構文情報を使ったリード文の具体化による要約支援

田中 英輝　木下 明徳　小早川 健　熊野 正　加藤 直人
NHK 放送技術研究所
{tanaka.h-ja, kinoshita.a-ek, kobayakawa.t-ko, kumano.t-eq, katou.n-ga}@nhk.or.jp

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支援の研究を行っている. この構造とは,「リード」,「本記」,「追記」の 3 要素からなる構造を指す（井上　81）.
著者らの聞き取り調査によると, ニュースの要約者はリードを要約の出発点として, 字数の制限に合わせて, これを具体化している.
著者らはこの作業を支援する研究を進めている. これには, ニュース記事の各文に「リード」「本記」「追記」のタグを認定する処理, リードの表現を, 本記の表現で具体化する処理が必要である. すでに前者のタグの認定処理については決定木を使った手法を提案している（田中他　07）. 今回, リードの表現を本記の文で具体化する手法を提案し, 実験したので報告する.

2　放送ニュースの要約支援

典型的な放送ニュース記事は, 全体の要約である「リード」, その詳述である「本記」, これらの補足である「追記」からなる. またニュース要約者への聞き取りにより以下のことがわかっている（田中他　05）.

1) ニュースを要約する場合, リードを要約文の出発点とし, 字数の余裕に合わせて, 本記の表現でリードを具体化する. これは, リードには固有名詞や日付がないことが多く, 情報が不足するからである.
2) 要約者は時間がない中で作業をするため, 記事を頭から読むのではなく, リードの各表現に対応する表現を他の文の中に見つけようとしながら読む.
3) 複雑な編集を行う時間がないことから, いわゆるコピーアンドペーストによる編集を行う.

著者らはこのような要約作業を支援するため, リードに対する可能な具体化を提案するシステムを検討した. このシステムは, コピーアンドペースト作業を想定し, 文節連続からなる, 挿入・置換候補を提示する. 要約者はシステムの提示する候補の中から最適な候補を選択する. これにより著者らは要約者の作業の軽減が期待できると考えた. また, 1) 文法的に正しく 2)リードの具体性が増す候補を提案することを目標にした. 可能なら, 要約に最もふさわしい候補のみを出力させたいが, 現時点でこれは難しいと考え, 上記を目標とした.

3 編集候補提示手法の概要

3.1 人手による表現編集実験

システムのアルゴリズムを検討するため, 2004 年 1 月 19 日と 20 日から選んだ 15 記事を対象に, リードを本記の表現で具体化する実験を人手で行った.

表 1 人手編集の例

リード	係り先	本記	係り先
	感染の	コイの大量死の原因となった「コイヘルペス」の	感染の
一部の旅客機~~で~~	見つかった	旅客機, MD811 型機~~の~~	一部の
ウルズガン州で	起きて	ウルズガン州のチャー・チノという村で	起きたもので
	会談し	中国国営の新華社通信によりますと	会談しました

編集操作:コピーアンドペーストによる表現の挿入と置換とした. 本稿ではこの操作を編集と呼ぶ.
表現:基本的には連続した文節群とする. ただし, 最終文節の助詞を除く操作は許す. これは, 助詞を除いた名詞句や動詞句の交換が可能な場合に対応するものである.

テキストだけを見ながら編集を行うのは時間がかかり効率が悪い. そこで, リードの文節と同一の文節が本記に出現していればそれをハイライトするインターフェースを作成し, これを参考に作業した.
表 1 に得られた編集結果の一部を示す. これには, 編集の対象となったリードと本記の表現と, それらの係り先文節を示している. 表 1 でリードが空の場合は挿入で, そうでなければ置換である. 挿入の場合の係り先とは, 表現をリードに挿入した場合の最も適切な係り先文節である.
このような編集の例が 34 例見つかった. このうち助詞を取り除いた編集は 5 例であった. 助詞を取り除いて名詞句になった例を表 1 の第 2 行に示す. この結果から大半は表現そのままの編集であったことがわかる.
次に, リードと本記の編集対象となった表現の係り先文節に着目し, その一致程度と, 挿入, 置換の数をまとめた結果を表 2 に示す. 自立語一致とは, 表 1 の「会談し」と「会談しました」のような例である. また, 異文節とは「見つかった」と「一部の」のような例である. 異文節の先の係り文節を見ると同一の文節や, 意味が類似した文節に係ることがわかった. 例えば「一部の」という文節は「一部の→機体に→見つかった」と最終的には同一の「見つかった」に係っていた.

表 2 リードと本記の係り先文節の一致と編集

	挿入	置換	計
完全一致	9	6	15
自立語一致	6	6	12
異文節	1	6	7
計	16	18	34

また,「到着しました」と「入りました」(「サマーワに入りました」)のような同義語が係り先になる場合もあった. 以上の観察から編集に使うことのできる表現は, 意味が“同一”の係り先を持つと仮定した.

今回の調査では, 置換は比較的容易に見つけられたが挿入を見つけるのに苦労した. 置換表現を見つけるには, リードと本記にある類似表現を対応させればよいことが多い. これに対して, 挿入表現は, リードにない表現を本記から探さなければならない. 挿入の方が表現の発見に使える手がかりが少ない問題がある.

3.2 編集候補の発見方針

システムを作るにあたっての課題は, 要約者に提示する編集候補をいかに系統的に探すかである. これに対して,「編集に使える表現は“同一”の係り先を持つ」と仮定したことから, リードと本記にあるすべての“同一”の文節を探し, そこに係っている表現から編集候補を探す方針を立てた. この手法は, 置換, 挿入の表現を探すのに利用でき, 特に, 手がかりの少ない, 挿入表現の発見に使える利点がある.

3.3 編集候補提示アルゴリズム

本説では編集候補提示アルゴリズムを示す. まず, 本稿で使う述語を定義する.

トリガー文節
前節で述べた「同一文節」をトリガー文節と呼ぶ. この「同一文節」は編候補発見の契機となることからこのように呼ぶ. 具体的な同一性判定法は後に示す.

単位句
トリガー文節に直接係っている文節が支配する係り受け構造を単位句と呼ぶ. これは文節の連続となる. 本アルゴリズムはこの単位句を編集候補の単位とする. なお, 本稿では文節の連続を句と呼ぶ.

以下, 動作概要を示す(アルゴリズム 1, 表 3). また図 1 を使って具体的に説明する.

(3-5 行)T の各トリガーペアに対して, トリガー文節の単位句を認定する. 図 1 はトリガー文節「到着しました」を処理している例で, それぞれ 2 つの単位句が認定されている.

(6 行)類似した単位句を対応付ける. 例えば, 図 1 では「IAEA のチームが」と「IAEA の査察官 5 人が」が対応付けられ, それ以外は対応付けられていない.

アルゴリズム 1

0: 入力:係り受け解析されたリードと本記の文
1: 与えられた 2 文のトリガー文節ペアを探す
　　それらをテーブル T に格納
2: 　{ l(リード), b(本記) トリガー文節 }
3: for all (l, b) ∈ T do
4: 　l の単位句群を求める
5: 　b の単位句群を求める
6: 　リードと本記の単位句群を対応付ける
　　結果をテーブル A に格納
7: 　for all (L, B) ∈ A do
8: 　　{ L(リード), B(本記) 単位句 }
9: 　　L, B の状況により表 3 に示す処理を行う
10: 　end for
11: end for

表 3 単位句対応結果と処理

		本記	
		$B=\varnothing$	$B\neq\varnothing$
リード	$L=\varnothing$		1:挿入処理
		3:操作せず	2:置換処理

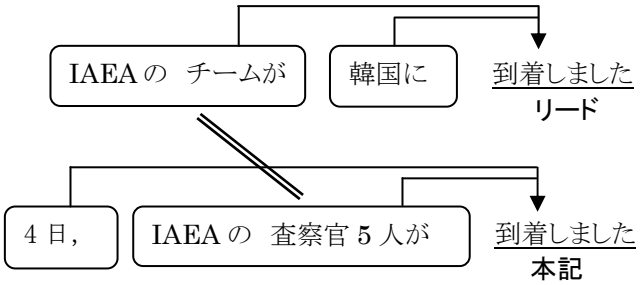

図 1 編集候補提示アルゴリズムの概念図

(9 行, 表 3)対応付け結果に応じて次の処理を行う.

表 3-1 挿入処理:本記だけにある句はリードへ情報を追加できるので, 挿入候補として処理する.

表 3-2 置換処理:対応している句同士は類似しており, 置換候補として処理する(置換せずに, 挿入すると内容が重複することに注意).

表 3-3:リードだけにある句は, そのままとする.

以上から, 本記の単位句は, 対応付けのあと, 挿入候補か置換候補となる.

図 1 の場合で表 3 に従って置換, 挿入を実行すると,「4 日, 」をリードに挿入して,

「4 日, IAEA のチームが韓国に到着しました」

という候補と,「IAEA のチームが」と「IAEA の査察官 5 人が」を置換して

「IAEA の査察官 5 人が(IAEA のチームが)韓国に到着しました」

の 2 つの候補が得られる. システムはこれらの候補を一度に表示せず個別に表示する. なお, ここでの置換の説明は簡略化しており, 正しくは 4.4 節を参照のこと.

4 編集候補提示アルゴリズムの詳細

本節ではアルゴリズム 1 のトリガー文節認定処理，句対応処理，挿入処理，置換処理を詳述する．まず，この後使う 2 つの基本的な評価指標を定義する．

文節類似度 $t(x,y)$

文節 x,y の類似度 $t(x,y)$ は，文節を自立形態素の集合とみなして計算した Dice 係数とする．付属語を含めた文節類似度は $s(x,y)$ と記す．

句の含有率 $D(X,Y)$

句 X,Y に対して，X がどれくらい Y に含有されるかの量である．文節類似度を用いて以下のように計算する．絶対値記号は自立形態素の数を示す．

$$D(X,Y)=\frac{1}{|X|}\sum_{x\in X}\max_{y\in Y}(t(x,y))$$

4.1 トリガー文節認定処理

今回，基本的には表層が同一の文節をトリガーとする方針を立てたが，形式動詞「する，なる」，形式名詞「こと…」機能動詞「与える，受ける，始める，かかる…」などからなる文節（形式文節）の一致では不正な編集が多数発生した．形式文節はそれだけでは意味を持たないため，一つのニュース記事の中であっても異なる事実を示すことがあるためである．形式文節の意味は，係っている格要素が決める．そこでトリガーの資格を検査するには，係っている格要素の一致を評価することとし，複数のルールを作成した．

また，表層が完全一致しないトリガーの検査については，動詞の連用形と終止形の違いのみ一致と判定した．例えば，「会談し」と「会談しました」である．

4.2 単位句の対応付け処理

次に示す単位句対応率を使い，どん欲法でこの値の高い順に対応付ける．対応率が閾値以下になれば対応付けをやめる．

単位句対応率 $P(X,Y)$

X を単位句 X,Y のうちの短い方とする．また x,y を単位句 X,Y の最終文節とする．

$$P(X,Y)=\alpha D(X,Y)+(1-\alpha)s(x,y),\quad (0\le\alpha\le 1)$$

この式は単位句全体としての含有率（第1項）と，最終文節の付属語を含めた類似度（第2項）の加重平均である．付属語を含めて，最終文節を別途評価しているのは，これが句の文法的な性質を決める上で重要だからである．

4.3 挿入処理

挿入候補に対して，「禁止語」と「重複率」の検査を行い，これを満たせば，挿入位置を決定して候補として提示する．

禁止語検査

例えば指示語で始まる句をリードに挿入すると，多くの場合，意味が不自然となる．なぜなら指示語は前の文脈を必要とするが，記事の冒頭にあるリードに挿入すると，適切な文脈が得られないからである．そこで，指示語のように前の文脈を必要とする語を禁止語リストにまとめ，このリストにある語で始まる句の挿入を禁止した．このリストは，「こそあど」や「接続詞」のほか，「その後」「今回」「他の」「同じ」などを含んでいる．

重複検査

挿入しようとする本記の単位句と類似した句がリードにあると，結果として情報が重複する．そこで，挿入する単位句とリードの重複度合いを次式で測り，閾値を超えた場合は挿入を禁止した．

重複率 $O(X,Y)$

挿入しようとする単位句を X，リードを Y として含有率を計算し，これを重複率とする．

$$O(X,Y)=D(X,Y)$$

挿入位置決定処理

単位句対応付けの後，本記の単位句は挿入，置換候補のどちらかになっている．そこで，挿入する句の右隣の句に着目し，その行き先の左を挿入位置とした．これは本記での句の順序を反映させるためである．図 1 の「4 日，」は，右隣の句の「IAEA の査察官…」の置換先，「IAEA のチーム…」の左，すなわち先頭となる．

4.4 置換処理

置換候補の単位句は，「完全一致文節」があれば分割し，次に具体性の検査を行う．その後，挿入処理に準じて禁止語検査と重複検査を行い，これらを満たせば置換候補として提示する．ここでは置換処理に特有な「句の分割」と「具体性検査」について説明する．

句の分割処理

対応句の中に完全一致する文節があれば，トリガーとこの間を置換候補とする．なければ対応句全体を置換候補とする．図 1 の置換では「IAEA の」が完全一致文節のため「チームが」と「査察官 5 人が」が置換候補となる．「IAEA の」の前に異なった句が係っていれば，「IAEA の」をトリガーとした処理が行われる．本処理は包含関係を持つ候補の提示を避けるのが目的である．

具体性検査

本記の挿入候補句とリードの対応句の文字列長を比較して，本記の挿入候補が長ければ挿入可能とする．多くの場合，長い表現の方がより具体的であるという経験則による．

5 主観評価実験

5.1 実験要領

提案手法の効果を調べる主観評価実験を行った．また結果に対し，係り受け解析の誤りによる影響の調査と，リードの編集を，本記と行う場合と，追記と行う場合の比較を実施した．実験要領は以下の通りである．

対象：2004 年 1 月 20 日，4 月 20 日，7 月 20 日のニ

ュース記事 257 記事. mecab と cabocha(標準)で解析した.

評価者:日本語母国話者 1 名

評価インターフェース:専用の評価インターフェースを作成した. これはリードとそれ以外の文の間で発生する単一の置換, あるいは挿入を提示する. 置換では削除と挿入が同時に発生するため, 削除される表現を括弧に囲んで表示する. このため評価者は両方を見比べて評価することができる.

評価項目

それぞれの編集候補に対して以下の 4 項目を評価した.

係り受け		事実関係	
0	誤り	0	不整合
1	正しい	1	整合

具体性		修正の程度	
0	減少	0	大幅な修正が必要
1	変化せず	1	軽微な修正が必要
2	増加	2	修正不要

これらの項目を設定した狙いを示す.

係り受け(前提):本手法は正しい係り受けを前提としているので, これを評価した. 評価したのは, トリガーペアの各単位句の係り受けなど, 編集に関わる, 部分的な係り受けである.

事実関係(必要条件):本記の表現を置換・挿入した結果が意味を持つには, 最低限, 事実関係がリードに整合しなければならない. そこで各編集について, この整合性を評価した.

以下の 2 項目は事実関係が整合した結果についてのみ評価した.

具体性(意味的効果):編集による具体性向上の効果を評価した. ここでは文法的な正しさは問わない.

修正の程度(文法性):編集結果の文法性を測るのに修正にかかる手間で評価した. 助詞の変更や, 挿入位置の変更のように, 一つの操作で修正が終わるものを「軽微な修正」とし, それ以上必要な場合を「大幅な修正」とした.

5.2 結果

評価の例を示す. 下線が挿入で()内が削除された句である.

具体性(2)修正(2)
民間団体の「コリア・ソサエティ」などが共催する「朝鮮半島平和フォーラム」に(催しに)出席する

具体性(1)修正(2)
部品に亀裂が入っているのが()見つかった

具体性(2)修正(0)
ヘリコプターから地上二十メートルの高さから()落下し死亡しました.

表 4 評価結果

		編集数	事実	具体	修正
係受正解	本記	359	324	1.86	1.45
	追記	64	29	1.66	0.72
係受誤り	本記	155	140	1.87	0.78
	追記	25	9	1.67	0.22

評価結果を表 4 に示す. まず, 係り受け解析の誤り率は約 30%(180/603)であった. この影響をみるため, 本手法の想定である, 本記による編集の場合に着目すると, 事実関係の認定に与える影響は見られなかった(324/359, 140/155). 調査したところ, トリガーペアに対して同じ係り受け間違いをしている場合は, 誤りの影響が相殺されること, 単位句の短い修飾に関する係り間違いであれば影響が少ないことなどが原因であった. また, 具体化についても(1.86, 1.87)と同程度の効果があった. 一方, 修正程度は(1.45, 0.78)とかなり異なった. 1 が軽微な修正なので, 係り受けが不正であれば, 大きく修正することが必要なことがわかった. 係り受けの精度向上はこの面で重要である.

次に係り受けが正しい場合の, 本記と追記による編集の違いを見る. 事実関係が正しい割合は(324/359, 29/64)と追記は低い. 追記には, リードにない情報が書かれているので, 編集に使えない単位句が多いことが原因だと思われる. また, 修正も(1.45, 0.72)と追記ではかなり必要になる. 著者らの提案どおり追記の情報は具体化に使わないのがよいと結論できる. なお, 係り受け解析の正解と誤りを合わせた本記での編集の結果は, (編集数, 514)(事実, 464)(具体, 1.86)(修正, 1.25)であった. 今後は文法性(修正), 特に句の挿入位置の改善が必要と考えている.

6 おわりに

リード文を本記の表現で具体化する作業を支援する手法を報告した. 本稿ではアルゴリズムを中心に報告したが, 著者らは, すでに, ニュース構造を認定するシステムと組み合わせたシステムを作成している. この概要, および関連研究については稿を改めて報告したい.

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. また安達展江さんには評価実験の協力をいただいた. ここに記して感謝する.

参考文献

井上 1981. ニュース文章は変えうるか. 文研月報 12 月号, NHK

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

田中他 2007. ニュース要約のための簡易文脈解析, 自然言語処理研究会, 2007-NL-182, 75-80