Vol.2015-NL-224 No.5
2015/12/3

# 大規模要約資源としての New York Times Annotated Corpus

菊池 悠太[1,a)]　渡邉 亮彦[1,b)]　高村 大也[1,c)]　奥村 学[1,d)]

**概要**：大量の訓練データによる機械学習は，自然言語処理の様々なタスクにおいて多大な貢献をしてきた．しかし文書要約においては，大規模訓練データを有効に利用することに成功していない．New York Times Annotated Corpus (NYTAC) には，約 65 万の文書-要約対が含まれるが，それらの要約の間には，長さの違い，抽出および非抽出の違いなど様々な違いが存在しうるため，単純に訓練データとして利用するだけでは十分な効果が期待できない．本研究では，NYTAC を要約器の訓練用データとして有効に活用する方法を提案する．

Yuta Kikuchi[1,a)]　Akihiko Watanabe[1,b)]　Hiroya Takamura[1,c)]　Manabu Okumura[1,d)]

## 1. はじめに

他の様々なタスクと同様に，文書要約においても機械学習が何らかの形で用いられる場面は増えてきている [1], [8], [13], [22], [25], [26], [29]．しかしながら，文書要約において機械学習を用いる上での共通の問題は，訓練事例（要約-文書対）の不足である．これは，単純な分類問題と比較し人手による要約作成はコストが高いことに由来している．そのため，従来のほとんどの研究は数百事例という小規模なデータによる学習を余儀なくされており，他の主要なタスクと比較しても十分とはいえない[*1]．この現状は，複雑な手法の開発や有効な素性の構築，人手による詳細な分析などを困難にしている．直近の研究では西川ら [22] が日本語報道記事とその要約 13,000 事例を用いた学習を行い，文書要約においても訓練事例数の増加が性能向上に寄与することを報告しており，訓練データの充実は更なる文書要約研究の発展に不可欠であると言える．

現在，このような現状の打破に寄与すると最も期待できるコーパスとして，New York Times Annotated Corpus (NYTAC)[24] が存在する．NYTAC を構成する約 180 万記事のうち，およそ 65 万もの記事に専門家による要約が付与されている．これは，潜在的には単一文書要約において最も大規模な訓練データとみなすことができる．しかしながら，これまで NYTAC の人手要約を何らかの形で利用した研究 [10], [28] は存在するが，直接的に大規模な訓練事例として利用したという報告はなされていない．NYTAC を直接的に訓練事例として利用する上で留意すべき[*2]は，そこに含まれる人手要約が特定の目的のもと統制されて作成および整備されたわけではないという点である．文書要約にはその作成方法や使途に複数の種類が存在しており，それぞれ特性や必要となる技術が異なるが，NYTAC においてはそれらが明示的に区別されることなく混在している．そのため，特定の要約器，例えば現在最も標準的なアプローチである文抽出に基づく要約器を訓練する際に，全ての事例を機械的に利用することが必ずしも有効であるかは疑問である．

本稿における我々の目的は，NYTAC を単一文書要約における大規模な訓練事例として利用する上での効果的な手法を示すことである．訓練のどの段階でどのような形で NYTAC を利用するべきか，全ての事例を用いるべきか何らかの基準でその部分集合に限って利用するべきか，実験を通して明らかにする．

評価の際は，単一文書要約において評価に用いられている複数のテストセット（ターゲットデータ）を用意し，そ

[1] 東京工業大学
Tokyo Institute of Technology
[a)] kikuchi@lr.pi.titech.ac.jp
[b)] watanabe@lr.pi.titech.ac.jp
[c)] takamura@pi.titech.ac.jp
[d)] oku@pi.titech.ac.jp

[*1] 例えば, 機械翻訳における古典的なコーパスの一つである Hansard コーパス [4] には約 130 万もの事例が収録されている．

[*2] これは同時に，従来 NYTAC が要約器の大規模訓練データとして用いられて来なかった理由であると考えられる．

れら全てにおいて NYTAC の利用が有効であるかを確かめる．

我々が扱う問題は，ドメイン適応 [15] の一つのケースだと考えることができる．ドメイン適応では，両ドメイン間で学習されたパラメータの線形補間や適用元ドメインにより訓練された分類器を素性として利用するなど，シンプルでありながら強力ないくつかの手法が存在しており [6]，今回のケースにも容易に適用が可能である．さらに，ドメインにあわせた事例選択 (instance selection, instance weighting) も同様にドメイン適応の方法として知られている [2], [27]．本研究では，ドメイン適応の分野を参考にいくつかの標準的な手法を用意し比較することで，NYTAC を利用するにあたり有効な方法を確かめる．

実験の結果，要約器を NYTAC で一度訓練したあと，そのパラメータを初期値として所望のターゲットデータで追加的に訓練する手法が有効に働くことが分かった．加えて，使用する要約器の特性に併せた事例選択を事前に行うことで更に精度が向上することを確認した．

## 2. New York Times Annotated Corpus

NYTAC は，約 180 万記事の New York Times 紙の新聞記事によって構成されており，そのうちの約 65 万記事には人手で生成された要約が付与されている．表 1 に主な統計量を，今回の実験で使用するテストセットと共に示す．1 節で述べたように，このコーパスには様々な種類の要約が混在している．たとえば，文抽出を行うことで作成されている要約もあれば，文抽出に加えて文圧縮や文融合などを行うことにより作成されている要約も存在する．加えて，言い換えなどの非抽出的な操作を行うことで作成されている要約や，メタな視点に立って書かれた説明口調な要約も存在する．このように異なる種類の要約が存在することは，文圧縮，文融合や非抽出的なアプローチに基づく要約など，異なる要約技術の発展に有益である [11], [21]．

しかしながら，各要約操作を計算機に行わせるためにはそれぞれ異なった技術が必要であり，それぞれにとって有効な訓練事例は異なることが予想できる．たとえば，文抽出による要約手法を用いる場合は，原文書からの表層的な抽出にこだわらない方法で作成された要約は訓練事例として適切ではないと考えられる．従って，本研究では NYTAC を訓練事例として利用するにあたり，訓練する要約器に合わせた事例の選択が必要であると考える．本稿では，現在の文書要約において最も広く用いられている文抽出に基づくアプローチを対象とする．そこで，適切な文抽出により人手の正解要約（参照要約）を再現できる度合いを基準とし，訓練事例の選別を行い，その有効性を確かめる．

*3 NYTAC に含まれる要約の総数は 658,874 であるが，原文書や要約文書が空であるなどのノイズを除去すると，最終的には 524,216 事例となった．

## 3. NYTAC の利用した要約器の訓練手法

本節では，今回用いる要約器とそのパラメータ推定方法，そして訓練に NYTAC を利用する手法を説明する．NYTAC の利用法として，訓練方法そのものを規定する 5 つの手法と，それらの前処理として要約器の特性に合わせた事例選択の方法を提案する．

### 3.1 ナップサック問題による文抽出型要約器

本稿では，ナップサック問題に基づく要約器を用いる．すなわち，原文書中の各文に要約として選択した場合の利得を付与し，与えられた最大単語数以下で利得の総和を最大化する文の組み合わせを選択する．ナップサック問題による定式化は，単一文書要約を組合せ最適化問題として定式化する上で最も自然な方法の一つである [19], [31]．近年提案されている単一文書要約を対象とした多くの手法はナップサック問題の拡張であり，そのためベースラインとして頻繁に登場している [9], [22]．具体的には，以下のように定式化できる:

$$\begin{aligned} \max. \quad & \textstyle\sum_i^n b_i x_i \\ \text{s.t. } & \textstyle\sum_i^n c_i x_i \le L; \qquad (1) \\ & x_i \in \{0,1\}; \quad \forall i. \qquad (2) \end{aligned}$$

$x_i$ は文 $i$ を要約として選択した場合に 1 となる決定変数であり，$b_i$ は文 $i$ を要約として選択した場合に得られる利得である．また，$c_i$ は文 $i$ を要約として選択した場合のコスト（文 $i$ の単語数）であり，$L$ が要約に含めることのできる最大の単語数である．ここで，文の利得 $b$ を素性関数 $\phi$ とその重みベクトル $\mathbf{w}$ として定義し，$\mathbf{w}$ を訓練により推定する: $b_i = \mathbf{w} \cdot \phi(s)$.

重みベクトル $\mathbf{w}$ を推定する方法として，本稿では教師あり構造学習の枠組みを採用する．具体的には Passive Aggressive 法 [5] の一種である PA-II 法に基づき，重みベクトルの更新を以下の最適化問題の解として定義する:

$$\begin{aligned} \mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \quad & \frac{1}{2}||\mathbf{w} - \mathbf{w}_t||^2 + C\xi^2 \qquad (3) \\ \text{s.t.} \quad & loss(\mathbf{w}) \le \xi. \qquad (4) \end{aligned}$$

$\mathbf{w}_t$ は $t$ ステップ目の重みベクトルである．また，$\xi$ は制約を破った場合のペナルティを意味するスラック変数であり，$C$ はペナルティの影響を決めるパラメータである．ここで，損失関数について $loss(\mathbf{w}) = 1 - ROUGE(o, s)$ と定義する．$s$ は重みベクトル $\mathbf{w}$ に基づき要約器が生成*4した要約である．$o$ は，参照要約により求めた文抽出オラクル要約であり，次節で詳しく説明する．$ROUGE(o, s)$ はオラクル要約 $o$ を正解としたときのシステム要約 $s$ の ROUGE

*4 要約のデコーディングは標準的な 0-1 ナップサック問題において用いられる動的計画法と同様の方法により求めることができるため，具体的なアルゴリズムについては割愛する．

**表 1** 本稿で扱うコーパスとその統計量．corpus size を除く全ての数値は平均値である．また，下から 3 つのコーパスは実験で用いるターゲットデータである．

| | corpus size | summary | | document | | compression rate |
|---|---|---|---|---|---|---|
| | | # sentence | # word | # sentence | # word | |
| NYTAC | 524,216[*3] | 2.42 | 38.79 | 29.36 | 607.62 | 0.128 |
| DUC2002 | 533 | 5.62 | 101.09 | 27.40 | 549.61 | 0.363 |
| $\text{RSTDTB}_{\text{long}}$ | 30 | 9.57 | 186.10 | 116.77 | 930.23 | 0.246 |
| $\text{RSTDTB}_{\text{short}}$ | 30 | 3.5 | 39.43 | 116.77 | 930.23 | 0.093 |

(Recall-Oriented Understudy for Gisting Evaluation) [16] 値である．この最適化問題は，ラグランジュの未定乗数法により以下のような閉形式で表すことができる:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \tau_t(\phi(o_t) - \phi(s_t)), \quad (5)$$

$$\tau_t = \frac{loss(\mathbf{w}_t)}{||\phi(o_t) - \phi(s_t)|| + \frac{1}{2C}}. \quad (6)$$

$\phi(\cdot)$ は文書を素性ベクトルへ変換する関数であり，与えられた文書中に含まれる各文の素性ベクトルの和を取ったベクトルである．

$\phi(\cdot)$ の構築に用いる文単位の素性として，以下のものを用いる:

(a) 文に含まれる内容語の割合,
(b) 頻度上位 1 万語の bag-of-words,
(c) 文に含まれる単語数の対数,
(d) 文に含まれる単語数の，文書全体の単語数に対する相対的な長さ,
(e) 文の，文書における絶対位置の対数,
(f) 文の，文書における相対的な位置,
(g) 文が文書の前半 20%以内に位置している場合に 1，それ以外に 0 となる二値素性,
(h) 文による文書中のユニグラム被覆率,
(i) 訓練済みの $NytOnly$ モデルが出力した文の利得（本素性は 3.2 節における $Featurize$ においてのみ利用される）.

また，訓練の際は Zhao らの報告に基づき並列化を行った [30].

#### 3.1.1 文抽出オラクル要約

オラクル要約とは，ある評価指標において，人手による参照要約に対して要約器が獲得できる評価値の上限となる要約である．本稿では評価指標として ROUGE-2 [16] を用いた．すなわち，文抽出型要約においては参照要約に対する ROUGE-2 が最も高くなる文の組み合わせを選択することでオラクル要約を生成する．

オラクル要約の評価値は要約器が獲得できる評価値の上限値であるため，その値が高い事例は，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．

本研究では，NYTAC の参照要約-原文書対から計算されたオラクル要約を，二つの異なる目的で利用する．

- 一つ目の目的は，要約器の訓練における教師信号，すなわち式 (6)-(**??**) における $o_t$ に，参照要約に代わって利用することである．構造化パーセプトロンに基づく学習手法においては，正解となる素性ベクトルを正解ラベルから構築する必要があるが，参照要約をそのまま変換するだけでは，文の位置など一部の素性が適切に学習されない．
- もう一つの目的は訓練時の事例選択への利用である．オラクル要約の獲得する ROUGE-2 値が低い場合，要約器がどのような文の組み合わせを選んだとしても高い評価値が得られない．そのため，文抽出要約器の訓練事例としてのふさわしさの指標としてオラクル要約の評価値を利用する．より詳細な説明は 3.3 節にて述べる．

本稿では，参照要約に含まれるバイグラムを被覆対象とする最大被覆問題を解くことでオラクル要約を作成する:

$$\begin{aligned} \max. \quad & \textstyle\sum_j^m z_j \\ \text{s.t.} \quad & \textstyle\sum_i^n c_i x_i \leq L; & (7) \\ & \textstyle\sum_i^n a_{ij} x_i \geq z_j;\ \forall j & (8) \\ & x_i \in \{0,1\}; \quad \forall i & (9) \\ & z_j \in \{0,1\}; \quad \forall j. & (10) \end{aligned}$$

ここで $z_j$ は，要約として選択した文が参照要約中に出現する $j$ 番目のバイグラムを含む場合に 1 となる決定変数である．また，式 (8) は文と被覆対象の整合性を取るための制約式であり，$a_{ij}$ は文 $i$ がバイグラム $j$ を含む場合に 1 で，そうではない場合に 0 となる．なお，オラクル要約の計算時には全ての文書に対して，英数字以外の全ての文字を削除した上で語幹抽出を行った．これは，ROUGE の公式評価スクリプトをオプションは”-m -r 2”で実行した時と同じ前処理である．このように，人手により自由に作成された参照要約から関連する原文書の抜粋を構築するという試みは構造学習に基づく要約器の学習 [25], [26] の際に行われているほか，より一般的な問題として Marcu らによる試みがある [18].

### 3.2 NYTAC を利用した要約器の訓練

本節では NYTAC を訓練に利用する 5 つの手法について述べる．ターゲットデータを訓練に用いる際は，五分割交差検定を利用している．なお，ターゲットデータによる五分割交差検定のみで構築されたモデルを $TrgtOnly$ とし，

以下に述べる全ての手法に対するベースラインとする．

*NytOnly* は NYTAC のみで訓練を行い，ターゲットデータの情報を一切利用しない手法である．そのため，この手法で訓練されたモデルのスコアは純粋に大量の訓練事例による効果を表しているといえる．

*Mixture* は，NYTAC の事例とターゲットデータの交差検定における訓練セットを混合した事例集合を用いて訓練を行う手法である．本手法は単純に訓練事例のサイズを増加させる目的で NYTAC を利用しているが，以下の 3 手法は訓練済みの *NytOnly* モデルを用いる．

*LinInter* は，訓練済みの *TrgtOnly* モデルと *NytOnly* モデルの重みベクトルの線形補間により作成した重みベクトルを用いて予測を行う手法である．重みパラメータは交差検定における検証セットを利用して決定する．

*Featurize* は，訓練済みの *NytOnly* モデルが出力する文の利得を追加の素性として，ターゲットデータで訓練を行う手法である．

*FineTune* は，訓練した *NytOnly* モデルの重みベクトルを初期値として利用し，そのモデルをターゲットデータで学習する手法である．NYTAC で学習した重みベクトルを初期値とすることは，PA 法をはじめとするオンライン学習手法で標準的なゼロベクトルによる初期化と比較してより良いパラメータの探索が可能となると期待できる．

### 3.3 オラクル要約による事例選択

2 節で述べた通り，NYTAC には多くの種類の参照要約が含まれる．ある要約は抽出型である，すなわち原文書から適切なテキストスパンを抽出することで生成されているが，ある要約は生成型，すなわち言い換えや一般化などにより原文書に存在しない表現を含んだ形で作成されている．本稿では文抽出に基づく要約器を用いるため，前者のような抽出に基づく要約事例を訓練事例として用いることは有用であると考えられる一方で，後者の要約を訓練事例として用いるのは適切ではないと考える．より一般的には，もし我々が特に統制のない状態で作成された大量の事例を手に入れた場合，その時の利用目的，今回は文抽出に基づく要約器の訓練をする上で有用であると思われる部分集合を自動で選択するという処理が必要になる．このような考えはドメイン適応の分野においても知られている [3], [23].

本稿では，事例選択の基準として 3.1.1 節で述べた文抽出オラクル要約の ROUGE-2 値を用いる．オラクル要約の ROUGE-2 値が十分に高い場合は，原文書から適切な文の組み合わせを選択することで参照要約の大部分を再現できる．反対に，オラクル要約の ROUGE-2 が低い場合はどのような文の組み合わせを選択したとしても参照要約とは表層的に異なる要約しか生成することが出来ない．このようなケースでは，言い換えや一般化など，より複雑な操作が必要となる．

具体的な事例選択の方法としては，まず NYTAC 中の全ての要約-文書対に対しオラクル要約を計算し，その ROUGE-2 値が閾値 $thr$ を超えた事例のみで構成された部分集合を選択する．実験を行う際は，$thr$ 毎に *NytOnly* モデルを訓練しておき，交差検定の検証セットにおいて，どのモデルを利用するかを決定する．

表 2 に $thr$ ごとの事例数を示す．表により，NYTAC の中の 2,430 事例 ($thr = 0.8$) は，原文書中の文を適切に選択することで人手により作成された要約の 8 割以上を被覆した要約文書が作成可能であることが分かる．

## 4. 実験

本稿では，ターゲットデータとして，単一文書要約におけるテストデータとして用いられることの多い三種類のコーパスを用いる．以下，各データを DUC2002, RSTDTB$_{long}$, RSTDTB$_{short}$ と呼び，それぞれの統計量を表 1 に示す．DUC2002[7] は評価型ワークショップである Document Understanding Conference (DUC) において単一文書要約が共通課題として設定された際のテストデータである[*5]．567 個の原文書-参照要約対で構成されており，各参照要約の長さは 100 単語以下[*6]となるように生成されている RSTDTB$_{long}$ および RSTDTB$_{short}$ は，Rhetorical Structure Theory Discourse Treebank (RSTDTB，LDC2002T07) と呼ばれる，修辞構造理論 (RST) に関するコーパスに含まれている要約データである．RSTDTB は Penn Treebank 中の 385 記事の Wall Street Journal 記事によって構成されており，各記事には RST に基づく談話構造が人手でアノテーションされている．さらに，そのうち 30 記事には人手で生成された参照要約が付与されている．RSTDTB$_{long}$ と RSTDTB$_{short}$ では参照要約の長さが異なる．具体的には，RSTDTB$_{long}$ では要約者は原文書の 25%の長さとなるよう要約を作成するよう指示されており，RSTDTB$_{short}$ は 2,3 文と指示されている．RST では，文書を文よりも細かいおおよそ節に相当する elementary discourse unit (EDU) に分割し，談話構造を考える上での最小単位としている．そのため，RSTDTB を利用した要約研究はこの EDU を抽出単位とすることが多く，本研究においても同様に文ではなく EDU の抽出を行う．

要約器に与える制限要約長 $L$ は，DUC2002 を用いた実験においては 100 単語とし，RSTDTB$_{long}$ および

[*5] 2002 年は単一文書要約が共通課題として設定された最後の年度である．
[*6] 評価の際には，全ての要約を 100 単語に切り詰める．

**表 2** *thr* 毎に選択される訓練事例の数

| *thr* | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| datasize | 524,216 | 403,222 | 285,628 | 203,813 | 134,068 | 74,595 | 31,181 | 10,472 | 2,430 | 299 | 71 |

$\text{RSTDTB}_{short}$ を用いた実験ではそれぞれの参照要約の単語数と等しくなるように設定した．

評価尺度としては，単一文書要約の評価によく用いられている ROUGE-1 を用いる．この際，DUC2002 において人手との相関が最も高かった設定 [16] である，ストップワードを削除した上で語幹の抽出を行った上での数値[*7]を報告する．また，参考のために同様に人手評価との高い相関を示した ROUGE-2 による評価結果も付与する．ROUGE-2 においては，ストップワードの有無による差は報告されていないため，語幹抽出のみを行った上で評価を行う[*8]．

ここで，三種類のテストセットの参照要約の使途の違いについて言及しておく．DUC2002 および $\text{RSTDTB}_{long}$ の参照要約は報知的要約と呼ばれる要約であり，対して $\text{RSTDTB}_{short}$ は指示的要約である[*9]．本稿で用いる要約器を含め，先行研究における多くの要約手法は報知的要約を対象としている．そのため，まず DUC2002 および $\text{RSTDTB}_{long}$ に対して提案手法の評価を行い，次に $\text{RSTDTB}_{short}$ おける実験を行うことにより NYTAC が指示的要約の生成についての学習にも寄与するかを確認する．

### 4.1 Result: DUC2002

DUC2002 による評価結果を表 3 に示す．まず，*Featurize* を除くすべての提案手法が *TrgtOnly* を有意に上回る結果となった[*10]．これは，NYTAC を追加の訓練データとして用いることの有効性を示している．*NytOnly* は訓練時に DUC2002 の情報を一切用いていないにもかかわらず，ROUGE 値が *TrgtOnly* を有意に上回っていることは興味深い結果であり，訓練データの規模の重要性を示唆している．5 つの提案手法間で比較すると，*FineTune* が NYTAC を用いる上で最も有効な手法であることが分かる．

名前の末尾に $_{slct}$ がついた結果は，3.3 節で提案した事例選択を事前に行った上で訓練した場合の結果である．その差は有意ではないものの，事例選択を行う前と比較して僅かな向上が見られた．また，$\textit{FineTune}_{slct}$ は事例選択を行わない場合と同様に他の手法を有意に上回っている．なお，五分割交差検定によって選択された *thr* の値は，それぞれ 0.2, 0.3, 0.3, 0.3, 0.3 となった．

**表 3** DUC2002 における各手法の ROUGE 値．表は便宜上横線により 4 つの区画に分けられており，それぞれの区画で最も高い ROUGE-1 値を獲得したシステムを太字で示している．

| | ROUGE-1 | ROUGE-2 |
|---|---|---|
| *TrgtOnly* | **0.400** | **0.218** |
| *NytOnly* | 0.409 | 0.217 |
| *Mixture* | 0.411 | 0.219 |
| *Featurize* | 0.404 | 0.220 |
| *LinInter* | 0.409 | 0.217 |
| *FineTune* | **0.415** | **0.221** |
| $\textit{NytOnly}_{slct}$ | 0.411 | 0.224 |
| $\textit{Mixture}_{slct}$ | 0.412 | 0.224 |
| $\textit{Featurize}_{slct}$ | 0.403 | 0.219 |
| $\textit{LinInter}_{slct}$ | 0.411 | 0.224 |
| $\textit{FineTune}_{slct}$ | **0.418** | **0.225** |
| LEAD | 0.413 | 0.224 |
| TextRank | 0.409 | 0.206 |
| s21 | 0.416 | 0.223 |
| s27 | 0.401 | 0.212 |
| s28 | **0.428** | **0.228** |
| s29 | 0.400 | 0.213 |
| s31 | 0.393 | 0.203 |

最後に，いくつかの既存手法と提案手法を比較する．**LEAD** は原文書の冒頭から 100 単語を機械的に抽出する手法である．LEAD は単純な手法であるにもかかわらず，単一文書要約において非常に強力なベースラインであることが知られている．**TextRank** [20] も単一文書要約において知られた手法であり，PageRank アルゴリズムに基づくグラフベースの重要文ランキングを行う．**s21-31** は DUC2002 の単一文書要約タスクに参加したシステムのうち上位 5 件のシステムである．$\textit{FineTune}_{slct}$ は，LEAD, s27, s29, s31 を有意に上回る性能を示している一方で，s28 を有意に下回る結果となっている．

### 4.2 Result: $\text{RSTDTB}_{long}$

$\text{RSTDTB}_{long}$ による評価結果を表 4 に示す．概ねの結果は DUC2002 と同様である．提案手法の中では *FineTune* が最も高い評価値を得ており，その評価値は *TrgtOnly* をはじめ，他の提案手法を有意に上回っている．加えて，事例選択により更なる向上が見られる．

ここで，比較する先行研究について説明する．$\textbf{LEAD}_{EDU}$ は要約長 $L$ に至るまで原文書の冒頭の $K$ 個の EDU を抽出する手法である．**Marcu** は原文書の RST に基づいて EDU の重要度をランキングすることで要約を生成する手

[*7] 具体的には，ROUGE の公式評価スクリプト（バージョン 1.5.5）においてオプション “-a -x -n 1 -m -s” で実行した時のスコアである．
[*8] 具体的には，ROUGE の公式評価スクリプト（バージョン 1.5.5）においてオプション “-a -x -n 2 -m” で実行した時のスコアである．
[*9] これら二つの要約についての具体的な説明は，例えば Nenkova による著作 [21] では以下のように記述されている: *A summary that enables the reader to determine about-ness has often been called an indicative summary, while one that can be read in place of the document has been called an informative summary.*
[*10] 検定にはウィルコクソンの符号順位検定 ($p \leq 0.05$) を用いた．

**表 4** RSTDTB$_{long}$ における各手法の ROUGE 値．表は便宜上横線により 4 つの区画に分けられており，それぞれの区画で最も高い ROUGE-1 値を獲得したシステムを太字で示している．

| | ROUGE-1 | ROUGE-2 |
|---|---|---|
| $TrgtOnly$ | **0.324** | **0.121** |
| $NytOnly$ | 0.313 | 0.128 |
| $Mixture$ | 0.320 | 0.132 |
| $Featurize$ | 0.359 | 0.150 |
| $LinInter$ | 0.313 | 0.128 |
| $FineTune$ | **0.385** | **0.158** |
| $NytOnly_{slct}$ | 0.320 | 0.134 |
| $Mixture_{slct}$ | 0.323 | 0.133 |
| $Featurize_{slct}$ | 0.376 | 0.156 |
| $LinInter_{slct}$ | 0.320 | 0.134 |
| $FineTune_{slct}$ | **0.408** | **0.173** |
| LEAD$_{EDU}$ | 0.323 | 0.128 |
| Marcu | 0.362 | 0.155 |
| Hirao | **0.405** | **0.161** |

**表 5** RSTDTB$_{short}$ における各手法の ROUGE 値．表は便宜上横線により 4 つの区画に分けられており，それぞれの区画で最も高い ROUGE-1 値を獲得したシステムを太字で示している．

| | ROUGE-1 | ROUGE-2 |
|---|---|---|
| $TrgtOnly$ | **0.258** | **0.086** |
| $NytOnly$ | 0.272 | 0.085 |
| $Mixture$ | 0.272 | 0.085 |
| $Featurize$ | 0.264 | 0.087 |
| $LinInter$ | 0.272 | 0.085 |
| $FineTune$ | **0.283** | **0.094** |
| $NytOnly_{slct}$ | 0.265 | 0.087 |
| $Mixture_{slct}$ | 0.264 | 0.088 |
| $Featurize_{slct}$ | 0.246 | 0.075 |
| $LinInter_{slct}$ | 0.265 | 0.087 |
| $FineTune_{slct}$ | **0.286** | **0.092** |
| LEAD$_{EDU}$ | 0.240 | 0.082 |
| Marcu | 0.272 | 0.094 |
| Hirao | **0.321** | **0.106** |

法である [17]．**Hirao** は，平尾らの提案した RST の依存構造木からの刈り込みに基づく要約手法であり，EDU 抽出による要約手法の中で state-of-the-art な手法である．

これらの手法と比較しても，$FineTune_{slct}$[*11]が高い評価値を得ていることが分かる．Marcu および Hirao は人手でアノテーションされた原文書の修辞構造を利用した手法であり，そのような情報を一切用いずに state-of-the-art な評価値を獲得したことは，大規模な訓練データの重要性を物語っているといえる．

### 4.3 Result: RSTDTB$_{short}$

RSTDTB$_{short}$ による評価結果を表 5 に示す．$FineTune$ が他の手法に比較して高い評価値を獲得しているという点では DUC2002, RSTDTB$_{short}$ と同様だが，いずれの差も有意では無いという点で異なる．事例選択[*12]による影響を見ると，$FineTune_{slct}$ を除き有効な効果は得られなかった．

RSTDTB$_{short}$ における性能向上は DUC2002 や RSTDTB$_{long}$ と比較して大きなものではなかったが，指示的要約と報知的要約という両者の違いを考慮すれば妥当な結果といえる．

## 5. 関連研究

他の様々なタスクと同様に，文書要約においても様々な形で機械学習が利用されている．いくつかの手法では文の利得を推定するために，ユニグラム，バイグラムあるいは係り受けエッジの重要度を学習している [1], [10], [13], [29]．また他には，文全体の重要度を推定するもの [8], [20] や，構造学習を利用して要約としてのスコアを直接学習する試みも行われている [22], [25], [26]．本研究では標準的なナップサックモデルを構造学習により訓練したが，今後他の様々な機械学習に基づく手法に NYTAC が有効であるか確かめる必要がある．

NYTAC の要約データを活用する試みはいくつかの先行研究に見受けられる．いくつかの手法は，NYTAC の要約によって訓練された言語モデルを利用している [10], [12]．Li and Nenkova[14] は NYTAC を用いて文の特異性を推定している．また Yang and Nenkova [28] は，文書の第一段落がその文書全体にとって重要な情報を含んでいるかを判定するために NYTAC を利用している．本研究は NYTAC を直接的に訓練データとしているが，これらの研究は NYTAC を補助的な情報として利用している．

## 6. おわりに

本研究では，文抽出に基づく要約器の訓練に NYTAC を活用するため 5 つの手法を提案し，3 つのテストデータをターゲットとした実験により比較を行った．また，要約器の特性に併せた事例選択を行い，その効果を確かめた．

実験の結果，すべてのターゲットデータにおいて，NYTAC により事前に学習した要約器のパラメータを初期値として追加的に学習する手法（$FineTune$）が最も学習に寄与することが分かった．さらに，事例選択による更なる精度向上は，RSTDTB$_{long}$ においては state-of-the-art な ROUGE 値を獲得した．

今後は，大規模な訓練事例の活用による，より柔軟で効果的な素性の設計を行う．また，標準的なナップサックモデル以外の文書要約モデルに対する NYTAC の有効性を検証する必要がある．

[*11] 五分割交差検定によって選択された $thr$ の値は，それぞれ 0.1, 0.1, 0.1, 0.1, 0.1 であった．
[*12] 五分割交差検定によって選択された $thr$ の値は，それぞれ 0.3, 0.6, 0.3, 0.3, 0.6 であった．

## 参考文献

[1] Almeida, M. and Martins, A.: Fast and Robust Compressive Summarization with Dual Decomposition and Multi-Task Learning, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 196–206 (2013).

[2] Axelrod, A., He, X. and Gao, J.: Domain Adaptation via Pseudo In-domain Data Selection, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 355–362 (2011).

[3] Biçici, E.: Domain Adaptation for Machine Translation with Instance Selection, *The Prague Bulletin of Mathematical Linguistics*, Vol. 103, pp. 5–20 (2015).

[4] Consortium, L. D.: Hansard Corpus of Parallel English and French, *Linguistic Data Consortium*, http://www.ldc.upenn.edu/ (1997).

[5] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S. and Singer, Y.: Online Passive-Aggressive Algorithms, *The Journal of Machine Learning Research*, Vol. 7, pp. 551–585 (2006).

[6] Daumé III, H.: Frustratingly Easy Domain Adaptation, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pp. 256–263 (2007).

[7] DUC: Document Understanding Conference., *ACL Workshop on Automatic Summarization* (2002).

[8] Hirao, T., Isozaki, H., Maeda, E. and Matsumoto, Y.: Extracting Important Sentences with Support Vector Machines, *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Vol. 1, pp. 1–7 (2002).

[9] Hirao, T., Yoshida, Y., Nishino, M., Yasuda, N. and Nagata, M.: Single-Document Summarization as a Tree Knapsack Problem, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1515–1520 (2013).

[10] Hong, K. and Nenkova, A.: Improving the Estimation of Word Importance for News Multi-Document Summarization, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 712–721 (2014).

[11] Jing, H. and McKeown, K. R.: Cut and Paste Based Text Summarization, *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL)*, pp. 178–185 (2000).

[12] Li, C., Liu, Y. and Zhao, L.: Using External Resources and Joint Learning for Bigram Weighting in ILP-Based Multi-Document Summarization, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 778–787 (2015).

[13] Li, C., Qian, X. and Liu, Y.: Using Supervised Bigram-based ILP for Extractive Summarization, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1004–1013 (2013).

[14] Li, J. J. and Nenkova, A.: Fast and Accurate Prediction of Sentence Specificity, *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, pp. 2281–2287 (2015).

[15] Li, Q.: Literature Survey: Domain Adaptation Algorithms for Natural Language Processing, Technical report, Department of Computer Science. The Graduate Center, The City University of New York (2012).

[16] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, *Proceedings of the Workshop on Text Summarization Branches Out*, pp. 74–81 (2004).

[17] Marcu, D.: Improving summarization through rhetorical parsing tuning, *Proceedings of Sixth Workshop on Very Large Corpora*, pp. 206–215 (1998).

[18] Marcu, D.: The Automatic Construction of Large-scale Corpora for Summarization Research, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 137–144 (1999).

[19] McDonald, R.: A Study of Global Inference Algorithms in Multi-document Summarization, *Proceedings of the 29th European Conference on IR Research (ECIR)*, pp. 557–564 (2007).

[20] Mihalcea, R. and Tarau, P.: TextRank: Bringing Order into Texts, *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pp. 404–411 (2004).

[21] Nenkova, A. and McKeown, K.: Automatic Summarization, *Foundations and Trends® in Information Retrieval*, Vol. 2-3, pp. 103–233 (2011).

[22] Nishikawa, H., Arita, K., Tanaka, K., Hirao, T., Makino, T. and Matsuo, Y.: Learning to Generate Coherent Summary with Discriminative Hidden Semi-Markov Model, *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING)*, pp. 1648–1659 (2014).

[23] Remus, R.: Domain Adaptation Using Domain Similarity- and Domain Complexity-based Instance Selection for Cross-domain Sentiment Analysis, *Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW 2012) Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE)*, pp. 717–723 (2012).

[24] Sandhaus, E.: The New York Times Annotated Corpus, *Linguistic Data Consortium*, https://catalog.ldc.upenn.edu/LDC2008T19 (2008).

[25] Sipos, R., Shivaswamy, P. and Joachims, T.: Large-Margin Learning of Submodular Summarization Models, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 224–233 (2012).

[26] Takamura, H. and Okumura, M.: Learning to Generate Summary As Structured Output, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1437–1440 (2010).

[27] Xia, R., Zong, C., Hu, X. and Cambria, E.: Feature Ensemble Plus Sample Selection: Domain Adaptation for Sentiment Classification, *IEEE Intelligent Systems*, Vol. 28, No. 3, pp. 10–18 (2013).

[28] Yang, Y. and Nenkova, A.: Detecting Information-Dense Texts in Multiple News Domains, *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1650–1656 (2014).

[29] Yih, W., Goodman, J., Vanderwende, L. and Suzuki, H.: Multi-document Summarization by Maximizing Informative Content-words, *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI)*, pp. 1776–1782 (2007).

[30] Zhao, J., Qiu, X., Liu, Z. and Huang, X.: Online Distributed Passive-Aggressive Algorithm for Structured Learning, *Proceedings of Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data - 12th China National Conference, CCL*, pp. 120–130 (2013).

[31] 平尾　努，鈴木　潤，磯崎秀樹：最適化問題としての文書要約，人工知能学会論文誌，Vol. 24, No. 2, pp. 223–231 (2009).