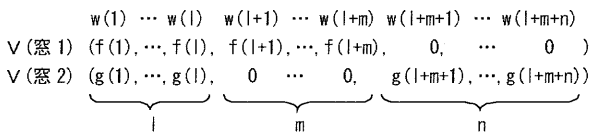

図3 頻度ベクトルV(A)とV(B)の関係

図3で、lは窓1と窓2に共通に現れる単語の数、mは窓1のみに現れる単語の数、nは窓2のみに現れる単語の数に対応する。ここでベクトルの基底となる単語の頻度に対して、次の処理を施す。

(1) g(i)の値が 0 に対応する単語(w(l+1)～w(l+m))に対して、g(j)の値が 0 以外に対応する単語(w(1)～w(l)または w(l+m+1)～w(l+m+n))との相互情報量が最大の単語を取り出す。相互情報量最大の単語が複数ある場合は、最も頻度値の高い(g(j)の値が大きい)単語を選択する。

(2) 上記(1)の処理によりw(p1)に対して、w(q1)が選択された場合、w(p1)と w(q1)の相互情報量の値MI(w(p1),w(q1))を用いて、次の式に従いV(窓1)の頻度値を補正する。

$$f(ql) \leftarrow f(ql) + \alpha * MI(w(pl), w(ql)) * f(pl) \quad (式1)$$

$$f(pl) \leftarrow 0$$

(αは設定可能な係数)

(3) V(窓2)に対しても、上記(1)、(2)と同様な処理により、頻度値を補正する。

(4) V(窓1)とV(窓2)の余弦値より結束度を求める。

4. 実験

以下の手順に従い実験を行った。

(1) 1995年1月～1999年12月の日本経済新聞の記事(約 160 万記事)と「あすを読む」の両方に現れる単語に対し、2つの単語が同一の記事に出現する確率から相互情報量を求める。このうち、共起回数が1000回以上、かつ相互情報量の値が1.0以上のデータ(8416対)を補正に利用した。

(2) 「あすを読む」の形態素解析結果から句読点や機能語を取り除き、内容語を抽出する。窓の大きさを50語とし、第3節の処理に従い、基準点を1単語(内容語)ずつ移動して結束度を求める。実測した相互情報量の最大値が6.67であることから、式1のαの値は0.1と設定した。

(3) 各基準点の結束度(C_n)を、次の式によって前後の結束度との平均値に置き換え、結束度の値の変化に対するスムージングを行う。

$$C_n = (C_{n-1} + C_n + C_{n+1})/3$$

(4) 結束度の値が極小となる点に対し、結束度の変動の大きさ(depth-score)を求める。基準点の結束度の値をC_{mp}、基準点に対し左側の結束度極大点の値をC_{lp}、右側の結束度極大点の値をC_{rp}としたとき、次の式によってdepth-scoreを求める。

$$d = (C_{lp} - C_{mp}) + (C_{rp} - C_{mp})$$

(5) depth-scoreが、次の式で求まる閾値以上である極小点を話題の境界と判定する。

$$h = \overline{C} - \sigma/2$$

($\overline{C}$は結束度の平均値、σは結束度の標準偏差)

表1に、11の書き起こし原稿に対して実験を行った結果を示す。人手による判定を正解とし、共起知識による補正のある／なしで、正解に対するもれの少なさ(再現率)とごみの少なさ(適合率)を比較した。表の±0 は、本処理で判定した話題の境界を含む文の前後が、人による判定と一致したときの値、±1は1文のずれを許容したときの値である。

表1 実験結果

	補正なし		補正あり	
	再現率	適合率	再現率	適合率
±0	0.36	0.29	0.33	0.34
±1	0.58	0.57	0.47	0.62

補正処理により、適合率は上昇したが再現率が低下した。これは、極小点付近の結束度が上がり、閾値以上のdepth-scoreを持つ極小点が減ったためである。顕著な改善が得られたとは言い難く、その理由として以下が考えられる。

(1) 元々話題の境界における単語の出現傾向の変化がない場合(例：裁判の一審・二審の判決の話題から、最高裁の判決の話題に移った場合)には、本手法の効果は現れない。

(2) 抽出した共起知識には複合語関係(例：金融と機関)や係り受け関係(例：金利と低下)が多い。近接して出現するため、窓間に跨る結束度測定に有効でなく、双方が共に特定の話題に特有の単語でない場合には、逆にノイズとして働く場合がある。

5. おわりに

本稿では、講演文を対象とし、新聞記事から抽出した相互情報量を利用して、テキストセグメンテーションを行う手法を提案した。今後、表層表現や音声情報等の手掛かりも利用して処理の精度を上げ、自動要約技術への適用を図る予定である。

参考文献

[1]望月ほか:複数の表層的手がかりを統合したテキストセグメンテーション, 自然言語処理, Vol.6, No.3, pp.43-58(1999).

[2]Hearst, M.A.:" Multi-paragraph segmentation of expository text ", Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics, pp.9-16(1994).