3C9-4

サブワードユニット上の文字列カーネルを用いたタスク非依存なトピック分割

Domain-independent topic segmentation
using a string kernel on recognized sub-word sequences

佐土原健 Ken Sadohara　李時旭 Shi-wook Lee　児島宏明 Hiroaki Kojima

産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

The goal of the present paper is to explore the feasibility of a topic segmentation method without using Large Vocabulary Continuous Speech Recognition (LVCSR). The proposed method is domain-independent in the sense that it is not constrained by vocabulary and does not require training data. For a sequence of sub-word units obtained using a continuous sub-word recognizer, the proposed method merges similar adjacent parts of the sequence in an agglomerative manner to produce a hierarchical cluster tree. The proposed method uses a string kernel to efficiently compute the similarity between two strings of sub-word units based on the frequencies of any sub-strings appearing in the strings. By carefully excluding the influence of the sub-strings that are irrelevant to the topic of interest, topically coherent clusters are formed without linguistic knowledge. An empirical study on a Japanese news speech corpus shows that the method performs better than a topic segmenter using LVCSR.

1. はじめに

音声を含むマルチメディアコンテンツが大量に利用可能になった今日, コンテンツを構造化したり索引化することにより, 概要を素早く把握したり, 欲しい情報に素早く到達することを可能にするコンテンツの資源化技術が切実に求められている. そのための基礎技術として, 音声を意味的に等質な部分に分割する, トピック分割がこれまでに研究されてきた [Allan 02].

これまでの研究においては, 韻律的な情報等, 音声に特有の特徴を用いた研究が一部存在するものの, ほとんどの研究は, 大語彙連続音声認識システムを用いて音声をテキストに変換し, テキストのトピック分割問題に帰着させている. しかし, 会議音声等の自由発話の音声認識精度は未だ十分ではなく, タスクに合せて言語モデルを作り込まない限りトピック分割を可能にする水準のテキストを生成することは難しい. 特に, 未知語の問題は重要であり, 辞書に登録されていない単語の存在により認識精度が劣化してしまうだけでなく, トピックの形成に寄与する可能性の高い固有名詞等が辞書に登録されていない場合は, トピック分割性能に悪影響を与えてしまう. 実際, 本稿では, トピックに重要な単語をわずかに辞書から取り除くだけでも, トピック分割性能が大きく劣化してしまうことを示す.

本稿で提案するトピック分割手法は, 大語彙連続音声認識に依存しない. 音声を, 音素や音素片等のサブワードユニットの列として認識した上で, この記号列を直接トピック分割する. 音声を単語列に変換しないことにより, 言語的な知識をトピック分割に用いることができないが, 単語が誤認識される場合であっても単語が持っていた音韻的な特徴が残存する可能性が大きくなる. 単語列を意味的に等質な部分に分割する従来の研究においては, 単語の出現分布が重要な手掛りとなっているが, そうであるならばサブワードユニットの分布も重要な手掛りとなるはずであり, 言語知識を用いた非内容語の除去等を別の手段で代替した上で, サブワードユニットの分布に基づいたトピック分割を行なうことは可能であると考えられる.

このようなトピック分割法は, タスクに依存した言語知識を用いないだけでなく, トピックやトピック境界のモデルを事前に学習することも必要としない. 語彙やトピックに関する事前知識を要求しないこのようなトピック分割は, 日常的な小会議の会議録の構造化等に応用することが可能であると考えられる. もちろん, 単独で会議録の構造化を自動化することは困難であるが, 数時間の会議音声の中で, 意味的に等質な部分を視覚的に提示するだけでも, 会議の概要を把握したり, 人手による構造化を省力化することが期待できる.

本研究では, 語彙やトピックに制約されないこのようなトピック分割法が, 意味的に等質な部分を実際に同定しうるか否かについて, ニュース音声コーパスを用いて検討する. コーパスに含まれるニュース読み上げ記事をランダムに結合して得られる音声を入力として, この音声をトピック分割法により複数の部分に分割して, 元のニュース記事の境界をどの程度正確に復元できているかを評価する.

連絡先: 佐土原 健, E-mail:ken.sadohara@aist.go.jp

2. トピック分割アルゴリズム

本研究では, サブワードユニットとして, 音素と音素片 [Lee 05] を検討した. 音素片は, 調音結合を考慮し音素の遷移部分も独立に符号化した時間的に細分化された符号である.

入力音声中の任意の発声区間は, このようなサブワードユニットの列に変換されるが, この際, 有効な分析のために, 一つの列が少なくとも W 個の記号を含むように隣接する列と併合される. このようなサブワードユニット列を原子クラスターと呼ぶ. こうして得られた原子クラスタ列に対して, 隣接するクラスタ対の中で最も類似性の大きなクラスタ対をボトムアップに併合することを繰り返すことで, 意味的に等質な部分にまとめ上げる. 以下, ここで用いる類似性の計算法についてより詳細に説明する.

任意の原子クラスタ s_i に対して, ある表現ベクトルを対応させるが, その成分は, 長さ D 以下の任意の文字列 u に対する以下のような素性 $\phi(u, s_i)$ とする.

$$\phi(u, s_i) = \begin{cases} 0 & p(u) = 0 \text{ or } \frac{p(u|s_i)}{p(u)} \leq 1 \\ \log \frac{p(u|s_i)}{p(u)} & \frac{p(u|s_i)}{p(u)} > 1 \end{cases},$$

ここで, $p(u \mid s_i)$ は, u の s_i における生起確率であり, $p(u)$

は, 全ての原子クラスタにおける u の生起確率である. このような定義は, どこにでもほぼ等確率で生起する助動詞などの非内容語の影響を小さくする効果を持つ. ただし, これらの生起確率は, 実際には, 以下の式で近似する.

$$\hat{p}(u \mid s_i) = \frac{o(u, s_i)}{|s_i| - |u| + 1},\ \hat{p}(u) = \frac{\sum_{i=1}^{\ell} o(u, s_i)}{\sum_{i=1}^{\ell} (|s_i| - |u| + 1)},$$

ここで, $o(u, s)$ は, s における u の出現回数であり, $|s|$ は, s の文字数, ℓ は原子クラスタの数である.

任意の 2 つの原子クラスタ s_i と s_j に対して, その表現ベクトルの内積 $\boldsymbol{K}_{i,j}$ を計算し, 行列 $\boldsymbol{K}$ を得る. トライを用いた文字列カーネル [Leslie 03] を用いることで, $\boldsymbol{K}_{i,j}$ は, 陽に表現ベクトルを生成することなく, $p(u)$ の計算を別にすれば $O\left(D(|s_i| + |s_j|)\right)$ で計算可能である. こうして計算した内積 $\boldsymbol{K}_{i,j}$ に基づいて, 原子クラスタ間の類似性を表現ベクトルの余弦で定義するが, 内積の計算に先立って, 表現ベクトル全体をセンタリングすることが効果的である. センタリングは, どの原子クラスタでもほぼ同じ重みを持つサブワードユニット列の重みを零にする効果を持つからである. センタリング後の内積は, やはり表現ベクトルを陽に計算することなく, $\boldsymbol{K}$ を用いて以下のように計算可能である.

$$\tilde{\boldsymbol{K}} = \boldsymbol{K} - \boldsymbol{U}\boldsymbol{K} - \boldsymbol{K}\boldsymbol{U} + \boldsymbol{U}\boldsymbol{K}\boldsymbol{U}, \quad (\boldsymbol{U}_{i,j} = \frac{1}{\ell})$$

連続する複数の原子クラスタを単にクラスタと呼び, クラスタを構成する原子クラスタの表現ベクトルの和を, そのクラスタの表現ベクトルを定義する. また, 任意の 2 つのクラスタ $C_i = s_{i_1}, \ldots, s_{i_N}$, $C_j = s_{j_1}, \ldots, s_{j_M}$ の類似性もその表現ベクトルの余弦で定義し, $\tilde{\boldsymbol{K}}$ を用いて

$$\frac{\sum_{n=1}^{N}\sum_{m=1}^{M} \tilde{\boldsymbol{K}}_{i_n, j_m}}{\sqrt{\sum_{n=1}^{N}\sum_{m=1}^{N} \tilde{\boldsymbol{K}}_{i_n, i_m}}\sqrt{\sum_{n=1}^{M}\sum_{m=1}^{M} \tilde{\boldsymbol{K}}_{j_n, j_m}}}$$

のように計算できる.

このような類似性に基づいて, 最も類似性の大きい隣接するクラスタを順に併合することにより, 原子クラスタを意味的に等質なクラスタにまとめ上げて行く. この過程において, クラスタの表現ベクトルが陽に計算されることはなく, 原子クラスタの内積を格納した行列 $\boldsymbol{K}$ のみを用いてクラスタの類似性が計算される. 表現ベクトルの次元は非常に高次元であるので, これを陽に計算しないことは計算量の低減に寄与する.

3. 実験

上述したアルゴリズムの性能を RWCP ニュース音声コーパス [*1] を用いて評価した. この音声データは, ニュース放送用に作成された原稿を 6 人のプロのアナウンサーが読み上げて録音したものであり, 各話者は, 平均 41 本のニュース記事をおよそ 1 時間読み上げている.

各話者ごとに, ニュース記事をランダムに並びかえた後に連結して 6 個のデータセットを作った. 各データセットのトピックの境界は, ニュース記事の境界とした. このトピック境界を正解として, トピック分割アルゴリズムが出力するトピック境界の誤り確率を評価する. 誤り確率とは, w 秒離れた任意の 2 つの時点が誤って分割される確率のことで, w としては平均トピック長の半分が良いとされている. 表 1 は, 分割数を変化させながら誤り確率の最小値を測定し, その平均をトピック分割アルゴリズム毎に調べたものである.

	誤り確率
Baseline	0.164
Baseline(キーワードを除去)	0.201
提案手法 (音素)	0.149
提案手法 (音素片)	**0.121**

表 1: 実験結果.

表 1 の中で, 'Baseline' は, 大語彙連続音声認識システム Julius で音声をテキストに変換し, テキストのトピック分割アルゴリズム [内山 01] を用いてトピック分割を行った結果である. また, 'Baseline(キーワードを除去)' は, 各ニュース記事において TF-IDF 尺度に基づいて最も重要な単語を辞書から取り除いた場合の結果である. この結果から, わずかなキーワード (173 種類, 229 エントリ) を辞書 (60,250 エントリ) から取り除くだけで, 誤り確率が 0.037 ポイント増加することが分かる. また, 音素あるいは音素片の連続認識に基づく提案手法が, 大語彙連続音声認識とテキスト分割を用いたトピック分割法に比べて誤り確率が小さいことも分かる. なお, 音素片の場合 $D = 20$, $W = 200$ とし, 音素の場合, 1 音素が平均 2.2 SPS に相当するので $D = 9$, $W = 91$ とした. また, これらパラメータを妥当な範囲で変化させてもほぼ同様な結果が得られることを確認した.

4. おわりに

大語彙連続音声認識を用いることなく, 音声を音素や音素片等のサブワードユニットの列として認識した上で, この記号列を意味的に等質な部分に分割するタスクに依存しないトピック分割手法を提案した. また, ニュース音声コーパスから生成した人工的なデータを用いた実験により, 単語の分布に基づく手法と少なくとも同程度の性能を, 言語的な知識を用いないサブワードユニットの分布に基づく手法で実現可能であることを確認した. 今後は, 貪欲法でないアルゴリズムの開発や, 会議音声コーパスを用いた評価を行っていきたい.

参考文献

[Leslie 03] C.Leslie et al.: Fast kernels for inexact string matching, in *Proc. of COLT*, pp. 114–128 (2003)

[Allan 02] J.Allan, ed.: *Topic detection and tracking: event-based information organization*, Kluwer Academic Publishers (2002)

[Sadohara 06] K.Sadohara et al.: Domain-independent topic segmentation using a string kernel on recognized sub-word sequences, in *Proc. of SLT* (2006)

[Lee 05] S.Lee et al.: Combining multiple subword representations for open-vocabulary spoken document retrieval, in *Proc. of ICASSP*, pp. 505–508 (2005)

[内山 01] 内山 将夫 et al.: 統計的手法による分野非依存のテキスト分割, 自然言語処理, Vol. 8, No. 4, pp. 19–36 (2001)

*1 http://unit.aist.go.jp/itri/itri-spg/rwc-db.htm