語彙的連鎖とトピックモデルに基づくテキストセグメンテーション

山村　崇，　嶋田　和孝　（九州工業大学大学院）

背景・目的

■複数人対話における要約手法

◆対話文中には，様々な話題（トピックが頻出）
◆対話要約において，トピックのまとまり毎に
対話文を分割するテキストセグメンテーションが重要
⇒対話要約の前処理として分割することで，対話文中のトピックを適切に捉えたより良い対話要約が可能

■本研究の目的

◆語彙的連鎖とトピックモデルの情報を統合して
利用する分割手法の提案

先行研究

■代表的なセグメンテーション手法

◆**TextTiling**
◆テキストセグメンテーションの古典的な手法
◆「意味的に関連する部分には同一の語が繰り返し 出現する」
という性質を利用

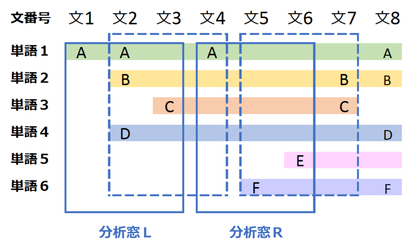
■TextTilingの改良手法

◆**Lexical Cohesion Segmentation（LCseg）**
◆**語彙的連鎖**に基づくTextTiling
◆**TopicTiling**
◆**トピックモデル**（LDA）に基づくTextTiling

提案手法

■LCseg

◆語彙的連鎖を用いて語彙的結束性を計算することで，
単語スパースネス問題を改善
◆**語彙的連鎖**
◆文書の意味的に関連する深い部分には同一の語が
繰り返し出現する性質を利用
◆**語彙的結束性の計算**
◆ある単語t_i の連鎖 R_i が2つの分析窓にオーバラップ
するかという情報を用いてコサイン類似度を計算



$$cosine_{LCseg}(L,R) = \frac{\sum_i w_{i,L} \cdot w_{i,R}}{\sqrt{\sum_i w_{i,L}^2 \sum_i w_{i,R}^2}}$$

$$w_{i,\Gamma} = \begin{cases} score(R_i) & if\ R_i\ overlaps\ \Gamma \in \{L,R\} \\ 0 & otherwise \end{cases}$$

■TopicTiling

◆トピックベクトルのコサイン類似度を計算することで，
単語スパースネス問題を改善
◆**トピックモデル**
◆潜在的な意味関係を数学的にモデル化
・Latent Dirichlet Allocation（LDA）
◆**トピックベクトルの類似度の計算**
◆LDAによって疎でないベクトル表現が得られる

$$cosine_{topic}(L,R) = \frac{\sum_n w_{i,L} \cdot w_{i,R}}{\sqrt{\sum_n w_{n,L}^2 \sum_n w_{n,R}^2}}$$

$w_{i,L}$と$w_{i,R}$は，左窓L，右窓Rにおける
トピックn の確率分布

■Merge（LCsegとTopicTilingの統合手法）

◆LCsegとTopicTilingでそれぞれ求めた文間の結束度を
sum_ratioの比率で足し合わせ，新たな文間の結束度を計算

$$\boldsymbol{cosine_{merge}(L,R) = sum\_ratio \times cosine_{LCseg}(L,R) + (1 - sum\_ratio) \times cosine_{topic}(L,R)}$$

→ $\boldsymbol{sum_ratio}$が１に近づくほどＬＣｓｅｇを重視，０に近づくほどＴｏｐｉｃＴｉｌｉｎｇを重視

■対話文の特徴を考慮したルールの追加

◆それぞれの分割手法に対話文の特徴を考慮したルールを適用
■R_1：相槌など短い発話に関するルール
→ 直前の発話と同一トピックと仮定
■R_2：話者系列に関するルール
→ 強く繰り返す話者の並びを同一トピックと仮定
■R_3：話者の切り替わりに関するルール
→ 発言者の切り替わりをトピックの切り替わりと捉える

評価実験

■実験データ

◆本研究室が開発した複数人対話コーパス（Kyutechコーパス）
◆4名１組によるグループディスカッション　：　9対話（開発用1対話）
◆各発話にアノテーションされたトピックタグの境界を，分割手法の
正解境界とし，評価実験

■2種類の評価基準

1. 分割位置の**完全一致 / 前後許容**のF値（精度と再現率の調和平均）
2. テキスト中の2つの文に対する誤分類（**2文評価**）

完全一致のF値

手法	（1）ルール追加なし	（1）＋R_1	（1）＋R_2	（1）＋R_3
LCseg	0.193	0.186	0.144	0.192
TopicTiling(10)	0.118	**0.124**	0.113	0.117
TopicTiling(20)	0.121	**0.127**	0.110	0.120
TopicTiling(30)	0.125	0.125	0.098	0.119
Merge(10,0.3)	0.126	**0.132**	0.115	**0.127**
Merge(10,0.7)	0.179	0.179	0.143	**0.180**
Merge(20,0.3)	0.149	**0.158**	0.121	0.149
Merge(20,0.7)	0.187	0.177	0.144	0.184
Merge(30,0.3)	0.152	0.152	0.108	0.141
Merge(30,0.7)	0.185	0.178	0.135	**0.188**

前後許容のF値

手法	（1）ルール追加なし	（1）＋R_1	（1）＋R_2	（1）＋R_3
LCseg	0.405	0.370	0.305	0.405
TopicTiling(10)	0.374	0.349	0.328	0.373
TopicTiling(20)	0.366	0.358	0.324	**0.371**
TopicTiling(30)	0.366	0.289	0.266	**0.328**
Merge(10,0.3)	0.363	0.341	0.313	**0.366**
Merge(10,0.7)	0.412	0.377	0.320	**0.414**
Merge(20,0.3)	0.374	0.364	0.295	**0.378**
Merge(20,0.7)	0.394	0.362	0.295	**0.398**
Merge(30,0.3)	0.368	0.333	0.276	0.353
Merge(30,0.7)	0.379	0.337	0.268	**0.392**

2文評価

手法	K = 2	k = 4	k = 6	k = 8	k = 16
LCseg	0.907	0.782	0.709	0.668	0.646
TopicTiling(10)	0.861	0.681	0.605	0.577	0.616
TopicTiling(20)	0.876	0.714	0.638	0.606	0.629
TopicTiling(30)	0.889	0.734	0.655	0.617	0.624
Merge(10,0.3)	0.871	0.698	0.619	0.591	0.629
Merge(10,0.5)	0.884	0.724	0.644	0.611	0.638
Merge(10,0.7)	0.903	0.771	0.700	0.663	**0.654**
Merge(20,0.3)	0.888	0.737	0.661	0.626	0.643
Merge(20,0.5)	0.902	0.763	0.685	0.645	0.648
Merge(20,0.7)	0.910	0.785	0.710	0.668	0.648
Merge(30,0.3)	0.898	0.756	0.679	0.641	0.645
Merge(30,0.5)	0.906	0.772	0.694	0.653	0.650
Merge(30,0.7)	**0.912**	**0.787**	**0.712**	**0.669**	0.644

■実験結果

◆今回の実験では，TopicTilingが有効に機能していなかったため
有効に統合できたとはいえない
◆提案手法は，LCsegと有意差は確認できなかったが，前後許容の結果や
2文評価では最も良い結果となった
◆対話文の特徴を考慮したルールを適用することで，一部精度が上昇した