2007－NL－182 (11)
2007／11／20

ニュース要約のための簡易文脈解析

田中英輝 木下明徳 後藤功雄 熊野正 加藤直人
NHK 放送技術研究所
tanaka.h-ja@nhk.or.jp

あらまし

本稿では放送ニュースへの「リード」,「本記」,「追記」のタグ付け作業の試み，およびこのデータを使ったタグの予測実験結果を報告する．タグ付け作業の結果によれば，これらのタグからなるニュースの 3 段構造は普遍的な構造であることがわかった．また，重要文抽出と類似の設定で行った C4.5 を用いたタグの予測実験の結果，最小の誤り率 8.325%を得た．ここでは，位置情報，文間類似度，手がかり表現を用いており，効果が高かったのは位置情報と文間類似度であり，統計的に抽出した手がかり表現の効果は限定的であった．

Simple Discourse Analysis for Japanese Broadcast News Summarization

Hideki Tanaka, Akinori Kinoshita, Isao Goto, Tadashi Kumano and Naoto Kato
Science and Technical Research Laboratories of NHK

Abstract

We describe our work of broadcast news tagging with the discourse tags of lead, body and supplemental information. In the tagged data, we found that the three-fold discourse structure represented by the tags were highly common. We then report our tag prediction experiments with the C4.5 decision tree algorithm in which we obtained the minimum error rate of 8.235%. The sentence position and inter-sentential similarity found to be effective for the prediction but the statistically derived clue expressions of each tag showed quite limited effect on the prediction.

1 はじめに

著者らはニュースの談話的な構造を利用した要約システムを研究している．この構造とは，ニュースの「リード」「本記」「追記」の 3 要素からなる構造（井上 1981）を指しており，本研究ではニュースの「3 段構造」と呼ぶ．

これを利用した著者らの要約は，まず上記の 3 要素を認定した上で，1）リード文や追記を抽出して基本的な要約とする，2）リード中の抽象的な表現を本記の具体的な表現で置換する，という 2 段階を踏む（田中他 2005）．

この 2 段階要約は，ニュースの 3 段構造が普遍的で，さらに自動的に認定できることが前提である．しかし普遍性は明らかでなく，予測も自明な問題ではない．そこで著者らは一定量のニュース記事に「リード」「本記」「追記」のタグ付与し，調査を行うとともに，その予測実験を行った．以下，2 節ではタグ付け作業と結果の分析を報告し，3 節でタグの予測実験を報告する．さらに 4 節で考察を行い 5 節でまとめを行う．

2 ニュース記事への構造タグ付与

2.1 ニュースの構造

ニュースは先に述べた 3 つの要素からなると言われている（井上 81）．本稿ではこれらの 3 要素を構造タグと呼ぶ．構造タグは，それぞれ次のような特徴を持つ．

リード

多くの場合，冒頭の文である．ニュースの最も重要な内容を簡潔に記述する．ここでは，具体名を使わず抽象的な表現が使われることがある．たとえば社名の代わりに「大手生命保険会社」が使われるといった具合である．

本記

リードの内容を詳述する．いわゆる 5W1H に相当する情報を補足する．リードで抽象的に述べられた内容もここで具体的に記述する．

追記

リードや本記で述べられていない情報を必要に応じて追加する．実際の原稿では今後の動きや関連情報などが記述されている．

以上のタグからなる 3 段構造がニュース記事にどれだけ反映されているかは，これまで調査されていない．そこで著者らは次に述べるように，タグ付け作業を行って調査することとした．

2.2 タグ付け作業の内容

作業対象記事

タグ付け対象は 2004 年の各月から選んだ 44 日分の記事 3,563 本である.

作業者

作業者は女性の日本語母国語話者 1 名で，特別な言語的な訓練は受けていない.

教示内容

前節で説明したリード，本記，追記は必ずしも定義が明確でなく作業が困難なことが予想された．そこで，著者らの目的に合わせて構造タグの役割を定義し直し，それを作業者に教示した．概要は以下の通りである.

リード

冒頭付近の文で，記事全体を要約している．その他の文との語彙の重なりが多い

本記

リードを詳説している．リードとの語彙の重なりが多い

追記

リード，本記の補足情報を記述している．リード，本記との語彙の重なりが少ない．リードと並べて要約としたときに，談話的な違和感がない．すなわち，追記は先行詞が欠落して不自然となる照応詞を含まないなどの条件を満たす

構造タグの定義にあたって語彙的な重なりを利用したのは，完全ではないものの内容の重なりの判定の一助になると考えたからである．また，追記に新たに課した「リードと並べたときの談話的な整合性」は，著者らの要約の目的を反映した結果である.

ユーザインターフェース

タグ付与作業にあたっては，文間の形態素の重なりを観察できるツールを提供した．このツールには，形式が不完全と思われる記事，レポートのような特別な記事にはコメントを記述できるようになっている.

作業中に判断に迷う場合も出てきたが，これらはその都度判断基準を変更しながら進めた．例えば，本記以外の情報を含んでいるため追記と思われる文であるが，一部が本記と重なっている場合である．この時は，その文の主たる情報が重なっているかどうかで判断することとした．このような判断基準の追加をしばしば行ったため，最後に全体の見直しを行った.

2.3 タグ付け結果の分析

現場レポートやインタビューなどの会話調の原稿はそもそも 3 段構造を持っていないため，タグ付けできなかった．このような対象外記事を除くと 3,481 記事にタグ付けできた．図 1 に記事の文数分布を示す．4 文ないし 7 文の記事に分布が集中していることが分かる.

2004 年 1 月 19 日の 22 記事 127 文を対象に，筆者と作業者のタグの一致度を評価した．表 1 に結果を示す.

図 1　記事の文数分布

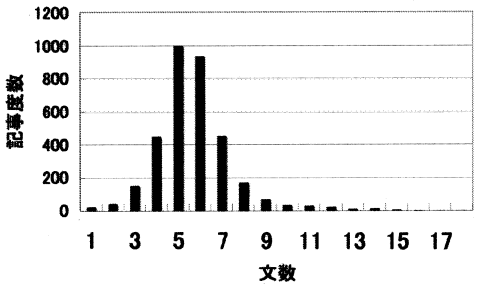

表 1　タグの一致状況

	リード	本記	追記
リード	24	0	0
本記	0	69	10
追記	0	2	22

この表をもとに κ 値を計算したところ 0.834 であった．これはほぼ完全な一致 (near perfect) という評価範囲に属し(野本 97)，全体として一致が良好だったことがわかる.

リードはその他のタグとの不一致を起こしていない．一方，本記と追記には 2 者の不一致がかなりある．文が記事全体に対して補足的かどうかの判断が難しく，今回の基準ではまだ不十分だったと思われる.

記事の構造を調査するために，文数が 6 の記事の，リード文と追記の文の数を調査した．結果を表 2 に示す．典型的にはリード文が 1 つで追記が 1 という構造をしていることがわかり 3 段構造はおおむね普遍的な構造だと考える．ただしリード文はすべての記事にあるが，追記を持たない記事は少なからずある点には注意が必要である.

なお，この表にリードが 2，本記が 1，追記が 3 という記事が 1 例ある．これは，台風のニュースで，リードはあるものの，その後は一文を除いて

表２　６文記事の構造分布

	リード文数１	リード文数２
追記文数０	121	53
追記文数１	464	4
追記文数２	203	5
追記文数３	51	1
追記文数４	4	0

各地の状況の羅列となっていた．このような特殊な構造をとるニュースも一部あった．

また文数の多い長い記事には３段構造で十分に表現できないものがあった．長い記事は複数のトピックを扱う傾向が強い．

例えば，台風の記事で，雨量に関する記述の後，被害の記述が続くような記事である．これは２つの独立した記事が直列結合した構造を持ち，それぞれにリード，本記，追記の構造が見られる．長い記事の構造にはどのようなものがあるか，またそれらの構造を記述するにはどのような情報が必要か，今後検討したい．

3 タグ予測実験

3.1 予測手法と利用属性

前節で作成したデータを使った構造タグの予測実験を行った．著者らはこの問題を解くのに，要約における重要文抽出問題と同じく，文の分類問題として扱うことにした．これは，本稿のリードと追記が従来の重要文に近いと考えられることから，重要文抽出で使われる特徴（属性）が本問題でも有効に使えると考えたからである．ただし，重要文抽出では文の重要・非重要の分類問題になるが，本研究ではリード，本記，追記の分類問題となる．

重要文抽出には，これまでさまざまな特徴が提案されている（奥村，難波 99）．著者らはこれらの知見を元に予備実験を行い，以下の特徴を利用することとした．また，分類器には内製の C4.5（Quinlan 93）相当のシステムを利用した．

位置属性

リードは記事の先頭付近に，追記は末尾付近に現れる．そこで，次の２つの２値属性(1/0)を利用した．

(a) 先頭文

先頭文のときに１それ以外は０をとる

(b) 最終文

最終文のときに１それ以外は０をとる

手がかり表現（特徴語）

重要文抽出問題では「まとめると」などの手がかり表現を使うと有効であることから，著者らもこれを使うことにした．ただし，どのような手がかり表現が有効かは分からないため，後述する手法で，構造タグ「リード」「本記」「追記」と正の関連を持つ文節と負の関連を持つ文節を抽出して属性とした．

(c) 正の特徴語属性

リード，本記，追記の各タグと正の関連を持つ文節の数を属性値とする

(d) 負の特徴語属性

同上，負の関連をもつ文節数を属性値とする

ここで，文節の有無ではなく「数」を属性値にしたのは，タグによっては呼応する複数の文節が手がかりになる可能性を考えたからである．

類似度

本記はリードを詳述するものであるから，これらの文間の類似性は高いと考えられる．また，追記はリードや本記との内容の重なりが少ない文と定義しており，これらの文間の類似性は低いと考えられる．そこで各文に対して，次の２種類の類似度を属性として利用した．

(e) 先頭方向最大類似度（B 最大類似度）

記事の先頭方向の各文との類似度を測り，その最大値を属性値とする（B: backward）

(f) 最終文方向類似度（F 最大類似度）

記事の最終文方向にある各文との類似とを測りその最大値を属性値とする（F: forward）

２文間の類似度は形態素単位でアラインメントを行って共通単語を抽出し，これを使った余弦値とした（田中他 07）．このため(e)と(f)の属性値は０から１までの値となる．

ここで，(e)と(f)の類似度には位置情報が含まれることに注意が必要である．記事の先頭文ではB最大類似度を計算できない．同様に，最終文ではF最大類似度を計算できない．そこでこれらの場合にはN.A.（not available）という属性値を与えた．この結果，類似度属性は数値と離散値（N.A.）を含むことになり，値 N.A.は先頭，最終文を示すことになる．これに伴って，決定木アルゴリズムも，離散値と数値を同時に扱えるように変更した．

3.2 手がかり表現抽出法

タグと文節の関連性を検出するのに，確率変数の

独立性検定の手法を採用した．具体的にはχ二乗検定と対数尤度比検定(Dunning 93)である．両者とも表3に示す2分割表を元に計算する．tは着目しているタグで，wはある一つの文節とする．

表3　2×2テーブル

	w	not w
t	a	b
not t	c	d

記号 a, b, c, d はタグと文節の共出現頻度を表している．例えば a はタグ t のついた文に w が出現した文の頻度を表す．b はタグ t が出現して w が出現しなかった文の頻度を表す．χ二乗値はこの表を使って，$n = a + b + c + d$ とすると

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (1)$$

と計算できる．この値が大きければ独立性の帰無仮説が棄却されて $p(t,w) \neq p(t)p(w)$

という結論を得る．

このとき $p(t,w) > p(t)p(w)$，すなわち共出現に正の関連があるときと $p(t,w) < p(t)p(w)$，すなわち負の関連があるときに分けられる（Munteanu and Marcu 96）．また，これらはそれぞれ $ad - bc > 0$ と $ad - bc < 0$ に対応することが計算できる．以上から(1)式の代わりに

$$\chi'^2 = \mathrm{sign}(ad - bc)\chi^2, \quad (2)$$

ただし

$$\mathrm{sign}(x) = \begin{cases} +1 & (x > 0) \\ -1 & (x < 0) \end{cases},$$

と符号付きにすると，符号によって正負のどちらの関連があるのかわかる．またその度合いは符号を除いた値で評価できる．

本研究ではこの符号付きの χ'^2 を使って，正の特徴語と負の特徴語を次のように定義した．ここで W は抽出対象コーパスの全文節集合である．

タグ t の正の特徴語

文節 $w \in W$ のうち，t と計算した χ'^2 が正の閾値以上であり，かつ t 以外のタグと計算した χ'^2 の値がゼロ以下となるもの．

タグ t の負の特徴語

文節 $w \in W$ のうち，t と計算した χ'^2 が負の閾値以下であり，かつ t 以外のタグと計算した χ'^2 がゼロ以上となるもの

閾値には±3.84，±6.63，±7.87，±10.82 を使用した．これらは自由度1のχ二乗分布のパーセント点 0.050，0.010，0.005，0.001 を与える値である．以上の議論は対数尤度比にも当てはまり，これを使った正負特徴語抽出も行った．また文節の中で固有表現を抽象化した場合も実験した．

表4に上位の正の特徴語を示す．これは文節そのものにχ二乗を指標として得られたものである．（これらは次節で説明する 2,834 記事から抽出した．）

表4　正の特徴語の一部

リード	きょう，なりました，事件で，大リーグ，イラクで，問題で，15日，わかりました，会談し，ことが，わかり，イラクの，25日，開かれ，きょうから，19日，疑いで，5日，逮捕された
本記	この，ものです，よりますと，もので，この中で，それに，うち，そして，その上で，その，これは，結果，この中で，調べに，およそ，事件は，その上で
追記	話しています，しています，一方，今回の，現場は，今後，狙いが，原因を，ものと，ことで，初めてだという，今後，さらに，なります，マリナーズが，片側，一週間を，メッツが，見られます，試合は，予定です

この表から次のような特徴が観察できる．リードは「きょう」などの日付に関わる表現や「なりました」など過去を表す述語の利用が顕著である．本記は「もので（す）」といった説明の述語，さまざまな接続表現が特徴的である．追記は「しています」「見られます」などの現在の状態を表す述語の利用，未来を表す語，「今後」「さらに」「予定です」などの利用などが特徴的である．

3.3 実験手法

実験対象記事

4文ないし7文からなる 2,834 記事を実験に利用した．これらの記事を使ったのは，頻度が集中しており，かつ今回の想定の3段構造を持つと考えたからである．

実験手順

2,834 記事からランダムに抽出した 500 記事から 3.2 節に述べた手法で特徴文節を抽出した．残りの 2,334 記事を決定木によるタグ認定実験に利用した．実験は5回の交差確認法を使い，平均誤り率で評価した．

表5　実験データ中の構造タグ分布

リード	2,538	19.80%
本記	7,950	62.07%
追記	2,320	18.11%
合計	12,808	100.0%

3.2節で述べた属性の組み合わせごとに，平均誤り率を計算した．ただし，次の属性は常に組み合わせて利用することとし，個別には実験しなかった．

- 先頭文と最終文　(a), (b)
- B最大類似度とF最大類似度　(e), (f)
- リード，本記，追記の正特徴語 (c)
- リード，本記，追記の負特徴語 (d)

正負の特徴語の抽出は，以下の組み合わせの16（2×2×4）種類で行った．

- 検定手法：χ二乗法，対数尤度比法
- 語単位：文節，抽象化文節
- 閾値：±3.84, ±6.63, ±7.87, ±10.82

3.4　タグ予測結果

表5に2,334記事に含まれたタグの分布を示す．

これより基準誤り率は「本記」を常に出力する場合の37.93% (100-62.07)となる．

すべての属性の組み合わせで実験を行った結果，最小誤り率は8.325%であり，基準誤り率からは大きな減少が得られた．このとき使った属性は位置（先頭，最終文），正の特徴語（リード，本記，追記），類似度（B最大類似度，F最大類似度）である．また正の特徴語は文節を単位とし，χ二乗の閾値3.84を使ったときに得たものである．特徴語の数は（リード，2,928）（本記，44）(追記，3,846)であった．抽出対象の500記事の異なり文節数（$|W|$）は50,522であった．

最小誤り率のときのタグごとの平均誤り率は表6に示す通りである．この表から，追記の平均誤り率が大きいことがわかる．人間のタグ付けの一致もこの部分が悪かったことと共通する．

次に各属性の効果を調査した．上記の最小の平均誤り率となった属性群から，各属性を取り除いた属性で実験して得られた平均誤り率を表7に示す．効果の大きい属性ほど，除去したときに平均誤り率が上昇する．

表6　各タグの平均誤り率

タグ	平均誤り（%）（標準偏差）
リード	4.369 (0.037)
本記	4.691 (0.037)
追記	25.06 (0.185)

表7　最高性能と各属性の貢献

条件	平均誤り(%)	標準偏差
最小誤り率	8.325	0.032
位置削除	9.645	0.011
類似度削除	17.97	0.009
正特徴語削除	8.380	0.031

表8　個別属性による平均誤り率

条件	平均誤り(%)	標準偏差
位置情報のみ	18.02	0.008
類似度情報のみ	9.676	0.011
正特徴語のみ	37.27	0.046

表7から次のことが観察できる．正の特徴語を削除しても平均誤り率が上昇しておらず，効果はほとんどなかった．位置情報を削除したときに，わずかに平均誤り率が上昇した．先に述べたように，本研究の類似度には位置情報も含まれているため，基本的には平均誤り率は同じ程度になると予想していた．しかし，属性の組み合わせの効果などから，両者に違いが生じたようである．結局，最も効果的だったのは類似度であった．

次に個別の属性を使った場合の平均誤り率を表8に示す．

表8からも，正の特徴語を使っても基準誤り率(37.93%)からほとんど改善しないことがわかる．一方，位置，類似度は効果があったことがわかる．

以下，特徴語についてさらに調査した結果を報告する．まず，抽出する特徴語の数の効果を調べた．抽出される特徴語は閾値の絶対値を増加させると減少する．そこでこれを増加して特徴語を減少させたときの分類性能を調べた．χ二乗法，対数尤度法のそれぞれで閾値を変えたときの，正の特徴語数，平均誤り率（正の特徴語だけを利用した実験結果）を表9と表10に示す．

これらの表から，閾値の増加に伴って特徴語数が減少するが，平均誤り率に大きな変化はないことがわかる．

次に，特徴語を抽出する対象記事の大きさの効果を調べた．3.3節の実験は2,834記事の中から

表９　χ二乗法：閾値変化の効果

閾値	3.84	6.27	7.87	10.82
特徴語数	6,818	427	400	97
平均誤り(%)	37.27	37.37	36.99	37.23

表１０　対数尤度比法：閾値変化の効果

閾値	3.84	6.27	7.87	10.82
特徴語数	553	264	119	61
平均誤り(%)	37.57	37.00	37.23	37.44

500 記事を特徴語抽出データとしたが，これが小さく必要な特徴語が漏れた可能性がある．

そこで特徴語抽出記事を 1,400 に増加して，残りの 1,434 記事で決定木作成実験を行った．この結果，データ量の減少によるためと思われる，全体的な誤り率の増加があったが（最小誤り率は 9.023%，標準偏差は 0.017）特徴語の効果はやはり見られなかった（正の特徴語を除いた時の誤り率は 9.074%，標準偏差は 0.017）．

以上の調査は固有表現を抽象化した文節でも行ったが結果は同様であり，特徴語の強い効果は見られなかった．

4　関連研究と考察

今回の構造タグ予測実験では特徴語の効果を確認できなかった．表４に示した上位の特徴語群からは予測効果が期待されるため，著者らの特徴語の使い方はまだ不十分だと考える．これに関連して，渡辺ら（Watanabe 96）は重回帰分析を用いた重要文抽出を報告している．この中では時制，文タイプ，修辞関係などを表す表層の表現を説明変数に使っている．著者らの特徴語はノイズも含めて多種の文法機能が混在している．これらを渡辺らのような形に整理，分類の上使うと効果が得られると考えている．

本研究に関連して，野本ら（野本，松本 97）は 112 人の学生に新聞の重要文抽出を行わせる実験を報告している．ここでの一致度は本研究よりも低く，タスクがかなり難しかったことが伺える．また，C4.5 を使った重要文抽出実験によれば人間の一致が高いほど予測性能が高くなったことを示している．これは著者らの，人間の一致の悪かった追記の予測性能が一番悪かった事実と一致する．

5　おわりに

ニュース構造を表すリード，本記，追記のタグをニュース記事に付与し，その結果からこれらの３つのタグはほぼ安定してニュース記事に見られることを示した．またこのタグを C4.5 決定木アルゴリズムで予測する実験を行った結果，最小誤り率 8.325%を得た．また位置情報，文間類似度が予測に有効であり，統計的に抽出した手がかり表現の効果はほとんどなかったことを示した．特徴語の使い方についてはさらに検討を重ねたい．さらに，今後はタグの予測と表現置換を合わせた要約システムを構築する予定である．

謝辞

本研究を進めるにあたり，ユージンソフトの脇隆三氏にはソフトウエア開発でお世話になった．ここに記して感謝する．

参考文献

Dunning, Ted. 1993. Accurate Method for the Statistics for Surprize and Coincidence. *Computational Linguistics*, 19(1), pages 61-74

Munteanu, S. M. and D. Marcu. 2006. Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora. *The 21st International Conference on Computational Linguistics and 44th Annual meeting of the ACL*. pages 81-88, Sydney

Quinlan J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

Watanabe, H. 1996. A Method for Abstracting Newspaper Articles by Using Surface Clues. The 16th International Conference on Computational Linguistics, pages 974-979.

井上 1981. ニュース文章は変えうるか. 文研月報 12 月号, NHK 総合放送文化研究所, pages 12-21.

奥村，難波 1999. テキスト自動要約に関する研究動向（巻頭言に代えて），自然言語処理，vol.6(6), pages 1-26.

田中他 2005. ニュース要約の実態調査と要約モデルの検討．情報処理学会自然言語処理研究会，2005-NL-170, pages 115-120.

田中他 2007. 放送ニュースのための表現置換．言語処理学会第 13 回年次大会, pages 306-309.

野本，松本 1997. 人間の重要文判定に基づいた自動要約の試み，NL 研資, 120-11, pages 71-76.