5M-6

概念ベースを用いたニュース音声のトピックセグメンテーション

黒田 雅友 † 西崎 博光 ‡ 関口 芳廣 ‡
山梨大学大学院医学工学総合教育部 † 山梨大学大学院医学工学総合研究部 ‡

1. はじめに

近年，情報通新技術の発達に伴い，大量の電子化されたマルチメディアデータを誰もが容易に送受信できる環境が整っている．しかし，膨大な量のマルチメディアデータが存在するため，MPEG-7 等それらを整理するためのメタデータ規格化や付与技術が注目を集めている．一般的にマルチメディアデータに対して情報検索等で利用するためのメタデータを付与するには，マルチメディアデータがある程度の意味的なまとまりに分割されている必要がある．マルチメディアデータを意味的なまとまりに分割するトピックセグメンテーションを行う試みは世界中で研究され，米国 連邦標準技術局 (NIST) 主催の TRECVID プロジェクト等§のコンペティションが催されている．セグメンテーションを行うには、画像に含まれる様々な情報を利用することができるが，本稿では特に音声データに着目し，その音声認識結果に含まれる語彙情報を利用する．語彙情報を用いたセグメンテーションは，同一セグメントには関連性の高い語彙が出現しやすいという情報より，同一語彙や関連語彙の出現情報から算出された語彙的結束度を用いる．関連性のある語彙情報として，複数の国語辞書より作成された汎用知識ベースである概念ベース [1] を利用し，ニュース音声を対象にこれを用いたセグメンテーション法を提案，その評価を行なっている．

2. 概念ベースを用いたトピックセグメンテーション

2.1 概念ベース

概念ベースとは，ある概念語に対し意味的な特徴を表す属性と概念語に対するその属性の重要度を表す重みの対により表された知識ベースである．現在では，概念ベースの構築方式に関する研究がいくつか進められている [2]．本稿の手法では，まず，国語辞典¶から概念として見出し語を抽出し，見出し語の説明文中の単語を属性語とし，この出現頻度を元に重みを計算し概念ベースを自動構築した．さらに，属性語やその重みの信頼度を上げるために，精錬と呼ばれる処理を行っている [1].

表 1 に概念「資産」に対する概念ベースの例を示す．概念ベースの機能として，語彙に対し関連性のある語を連想し，属性が持つ重みを利用することにより，語間の意味的関係の強さを定量化することができる．トピックセグメンテーションにおいて，セグメント対象となるテキスト中の各語彙に，概念ベースの情報を対応付けて得られる属性群は，セグメント対象テキスト中の語彙の変遷をより精密に表すことができると考えられるので，この属性群の変化を利用することにより高精度なセグメンテーションが期待できる．

表 1: 概念ベースの例

概念	(属性，重み)
資産	(資産,0.12),(財産,0.11),(担保,0.08),(債務,0.07), (身代,0.05),(子孫,0.05),(金銭,0.05),(資本,0.04), (残す,0.04),(建物,0.03),(土地,0.03),(法律,0.03), (できる,0.03),(動産,0.02),(債権,0.02),(顧客,0.01),…

2.2 セグメンテーション処理

概念ベースを用いた語彙間の意味的な結束度算出の概要を図 1 に示す．

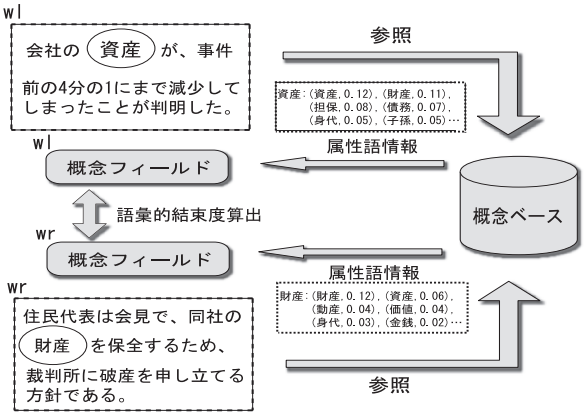

図 1: 概念ベースを用いた語彙的結束度の算出

本稿では，トピックセグメンテーションを行うため，音声認識結果の各発話境界における境界の前の単語列と後の単語列の語彙的結束度に基づいてトピック境界を抽出する．Hearst ら [3] により提案された，境界周辺の単語共起頻度に基づく単語列の語彙的結束度は，下記の式で求められる．

$$S(wl, wr) = \frac{\sum_t F(t_{wl})F(t_{wr})}{\sqrt{\sum_t F(t_{wl})^2 F(t_{wr})^2}} \quad (1)$$

ここで，wl と wr は，それぞれ左窓と右窓であり，$F(t_{wl})$ と $F(t_{wr})$ は，それぞれ，単語 t の左窓，右窓における出現頻度である．一方，概念ベースを用いた場合，単語列の語彙的結束度は下記の式で求めることができる．

$$S(wl, wr) = \frac{\sum_{tc} M(tc_{wl})M(tc_{wr})}{\sqrt{\sum_t F(t_{wl})^2 F(t_{wr})^2}} \quad (2)$$

ここで，単語 tc は，概念ベースに登録されている単語 t の属性語であり，$M(tc_{wl})$，$M(tc_{wr})$ は，単語 t の出現

Topic Segmentation for Broadcast News Speech Using Concept-Base

†M.Kuroda, Interdisciplinary Graduate School of Medicine and Engineering,University of Yamanashi

‡H.Nishizaki,Y.Sekiguchi,Interdisciplinary Graduate School of Medicine and Engineering,University of Yamanashi

§http://www-nlpir.nist.gov/projects/tredvid/

¶「学研国語大辞典」を利用した

頻度 $F(t)$ と単語 t が持つ属性語 tc の重みを乗算し，窓内に出現する属性語 tc の値の総和である．但し，類似度の微妙な揺れを無視するため，極小点 j の類似度 S_j と，左側の極大点 l における類似度 S_l，右側の極大点 r における類似度 S_r の差を考慮し，式 (3) により語彙的結束度の谷の深さ D を求める．D の値が大きいほど，語彙的結束度の極小点 j は話題境界の可能性が高いと考えることができる．

$$D(j) = (S_r - S_j) + (S_l - S_j) \quad (3)$$

Hearst 法によるトピックセグメンテーションは，単語境界の前後に一定の窓幅の窓をとり，左右の窓に含まれる単語の出現頻度を用いて語彙的結束度を計算するが，セグメント対象となるトピックの長さが異なり，それに従い最適な窓幅も変化することが予想される．しかし，予め任意の話題境界に対して最適な窓幅を設けることは難しいため，本稿では，複数の窓幅を用い，各窓幅を用いて算出した話題境界候補点を用いて多数決的に最終的な話題境界候補点を算出することにより，高精度なトピックセグメンテーションを目指している．

3. 評価実験

評価データとしてニュース音声を用い，概念ベースを用いたトピックセグメンテーションの評価を行った．

3.1 評価用音声データ

評価用音声データとして，NHK ニュース「ニュース 10」の 2005 年 6 月 22 日から 7 月 6 日放送分の 10 本 (10 日分) を用いた．評価データ全体での発話数は約 5600，総自立語数は約 38000 である．

3.2 音声認識

音声認識エンジンには，Julius ver.3.4.2 を用いた．音響モデルは，5 状態 3 ループ，性別依存 (男性)，対角共分散，16kHz サンプリング，25ms ハミング窓，フレーム周期 10ms，音素 triphone モデル (16 混合，総状態数 3000)，特徴ベクトル MFCC(12 次元)+ΔMFCC+ΔPOW(計 25 次元) である．言語モデルには，毎日新聞の記事データ 75ヶ月分 (1991 月 1 月〜1994 月 9 月，1995 年 1 月〜1997 年 6 月) より作成された 2 万単語の trigram を用いる．音声認識実験を行ったところ，単語正解率は 47.0%，単語正解精度は 37.7%であった．

3.3 トピックセグメンテーション実験

トピックセグメンテーションの実験において，語彙的結束度を算出するために用いる単語の品詞として，一般名詞，固有名詞，サ変名詞を用いた場合 (名詞組) と自立語全てを用いた場合の 2 通りを試みた．Hearst 法により境界周辺の同一単語の共起頻度に基づいた語彙的結束度の変遷を用いた場合と概念ベースに基づいた語彙的結束度の変遷を用いた場合の実験結果の比較も行った．窓幅には，評価データの 1 トピックの平均発話数の 5 分の 2 程度の長さ (25 発話) の窓幅をとった場合を表 2 に示し，窓幅に 15 発話から 50 発話までを 5 発話刻みで窓幅に用いて導出したトピック境界候補点の多数決をとった場合を表 3 に示す．

正解のトピック境界区間は，先行トピックの終端時刻と後続トピックの開始時刻の間の区間とする．実験結果には，導出された話題境界候補点が正解区間に含まれる場合 (No Margin) と正解区間から前後 3 秒までの区間を許容した場合 (3s Margin) の結果を示した．

表 2: トピックセグメンテーション結果 (一定窓幅)
(R.:Recall, P.:Precision, F.:F-measure)

Method	POS	No Margin			3s Margin		
		R.	P.	F.	R.	P.	F.
単語	名詞組	.155	.166	.160	.228	.243	.235
共起	自立語	.184	.199	.191	.225	.277	.265
概念	名詞組	.225	.225	.224	.290	.288	.288
ベース	自立語	.214	.216	.214	.259	.256	.256

表 3: トピックセグメンテーション結果 (複数窓の多数決)
(R.:Recall, P.:Precision, F.:F-measure)

Method	POS	No Margin			3s Margin		
		R.	P.	F.	R.	P.	F.
単語	名詞組	.171	.192	.181	.254	.289	.269
共起	自立語	.192	.209	.193	.255	.279	.264
概念	名詞組	.230	.288	.255	.324	.392	.354
ベース	自立語	.218	.297	.231	.269	.303	.282

表 2 の一定の窓幅を設定した実験結果において，概念ベースに基づく語彙的結束度の変遷を用いた場合と同一単語の共起頻度に基づいた語彙的結束度の変遷を用いた場合と比較すると，トピックセグメンテーションの性能は名詞組を用いた場合で 6.4%，自立語を用いた場合で 2.3%の精度の向上がみられた．自立語の場合と比較し名詞組を用いた場合の精度の改善率が大きくなっているのは，概念ベースに登録されている自立語の概念と属性の意味的繋がりの信頼度が，名詞組の語彙よりも低いためと考えられる．表 3 の多数決法を用いた実験結果でも同様に，概念ベースを用いた場合と同一単語の共起頻度を用いた場合と比較すると，名詞組を用いた場合 7.4%，自立語を用いた場合 3.8%の精度の向上がみられた．

4. まとめ

本稿では，概念ベースを利用したニュース音声のトピックセグメンテーション手法を示した．本手法により，同一単語の共起頻度情報に基づいた場合と比較してトピック境界検出精度の向上が確認できた．国語辞書より作成された基本概念ベースを用いたが，新聞記事データの単語共起情報を用いるなど，属性信頼度を向上させた概念ベースを用いることにより，さらなる精度の向上が望めるのではないかと考えられる．今後は，同一単語の共起頻度情報と概念ベース情報の併用や話者情報などの音響情報との組み合わせによりトピック境界検出精度の向上を図ることを考えている．

参考文献

[1] 渡辺他，国語辞書を利用した日常語の類似性判別，情処論文誌, Vol.38, No.7, pp.1272–1283, 1997.
[2] 小島他，連想システムのための概念ベース構成法，第 2 回情報科学技術フォーラム, E-049, pp.197–199, 2003.
[3] M.A.Hearst, TextTiling Segmenting Text into Multi-paragraph Subtopic Passages, Computational Linguixtics,Vol.23, pp.33–64, 1997.