2005－DBS－137（Ⅰ）（22）
2005／7／13

テキストストリームのオンライン・セグメンテーションとその応用

馬　　強[†]　　田　中　克　己[††,†]

本稿では，受信中の字幕データのようなテキストストリームのオンライン・セグメンテーション手法の提案と比較を行う．従来手法の多くは，トップダウン的なアプローチであり，データ全体のアクセスを前提としている．本稿では，受信中の字幕データの類似関係と語の共起関係を考慮して，同じストーリーや話題について述べている字幕をまとめていく，ボトムアップ的なセグメンテーション手法をいくつか提案し，その比較を行う．また，これらの手法を用いたアプリケーションの紹介も行う．

Online Segementation of Text Stream and Its Application

Qiang Ma ,[†]　and Katsumi Tanaka[††,†]

In this paper, we propose some ways of online segmentation of text stream. Conventional text stream segmentation methods of which most are top-down approaches needing to scan the whole data. In contrast, we propose the bottom-up methods which incrementally identify the story boundary and need not scan the whole data. We also show some comparative experimental results and the application to valid the proposed methods.

1. はじめに

我々は，放送と Web のコンテンツ融合について研究を行ってきた [3)~5)]．特に，番組の内容を補う Web ページを検索する手法 [3)] を提案した．字幕データ処理がこれらの手法の基本である．連続受信される，切れ目のない字幕データのセグメンテーション手法が，番組の索引付けおよびそれに基づく情報統合の精度・パフォーマンスに大きく影響を及ぼす [3)]．

本稿では，受信中の字幕データのセグメンテーション手法をいくつか提案し，その比較を行う．提案する手法は，隣接する字幕データの間の類似度と語の共起関係を考慮して，ボトムアップ的に同じストーリーや話題について述べている字幕データをまとめる．基本的に，同じストーリーや話題について述べている字幕データの類似度と語の共起関係が高いと考え，類似度と共起関係の値およびその変動幅に注目して，セグメントの切れ目を決める．

- 共起関係
  - 共起関係の値：連続受信した字幕データの語の共起関係が強ければ，これらの字幕データの同じストーリーについて述べている可能性が高い．
  - 共起関係の値の変動幅：連続受信した字幕データの語の共起関係を計算し，過去に計算された字幕データ間の共起関係と比較する．共起関係の変動幅が大きければ，この二つの字幕データが別のストーリーについて述べている可能性が高い．
- 類似関係:
  - 類似度の値：連続受信した字幕データの類似度を計算する．類似度が高ければ，同じストーリーについて述べている可能性が高い．
  - 類似度の変動幅：字幕データ間の類似度を計算して比較する．類似度の変動幅が大きいところは，セグメントの境である可能性が高い．

本稿では，提案手法の比較を行い，オンライン・セグメンテーション手法の応用システム，WebTelop[4)] の紹介を行う．

以下，本論文の構成を示す．2 節では，提案するオンライン・セグメンテーション手法について述べる．3 節では，提案手法の比較を行う．応用システムについては，4 節で述べる．5 節では，本論文のまとめと今後の研究課題について述べる．

† 独立行政法人　情報通信研究機構
†† 京都大学大学院 情報学研究科

<SYNC Start=001507390><P Class=JACC ID=Source0347>さて、</P></SYNC>
<SYNC Start=001508984><P Class=JACC ID=Source0348>卓球の愛ちゃん。</P></SYNC>
<SYNC Start=001512187><P Class=JACC ID=Source0349>福原愛選手、来年の春から</P></SYNC>
<SYNC Start=001513390><P Class=JACC ID=Source0350>卓球女子では世界最高峰の</P></SYNC>
<SYNC Start=001515796><P Class=JACC ID=Source0351>　</P></SYNC>
<SYNC Start=001518796><P Class=JACC ID=Source0352>中国のスーパーリーグに参戦することを前向きに検討していること</P></SYNC>
<SYNC Start=001523406><P Class=JACC ID=Source0353>がわかりました。アテネオリンピックには</P></SYNC>
<SYNC Start=001527812><P Class=JACC ID=Source0354>日本の卓球史上最年少の１５歳で出場した福原選手。</P></SYNC>

図 1　字幕データの例

2. オンライン・セグメンテーション

2.1 字幕データ

本研究では，受信した字幕データに対する，文（sentence）の識別を行わず，一回分の受信データをブロック (block) と呼び，セグメンテーション処理の最小単位とする．

本研究で想定している字幕データの例は，図 1 で示される．一回分の受信データは，タグ $< P >$ で囲まれている．字幕データの受信時間なども記録されている．つまり，番組の字幕データは，$(time, CC)$ のペアの系列である．TDT[1] や Informedia[2] で提案されているインクリメンタル・クラスタリング手法は，２つ以上の文を一つのブロックとしており，セグメンテーション処理の前に，文の識別作業およびクラスタ（セグメント）の数を設定する必要がある．例で示されているように，日本でリアルタイムに放送されている番組（ニュース 7 など）では，一回分の受信できるデータは，一文の断片である場合が多く，クラスタリングベースのセグメンテーション手法の適さない可能性が高い．

2.2 セグメンテーション手法

2.2.1 共起関係に基づくセグメンテーション

本研究では，隣接するブロック間の語の共起関係を調べ，共起関係の弱いキーワードが混じりだしたところで，ブロック（字幕データ）をまとめる操作を打ち切るという基本的な考えに基づいて，共起関係の計算対象および共起関係の弱いキーワードの混じりだす時間を判断する基準のバリエーションを考慮して，いくつかのセグメンテーション手法を提案する．

本研究では，あるテキストコレクション☆において，語 w_1 と w_2 が同時に出現されるテキストが多いほど，この二つの語の共起関係が強いと言う．語 w_i と w_j の共起度 $cooc(w_i, w_j)$ を次のように定義する．

$$cooc(w_i, w_j) := \frac{df(\{w_i, w_j\})}{df(\{w_i\}) + df(\{w_j\}) - df(\{w_i, w_j\})} \quad (1)$$

ただし，$df(\{w_i\})$ は，テキストコレクションにおける，語 w_i を含むテキストの数である．$df(\{w_i, w_j\})$ は語 w_i と w_j を同時に含むテキストの数である．

2.2.1.1 ブロック群におけるキーワードの共起関係に基づくセグメンテーション

受信された字幕データの中で共起関係の強いキーワードペアが多いほど，それらの字幕データは一つの話題を述べている可能性が高いのである．ノイズとなる字幕データの混じりだす時間の判定基準として，共起関係の強いキーワードペアの割合およびその割合の変動幅の二つがある（図 2）．

(C1) 割合の値に基づくセグメンテーション手法

共起関係の強いキーワードペアの割合に基づくセグメンテーションの手順を以下に示す．ここでは，K_i を時間 t_i におけるキーワード集合とする．ST と ET は，それぞれ抽出されるセグメントの開始・終了時間である．

(1) $K_0 := \emptyset, K := \emptyset, ST := 0, i := 1$ とする．

(2) 字幕データを受信する．データがなければ，終了する．

(3) 時点 t_i $(i \geq 1)$ でデータを受信したら，キーワード集合 K_i を受信された字幕データから抽出する．

(4) $K := K \cup K_i$ とする．

(5) K におけるすべてのキーワードペアの中に，共起関係の強いキーワードペアの割合 $cwf(t_i)$ を計算する．ここでは，共起関係の強いキーワードペアとは，共起度がある閾値 θ より大きい二つのキーワードのことである．m は K におけるキーワードの数である．

$$cwf(t_i) := \sum_{j=1, k=j+1}^{j=m-1, k=m} cr(w_j, w_k) / \frac{m \cdot (m-1)}{2}$$

$$cr(w_j, w_k) := \begin{cases} 1, & cooc(w_j, w_k) \geq \theta \\ 0, & cooc(w_j, w_k) < \theta \end{cases}$$

(6) $cwf(t_i) \geq \Theta$ であれば，8 へ．でなければ，次へ．ただし，Θ は予め定義された閾値である．

(7) $K := K_i, ET := t_i$ とする．ET に対するブロックをセグメントの境とする．

(8) $ST := t_i$ とする．

(9) $i := i + 1$．字幕データを受信する．これ以上のデータがなければ，終了する．でなければ，3 へ．

(C2) 割合の変動幅に基づくセグメンテーション手法

手法 C2 は，C1 の変形である．共起関係の強いキーワードペアの割合の値ではなく，前回の計算された割

☆ 本稿では，一定期間内のすべての話題に対応するすべてのテキストの集合とする．

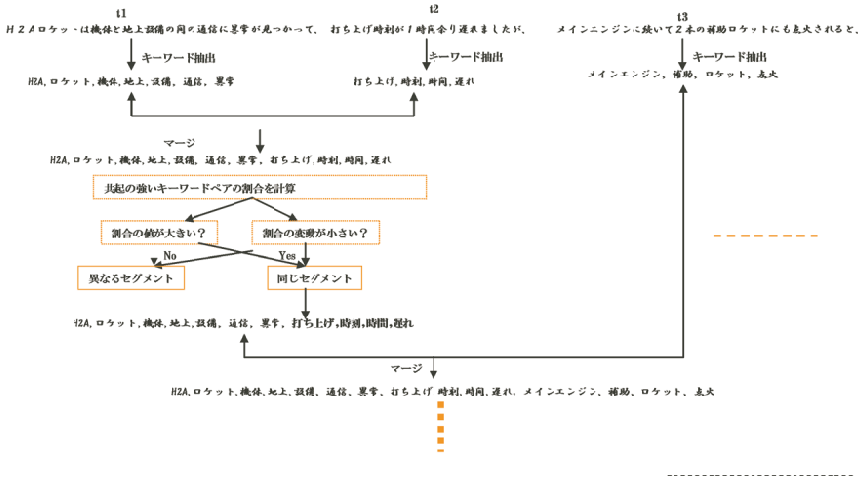

図 2　ブロック群におけるキーワードの共起関係に基づくセグメンテーション

合との差を計算して，この差分が閾値 δ より大きければ，新しいブロックをセグメントの境とする．

2.2.1.2　隣接ブロックにおけるキーワードの共起関係に基づくセグメンテーション

隣接する二つのブロックをマージして，その中にあるキーワード間の共起関係を調べる．共起関係の強いキーワードペアが多ければ，その二つのブロックは，同じ話題・ストーリーについて述べているとする．そうでなければ，その二つのブロックは，異なる話題・ストーリーについて述べている．**図 3** では，その処理の流れを示している．

(C3) 割合の値に基づくセグメンテーション手法

共起関係の強いキーワードペアの割合に基づくセグメンテーションの手順を以下に示す．ここでは，K_i を時間 t_i におけるキーワード集合とする．ST と ET は，それぞれ抽出されるセグメントの開始・終了時間である．

(1)　$K_0 := \emptyset, ST := 0, i := 1$ とする．

(2)　字幕データを受信する．データがなければ，終了する．

(3)　時点 t_i $(i \geq 1)$ でデータを受信したら，キーワード集合 K_i を受信された字幕データから抽出する．

(4)　もし $K_{i-1} = \emptyset$ であれば，8 へ．でなければ，$C := K_{i-1} \cup K_i$.

(5)　C におけるすべてのキーワードペアの中に，共起関係の強いキーワードペアの割合 $cwf(t_i)$ を計算する．m は C におけるキーワードの数である．

$$cwf(t_i) := \sum_{j=1,k=j+1}^{j=m-1,k=m} cr(w_j,w_k) / \frac{m\cdot(m-1)}{2}$$

$$cr(w_j,w_k) := \begin{cases} 1, & cooc(w_j,w_k) \geq \theta \\ 0, & cooc(w_j,w_k) < \theta \end{cases}$$

(6)　$C := \emptyset$．もし $cwf(t_i) \geq \Theta$ であれば，8 へ．でなければ，次へ．

(7)　$ET := t_i$ とする．ET に対するブロックをセグメントの境とする．

(8)　$ST := t_i$ とする．

(9)　$i := i + 1$．字幕データを受信する．これ以上のデータがなければ，終了する．そうでなければ，3 へ．

(C4) 割合の変動幅に基づくセグメンテーション手法

C4 が C3 の変形である．共起関係の強いキーワードペアの割合の値ではなく，前回の割合と比較して割合の差分を計算する．差分が閾値 δ より大きければ，新しいブロックをセグメントの境とする．

2.2.1.3　新しいブロックと直前のブロック群のキーワード共起関係に基づくセグメンテーション

同じセグメントであると判断された直前のブロック群と新しいブロックの間の語の共起関係を調べ，共起関係の強いキーワードペアの割合または割合の変動幅でセグメントの境を決める．図 4 では，処理の流れを示している．

(C5) 割合の値に基づくセグメンテーション

(1)　$K_0 := \emptyset, ST := 0, i := 1$ とする．

(2)　字幕データを受信する．データがなければ，終

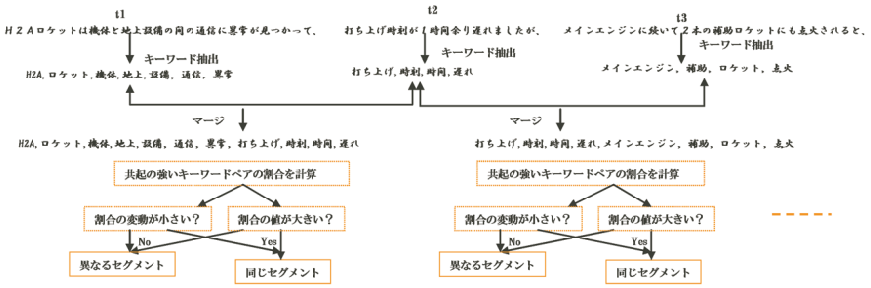

図 3　隣接ブロックにおけるキーワードの共起関係に基づくセグメンテーション

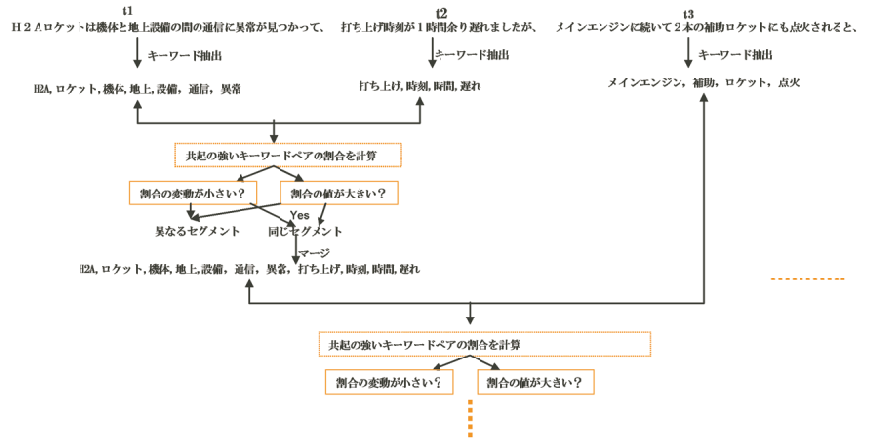

図 4　新しいブロックと過去のブロックのキーワード共起関係に基づくセグメンテーション

了する．

(3)　時点 t_i $(i \geq 1)$ でデータを受信したら，キーワード集合 K を受信された字幕データから抽出する．

(4)　$K_{i-1} = \emptyset$ であれば，$cwf(t_i) := 1$ とし，次へ．でなければ，任意の $w \in K_{i-1}, w' \in K$ のペア (w, w') の共起関係を調べ，共起の強いペアの割合 $cwf(t_i)$ を計算する．以下，m, n は，それぞれ，K_{i-1} と K のサイズとする．

$$cwf(t_i) := \sum_{j=1,k=1}^{j=m-1,k=n} cr(w_j, w'_k)/(m \cdot n)$$

$$cr(w_j, w'_k) := \begin{cases} 1, & cooc(w_j, w'_k) \geq \theta \\ 0, & cooc(w_j, w'_k) < \theta \end{cases}$$

(5)　$cwf(t_i) \geq \Theta$ であれば，$K_i := K_{i-1} \cup K$, 8 へ．でなければ，次へ．

(6)　$ET := t_i$ とする．ET に対するブロックをセグメントの境とする．

(7)　$K_i := K, ST := t_i$ とする．

(8)　$i := i+1$．字幕データを受信する．これ以上のデータがなければ，終了する．でなければ，3 へ．

(C6) 割合の変動幅に基づくセグメンテーション

C6 は，C5 の変形であり，割合の値の変わりに，割合の変動幅に基づいてセグメントの境を決める手法である．つまり，前回の値との差を求めて，その差分が閾値 δ より大きければ，このブロックをセグメントの境目とする．

2.2.1.4　隣接ブロック間の語の共起関係に基づくセグメンテーション

隣接する二つのブロックの間の語の共起関係を調べる．もしこの二つのブロックの語の共起関係が強ければ，この二つのブロックが同じストーリーまたは話題について述べているとする．もし語の共起関係が弱ければ，この二つのブロックは，別のストーリー・話題について述べている可能性が高い（図 5）．

共起関係の強さを判断するには，前述の手法と同じく，共起の強い語のペアの割合およびその割合の変動幅を用いることが可能である．

(C7) 割合の値に基づくセグメンテーション

(1)　$K_0 := \emptyset, ST := 0, i := 1$ とする．

(2)　字幕データを受信する．データがなければ，終

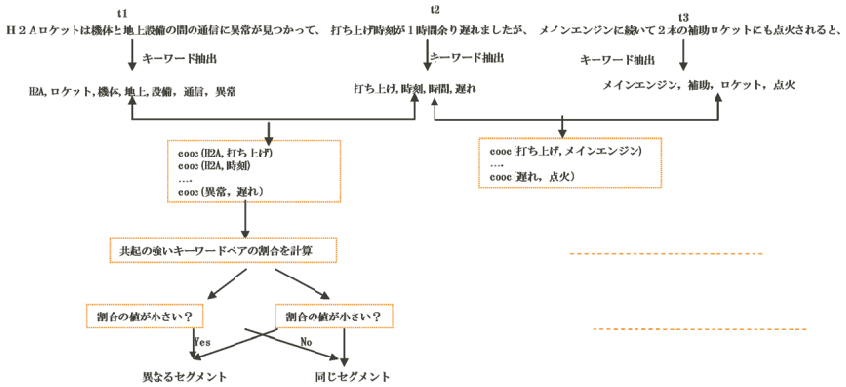

図 5　隣接ブロックの語の共起関係に基づくセグメンテーション

了する．

(3)　時点 t_i $(i \geq 1)$ でデータを受信したら，キーワード集合 K_i を受信された字幕データから抽出する．

(4)　$K_{i-1} = \emptyset$ であれば，$cwf(t_i) := 1$ とし，次へ．でなければ，任意の $w \in K_{i-1}, w' \in K_i$ のペア (w, w') の共起関係を調べ，共起の強いペアの割合 $cwf(t_i)$ を計算する．以下，m, n は，それぞれ，K_{i-1} と K_i のサイズとする．

$$cwf(t_i) := \sum_{j=1,k=1}^{j=m-1,k=n} cr(w_j, w'_k)/(m \cdot n)$$

$$cr(w_j, w'_k) := \begin{cases} 1, & cooc(w_j, w'_k) \geq \theta \\ 0, & cooc(w_j, w'_k) < \theta \end{cases}$$

(5)　$cwf(t_i) \geq \Theta$ であれば，8 へ．でなければ，次へ．

(6)　$ET := t_i$ とする．ET に対応するブロックをセグメントの境とする．

(7)　$ST := t_i$ とする．

(8)　$i := i + 1$．字幕データを受信する．これ以上のデータがなければ，終了する．でなければ，3 へ．

(C8) 割合の変動幅に基づくセグメンテーション

C7 の変形として，割合の値ではなく，割合の変動幅に基づいて境目を決める手法 C8 も考えられる．つまり，隣接する二つのブロックの間の共起関係の強い語のペアの割合を計算し，前回の値との差分を求める．この差分が閾値 δ より大きければ，このブロックはセグメントの境である．

2.2.2　類似度に基づくセグメンテーション

隣接ブロック間の類似度を計算してセグメンテーションを行うことも考えられる．同様に，単に隣接ブロックの類似度を比較する方法と，同じセグメントと判断されたブロック群と新しいブロックの類似度を比較する方法の 2 通りがある．さらに，類似度の値または類似度の値の変動幅に応じて切れ目を決めることが可能である．

2.2.2.1　新しいブロックと過去のブロック群の類似度に基づくセグメンテーション

(S1) 類似度の値に基づくセグメンテーション

同じセグメントであると判定されたブロック群と新しいブロックの類似度を計算する．この類似度が高ければ，新しいブロックも過去のブロック群と同じセグメントに属する．逆に，この類似度が小さければ，新しいブロックが，セグメントの切れ目である．(図 6 を参照)

(1)　$K_0 := \emptyset, ST := 0, i := 1$ とする．

(2)　字幕データを受信する．データがなければ，終了する．

(3)　時点 t_i $(i \geq 1)$ でデータを受信したら，キーワード集合 K を受信された字幕データから抽出する．

(4)　$K_{i-1} = \emptyset$ であれば，$sim(t_i) := 1$ とし，次へ．でなければ，K_{i-1} と K の類似度 $sim(t_i)$ を計算する．

(5)　$sim(t_i) \geq \Theta'$ であれば，$K_i := K_{i-1} \cup K$, 8 へ．ただし，Θ' は予め定義された閾値である．

(6)　$ET := t_i$ とする．ET に対応するブロックをセグメントの境とする．

(7)　$ST := t_i$ とする．

(8)　$i := i + 1$．字幕データを受信する．これ以上のデータがなければ，終了する．そうでなければ，3 へ．

(S2) 類似度の変動幅に基づくセグメンテーション

S2 は，S1 の変形である．同じセグメントと判定されたブロック群と新しいブロックの類似度を計算し，この類似度と前回計算された類似度との差分を求める．

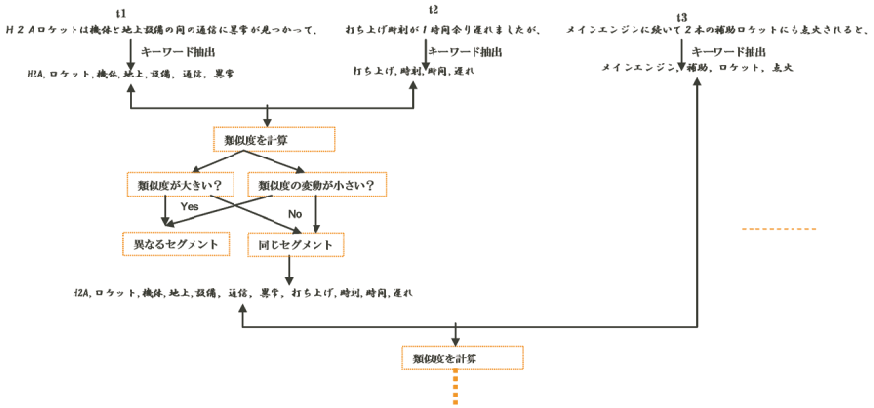

図 6 新しいブロックと過去のブロック群の類似度に基づくセグメンテーション

この差分が閾値 δ' より大きければ，新しいブロックをセグメントの切れ目とする（図 6）．

2.2.2.2 隣接ブロック間の類似度に基づくセグメンテーション

隣接する二つブロック間の類似度を計算する．この類似度または前回の類似度との差に基づいて，セグメントの境を決める（図 7）．

(S3) 類似度の値に基づくセグメンテーション

(1) $K_0 := \emptyset, ST := 0, i := 1$ とする．
(2) 字幕データを受信する．データがなければ，終了する．
(3) 時点 t_i $(i \geq 1)$ でデータを受信したら，キーワード集合 K_i を受信された字幕データから抽出する．
(4) $K_{i-1} = \emptyset$ であれば，$sim(t_i) := 1$ とし，次へ．でなければ，K_i と K_{i-1} の類似度 $sim(t_i)$ を計算する．
(5) $sim(t_i) \geq \Theta'$ であれば，8 へ．
(6) $ET := t_i$ とする．ET に対応するブロックをセグメントの境とする．
(7) $ST := t_i$ とする．
(8) $i := i + 1$．字幕データを受信する．これ以上のデータがなければ，終了する．そうでなければ，3 へ．

(S4) 類似度の変動幅に基づくセグメンテーション

S3 では，隣接する二つブロック間の類似度の値に基づいてセグメントの切れ目を決めている．それに対して，S4 では，この類似度の前回計算された類似度との差分を計算する．この差分が閾値 δ' より大きければ，新しいブロックはセグメントの切れ目である．

表 3 類似関係に基づくセグメンテーション手法の比較

	S1	S2	S3	S4
再現率	0.612	0.598	0.917	0.946
適合率	0.334	0.328	0.209	0.215
F 値	0.432	0.424	0.340	0.351
セグメント数	135	134	322	320
S 値	0.007	0.007	0.001	0.001
パラメター	$\theta' = 0.2$	$\delta' = 0.15$	$\theta' = 0.2$	$\delta' = 0.1$

3. セグメンテーション手法の比較

3.1 セグメンテーション手法の分類

本稿で提案するセグメンテーション手法は，基本的に，受信した字幕データの語の共起関係および類似関係に基づくものである．さらに，表 1 で示されているように，同じセグメントと判断されたブロック（群）のマージの有無，ブロックとブロック（群）の間の関係（類似，共起）の計算の有無によって，分類することができる．

3.2 比 較 実 験

提案手法の評価実験を行った．我々は，2002 年 9 月から 2004 年 12 月までの 28ヶ月間の NHK ニュース 7 の字幕データを利用して，共起度辞書を作成した．この共起度辞書を利用して，字幕データのセグメンテーションに必要となる語の共起度を調べた．共起度辞書に登録されてない語のペアの共起度は 0 とした．

本稿では，セグメンテーション手法の評価基準の一つとして，情報検索分野でよく利用されている F 値を用いる．正解の判断モデルは，図 8 に示されているように，Informedia で利用されているモデルを修正したものを利用する．システムが判定したセグメントの境（identified boundary）と，人間が判定したセグメントの境 (reference boundary) との距離（ブロッ

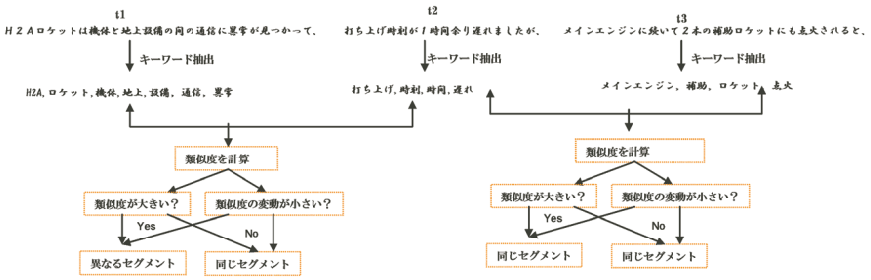

図 7 隣接ブロック間の類似度に基づくセグメンテーション

表 1 セグメンテーション手法の分類:表では,「マージ」が同じセグメントと判断されたブロックのマージを意味する.「比較」は，ブロックとブロック（群）の間の関係（類似または共起）の計算を意味する．”○”と”×”は，有と無をそれぞれ意味する．ただし,「値・差分」の”○”と”×”は，それぞれ，値と差分を意味する．

	C1	C2	C3	C4	C5	C6	C7	C8	S1	S2	S3	S4
マージ	○	○	○	○	○	○	×	×	○	○	×	×
比較	×	×	×	×	○	○	○	○	○	○	○	○
値・差分	○	×	○	×	○	×	○	×	○	×	○	×

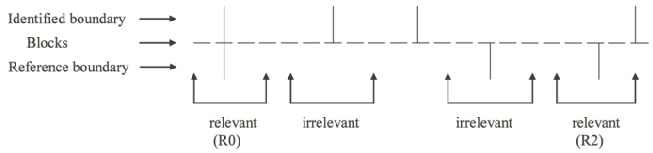

図 8 テキストストリームのセグメンテーションの判定モデル：Rx は，reference boundary と identified boundary の距離（ブロック数）が x 以下であれば，identified boundary を relevant と判定するのを意味する．

表 2 共起関係に基づくセグメンテーション手法の比較

	C1	C2	C3	C4	C5	C6	C7	C8
再現率	0.426	0.464	0.652	0.698	0.531	0.495	0.562	0.809
適合率	0.330	0.357	0.296	0.258	0.456	0.271	0.348	0.220
F 値	0.372	0.403	0.407	0.376	0.491	0.350	0.430	0.346
セグメント数	63	95	126	199	85	133	119	269
S 値	0.045	0.018	0.008	0.003	0.021	0.006	0.009	0.002
θ	0.2	0.2	0.1	0.2	0.15	0.15	0.2	0.15
Θ	0.25	-	0.2	-	0.1	-	0.1	-
δ	-	0.03	-	0.1	-	0.05	-	0.05

クの数）が 1 以内であれば，システムが正しく判定したと言う．ただし，この判定モデルを用いると，適合率，再現率および F 値が，セグメントの数に依存することになるので，本稿では，F 値とセグメント数を考慮した評価値 (S) を導入する．

$$S := \frac{F}{|seg_{ide} - seg_{ref}| + 1} \qquad (2)$$

つまり，F 値が大きく，提案手法と人間が判定したセグメント数の差が小さいほど，提案手法がよい．

評価実験では，二日間の字幕データ（NHK ニュース 7）を用いた．821 ブロックがあった．人間によって判定されたセグメントは 73 個あった．実験結果は，表 2 と表 3 に示されている．表で示されているパラメータは，一番いい結果（F 値）を得られた場合の値である．共起関係に基づく手法 C1,C5 と C2 がよい結果を得られたことがわかる．また，Informedia のインクリメンタル・クラスタリング手法の最善の F 値 $(0.367)^{2)}$ より，われわれの提案手法の方がいくつか改善されていることがわかる．

短い文章の類似度の計算がうまくいかない場合が多いため，共起関係ベースの手法は，類似度ベースの手法よりよい結果が得られている．値の変動幅より，値を用いた手法の方がよい結果が得られた．複数のブロック（> 2）を用いた手法が隣接ブロックのみを用

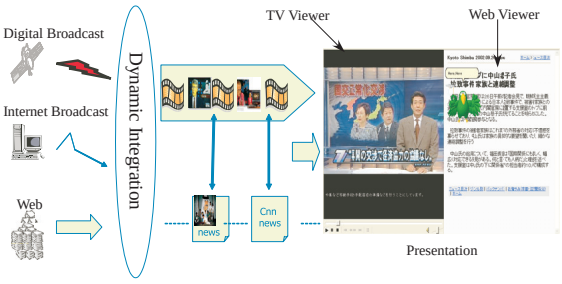

図 9 WebTelop の概念図

いた手法よりいい結果が得られた．これらのことから，できるだけ大きいサイズのブロック（長い文章）を用いることがよいと推測でき，類似度ベースのセグメンテーション手法（クラスタリング含む）がオンライン処理に適してないことが分かる．

4. 放送と Web の動的統合システム

放送と通信のインフラの融合に伴い，新しいタイプのコンテンツの生成・アクセス方法が求められる．我々は，話題構造に基づく補完情報の検索手法を利用して，放送コンテンツと Web ページの動的統合システム WebTelop を提案した [4]．WebTelop は，番組コンテンツのメタデータ（字幕）を利用して，リアルタイムにセグメンテーションを行い，番組の話題構造を抽出して，それに基づいて番組の内容を補う Web ページを検索して呈示する．

5. ま と め

我々は，受信データの類似および語の共起関係の解析に基づく，字幕データのオンライン・セグメンテーション手法をいくつか提案し，比較を行った．実験結果と応用システムの紹介を通して，我々の提案手法の検証を行った．今後，これらの手法のさらなる検証，特に，Informedia プロジェクトのインクリメンタル・クラスタリング手法との比較を行う予定である．また，これらの手法の我々の補完情報検索手法への影響を調べる予定である．

参 考 文 献

1) Allan J.,Carbonell J.,Doddington G., Yamron J. and Yang Y. Topic Detection and Tracking Pilot Study Final Report, *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, pp.194-218, 1998.
2) Huauptmann A., Chang J.C., Hu N.N., and Wang Z.R. Text Segmentation in the Informedia Project, *http://www-2.cs.cmu.edu/hnn/project/ML-project/ml-report.htm*.
3) Ma Q. and Tanaka K. 話題構造に基づく放送と Web コンテンツの統合のための検索機構，**情報処理学会論文誌：データベース**, *Vol. 45, No. SIG 10 (TOD23)*, pp. 18-36, 2004.
4) Ma Q. and Tanaka K.: WebTelop: Dynamic TV-content Augmentation by Using Web Pages, *Proc. of ICME2003 Vol.2*, pp.173–176 (2003).
5) Ma Q. and Tanaka K. Topic-structure-based complementary information retrieval for information augmentation, *Proc. of APWeb2004, LNCS3007*, pp. 608-619 (2004).