要約文の選定による用例利用型要約の可読性向上

牧野 恵　山本 和英
長岡技術科学大学 電気系
E-mail: {makino,ykaz}@nlp.nagaokaut.ac.jp

1　はじめに

近年、インターネットの普及や企業に対する e-文書法等の施行に伴い、我々の周りには膨大な電子化文書が存在するようになってきた。さらにここ数年間におけるウェブ上の文書数は加速度的に増え続けており、今後も我々が処理しなければならない文書量は年々増す一方である。そこで近年、これらの文書を自動的に要約する研究が盛んになってきている。

自動要約の既存研究としては入力文書からタイトルや文の位置、使われている単語の頻度情報を考慮して文に対して重要度を計算する重要文抽出 [1] や 1 文内で語に対して重要度を計算し、その重要度に基づいて 1 文を圧縮する文圧縮 [2] する手法などがある。しかし人間が要約を行う時にはどのような内容で要約したらよいのかという経験や文法等の様々な知識を使って要約する。そのため我々は人間と同じような重要度の設定は困難であると考える。

そこで以前から我々は重要度の設定を行わずに複数文の情報を含んだ 1 文要約を行う「用例利用型要約」手法を提案している [3]。入力は文書とし、複数の文から文節を抽出、さらにそれらを組合せることで要約を行っている。しかしどのような文節でも組合せることで日本語として読みやすい、適切な内容が含まれている要約文を作ることはできない。そこで我々は要約事例を用例として模倣することによって要約を行った。用例利用型要約は類似用例文の選択、類似用例文と入力記事間での文節の対応付け、そして文節の組合せの 3 つのステップから構成される。本稿ではこれらの各ステップに対して改良を行った。また考察から得られた結果を元に最終的な出力要約文の可読性向上を目指し、複数の要約文を出力した中からより可読性の高いものを選定する方法も述べる。

2　手法概要

図 1 に用例利用型要約の手法概要を示す。

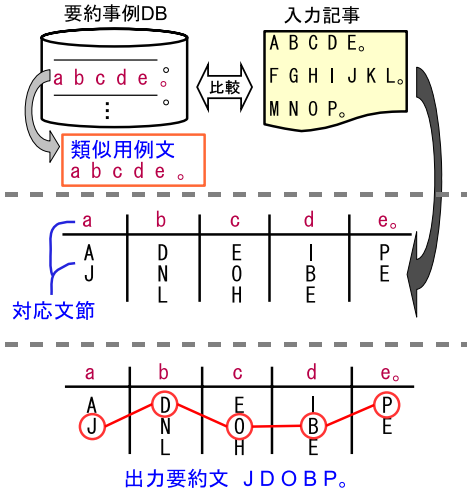

図 1　用例利用型要約の手法概要

まず初めに入力記事と要約事例データベースに含まれている用例文を比較し、入力記事と内容の似ている類似用例文を獲得する。要約事例データベースに含まれている用例文とは人間が作成した要約文である。つまり人間が要約する際にどのような内容で要約したらよいのかという経験や文法等の様々な知識を含んでいると考えることができる。このステップについては 3 章で述べる。続いて先ほど得られた類似用例文と入力記事の文節を比較し、類似している文節を対応付ける。このステップについては 4 章で説明する。そして最後に対応付けられた文節を組合せることによって要約文を得る。この組合せの部分については 5 章で述べる。

3　類似用例文の選択

類似用例文の選択では要約事例データベースに含まれている用例文の中から入力記事の内容に類似した用例文「類似用例文」を獲得する。ここで内容の類似性をどの部分に注目して測るかというところに問題があるが、本稿では述語と内容語の一致に着目し、類似用例文を獲得する。以前に提案した手法では類似度を述語と内容語の一致数の重み付き和で表していた。しかし重みの調整が困難であり、常に内容の似ている用例文を獲得することが難しかった。そのため本稿ではまず文の主題であり格を決める要素でもある述語に着目して用例文を獲得する。そして次に述語が一致した用例文の中でも内容語の一致率が高いものを類似用例文として選択する。類似用例文の選択方法を図 2 と図 3 を用いて説明する。

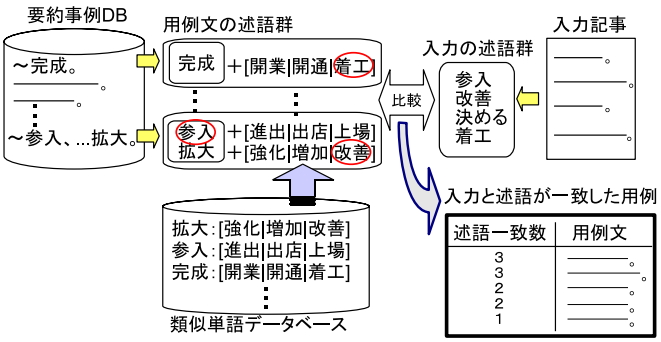

図 2　述語に注目した類似用例文の選択

類似用例文の選択ではまず図 2 に示すように入力記事の述語と要約事例データベースに含まれる用例文の述語で一致数をカウントする。ここで述語が完全に一致しない場合でも使われ方が似ている単語が一致した場合、それらの記事は類似していると判断できたため、これらを述語一致数にカウントできるよう用例文の述語群を拡張した。これには類似単語データベースを使用した。このデータベースについては 4.3 節で説明を行う。

続いてこの入力と述語が一致した用例文の中からさらに内容語に注目して類似用例文の選択を行う。これについては図 3 に示す。

内容語の一致では、図 3 に示すように入力記事の内容語と先程述語の一致で得られた用例文との間で内容語の一致率を計算する。そして述語の一致数が多く、さらに内容語の一致率が最も高かったものを類似用例文として獲得する。

4　文節の対応付け

得られた類似用例文と入力記事間で文節の対応付けを行う。対応付けの単位としては形態素単位が考えられるが、あまりにも小さい単位では対応が取れない場合が存在する。そのため本稿では対応付けの単位として文節に注目した。但し注目している文節に対して連体修飾部が存在する場合にはその連体修飾部は被修飾文節と連結させる。

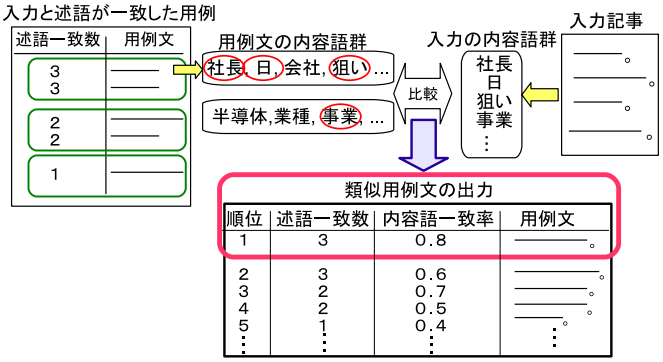

図 3　内容語に注目した類似用例文の選択

なお文節の対応付けには助詞の一致、固有表現タグの一致、単語間類似度の 3 つの対応付け尺度を用いて類似用例文の文節一つに対して、入力記事の文節複数個を対応付ける。つまり一対多の関係で対応 をとる。次節からこの尺度について説明する。

4.1　助詞の一致

助詞の一致では類似用例文と入力記事の文節で助詞が一致したものを対応付ける。これは主語や目的語など使われ方の似た文節を対応付けるために用いた。

4.2　固有表現タグの一致

固有表現タグの一致では構文解析ツール CaboCha[1] の固有表現タグの出力を使用し、類似用例文と入力記事との間で固有表現タグが同じである語を含む文節を対応付ける。これは固有表現が持つ意味のまとまりが同じであるものを対応付けるために用いた。

4.3　単語間類似度

単語間類似度では類似単語データベースを用いて類似用例文と入力記事との間で使われ方の類似した語、意味の類似した語を含む文節を対応付ける。まず類似単語データベースを説明する。

類似単語データベース

本稿ではあらかじめ新聞コーパス 15 年分を用いて類似単語データベースの構築を行った。単語間の類似度を算出する方法は Lin[4] の手法を用い、統語的に似ている語に注目した。本稿ではデータベース構築の際に使用した新聞コーパスサイズを以前 [3] より大きく変更し、さらに格納する類似単語を類似度上位 20% に絞った。データベース例を例 1 に示す。

例 1)
事務所 ::
　支店 (0.22[*1])、室 (0.22)、センター (0.21)、･･･
値上げ ::
　引き上げ (0.24)、値下げ (0.22)、引き下げ (0.22)、･･･
スポーツ ::
　サッカー (0.16)、ゴルフ (0.15)、ビジネス (0.14)、･･･

単語間類似度による対応付けではこの類似単語データベースを用いて、類似用例文の文節 1 つに対して類似度の高かった入力記事の文節上位 3 つを対応付ける。なお文節の類似度をみるときには文節内に含まれる主辞に着目した。

5　対応文節の組合せ

前章で得られた類似用例文に対する対応文節を組合せて要約文を作成する。組合せにより得られる要約文の理想的な形を以下に示す。

- 類似用例文にしたがって要約文を作成するため、組合せにより得られる要約文は類似用例文に似た文節で構成されていること。
- 組合せにより得られる要約文は文節間の繋がりが良く、読みやすいこと。

上述のような要約文の獲得を目標として、文節の組合せを行う。この組合せの方法を図 4 を用いて説明する。

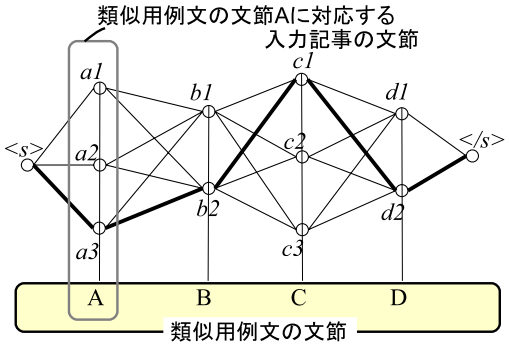

図 4　対応文節の組合せ図の例

図 4 のノード a_i は類似用例文の文節 A に対して得られた入力の対応文節を指す。なお組合せの際には初期状態と最終状態を明確にするため文頭記号 ⟨s⟩ と文末 ⟨/s⟩ を挿入する。

ここでノード n_i に対してのスコア $N(n_i)$ として類似用例文の文節にどれだけ類似しているかを与え、ノード間エッジのスコア $E(n_{i-1}, n_i)$ としてフレーズ間の繋がりの良さを与える。これにより類似用例文の文節により似たもので構成され、且つ日本語として連接の良い部分文節列を得るためにはこのノードとエッジのスコアの和を最大にするようなパスを求める問題に帰着できる。さらに図 4 では文頭から文末に向かう全ての組合せを 2 次元空間に示したものであり、探索領域は限られている。そのためこの問題は動的計画法で効率的に解くことができる。続いて式を用いて具体的にどのような問題を解くのかを考える。

経路列 $W_p = \{n_0, n_1, n_2, \cdots, n_m\}$[*2] に対し、以下のスコアを最大にするような経路を求める問題を考える。このとき最適経路列 $\hat{W}_p$ は以下で与えられる。

$$\hat{W}_p = W_p \quad \text{s.t.} \quad \underset{p}{\text{argmax}}\, Path(W_p) \tag{1}$$

またスコア $Path(W_p)$ を次式で表す。

$$Path(W_p) = \sum_{i=0}^{m} N(n_i) + \sum_{i=1}^{m} E(n_{i-1}, n_i) \tag{2}$$

m は類似用例文の文節の最終番号を表す。以下にノード重みを定義する。

$$N(n_i) = \alpha \cdot particle(n_i) + \beta \cdot NEtag(n_i) + \gamma \cdot MI(n_i) \tag{3}$$

式 3 の $particle(n_i), NEtag(n_i), MI(n_i)$ は以下の式で表される。また α, β, γ は各スコアに対するバランスパラメータである。

$$particle(n_i) = \begin{cases} 1 & \text{ノード } n_i \text{が助詞の一致で} \\ & \text{対応付けされた場合} \\ 0 & \text{それ以外} \end{cases} \tag{4}$$

$$NEtag(n_i) = \begin{cases} 1 & \text{ノード } n_i \text{が固有表現タグの一致で} \\ & \text{対応付けされた場合} \\ 0 & \text{それ以外} \end{cases} \tag{5}$$

$$MI(n_i) = \begin{cases} sim(n_i, ph) & \text{ノード } n_i \text{が単語間の類似度で} \\ & \text{対応付けされた場合} \\ 0 & \text{それ以外} \end{cases} \tag{6}$$

[*1] 「事務所」と「支店」間の類似度を表す。

[*2] 図 4 の太線ならば $W_p = \{\langle s\rangle, a_3, b_2, c_1, d_2, \langle /s\rangle\}$ を通る経路。

表 1 要約文の可読性評価の指標

評価値	評価尺度
評価 1	要約文の文節をほとんど変更せずに流暢に読める。
評価 2	要約文に含まれる文節のうち 25%以上 50%未満 の割合で文節を変更することで流暢に読める。
評価 3	要約文に含まれる文節のうち 50%以上 75%未満 の割合で文節を変更することで流暢に読める。
評価 4	要約文に含まれる文節のうち 75%以上 100%以下 の割合で文節を変更しなくては流暢に読めない。

表 2 要約文の内容の適切性評価の指標

評価値	評価尺度
評価 1	要約文に必要だと考える内容が十分に含まれている。
評価 2	要約文に必要だと考える内容が 50%以上 75%未満で含まれている。
評価 3	要約文に必要だと考える内容が 25%以上 50%未満しか含まれていない。
評価 4	要約文に必要だと考える内容ほとんど含まれていない。

式 (6) 内の ph は類似用例文のある文節を示す。また $sim(n_i, ph)$ は類似用例文の文節 ph と対応付けられた入力の文節との類似度を表す。次にエッジ重みを以下に定義する。

$$E(n_{i-1}, n_i) = \begin{cases} \sigma \cdot \frac{1}{loc(n_i) - loc(n_{i-1}) + 1} \\ \qquad loc(n_i) >= loc(n_{i-1}) \text{ の場合} \\ 0 \qquad \text{それ以外} \end{cases} \quad (7)$$

エッジスコアは文節間の繋がりの良さを示す。本稿では様々な文の文節を組み合せることにより文を作成するが、1 文目 →5 文目 →2 文目 →10 文目の様に 1 つ 1 つの文節があまりにも様々な文を跨いで組合わさるようなものは多くの話題が混在することとなり、連接も悪くなる。そのため式 (7) では文節が存在する入力記事の文の位置 $loc(n_i)$ を考慮した。$loc(n_i)$ はノードつまり対応文節 n_i が入力したニュース記事の何文目に存在しているかという情報である。連接する文節 (n_{i-1}, n_i) がどれだけ離れているかということを $loc(\cdot)$ の差の絶対値を取ることで測っている。このとき文節が文頭に向かって (4 文目 →2 文目のように) 戻る場合は話題が戻ることとなり、連接をより悪くしてしまう可能性があるため、このような場合にはスコア 0 を与えている。上述の方法により出力する要約文を獲得する。

6 要約結果

本章では提案手法の有効性を確認するため従来法である Hori[5] の手法と比較実験を行った。

6.1 使用するデータ

要約事例データベース

要約事例データベース内の用例には日経ニュースメール Nikkei-goo[2] から配信されているニュースの要約文を用いた。このニュース要約文は人手で作成されているものであり、1999 年 12 月から 2007 年 12 月までに収集した 27036 件を用いた。1 文あたりの形態素平均数は 23.1 形態素、文節平均数は 6.6 文節である。

テストデータ

日本経済新聞 1998 年 [3] のデータ 100 件を用いた。この 100 件は 1 記事あたりの文数が 3 文以下で比較的短いものを中心にテストを行った。このデータの詳細は 1 記事当たりの平均文数が 2.5 文、平均形態素が 35.4 形態素、平均文節が 10.2 文節である。

6.2 評価方法

評価者 3 人に教示を与え、可読性の評価及び内容適切性の評価の 2 点について評価を行った。可読性の評価では評価者それぞれがシステムが出力した要約文を読み表 1 の指標に基づいて 4 段階評価を行った。

内容の適切性評価では可読性の評価を行った同じ評価者 3 人が入力のニュース記事とシステムが出力した要約文を読んで、表 2 の指標に基づいて内容の適切性評価を行った。

6.3 評価結果

可読性の評価結果

表 3 と表 4 に評価者が可読性評価した結果を示す。可読性の評価は 1-4 の 4 段階で 1 が良好であり、4 が不良である。表 3 は従来手法が出力した要約文を評価した結果であり、表 4 は本手法の結果を表している。

表 3 従来手法が出力した要約文の可読性評価

	評価者 A	評価者 B	評価者 C
評価値 1	50	17	51
評価値 2	22	38	16
評価値 3	12	35	21
評価値 4	16	10	12
平均値	1.94	2.38	1.94

表 4 本手法が出力した要約文の可読性評価

	評価者 A	評価者 B	評価者 C
評価値 1	66	79	88
評価値 2	26	17	10
評価値 3	6	3	2
評価値 4	2	1	0
平均値	**1.44**	**1.26**	**1.14**

表 5 において、評価者 3 人全員が評価値 1(良好) を付与したのが入力データ中にどれ程存在するか、また評価者 2 人以上が評価値 1 を付与したのがどれ程存在するか調査した結果を示す。

表 5 評価者の人数と可読性の評価値 1(良好) の件数

評価値 1 を付与した評価者の人数	従来手法	本手法
3 人共に評価値 1 を付与	11	**52**
2 人以上が評価値 1 を付与	34	**84**

表 5 より本手法において、評価者 3 人が共に評価値 1 を付与した入力データの件数は 52 件という結果が得られた。またいずれの場合も本手法は従来手法の結果よりも優位な結果が得られたことが分かる。評価者の過半数が最も良好である評価値 1 を付与したものを正解とすると本手法の正解率は 84%を獲得できたこととなる。

内容適切性の評価結果

表 6 と表 7 に評価者が内容適切性を評価した結果を示す。6 は従来手法が出力した要約文を評価した結果であり、表 7 は本手法の結果を表している。

続いて表 8 において、評価者 3 人全員が良好である評価値 1 を付与したのが入力データ中にどれ程存在するか、また評価者 2 人以上が評価値 1 を付与したのがどれ程存在するか調査した結果を示す。

表 8 より本手法において評価者 3 人が共に評価値 1 を付与した入力データの件数は 26 件いう結果が得られた。またいずれの場合も本手法は従来手法の結果よりも優位な結果が得られたことが分かる。評価者の過半数が最も良好である評価値 1 を付与したものを正解とすると本手法の正解率は 43%を獲得できたこととなる。また表 7 をみると評価者の多くは評価値 1 と評価値 2 の合計が 50%を超えている。そのため、正解率が 43%であると言っても不正解である 57%の多くは評価値 1 に近いことが分かる。

表 6 従来手法が出力した要約文の内容適切性の評価

	評価者 A	評価者 B	評価者 C
評価値 1	20	6	17
評価値 2	19	17	40
評価値 3	36	32	28
評価値 4	25	45	15
平均値	2.66	3.16	2.41

表 7 本手法が出力した要約文の内容適切性の評価

	評価者 A	評価者 B	評価者 C
評価値 1	48	33	71
評価値 2	25	33	17
評価値 3	19	26	4
評価値 4	8	8	8
平均値	**1.87**	**2.09**	**1.49**

表 8 評価者の人数と内容適切性の評価値 1(良好) の件数

評価値 1 を付与した評価者の人数	従来手法	本手法
3 人共に評価値 1 を付与	3	**26**
2 人以上が評価値 1 を付与	11	**43**

7 追加実験：要約文の選定

7.1 要約文の選定方法

本手法は入力と内容の類似した用例文に模倣して 1 文の要約文を作成するのだが、類似用例文 1 位のものから作成した要約文だけが良好なものなのか調査した。調査では類似用例文上位 x 件を獲得し、その各類似用例文からそれぞれ要約文を出力する。今回は上位 10 件の類似用例文を使用した。

この調査ではまず評価者 2 人に要約課題とシステムが出力した要約文 (上位 10 件) を与えた。そして評価者各々が可読性の評価で評価値 1(表 1 参照)、内容適切性の評価でも評価値 1(表 2 参照) を獲得した要約文にチェックを付けた。表 9 に 100 件の要約課題で各々上位 10 件まで要約文を出力したときの正解の有無を示す。

表 9 上位 10 件中での正解の有無

	評価値 1 となった件数 (100 件中)
評価者 A	100
評価者 B	69

これにより上位 10 件まで要約文を出力したときほとんどの場合は正解が含まれていることが分かる。そこで本稿では要約文を作成してから今まで用いたものと別の尺度で要約文の良さを測り、再ランキングを行う。今回用いた尺度を以下に示す。

$$Eval(s) = w_1 \cdot 2gram_{back}(s) + w_2 \cdot skip2gram_{back}(s) + w_3 \cdot pos2gram(s) \quad (8)$$

式 (8) での $2gram_{back}(s)$ は文末から見た後向き 2gram の平均値が上位 10 位の要約文のうち何位であるかという情報である。また $skip2gram_{back}(s)$ は後向き skip2gram の平均値の順位、$pos2gram(s)$ は品詞 2gram 平均値の順位である。後向き skip2gram では 2 つのギャップまでを許した 2gram としている。本稿では後向き 2gram や後向き skip2gram、品詞 2gram の各平均値に大きな差があるため値そのままは使わず、上位 10 位の要約文中何位であるかという順位を用いた。また式 (8) の w_1, w_2, w_3 は各項の重みである。この重みはあらかじめ用意したテストデータとは別の要約課題 100 件を用いて調整を行った。調整による結果では $2gram_{back}$ の重み w_1 が 0.2、$skip2gram_{back}$ の重み w_2 が 0.1、$pos2gram$ の重み w_3 が 0.5 であった。

7.2 評価実験

評価実験では 6.1 節と同じテストデータ 100 件を用いて、各々の要約課題に対して上位 10 位までの要約文を出力したものを再ランキングする。これら上位 10 位までの要約文には 7.1 節で述べたように評価者による正解要約文のチェックが存在する。評価者 A では上位 10 位まで要約文を出力した場合、100 件の要約課題中で 100 件全てに対して正解が存在するとしている。また評価者 B では 100 件中 69 件の正解が存在するとしている。

これらのデータを使用して再ランキングする前の上位 1 位の要約文の正解数と上位 10 位まで出力して再ランキングを行った後、1 位になった要約文の正解数を表 10 に示す。

表 10 再ランキング前後での正解数の変化

	再ランキング前	再ランキング後
評価者 A	36/100(36%)	54/100(64%)
評価者 B	23/69(33%)	38/69(55%)

表 10 に示した通り、再ランキングの後では正解数が増えたことが分かる。よって本手法による再ランキングの有効性を確認することができた。

8 まとめ

本稿では重要度の設定を行わずに要約文を作成することを目的とし、複数の文からの文節を組合せることで要約文を作成する方法を述べた。また複数文からの文節の抽出、組合せ方については要約事例データベースの中にある用例文に従った。これにより要約文として読みやすく、さらに内容も適切である要約文の作成を目指した。評価実験では従来法の 1 つ比較手法としてとりあげ、人手による評価を行った。この結果、どの評価者に対しても本手法の方が有効である結果が得られた。またさらに可読性の高い要約文を目指し、複数の要約文を出力し再ランキングする手法についても述べた。この再ランキングによりさらに精度を向上させることができることを示した。

謝辞

本研究の一部は、 科学研究費補助金 基盤 (A)「円滑な情報伝達を支援する言語規格と言語変換技術」課題番号 16200009 によって実施した。

使用したツール及び言語資源

1) 構文解析器 CaboCha, Ver.0.53, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/˜taku/software/cabocha/
2) 日経ニュースメール, NIKKEI-goo, http://nikkeimail.goo.ne.jp/
3) 日本経済新聞全記事データベース 1990-2004 年度版, 日本経済新聞社

参考文献

1] Klaus Zechner. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. In *Proc. of the COLING96*, pp. 986 – 989, 1996.

2] 小黒玲, 尾関和彦, 張玉潔, 高木一幸. 文節重要度と係り受け整合度に基づく日本語文簡約アルゴリズム. 自然言語処理, Vol. 8, No. 3, pp. 3–18, 2001.

3] Megumi Makino and Kazuhide Yamamoto. Summarization by Analogy: an Example-based Approach for News Article. In *Proc. of IJCNLP08*, pp. 739–744, 2008.

4] Dekang Lin. Automatic Retrieval and Clustering of Similar Words. In *Proc. of COLING-ACL98*, pp. 768–774, 1998.

5] Chiori Hori. *A Study on Statistical Methods for Automatic Speech Summarization*. PhD thesis, Tokyo Institute of Technology, 2002.