会話によるニュース記事伝達のための情報選択

高津 弘明[1] 福岡 維新[1] 藤江 真也[1,2] 林 良彦[1] 小林 哲則[1]

早稲田大学[1] 千葉工業大学[2]

{takatsu,fukuoka}@pcl.cs.waseda.ac.jp, shinya.fujie@p.chibakoudai.jp,
yshk.hayashi@aoni.waseda.jp, koba@waseda.jp

1 はじめに

インターネットの発展に伴い、インターネット上には大量の文書が蓄積されるようになった。文書には、多くの情報が含まれているが、それらは単なる情報の羅列ではなく、それぞれが役割を持ち、互いに関係付けられて配置された情報の塊とみることができる。この情報の塊を効果的に伝える方法として、文書から適当な情報を選び、会話的に伝えることが考えられる [1]。本稿では、ニュース記事を取り上げ、これを会話で伝えるときに必要となる情報選択手法について検討する。

文書閲覧のような視覚メディアの情報アクセスでは、読み手は必要な情報だけを効率的に選びながら読み進めていくことができる。これに対し、聴覚メディアにはこのような拾い読みに相当する手段がない。このため、未加工のまま情報を伝えたのでは、聞き手はたとえ興味がない内容であっても最後まで黙って聴くしかなく、フラストレーションが大きい。

そこで、要約した内容から発話計画を立て会話でインタラクティブに伝達する方法が考えられる [1]。適度に情報を落としながらも聞き手にニュースの面白さを伝えることができれば、興味を持った聞き手からの質問を誘発することができ、ユーザーのニーズに合わせた情報伝達を実現できる。しかし、伝える要約内容によっては、ユーザーは新たな情報 (興味を抱くかもしれない情報) に遭遇する機会を失ってしまう。そのため、会話での情報伝達を意図した要約では、聞き手が質問により情報を補完できることも想定して要約を作る必要がある。

そこで、会話でインタラクティブに伝えることを前提として要約コーパスを設計し、会話で主として伝える内容の選定を (1) 重要文抽出、(2) 整列、(3) 文圧縮の 3 段階で行った。それぞれの段階で人間の作業者が作成した要約に近い要約を生成する方式について検討し、評価実験を行った。

以下、2 章で設計した要約コーパスについて説明し、3 − 5 章で各段階における手法と実験結果を述べる。

2 要約コーパス

次のニつの観点で要約を作成した。一つは、ニュースの見出し的な内容またはニュース記事の核となる情報を 1〜4 文選択したものである (以降、要約 A)。もう一つは、ユーザーの関心を引けそうな内容または記事への理解を促進させる内容を要約 A に加えて 1〜3 文選択したものである (以降、要約 B)。文選択を行った後、作業者は選択した文から不要な情報を文節単位で取り除く。文節の区切りは KNP で与えた。100 トピックについて 3 人の作業者が要約を行った。ニュース記事は毎日新聞、産経新聞、読売新聞、NHK、nifty から収集したものを利用した。トピックは同じ話題について記述した 2 つ以上の記事から構成され、各記事は 5 文以上のものを選んだ。なお、このコーパスでは、作成した要約の口語化も行っている [2]。

	要約
記事1の文1 (A,B)	テニスのメキシコ・オープンは28日、メキシコのアカプルコで行われ、男子シングルス決勝で第1シードの錦織圭は第2シードのダビド・フェレールに3―6、5―7のストレートで敗れ、ツアー通算9勝目はならなかった。
記事1の文2 (A,B)	世界ランキング5位の錦織は、3月2日に発表される最新ランキングで自己最高の4位に浮上することが確定している。
記事3の文3 (B)	女子シングルス決勝はティメア・バシンスキーがキャロリン・ガルシアに6―3、6―0で勝ち、6シーズンぶりのツアー2勝目。

図 1: 要約例

要約例を図 1 に示す。このトピックでは、錦織圭がメキシコオープンで負けたことが主題であるが、女子シングルスの結果も新たな着眼点として含めている。

3 重要文抽出

ここでの目的は、要約に含めるべき文を抽出することである。近年、複数文書要約を最大被覆問題で定式化することが多い。最大被覆問題に基づく要約モデル (最大被覆モデル) は、被覆した概念 (ユニグラムやバイグラム) の重要度の和が最大となる文集合を選択する手法である。最大被覆問題は NP 困難な問題である。Filatova らは単語を概念とし、重要度として TF-IDF を与え、貪欲法で解いた [3]。Yih らは文書集合での出現頻度に比例する重みによる単語の重み付けと訓練データで学習したロジスティック回帰の確率値を利用した単語の重み付けを行い、スタック・デコーディングを用いて最大被覆問題を解いた [4]。高村らは最大被覆問題を整数計画問題として厳密に定式化し、様々なデコーディング方法を適用し、包括的な比較実験を行った [5]。Gillick らも同様に整数計画問題として定式化を行っている [6]。一方、文書要約をナップサック問題として定式化したモデルとして McDonald のモデルがある [7]。McDonald は文書要約を主題への関連性から冗長性 (文同士の類似度) を差し引いた目的関数で定式化した。

最大被覆モデルは、制限長内でできるだけ多くの内容を含めようとするモデルであり、自然と冗長性が削減される。しかし、文同士の関係や主題への関連度が

表 1: 変数の定義

$x_{(i,m)}$	文 (i,m) が選択されたかどうか
$y_{(i,m)(j,n)}$	文 (i,m) と文 (j,n) が両方選択されたかどうか
z_l	単語 l が被覆されたかどうか
K	最大要約長
$a_{(i,m)l}$	文 (i,m) が単語 l を含むかどうか
b_l	単語 l の重要度
$c_{(i,m)}$	文 (i,m) の長さ (要約長が文数の場合は 1)
$r_{(i,m)}$	文 (i,m) の主題への関連度
$s_{(i,m)(j,n)}$	文 (i,m) と文 (j,n) の類似度
M_i	記事 i の文の数
I	記事の総数
L	単語の総数

$$\max. \quad \sum_{l=1}^{L} b_l z_l \times \left(\sum_{i=1}^{I} \sum_{m=1}^{M_i} r_{(i,m)} x_{(i,m)} - \sum_{(i,m)<(j,n)} s_{(i,m)(j,n)} y_{(i,m)(j,n)} \right) \quad (1)$$

s.t.

$$\forall i,j,m,n,l: \quad x_{(i,m)} \in \{0,1\}, \ y_{(i,m)(j,n)} \in \{0,1\}, \ z_l \in \{0,1\} \quad (2)$$

$$\sum_{i=1}^{I} \sum_{m=1}^{M_i} c_{(i,m)} x_{(i,m)} \leq K \quad (3)$$

$$\forall l: \quad \sum_{i=1}^{I} \sum_{m=1}^{M_i} a_{(i,m)l} x_{(i,m)} \geq z_l \quad (4)$$

$$\sum_{i=1}^{I} x_{(i,1)} = 1 \quad (5)$$

$$\forall i,j,m,n: \quad y_{(i,m)(j,n)} - x_{(i,m)} \leq 0 \quad (6)$$

$$\forall i,j,m,n: \quad y_{(i,m)(j,n)} - x_{(j,n)} \leq 0 \quad (7)$$

$$\forall i,j,m,n: \quad x_{(i,m)} + x_{(j,n)} - y_{(i,m)(j,n)} \leq 1 \quad (8)$$

考慮されていない。そこで、最大被覆モデル (MCM) と McDonald のモデル (以降、関連性モデル;Rel) を組み合わせたモデル (MCM+Rel) を提案する (式 1)。各変数の説明を表 1 に示す。

制約として、ほぼ全ての作業者がニュース記事の一文目を要約に含めていたことから、複数記事の内いずれかの記事の一文目を必ず要約に含めるような制約 (式 5) を加える。最適化問題は分枝限定法で解く。

動詞、名詞、形容詞を対象として単語の重要度 b_l を計算する (ただし、形式名詞、副詞的名詞、“する”は除いた)。単語の重みは、頻度 (Freq) や TF-IDF、RandomForest で与える。実験では次の 2 つの観点で RandomForest のパラメータを学習した。一つは、作業者が対象単語を要約に含めたかどうかの二クラス分類問題として学習し、単語が要約に含まれる確率を単語の重みとしたもの (RFC)。もう一つは、複数の作業者が要約に含めた単語の重みが高くなるように回帰で学習したものである (RFR)。3 人の作業者が要約に含めた場合の単語の重みを 3/3 とし、2 人の作業者が要約に含めた場合の単語の重みを 2/3、1 人の作業者が要約に含めた場合の単語の重みを 1/3 とする。素性として、文書集合における単語の頻度や単語の TF-IDF 値、タイトルに含まれるかどうか、固有表現かどうかなどの情報を利用した。

主題への関連度 $r_{(i,m)}$ は、$r_{(i,m)} = 1/(\text{文 } (i,m) \text{ の出現位置}) + (\text{文 } (i,m) \text{ と文書集合の類似度})$ で与える。ここで、文 (i,m) は記事 i の m 番目の文を表す。文と文書集合の類似度と文間の類似度 $s_{(i,m)(j,n)}$ は、単語の Bag-of-Words(BoW)、単語の TF-IDF ベクトル (TF-IDF)、300 次元のパラグラフベクトル (PV-DM) [8] のいずれかを用いたコサイン類似度で与える。

実験では、要約 A,B のデータセットを使用し、ROUGE-1 による評価実験を行った。結果を表 2,3 に示す。ただし、最大要約長 K は作業者が選択した文数とする。また、MCM(RFC) と MCM(RFR)、および、MCM+Rel(RFC) と MCM+Rel(RFR) は、10 分割交差検定で単語の重みを計算し、各 ROUGE-1 の値を平均した値である。実験結果は、提案手法 (MCM+Rel) が最も高い ROUGE-1 を示した。提案手法の ROUGE-1 の値は、およそ 80%から 85%程度であり、このことから、提案手法を用いることで、要約に含めるべき内容の 80%から 85%を抽出できることが分かった。単語の重みの与え方に関する比較では、RFR の方が良い結果を示し、関連性モデル (Rel) における類似度の与え方に関する比較では、BoW が最も良い結果を示した。

表 2: 重要文抽出の ROUGE-1 による評価 (要約 A)

	作業者 1	作業者 2	作業者 3
MCM (Freq)	73.7	76.8	78.7
MCM (TF-IDF)	66.2	70.7	74.5
MCM (RFC)	76.2	81.9	82.9
MCM (RFR)	77.5	81.4	84.4
Rel (BoW)	73.6	80.3	82.9
Rel (TF-IDF)	72.6	80.8	82.6
Rel (PV-DM)	71.4	76.3	81.5
MCM+Rel (RFC,BoW)	78.7	82.0	84.8
MCM+Rel (RFR,BoW)	**79.3**	**83.1**	**84.9**

表 3: 重要文抽出の ROUGE-1 による評価 (要約 B)

	作業者 1	作業者 2	作業者 3
MCM (Freq)	76.2	78.6	80.4
MCM (TF-IDF)	74.5	76.0	79.8
MCM (RFC)	82.1	83.4	84.8
MCM (RFR)	82.6	83.7	85.3
Rel (BoW)	72.2	74.8	78.1
Rel (TF-IDF)	69.2	74.4	74.5
Rel (PV-DM)	69.9	74.7	76.0
MCM+Rel (RFC,BoW)	83.1	84.2	85.8
MCM+Rel (RFR,BoW)	**83.8**	**84.4**	**86.2**

表 4: 重要文整列実験の正解率

	要約 A			要約 B		
	一位	MRR	五位	一位	MRR	五位
作業者 1	95.0	96.0	97.0	62.0	73.8	94.0
作業者 2	94.0	96.5	99.0	82.0	87.6	97.0
作業者 3	100	100	100	97.0	98.0	100

4 整列

重要文抽出で抽出した文の提示順序を決定する。単一文書要約の場合、記事における文の出現順序をそのまま利用できるが、複数文書要約の場合、異なる記事から文が選択される可能性があるため、要約内での文の提示順序を決める必要がある。

文の整列問題は、主にテキスト生成の分野で取り組まれてきたが、近年では、自動要約の分野でも重要な課題として認識されている。岡崎らは記事が書かれた時間情報を利用して並び替える手法を提案した [9]。Lapata は文中での動詞や名詞とその他の係り受け関係を考慮した統計的なモデルに基づいて文の順序を決定する手法を提案した [10]。最近では、RNN 言語モデルを用いて文の整列を試みる研究も行われており、Lin らは文間の結束性を考慮した RNN 言語モデル (Hierarchical Recurrent Neural Network Language Model) を提案し、HRNNLM が文整列タスクにおいて、最大エントロピー法に基づく手法や再帰ニューラルネットワークに基づく手法よりも優れていることを示した [11]。

我々は、コーパスにおいて同じ記事から選ばれた文の順序関係が保たれていたことから、ある記事の文が別の記事のどの文に該当するかをコサイン類似度に基づいて計算し、最もスコアが高くなる順序を採用するという方法をとった。

$$sequence = \max_{seq \in S} Score(seq) \quad (9)$$

$$Score(seq) = \sum_{(i,m)\to(j,n)\in seq} f((i,m),(j,n)) \quad (10)$$

$$f((i,m),(j,n)) = \quad (11)$$

$$\begin{cases} g(\max_{l>m} \mathrm{sim}((i,l),(j,n)), \max_{l<n} \mathrm{sim}((i,m),(j,l))) & (12) \\ g\left(\mathrm{sim}((i,m),\mathrm{before}(j,n)), \mathrm{sim}(\mathrm{after}(i,m),(j,n))\right) & (13) \end{cases}$$

ここで、S は文の順序候補の集合、$f((i,m),(j,n))$ は文 (i,m) が文 (j,n) よりも前に来うるスコア、g,h は $\sum$ または max、sim は BoW のコサイン類似度を表す。また、$\mathrm{before}(j,n)$ は記事 j の文 n よりも前に出現する文の BoW を表し、$\mathrm{after}(i,m)$ は記事 i の文 m よりも後ろに出現する文の BoW を表す。制約として、いずれかの記事の一文目が最初になるように定めた。

要約 A,B のデータセットに対する実験結果を表 4 に示す。ここには、式 13、g,h として max を用いたときの結果を示すが、他の設定でもさほど結果に違いはなかった。要約 B における作業者 1 と作業者 2 の一位正解率は他と比べて低かったが、要約 A と要約 B の作業者 3 の正解率は 90%以上であった。

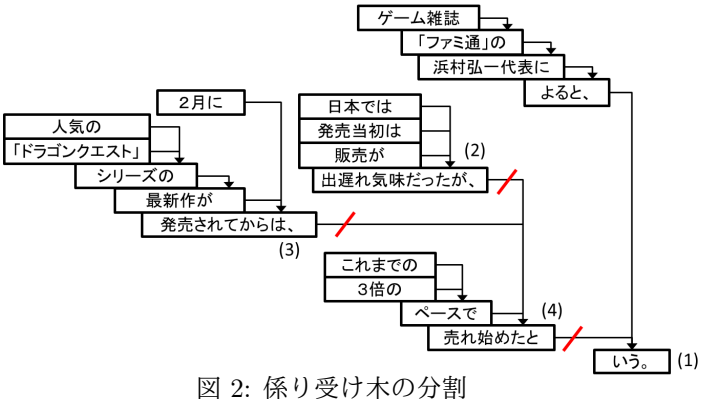

図 2: 係り受け木の分割

5 文圧縮

重要文抽出により、要約に含めるべき文は選択されたが、一文一文は長く会話で話す内容としては冗長である。ここでの目的は抽出された重要文から不要な箇所を取り除き、文自体を短く要約することである。

文圧縮の技術は、文字放送や字幕作成、限られた表示リソースしか持たない携帯端末でのテキスト表示を支援する技術として注目されている。字幕作成では、表示する機器の大きさや人間の読むスピードに制限があることから、体言止めや漢字熟語を多用した表現への書き換えなどが行われる [12]。しかし、本研究では、会話で話すための話し言葉への書き換えを目的としているため、このような書き換えは行わず、ここでは、不要な文節の削除のみを行う。

統計的な文圧縮の手法として、Knight と Marcu は、Noisy-Channel モデルと決定木による文圧縮を行った [13]。野本は、係り受け木の刈り込みにより要約文候補を生成し、CRF を用いて選別する手法を提案した [14]。この手法はあらかじめ妥当な要約文候補を生成しておくことで、文が不自然なところで切れたり、読解に必要な文法要素を欠落させてしまうリスクを軽減している。本研究でも、野本の手法を参考に CRF による文圧縮を行うが、彼が導入した Terminating Dependency Path (TDP) による要約文候補の生成では、正解要約文が生成されにくいという問題があったため、独自の手法で要約文候補を生成する。

要約文候補の生成

(i) 係り受け木を述語を中心とした部分木に分割する。例えば、図 2 のような係り受け木は (1)〜(4) の部分木に分割できる。ただし、(1) の「いう」の必須格「と」が (4) の述部に含まれているため、要約文候補に (1) を含めるときは (4) も含める。

(ii) 各部分木を複製し、規則に基づいて枝刈りする。例えば、部分木 (2) から「2 月に」という文節を刈り込む。ただし、必須格や唯一格は必ず含める。また、同格の文節は格助詞を継承して係り先を削除できるものとする (e.g. タクシー運転手／伊藤隆行容疑者が ⇒ タクシー運転手が)。

(iii) 枝刈りした部分木そのもの、または、それらを統合したものを要約文候補とする。例えば、(3) と (4) の組

表 5: 生成された要約文候補に正解が含まれる割合

	要約 A	要約 B
作業者 1	76.4	78.5
作業者 2	87.7	88.1
作業者 3	88.1	91.1

表 6: 文圧縮実験の正解率 (要約 A)

	全データ			正解を含むデータのみ		
	一位	MRR	十位	一位	MRR	十位
作業者 1	40.7	49.2	64.1	52.9	63.9	83.5
作業者 2	62.0	65.3	73.5	70.5	74.3	83.6
作業者 3	65.3	69.7	79.5	73.9	79.0	90.6

表 7: 文圧縮実験の正解率 (要約 B)

	全データ			正解を含むデータのみ		
	一位	MRR	十位	一位	MRR	十位
作業者 1	46.8	54.7	68.3	59.7	69.7	87.0
作業者 2	65.5	68.4	76.5	74.5	77.8	86.9
作業者 3	73.3	77.1	85.9	80.9	85.2	94.9

み合わせから、「「ドラゴンクエスト」／シリーズの／最新作が／発売されてからは、これまでの／ 3 倍の／ペースで／売れ始めた」という要約文候補が生成される。

この手法によって生成された要約文候補に正解が含まれる割合を表 5 に示す。正解要約文を生成できなかったものの多くは、係り受け誤りによるものであったが、中にはゼロ代名詞の推定を必要とするケースも存在した。

要約文候補のランキング

要約に含める文節のラベルを I、含めない文節のラベルを O として、要約 A,B の各コーパスで CRF のパラメータを学習する。素性として、文節が文頭か文末か、葉ノードかどうか、必須格を持つかどうか、文節の最初と最後の形態素情報、文節の主辞がタイトルに含まれているかどうかなどを利用した。要約に含める文節のスコアには I の確率を、含めない文節のスコアには O の確率を使用し、それらの積で、要約文候補のスコアを与える。その値でランキングし、10 分割交差検定で正解率を算出した結果を表 6,7 に示す。当然だが、全データを使用した場合よりも正解を含むデータのみを使用したときの方が良い結果を示した。また、要約 B の方が要約 A よりも高い正解率を示したが、これは要約 B の方がデータ数が多いためだと考えられる。

6 おわりに

まとまりのある情報を伝える場合、伝える情報が多すぎるとユーザーは興味がない内容を最後まで黙って聞いていなければならない。一方、伝える情報が少なすぎるとユーザーが新たな情報 (興味を抱くかもしれない情報) に遭遇する機会を奪ってしまう可能性がある。そこで、見出し的な内容だけでなく、ユーザーの関心を引けそうな内容も加味した要約コーパスを設計した。見出し的な内容からなる要約を要約 A とし、要約 A にユーザーの関心を引けそうな内容を加えたものを要約 B とする。このコーパスに対して自動要約手法を適用して結果を比較したところ、重要文抽出に関しては、要約 B のデータセットの方が要約 A よりも高い ROUGE-1 の値を示した。また、文圧縮に関しても、要約 B のデータセットの方が高い正解率を示した。一方、整列に関しては要約 A の方が高い正解率を示した。結果を総合的に判断すると、ユーザーの関心を引けそうな内容も加味した要約に関しても見出し的な内容のみからなる要約と同様の自動要約手法で同程度の結果が得られることが分かった。

ここで生成された要約は、口語化処理で話し言葉に書き換えられ [2]、音声対話システムの発話主計画として利用されている [1]。

参考文献

[1] 高津弘明, 福岡維新, 藤江真也, 林良彦, 小林哲則: “快適な情報享受を可能とする音声対話システム”, 言語処理学会第 22 回年次大会発表論文集, 2016.

[2] 高津弘明, 福岡維新, 藤江真也, 林良彦, 小林哲則: “会話によるニュース記事伝達のための口語化における述語の書き換え”, 言語処理学会第 22 回年次大会発表論文集, 2016.

[3] E.Filatova and V.Hatzivassiloglou: “A Formal Model for Information Selection in Multi-Sentence Text Extraction”, in Proceedings of the 20th International Conference on Computational Linguistics, pp.397403, 2004.

[4] W.Yih, J.Goodman, L.Vanderwend, and H.Suzuki: “Multi-Document Summarization by Maximizing Informative Content-Words”, in Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp.17761782, 2007.

[5] H.Takamura and M.Okumura: “Text Summarization Model based on Maximum Coverage Problem and its Variant”, in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL’09, pp.781-789, 2009.

[6] D.Gillick, and B.Favre: “A Scalable Global Model for Summarization”, in Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing, pp.1018, 2009.

[7] R.McDonald: “A Study of Global Inference Algorithms in Multi-Document Summarization”, in Proceedings of the 29th European Conference on Information Retrieval, pp.557564, 2007.

[8] Q.Le and T.Mikolov: “Distributed Representations of Sentences and Documents”, in Proceedings of The 31st International Conference on Machine Learning, pp.11881196, 2014.

[9] N.Okazaki, Y.Matsuo and M.ishizuka: “TISS: An Integrated Summarization System for TSC-3”, in Working Notes of the Fourth NTCIR Workshop Meeting, pp.436-443, 2004.

[10] M.Lapata: “Probabilistic Text Structuring: Experiments with Sentence Ordering”, in Proceedings of the 41st Meeting of the Association of Computational Linguistics, pp.545-552, 2003.

[11] R.Lin, S.Liu, M.Yang, M.Li, M.Zhou and S.Li: “Hierarchical Recurrent Neural Network for Document Modeling”, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.899-907, 2015.

[12] 江原暉将, 和田裕二, 福島孝博: “聴覚障害者向け字幕放送における自動要約”, 情報処理, テキスト自動要約特集, Vol.43, No.12, pp.1305-1309, 2002.

[13] K.Knight and D.Marcu: “Summarization beyond sentence extraction: A probabilistic approach to sentence compression”, Artificial Intelligence 139, pp.91-107, 2002.

[14] T.Nomoto: “A Generic Sentence Trimmer with CRFs”, in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Human Language Technologies, Columbus, pp.299-307, 2008.