語学番組検索システムにおけるシーン区切り検出手法

周　清楠[†]　渡辺　陽介[††]　勝山　裕[‡]　直井　聡[‡]　横田治夫[†,††]

[†]東京工業大学 工学部情報工学科　[††]東京工業大学 学術国際情報センター
[‡]株式会社　富士通研究所

1　はじめに

近年，NHK ゴガクル [1] やスペースアルク [2] など数多くの学習サイトが提供されるようになったが，学習したいフレーズの文字列や音声しか提供しておらず，実際の会話の雰囲気を把握しながら学習することができない．

語学力の向上や状況に合致した会話を身につけるために，テレビの語学番組を利用することは有用であるが，大量の語学番組の中から，自分が学習したいフレーズに関連ある会話が行われているシーンを探し出すには，非常に労力と時間がかかる．

これらの問題を解決するため，我々の研究グループは，テロップを用いて，ユーザーが入力したキーワードに関連する会話が行われているシーンを検索するシステムを開発している．テロップはクローズキャプションと違い，重要な場面や強調したい場面に出現するため，シーンの検索や区切りに有効であると考えられる．ただしテロップを抽出する際，誤認識や認識漏れなどの問題が存在するため，ニュース番組を対象にテロップ認識率向上手法が提案されている [3]．

語学番組検索システムを実現するには，論理的に繋がっている会話のシーンを見つけ出す必要がある．そこで本稿は，テロップの出現する時間差を利用したシーン区切りの検出手法を提案する．まずテロップ認識ツールを用いて語学番組からテロップを抽出し，Web 上の情報などを用いてノイズを除去する．次に，テロップの出現する時間差を利用し，論理的に繋がっているシーンの区切りを検出する．

2　提案手法

本稿では，NHK の英語番組を対象とし，(株) 富士通研究所により開発されたイメージ文字認識システム [4] を利用しテロップ認識を行う．処理の流れとしては，まず意味不明な文字列を除去し，そしてシーン区切りを検出する．

2.1　ノイズの除去

本稿におけるノイズとは，動画中テロップが出現していないが，背景の画像などを誤ってテロップとして認識し，認識結果に意味不明な文字列として出力されたものとする．例えば，「■口口■」，「いＩ」，「癖§ ソ」のような文字列がテロップ認識結果の約 6 割を占めている．そこで本稿はまず記号の除去を行い，次に短い文字列の除去を行う，最後に意味不明な文字列の除去を行う．

記号の除去　語学番組中，記号を含むテロップはとても少ない，なおかつ一つのテロップに複数個の記号が含まれることは極めて少ない．そこでテロップ中記号の割合が T_k 以上の場合はテロップを除去し，そうでない場合はテロップ中の記号を除去する．

短い文字列の除去　語学番組は，言語の正しい使い方を教えることを目的としているため，不完全なセンテンスや省略語などは少ない．そこでテロップの長さが T_l 以下の場合はテロップを除去し，そうでない場合は何もしない．

意味不明文字列の除去　意味不明かどうかの推測は機械とって困難であるため，本稿は YahooAPI[5] で取得したサーチエンジンのヒット数を用いて行う．まずすべてのテロップに N-gram を適用し，分割した文字列を空白で繋いで一つの問い合わせとする．そして，OR 条件で検索しヒット数を得る．ヒット数が T_m 未満の場合はテロップを除去し，そうでない場合は何もしない．

2.2　シーンの区切り検出

2.2.1　テロップ間の時間関係

実際の動画中，一つの画面に一つだけのテロップが出現するとは限らない，また同じ画面に出現するテロップの出現開始時刻と終了時刻が一緒とは限らない．テロップ間の時間関係はオーバーラップ (図 1) と非オーバーラップ (図 2) に分けることができる．

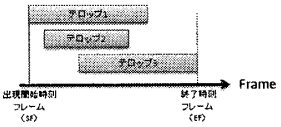

図 1: オーバーラップ

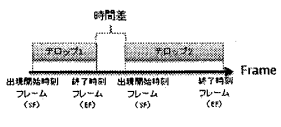

図 2: 非オーバーラップ

オーバーラップになる原因は二つ挙げられる．一つは，複数個のテロップが同時に表示されたことである．もう一つは，一つのテロップを出現時間の重なる別々のテロップとして認識したことである．これは認識ツールの仕様として，長時間にわたり出続けるテロップは，

A Scene Detection Method for Language Education Videos Retrieval System

Qingnan ZHOU[†], Yousuke WATANABE[††], Yutaka KATSUYAMA[‡], Satoshi NAOI[‡]and HaruoYOKOTA[†,††]
[†]Dept. of Computer Science, Faculty of Engineering, Tokyo Institute of Technology,
[‡]Fujitsu Laboratoies Ltd.
[††]Global Scientific Information and Computing Center, Tokyo Institute of Technology
[†]{seinan,watanabe}@de.cs.titech.ac.jp
[‡]{katsuyama,naoi.satoshi}@jp.fujitsu.com
[††]yokota@cs.titech.ac.jp

背景などに変化があると，別のテロップとして，再認識される．

非オーバーラップになる原因は二つ挙げられる．一つは，シーンとシーンの切れ目である．語学番組中のテロップは主に会話中の字幕やフレーズ解説中の例文などとして，テロップ情報が必要な場面に出現する場合が多い．そのため，シーンからシーンに切り替わる際，テロップは出現しないことが多い．もう一つは，論理的に繋がっているシーンの一部の認識漏れである．

2.2.2 検出処理の流れ

シーン区切り検出はテロップ間の時間関係を用いて行う．

まず，テロップ間の時間差を算出する．そして，オーバーラップか，それとも非オーバーラップかを判定する．

テロップ間がオーバーラップの関係の場合，必ず一つのシーンの一部であるため，連続したテロップをシーンとしてまとめる．

テロップ間が非オーバーラップの関係の場合，必ずしもシーンの区切りとは限らないため，テロップ間の時間差 D 及びしきい値 T_d を用いて，D が T_d 以上であった場合のみシーン区切りとみなす ($D \geq T_d$)．それ以外の場合一つのシーンとしてまとめる．

3 評価実験

3.1 実験の概要

本実験では，テロップ情報のみを用いてどれほどシーンを区切れるか検証する．また，ノイズ除去を適用した場合としない場合とで，シーン区切りに対する効果も検証する．

実験のデータは 2009 年 9 月～12 月に放送された NHK の 5 種類の英語番組について，各番組 3 個，計 15 個の動画を使用する．また，人間から見て論理的に繋がっている場面を一つのシーンとして，実験データから正解シーンを作成する．

実験の評価は $F-measure$ を用いて行う．また，正解シーンは人間が作成するため，必ずシーンの開始や終了にタイムラグが生じる．今回は許容範囲 T_c を用いて，システムが検出したシーンの開始時刻フレーム (SF) と正解シーンの開始時刻フレーム (SF') の差の絶対値が T_c 以下，なおかつシステムが検出したシーンの終了時刻フレーム (EF) と正解シーンの終了時刻フレーム (EF') の差の絶対値が T_c 以下の場合，検出したシーンは正しいと判定する ($|SF-SF'| \leq T_c \cap |EF-EF'| \leq T_c$)．

3.2 実験の結果

今回の実験では，$T_k = 0.3$，$T_l = 3$，$T_m = 1000000$，$N = 3$，$T_c = 240$，に固定した．T_d を 55，105，155，205，255 と変化させた場合に，システムが検出したシーンと正解シーンを比較し，提案手法を適用した場合の平均 $F-measure$ を算出した結果 (図 3) を示す．

縦軸は $F-measure$，横軸は T_d，青線の「NoiseFilter なし」はノイズ除去を適用前の結果，赤線の「NoiseFilter あり」はノイズ除去適用後の結果である．実験結果の考察については次節で述べる．

3.3 考察

「NoiseFilter あり」の方がはるかに良く，「NoiseFilter なし」との差は最大で約 4 倍である．その理由として，ノイズが本来のシーンとシーンの切れ目を繋いでしまったことが原因である．ノイズ除去することで，本来の切れ目の部分も明らかになり，適切にシーンを区切ることに成功した．しかし，今回はノイズ除去に必要なしきい値を固定して行ったため，どの値が最適のしきい値かについての調査は今後の課題となった．

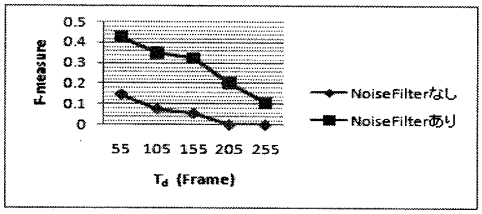

図 3: シーン検出手法適用の結果

T_d を 55 から 255 まで調整をした結果，「NoiseFilter」ありなし関係なく，$F-measure$ が下がる傾向が明らかになった．その理由として，今回の正解シーンは細かく区切ったことが挙げられる．そのため，T_d が小さいほど，本システムはシーンを細かく区切り，良い結果を得た．

この手法はテロップ情報のみ用いてシーン区切りをどこまで正しくできるかの手法であるため，テロップが出現する区間しか検出できない．しかし，実際の動画中，テロップ情報だけでは検出が不可能と思われるシーンは検出できないシーン中約 4 割ほど見られた．そのようなシーンを正しく検出するためには，テロップ情報だけでなく，他のどの情報を利用すればいいのかが今後の課題となった．

4 まとめと今後の課題

語学番組検索システムの第一歩として，ノイズの除去とシーン区切りの検出の提案，および実際の動画を用いた評価結果を報告した．本手法は語学番組シーン検索システム [6] で利用している．今後は，対象番組数を増やした実験やパラメータを変更した実験を行い，適切なパラメータに関する検討を行う予定である．

謝辞

本研究の一部は文部科学省科学研究費補助金特定領域研究 (#21013017) の助成により行われた．

参考文献

[1] NHK ゴガクル, http://gogakuru.com/index.html

[2] スペースアルク, http://www.alc.co.jp/

[3] ドウンゴフゥン，勝山裕，直井聡，横田治夫. Web サーチを活用した TV テロップ認識率向上手法. 信学技報, vol.108, no.93, DE2008-29, pp.163-168, Jun.2008.

[4] Y. Katsuyama, H. Bai, H. Takebe and K. Fujimoto. A study for caption character pattern extraction. IEICE Tech. Rep., vol. 107, no. 491, PRMU2007-239, pp. 143-148, Feb. 2008.

[5] YahooAPI, http://developer.yahoo.co.jp/

[6] 周清楠，渡辺陽介，勝山裕，直井聡，横田治夫. テロップと Web 情報を用いた語学番組シーン検索システム. DEIM, 2010.