重要箇所同定用コーパスの構築

–New York Times Annotated Corpusの文書要約資源化に向けて–

菊池 悠太 † 渡邉 亮彦 † 高村 大也 ‡ 奥村 学 ‡

† 東京工業大学 大学院総合理工学研究科，‡ 東京工業大学 精密工学研究所

{kikuchi, watanabe}@lr.pi.titech.ac.jp,{takamura, oku}@pi.titech.ac.jp

1 はじめに

文書要約は，与えられた単一あるいは複数の原文書 (集合) から，一つの要約を作成するタスクである．要約は，要約者がどの視点で作成したかによって大きく二種類に分けられる．以下に二つの具体例を示す[1]．

(a) Second annual Family Television Awards are presented by coalition of advertisers interested in stimulating production of prime-time television programsthat appeal to broader multigenera tional audiences.

(b) Daniel Gross Economic View column on economic advantages of congestion pricing, particularly as means of unclogging urban streets; some economists say now is time to expand concept; photo; graph.

要約例 (a) は，原文書の著者と同じ視点で要約が書かれている．つまり，要約そのものが原文書と同じ主張をしている．一方で，要約例 (b) は，原文書そのものの情報やその内容を客観的に説明するような論調で，原文書の著者の視点とは異なった第三者の視点 (しばしば要約者自身の視点) に立って書かれた要約である．本稿では，(a) のような要約を**概要的要約**，(b) のような要約を**説明的要約**と呼び両者を区別する．両者はそれぞれが強い特徴を持っており，各々の違いを意識して分析を行うことが重要である．また，それぞれに対し要約手法が担うべき技術も大きく異なる．しかしながら，文書要約に用いるコーパスには，これらの分類が明示的に付与されていることは少ない．

従来の文書要約について取り組んだ研究の多くは，概要的要約の作成であるといえる．このときしばしば議論される要約の分類は，**抽出型要約手法**と**生成型要約手法**という要約の作成方法に焦点を当てたものである．抽出型要約手法は顕在性スコアに基づいて原文書の一部を抜き出す．抜き出す単位は文や単語，句など様々な単位を考えることができる．生成型要約手法は原文書に存在しない表現を用いて高度な言い換えや内容の抽象化を行う．どちらも原文書の視点を保ちつついかに短くするかという共通の前提を持つ．

現在，人手による要約（参照要約）が付与された最も大規模なコーパスの一つは，New York Times Annotated Corpus[2](NYTAC) である．NYTAC を構成する 180 万記事のうち 64 万記事に専門家による参照要約が付与されている．この規模の要約付きコーパスは他に類を見ないが，本コーパスにも上述のような細分された参照要約の情報は付与されていない．具体的には，参照要約が概要的か説明的か，またその作成方法が抽出型か生成型かという情報である．

本研究の目的は，NYTAC の各参照要約に細分化された情報を付与することである．これにより，文書要約に関するより詳細な分析や信頼性の高い定量評価，大規模データによる構造学習などが可能となる．本稿では本研究の最初の段階として，本コーパスの一部から重要箇所同定に特化した訓練，評価用サブコーパスの構築を行う．

2 重要箇所同定用サブコーパス

本節では，本稿で我々が構築する重要箇所同定用のサブコーパスについて説明する．

2.1 対象とするNYTACの記事

NYTAC は米国の New York Times 社が過去に発行した記事にアノテートを行なった約 180 万記事からなるコーパスである．そのうち約 64 万記事には専門家が作成した要約が付与されている．NYTAC は非常に大規模なコーパスであるため，今回は一部の記事のみを取り扱う．具体的には，“types_of_material” という記事種別を表すメタ情報のうち，「記事の訂正」や「人

[1] 二つの要約はそれぞれ別の文書を要約したものであり，内容は一致しないことに注意されたい．

[2] https://catalog.ldc.upenn.edu/LDC2008T19

処理	処理前	処理後
抽出	The Turkish Government should be encouraging, not silencing, those who support a political solution to the Kurdish crisis.	says Turkish Govt should be encouraging, not silencing, those support political solution to Kurdish crisis.
圧縮	Over the next two weeks, the justices are to hear four cases that will help define the true dimensions of the law.	note that justices will hear four cases that will help define true dimensions of the law.
融合	[文 1] Studies consistently show that those obtaining Chapter 7 protection are truly in dire straits. [文 2] On the broader issue, there is scant evidence that bankruptcy abuse is rampant.	says there is scant evidence that bankruptcy abuse is rampant, and studies consistently show that those obtaining Chapter 7 protection are truly in dire straits.
改変	Editorial contends that China should be admitted to World Trade Organization if it is prepared to honor international fair-trade rules.	If China is prepared to honor international fair-trade rules, it should be admitted to the World Trade Organization.

図 1: 原文から抽出型処理により作成された要約文の例.

物の経歴など」特殊な記事種別を除いた上でその総数が最も多かった「社説 (Editorial)」を対象とした．将来的には，NYTAC の文書要約資源化に向けて，本研究により得られた知見から 64 万記事すべての要約に対し処理を行う．

2.2 文の種類

要約に含まれる各文は，以下のいずれかの処理により作成されている．**抽出**は，原文書中の文（原文）をほぼそのまま抜粋する．**圧縮**は，原文の不要な部分を削除し，冗長な表現を短縮する[3]．**改変**は，文内で完結した節の入れ替えや数単語のみを対象とした言い換えなど，比較的小規模な変更[4]を行う．**融合**は，原文中の複数の文を連結し一つの文を作成する．連結される各文は，事前に圧縮，改変の処理が施されている場合もある[5]．**生成**は，原文書に存在しない表現を用いて高度な言い換えや内容の抽象化を行う．**説明**は，原文書の内容を要約者自身の視点から客観的に説明する．上記の処理のうち，本稿では抽出，圧縮，改変，融合をまとめて抽出型処理と呼ぶ．図 1 に抽出型処理のそれぞれの具体例を示す．生成，説明の処理が施された文は抽出型処理による再現が困難であり，今回構築するサブコーパスには含まない．

2.3 記事の種類

社説に関する記事を調査した結果，**概要的要約**，**説明的要約**，**冒頭のみ説明的でその後が概要的な要約**，の三種類の要約が存在することを確認した．概要的要約は原文書を原文書の著者と同じ視点で短くまとめた要約である．そのため，要約には原文に含まれる表現が多く出現するため，抽出型処理により十分に再現できることが予想される[6]．ただし，時には要約者自身により原文書に存在しない表現が生成されることもある．説明的要約は原文書の内容を第三者の視点に立って客観的に説明するような論調で書かれている．そのため，原文書に存在しない表現が現れることが多い．説明的要約はしばしば 1,2 文のみの比較的少ない文数で構成されている．また，冒頭のみが説明的でその後が概要的な要約では，記事の冒頭に要約者による客観的な説明が記述された後，具体的な内容が概要的要約の形で現れているものも存在する．そのような場合，説明となっている冒頭部分以外は抽出型処理により再現が可能である．

2.4 本サブコーパスを作成する意義

単一文書要約における重要箇所同定は困難な課題の一つである．複数文書要約では重要な話題は文書横断的に原文書集合中の多くの文書に出現するという特性を手がかりとして利用できるが，単一文書ではそのような情報が利用出来ない．報道記事であれば冒頭に重要なまとめが存在する場合も多いが，必ずしもそうでないことも報告されており [4]，必ずしも文の位置に依存しない重要箇所の同定手法を構築することは有用である．ここで，参照要約が生成型アプローチで作成された概要的要約や，そもそも説明的要約であった場合，重要箇所の同定のみでは参照要約を再現できず，重要箇所同定に対する定量的な判断ができない．

そのため，本稿では，技術的な焦点を重要箇所同定に絞るために，抽出型アプローチで参照要約を再現可能な参照要約を収集し，**重要箇所同定用サブコーパス 1** とする．ただし，抽出型処理のなかでも文の圧縮や

[3]ただし，原文からの冠詞や副詞の削除は極めて高い割合であらゆる文に発生するので，圧縮とはみなさないこととする．
[4]当該箇所以外の大部分は抽出である．
[5]ただし，融合される文に後述の説明や生成による文が含まれている場合は融合に含まずにそれぞれ説明，生成に属することとする．
[6]ただし，要約文の冒頭に表 1 中の “says” のような特有の語句が挿入される場合が多いため，このような語句は原文からは必ずしも再現できないことに注意されたい．

融合など比較的複雑な処理を経て要約が生成されている場合は，重要箇所の同定の他に技術的に難しい側面が出てきてしまう[7]．そこで，サブコーパス 1 のうち特に純粋に文抽出のみで構築された**サブコーパス 2** を構築する．適切な文を選択することさえできれば，高い評価値が得られることが分かっているため，重要箇所同定の側面のみを定量評価できるようになる．

3 重要箇所同定サブコーパスの構築

3.1 サブコーパス 1 の構築

まず，今回対象となる 12996 記事から，サブコーパス 1 を構築する．構築には，参照要約側の文と原文書側の文に対し，それらの語幹バイグラム集合間の Simpson 係数を用いる:

$$simp(s_i, s_j) = \frac{|sb(s_i) \cap sb(s_j)|}{min(|sb(s_i)|, |sb(s_j)|)}.$$

ここで s_i は文，$sb(\cdot)$ は，文を受け取り語幹バイグラム集合を返す関数である．Simpson 係数は，単語数の少ない方の文から見た被覆率を意味することから，ある要約文 s_i に対し $simp(s_i, s_j)$ の高い文 s_j が存在する場合，その要約文は抽出型処理によって作成されたとみなせる．

以下の条件を満たす参照要約を，サブコーパス 1 の構成要素とした．ここで，参照要約側の文インデックスを i，原文書側の文インデックスを j，s_i が抽出型処理によって作成されたかを決定する閾値を thr_1 とする．

- 全ての文 i に対し $simp(s_i, s_j) \geq thr_1$ となる文 j が存在する．
- i が一文目の場合は，カンマ[8]で区切った最初のセグメント s_k も同様に $simp(s_k, s_j) \geq thr_1$ となる j が存在する．

二つ目の条件は，2.3 節で述べた冒頭のみに説明的な記述がされている要約を除外するための条件である．冒頭の文をカンマで区切った最初のセグメントが説明であった場合は，原文に対応する箇所が存在しないため，Simpson 係数の値が低くなり，サブコーパスから除外することが可能となる．

3.2 サブコーパス 2 の構築

サブコーパス 1 のうち，さらに文抽出による再現率が高くなるサブコーパス 2 を構築する．構築には，文抽出に基づくオラクル要約を手がかりとして利用した．オラクル要約とは，要約システムにより再現できる評価関数の上限値を持った要約である．評価関数には Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[1] を利用した．オラクル要約の評価値が高い要約は，適切な文を選択することで参照要約の大部分を再現できることを意味する．そこで，オラクル要約の ROUGE-2 値が閾値 thr_2 以上の参照要約を取り出し，サブコーパス 2 とする．オラクル要約は，参照要約に含まれる語幹バイグラム集合をできるだけ被覆する文集合を選択する問題として，以下の整数計画問題として定式化した:

$$\begin{aligned}
\text{max.} \quad & \textstyle\sum_j^m z_j & & \\
\text{s.t.} \quad & \textstyle\sum_i^n c_i x_i \leq L; & & (1)\\
& \textstyle\sum_i^n a_{ij} x_i \geq z_j; & \forall j & (2)\\
& x_i \in \{0,1\}; & \forall i & (3)\\
& z_j \in \{0,1\}; & \forall j. & (4)
\end{aligned}$$

z_j は，j 番目の被覆単位を要約に含めるときに 1 となる決定変数である．x_i は文インデックス i を要約に選択するときに 1 となる決定変数である．制約式 (1) は，要約に選択された文に含まれる単語数が要約長 L 以下になることを保証する．式 (2) は，文と文に含まれる概念単位の整合性を保つための制約である．

表 1: 構築したサブコーパスの統計値

	要約単語数	原文書単語数	要約文数	圧縮率
sub1	51	468	2.1	0.120
sub2	58	444	2.3	0.142

4 構築と評価

4.1 サブコーパス構築結果

3 節の方法でサブコーパス 1(sub1) およびサブコーパス 2(sub2) を構築した．このとき，thr_1，thr_2 をそれぞれ 0.5，0.7 とした．構築されたサブコーパス 1 および 2 の記事数はそれぞれ 4228 記事，166 記事となった．サブコーパスの各種統計値を表 1 に示す．表中の値は，全記事に対する平均値を示している．

4.2 アノテーションによる評価

作成したサブコーパスの信頼性を評価するために，それぞれのサブコーパスからランダムに 15 記事をサンプリングし，1 名の作業者による参照要約のアノテーションを行った．具体的には，サブコーパス 1 の記事については，要約中の全ての文が抽出型処理のみで作

[7] 文圧縮を含んだ文書要約手法は近年盛んに研究されているトピックの一つである．
[8] ただし，文頭から 10 単語以内のカンマは除く．

表 2: 構築したサブコーパスへのアノテーション結果

	正例	負例
sub1	13	3
sub2	7	8

表 3: オラクル要約と各要約手法の ROUGE-2 値

	Editorial	sub1	sub2
オラクル	0.318	0.410	0.762
LEAD	0.131	0.199	0.484
MCP	0.093	0.124	0.237
TR	0.066	0.087	0.159

成された文である（正例）かどうか，サブコーパス 2 の記事については，各記事の要約の全ての文が抽出により作成された文（正例）かどうかをアノテーションした．信頼性評価の結果をそれぞれ表 2 に示す．表 2 を見ると，サブコーパス 1 は大部分が意図通りの参照要約で構成されていることが確認できた．サブコーパス 2 における負例のほとんどは，2-3 単語の削除により文圧縮処理だと判断されたものの，圧縮箇所以外は抽出であることを確認したため，極めて抽出に近い要約が取り出せていると言える．詳しくは 6 節に述べるが，今後はより精度の高い分類を進める予定である．

5 構築したコーパスに基づく要約実験

構築したサブコーパス 1，サブコーパス 2，および原文書集合 (Editorial) を対象に，いくつかの文書要約手法についての評価実験を行なった．今回用いた要約手法について説明する．**LEAD** は，原文書の冒頭から逐次的に文を抽出することで要約とする手法である．**MCP** は，原文書内の語幹ユニグラムを概念単位とした最大被覆問題に基づく文書要約モデル [3] に基づく手法である．**TR** は，TextRank に基づく手法 [2] である．TextRank は，共通課題タスクとして単一文書要約が扱われた Document Understanding Conference (DUC) 2002 において最も性能が高かった要約手法と同等の性能を示した文書要約モデルである．表 3 にオラクル要約と各要約手法による平均 ROUGE-2 値を示す．表の sub2[9] の列から，LEAD による要約は文抽出における ROUGE-2 値の上限であるオラクル要約の約 6 割のスコアを獲得していることがわかる．しかしながら，sub2 における LEAD が獲得した ROUGE-2 値の度数分布（図 2）を見ると，その約 2 割は 0.1 未満である．LEAD による要約で十分な ROUGE-2 値が得られない文書は，重要箇所が記事冒頭に無いため

[9] 文抽出に特化したサブコーパス

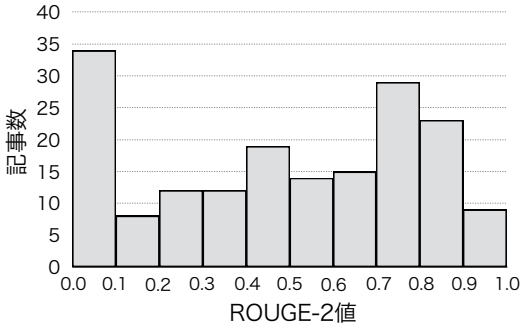

図 2: LEAD 手法のサブコーパス 2 における ROUGE-2 値の度数分布

重要箇所を同定する必要がある．表 3 で示したように今回用意したベースライン手法では十分な ROUGE-2 値が得られていないことから，より強力な重要箇所同定手法が必要である事がわかる．

6 まとめ

本稿では，NYTAC を大規模な要約資源として整備するための研究の第一段階として抽出型アプローチによる概要的要約のみで構成された参照要約を収集することで重要箇所同定用サブコーパスを構築した．

今回は，抽出型処理による要約とそれ以外の分類のみに焦点を当てた．しかしながら，説明的要約と概要的要約はその特性が大きく異なるため，両者を分離することができれば非常に興味深い分析が可能になる．今後は，特に説明的要約と，生成型アプローチによる概要的要約との違いについて焦点を当てて研究を行う．

今回は，語幹バイグラムの Simpson 係数により抽出型処理か否かの分類を行なったものの，抽出型処理のうちいずれの処理であるかの詳細な分類基準は作成していない[10]．今後は，要約文書と原文書内のテキストスパン同士のアライメントを取るなど，詳細な分析を行う．また，構築したサブコーパスは，英語話者による詳細なチェックを行うことで，本サブコーパスの評価セットとしての信頼性を担保する．

参考文献

[1] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, pp. 74–81, 2004.

[2] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. In *EMNLP*, pp. 404–411, Jul. 2004.

[3] Hiroya Takamura and Manabu Okumura. Text summarization model based on maximum coverage problem and its variant. In *EACL*, pp. 781–789, March 2009.

[4] Yinfei Yang and Ani Nenkova. Detecting information-dense texts in multiple news domains. In *AAAI*, pp. 1650–1656, Jul. 2014.

[10] バイグラムの Jaccrad 係数や被覆率などの他の指標，あるいはそれらの組み合わせが一定の分類基準になることは確認している．