4Q-03

語の結束度と感情を考慮したオンライン小説の段落分割手法の提案

伊藤　志暢†　　松村　敦‡　　宇陀　則彦‡

†筑波大学情報学群知識情報・図書館学類　　‡筑波大学図書館情報メディア系

1　はじめに

現在、オンライン小説のコンテンツ数は増加しており、メディアミックスなど様々な方法で受容されている。オンライン小説の長さは多様であり、100 万文字を超える長さの小説も少なくない。しかし、このような長大な小説の中には章分けがなされていないものも多く、限られた時間で読むときにはどこまで読むべきかの判断が難しい。そこで本研究ではオンライン小説を複数回に分けて読むための適切な範囲に区切る 2 つの手法を提案する。

本研究では、段落分割の手法でテキストデータを区切る。テキストデータに対して意味のある位置で区切ることを段落分割と呼び、今回は読者が読書を読書を中断するのに適切な位置で区切ることを目的とする。

2　関連研究

段落分割は以前より、文書理解や情報検索、文書要約などを目的として多くの研究が行われてきた。中野ら [1] による語の近接性を文間結束度と定義し、それに基づいて段落境界を推定する手法や、但馬 [2] による、離散型隠れマルコフモデルを使った段落分割手法などが挙げられる。これらは新聞の社説や、ウェブのニュース記事といったテキストに対して、違う記事との境界を推定することを目的とした手法である。しかし、これらの手法では似たようなテキストが連続する小説の章分けに対しての段落分割には対応できない。そこで本稿では、小説の章という区切りに着目して分割する手法を提案する。

3　提案手法

本研究では語彙結束度、感情結束度をそれぞれ利用する 2 つの手法を提案する。

3.1　語彙結束度

中野ら [1] の文間結束度を発展させたものとして、ある話とその前までの話群との間の結束度として語彙結束度を提案する。語句の近接する関係が話と話の結びつきを示すという前提に基づいた結束度である。

語彙結束度の求め方を示す。話$_i$ に含まれるある語句を $W_j(j = 1 \sim n)$ とした場合、話$_i$ と 話$_{i-1}$ までの話群との語彙結束度 P_i は式 (1) で表される。

$$P_i = \frac{\sum_{j=1}^{n} \frac{1}{l_{ij}}}{n} \tag{1}$$

ここで l_{ij} は 話$_{i-1}$ までで最も近くに出現する語句 W_j と同一語句までの距離である。ただし、距離が大きくなればなるほど結束度に与える影響は小さくなるため、距離には上限（5000）を設けた。

3.2　感情結束度

同一の感情語句が近傍語句に出ていれば結束度が強いと考えられるため、感情語句の近接性に基づいた結束度を感情結束度と定義する。感情語句には、寺崎ら [3] が感情状態尺度に示した「抑鬱・不安」、「敵意」、「倦怠」、「活動的快」、「非活動的快」、「親和」、「集中」、「驚愕」の 8 つの尺度毎に挙げられた項目の語幹を採用する。さらに感情語句を増やすため、国立国語研究所の分類語彙表-増補改訂版データベース- [4] で、分類番号、段落番号が同一の語句を追加する。小説の 話$_i$ のある 感情$_j$ の割合を P_{ij} とすると 話$_i$ と 話$_{i-1}$ までの話群との感情結束度 E_i は式 (2) で表される。ここで、E_i は 話$_i$ の前の話までの影響を考慮するため、$k < i$ とする。

$$E_{ij} = \sum_{j=1}^{8} \sum_{k=1}^{i} -\frac{|P_{ij} - P_{kj}|}{i - k + 1} \tag{2}$$

3.3　境界判定基準

3.1 又は 3.2 で求めた結束度が大きく落ち込んだ箇所が区切りの開始として適切だと考えられる。中野ら [1] は極大値 S_i から右隣接する 2 つ目の極小値 $m_i + 1$ と結んだ線分に対して、極小値 m_i から下ろした垂線の長さを d_i とし、d_i の値によって大きく落ち込んだ箇所を判定している。

Text segmentation method for online novels using cohesion scores of words and emotional words

†Shinobu Ito

College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba

‡Atsushi Matsumura, Norihiko Uda

‡Faculty of Library, Information and Media Studies, University of Tsukuba

表 1: 語彙結束度:結果

	提案手法		ベースライン	
	再現率	精度	再現率	精度
平均	0.45	0.12	0.26	0.07
標準偏差	0.16	0.08	0.05	0.05

表 2: 感情結束度:結果

	提案手法		ベースライン	
	再現率	精度	再現率	精度
平均	0.33	0.10	0.23	0.07
標準偏差	0.13	0.09	0.08	0.05

本研究でも同様の方法で d_i を求め、垂線の長さが大きいものからリスト化し、大きいものから順に区切りとして採用する。そして、精度と再現率の調和平均が最大になるところを小説の区切りとする。

4 評価と考察

4.1 実験

小説家になろう (https://syosetu.com/) の各ジャンル（ファンタジー、SF、恋愛、文芸）の評価が高い上位 50 作品のうち、10 以上の章分けをされているオンライン小説を対象に実験を行った。実験対象作品は 42 小説となった。

本研究の手法の有効性を検証するため、ランダムに区切りを作った際の再現率、精度をベースラインとした評価を行った。さらに、感情の結束度による話の区切りについて、利用者実験を行った。実験参加者 18 人に対し小説の区切りを読んでもらい、読後の感情を 8 個の感情から選んでもらった上で、1〜5 の 5 段階で読後感を尋ねた。対象とした小説区切りは、感情によって区切られた小説区切りから感情語の属性が高いものを 8 個の感情につき 2 個ずつ計 16 個を選んだ。16 区切りのそれぞれに対して、計 121 件の回答を得た。

4.2 結果と考察

語彙結束度の比較結果を表 1、感情結束度の比較結果を表 2、利用者実験の結果を表 3 に示した。語彙結束度については、精度と再現率ともに t 検定でベースラインよりも有意な差があると認められた。また、感情結束度については再現率で、t 検定においてベースラインよりも有意な差があると認められた。利用者実験については、提案手法により出ていた感情が利用者が感じた感情と一致した割合は、ランダムと変わらない結果になった。しかし、利用者全員が選んだ感情と本手法で導出した感情が一致している区切りもあり、区切りによっては本手法が適切に感情を付与できる可能性を示した。また、読後感については、平均が 3.64 で比較的心地よい読書区切りになった。

表 3: 利用者実験:結果

	読後感	正解割合	ランダム
平均	3.64	0.36	0.38
最大	4.71	1.00	
最小	2.25	0.13	

5 おわりに

本研究では、語および感情の結束関係からオンライン小説の区切りを作成する 2 つの手法を提案し、評価を行った。今後の課題は、より適切な長さの区切りの作成とより実際の読書に近い区切り、適切な感情に依拠した区切りを提案することである。

参考文献

[1] 中野滋徳, 足立顕, 牧野武則. 語の近接性に基づいた意味段落境界の判定手法. 情報処理学会研究報告自然言語処理（NL）, Vol. 2005, No. 22, pp. 23–30, 2005.

[2] 但馬康宏. 言語モデルの違いによる HMM を用いたテキストセグメンテーションの性能比較. 情報処理学会論文誌数理モデル化と応用（TOM）, Vol. 6, No. 1, pp. 38–46, 2013.

[3] 寺崎正治, 岸本陽一, 古賀愛人. 多面的感情状態尺度の作成. 心理学研究, Vol. 62, No. 6, pp. 350–356, 1992.

[4] 国立国語研究所 (2004). 分類語彙表増補改訂版データベース』(ver.1.0).