文書要約のための数理的手法

高村　大也

文書要約は，言語的な知識と数理的な知識の両方が必要とされる研究課題である．入力文書の内容をできる限り含んだ冗長性の少ない要約を生成するために，組合せ最適化が活躍する分野である．本稿では，文書要約の基礎的な知識を導入し，組合せ最適化がどのように役立つかに重点を置きつつ，文書要約のための数理的な手法を紹介する．また，近年急速に発展しつつあるニューラルネットワークに基づく手法についても紹介する．

キーワード：自然言語処理，文書要約，組合せ最適化，整数計画問題

1. はじめに

大量の文書を目の前にして途方に暮れ，とりあえずその概要だけ知りたいという状況を経験した方は多いのではないだろうか．ある案件についての大量の報告書を読むとき，ある事件に関してこれまでに書かれた新聞記事群を読むとき，情報を求めて巨大な電子掲示板のスレッドを読まなくてはならないとき，同様に大量のツイート[1]を読まなくてはならないときなど，枚挙に暇がない．このような状況で概要を自動的に生成してくれる技術が，**文書要約技術**である．

文書要約は自然言語処理分野の応用的研究課題の一つであり，長い文書や多数の文書を入力として，一つの簡潔で短い文書（すなわち要約）を生成する課題である．文書要約は，組合せ最適化が非常にわかりやすい形で活躍する研究課題でもあり，本稿では組合せ最適化に焦点を置きながら，文書要約課題について解説する．文書要約課題は，**単一文書要約**と**複数文書要約**に大別される．二つの違いは入力文書数であり，前者は入力が一つの文書であるが，後者は入力が複数の文書である．複数文書要約における入力文書は互いに無関係というわけではなく，すべての文書が同じ話題を記述している状況を想定することになっている．たとえば，ある事件に関する複数の新聞社の記事があり，この要約を生成するような状況である．

また，単一文書要約か複数文書要約かにかかわらず，文書要約課題に対する手法には，**抽出的手法**と**生成的手法**がある．抽出的手法は，入力文書の一部を選択して並べることで要約を生成する手法である．典型的な抽出的手法としては，文選択手法がある．この場合は，入力文書を文集合とみなし，この中から最もよい部分集合を選択することになる．つまり，この場合文書要約は，一種の組合せ最適化問題に帰着している．文選択手法には，要約の柔軟性は下がるものの，文レベルの文法性が保証されるという利点がある．一方，生成的手法は，入力文書で用いられている文や単語に限定せず，さまざまな言語表現を用いて要約を生成する手法である．生成的手法は柔軟性は高いが，文法的な文，自然な文章を生成することが難しい．それゆえ，要約研究の対象はかつて主に抽出的要約であった．しかし近年，ニューラルネットワークに基づいた，文法的な文，自然な文章を生成する技術の開発が進み，これに伴い生成的手法の研究が非常に盛んになってきている．

2. 文書要約研究の歴史

文書要約課題の性質や，その現在位置を把握するために，ここで文書要約研究の歴史を簡単に振り返ってみる．2000 年前後には，機械学習の分類器を用いて，各文が要約の一部として選ばれるか否かを分類することで，要約生成を行う手法が提案されている．ナイーブベイズ分類器 [1]，決定木 [2]，対数線形モデル [3]，サポートベクトルマシン [4] など，さまざまな分類器が，単一文書要約に適用された．しかし，入力文書の最初の箇所を選択する**リード法**という単純なベースラインが非常に強力で，これを上回る手法の開発に研究者達は苦労している．その後，単一文書要約の研究はやや下火となる．

一方，2005 年頃から複数文書要約の研究が盛んになってきた．それまで，要約手法は，通常は手続きとし

たかむら　ひろや
東京工業大学科学技術創成研究院
〒 226–8503 神奈川県横浜市緑区長津田町 4259
takamura@pi.titech.ac.jp

[1] https://twitter.com

て記述され，全体としてどういう要約がよいとされるかは明確に表現されていなかった．つまり，目的関数が明示的に与えられていなかった．しかし，この頃から，さまざまな目的関数が提案されるようになってきた．複数文書要約では，1節で述べたように，入力文書群の各文書は同じ話題を記述している．よって，重要な内容は入力文書中の多数の文書で記述されている．逆に言うと，多数の文書で記述されている内容が重要であることがわかる．近似的には，多数の文書で出現する単語が重要な単語であり，そのような単語を多く含む文が重要文となることが多い．このような単語の出現頻度による手がかりは強力であるが，単一文書要約では同様の強力な手がかりは存在せず，この点において，単一文書要約は複数文書要約より難しいといえる [5]．ただし，これは諸刃の剣であり，同内容が複数の文書で記述されているということは，類似した文を複数選択してしまう危険があるということである．当然ながら，そのような冗長な要約は望ましいものではない．2005 年頃からの複数文書要約の研究の焦点は，この冗長性の削減にあった．冗長度合いを目的関数とし，それを最小化する（あるいは非冗長度を最大化する）組合せ最適化問題を解くことで要約を生成する試みが多くなされた．それらの具体例を 3 節で紹介する．

3. 整数計画問題による複数文書要約モデル

組合せ最適化問題による文書要約モデルをいくつか紹介する．互いの違いは，冗長度あるいは非冗長度をどのように表現するか，という点である．

3.1 McDonald のモデル

McDonald は 2007 年に，文書要約を次のような最適化問題でモデル化することを提案した [6]:

$$\max. \quad \lambda \sum_i r_i x_i - (1-\lambda) \sum_{i<j} s_{ij} y_{ij} \tag{1}$$

$$s.t. \quad \sum_i c_i x_i \le K, \tag{2}$$

$$\forall i, \forall j, y_{ij} - x_i \le 0, \tag{3}$$

$$\forall i, \forall j, y_{ij} - x_j \le 0, \tag{4}$$

$$\forall i, \forall j, x_i + x_j - y_{ij} \le 1, \tag{5}$$

$$\forall i, x_i \in \{0,1\}, \tag{6}$$

$$\forall i, \forall j, y_{ij} \in \{0,1\}. \tag{7}$$

ここで，x_i は文 i が要約に含まれれば 1，そうでなければ 0 となる決定変数である．K は許容される要約長を表す定数であり，c_i は文 i の長さを表す定数である．つまり，$\sum_i c_i x_i \le K$ は，要約長が K 以下であるという制約である．y_{ij} は文 i と文 j が両方とも要約に含まれていれば 1，そうでなければ 0 となる決定変数である．y_{ij} のこのような意味は，式 (3), (4), (5) により与えられる．r_i は文 i と入力文書全体の類似度を表す定数である．一方，s_{ij} は文 i と文 j の類似度を表す定数である．どちらも，たとえば，bag-of-words ベクトル[2]の余弦値などで表すことができる．目的関数である式 (1) の第 1 項は，全体と類似した文が多く選ばれると大きくなる．しかし，これだけでは，冗長な要約が生成されてしまう可能性が高い．よって，第 2 項により，互いに類似している文が含まれているとペナルティが与えられるようにしている．定数 λ を調整することで，第 1 項および第 2 項のバランスをとる[3]．

3.2 最大被覆モデル

文書要約は最大被覆問題によってもモデル化されており [7–9]，**最大被覆モデル**と呼ばれる．3.1 節のモデルと同様に長さに制限があるので，より厳密には予算制約付きの最大被覆問題をもとにした整数計画問題であり，次のように定式化される：

$$\max. \quad \lambda \sum_i r_i x_i + (1-\lambda) \sum_k b_k z_k \tag{8}$$

$$s.t. \quad \sum_i c_i x_i \le K, \tag{9}$$

$$\forall k, \sum_i a_{ik} x_i \ge z_k, \tag{10}$$

$$\forall i, x_i \in \{0,1\}, \tag{11}$$

$$\forall k, z_k \in \{0,1\}. \tag{12}$$

最大被覆モデルにおける基本的な考え方は，なるべく多くの種類の単語を含むような要約を作るというものである．ただし，各単語 k にはスコア b_k が与えられているので，要約に含まれる単語のスコアの総和 $\sum_k b_k z_k$ を大きくすることを考える．ここで z_k は単語 k が要約に含まれるときに 1 となり，そうでなければ 0 となる決定変数である．z_k がこのような意味をもつことは，$\sum_i a_{ik} x_i \ge z_k$ により保証されている．ただし，a_{ik} は文 i に単語 k が含まれているときに 1 となり，そうでなければ 0 となるような定数である．3.1 節のモデルと同様に，目的関数に $\sum_i r_i x_i$ を導入し，定数 λ により第 1 項と第 2 項のバランスをとる．

最大被覆モデルは有効だが，一つの欠点がある．そ

[2] 各要素が単語に対応し，要素の値は，対応する単語の対象文（あるいは文書）における出現回数となっているようなベクトルのことである．出現回数の代わりに，出現していれば 1，そうでなければ 0 とする場合もある．

[3] McDonald の論文 [6] では，λ は導入されていないが，一般性をもたせるためにここでは導入した．

れは，入力文書の内容をなるべく被覆するという目的を，単語を被覆することに置き換えている点である．単語という言語単位は扱いやすい半面，文構造などを無視することになり，文の意味を表すためには不十分である場合がある．単語でなく，別の言語単位を使うこともできるが，文構造を充分に利用できないという点では同じである．

3.3 施設配置モデル

文の意味を単語の集合で表さなくてはならないという最大被覆モデルの欠点を補う可能性があるモデルとして，**施設配置モデル** [10] を紹介する．これは，文書要約を施設配置問題として定式化したものであり，より厳密には，需要点集合と施設の候補点集合が一致している予算制約付き p-メディアン問題である [11]:

$$\max. \quad \sum_{ij} e_{ij} y_{ij} \tag{13}$$

$$s.t. \quad \sum_{i} c_i x_i \leq K, \tag{14}$$

$$\forall i, \forall j, y_{ij} \leq x_i, \tag{15}$$

$$\forall j, \sum_{i} y_{ij} = 1, \tag{16}$$

$$\forall i, y_{ii} = x_i, \tag{17}$$

$$\forall i, x_i \in \{0, 1\}, \tag{18}$$

$$\forall i, \forall j, y_{ij} \in \{0, 1\}. \tag{19}$$

ここで，y_{ij} は文 i に文 j を割り当てるときに 1 となり，そうでなければ 0 となる決定変数である．ここでわれわれが期待するのは，文 i は要約の一部として選択されており，文 j の内容は文 i により被覆されているということである．たとえば，“Barack Obama is an American politician” という文が要約に入っていれば，“Barack Obama is a politician” という文は不要であり，この場合前者により後者を被覆すればよい．逆に言うと，このような文と文との割当が可能になるように，文を選択することでよい要約ができるはずだ，というのが基本的な考え方である．この割当のよさを表す定数が e_{ij} である．また，文 i に文 j を割り当てるときには，文 i は選択されている必要がある．このことを保証するために，$\forall i, \forall j, y_{ij} \leq x_i$ なる制約を課している．また，すべての文がいずれかの文に割り当てられることを保証するため，$\forall j, \sum_i y_{ij} = 1$ なる制約を課している．最後に，選択された文はそれ自身に割り当てられるものとするため，$\forall i, y_{ii} = x_i$ としている．

3.4 その他の話題

実際は，文集合を選択しただけでは充分でなく，選択した文を適切な順序に整列する必要がある．文選択問題と文の整列問題を同時に解く手法などが提案されている [12].

最大被覆モデルは，後に劣モジュラ関数の最大化問題による要約モデルとして一般化された [13]．最適化問題を解く際には，CPLEX などのソルバーに実装されている厳密解法を使う場合や，貪欲法を使う場合などがある．たとえば最大被覆モデルであれば，Khuller らによる貪欲法 [14] が用いられている．より一般に劣モジュラ関数の最大化に対しても，同様の貪欲法が用いられている．

4. 整数計画問題による文要約

前節では文選択による文書要約について紹介したが，各文そのものを短くしたいような状況もあるだろう．本節では，文書要約でなく，文要約について紹介する．つまり，入力が単一の文であるような要約である．

整数計画問題による文要約手法は，基本的に抽出的手法である．ただし，抽出単位は単語となる．入力文の一部の単語のみを残すことで，短い要約文を生成する．それゆえ，文圧縮と呼ばれることも多い．当然ながら問題となるのは要約文の文法性である．単語を削除することで，文は容易に非文法的になってしまう．これをどのように防ぐかが問題となる．ここでは，Clarke and Lapata の手法 [15] を紹介する．ただし，ここでは直感的な説明に留めるものとする．

目的関数は，trigram 言語モデルによる要約文の生成確率である．ここで trigram 言語モデルとは，文 $w_1 w_2 \cdots w_n$ の生成確率を $\prod_{i=1}^{n} P(w_i | w_{i-1} w_{i-2})$ で表すモデルである[4]．言語モデルは $P(w_i | w_{i-1} w_{i-2})$ というパラメータから成るが，これは通常の大規模テキストデータから推定する．これを目的関数にするということは，この言語モデルにより文が生成される確率という観点から，要約文がどのくらいもっともらしいかを考えるということである．ただし，trigram 言語モデルでは確率値を積算することで文全体の確率値を算出するので，短い文に高い確率値が与えられる．Clarke and Lapata のモデルでは，極端に短い文が生成されることを防ぐため，要約文がある長さ以上であることを制約として加えている．

言語モデルだけでは，非文法的な要約文を生成してしまう可能性が高い．よって，さまざまな制約を加えることで，可能な限り非文法的な文の生成を防ぐ．た

[4] w_0 と w_{-1} は文頭を表すダミー単語である．

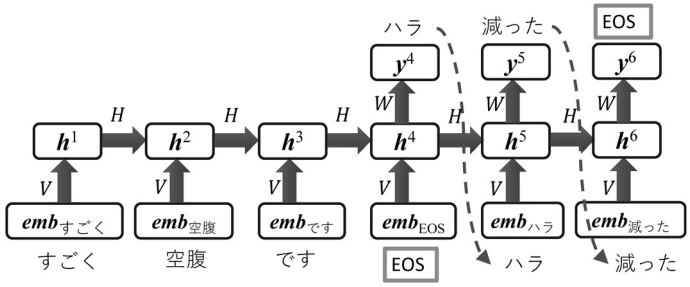

図 1　系列変換モデルによる文要約

とえば，“彼は那須に行った”の“那須”を削除してしまうと，“彼はに行った”という非文法的な文が生成されてしまう．これを防ぐために，“那須に”のような句において，名詞部分を削除するときは助詞も一緒に削除するという制約を加えている[5]．このような細かい文法的な制約を数多く課すことで，非文法的な文の生成を抑制している．ただし，文法性を保証するための数々の制約を導き出すための汎用的な方法が存在するわけではない．

要約文としてのよさは，要約文中の各単語のスコアの総和で与えている．各単語のスコアは，文構造においてその単語がどのくらい深い位置に出現したか，その単語がどのくらい珍しいかなどの情報を用いて計算される．また，教師付き学習を用いて要約文のよさを測る方法も提案されている．

この手法は文要約を単体で解くためのものであるが，文要約を文書要約へ組み込んだ手法も盛んに研究されている [16, 17].

5.　ニューラルネットワークによる文要約

3 節および 4 節で紹介した整数計画問題による文書要約および文要約の手法は，いずれも抽出的手法であった．しかし，ニューラルネットワークに基づいた生成的文要約手法が急激に発展しており，本節ではこれを紹介する．

5.1　系列変換モデルによる文要約

近年のニューラルネットワーク技術の発展が自然言語処理に及ぼした最も大きな影響は，言語生成技術の進化であろう．これまで，ある入力に対応した自然な文を生成することは，やや難しい課題であった．しかし，リカレントニューラルネットワークに基づく言語生成技術は，これを可能にしたと言ってよい．この技術は，**系列変換モデル** (sequence-to-sequence model) という形で，機械翻訳において発展し，対話における応答生成など多くの研究課題に適用され，文要約においても大きな成果を上げている [18–20][6]．ここでは，その最も基本的なモデルを紹介する．図 1 は，系列変換モデルで「すごく　空腹　です」を「ハラ　減った」に変換する様子を図示したものである．このように，入力文を一単語ずつ入力していき，内部状態を引き継ぐ．内部状態は数値を要素とするベクトルであるが，そこまでに入力された文の意味がなんらかの形で表現されていることが期待されている．各単語は，ベクトル表現（$\boldsymbol{emb}_{空腹}$ など[7]）に変換され，変換行列 V が積算されたうえで，$\boldsymbol{h}^2 = tanh\,(V\boldsymbol{emb}_{空腹} + H\boldsymbol{h}^1)$ のように内部状態 $\boldsymbol{h}^2$ の計算に利用される．ここで，$\boldsymbol{h}^1$ は直前の内部状態であり，これに変換行列 H が積算されて，$V\boldsymbol{emb}_{空腹}$ と足し合わされている．これを活性化関数 $tanh$ に入力することで新たな内部状態 $\boldsymbol{h}^2$ が得られる．文の最後を表す EOS というタグが入力されたら，一単語ずつ出力していく．その際，内部状態に入力文の情報がエンコードされており，それに対応する自然な文が出力される．内部状態の受け渡しは入力側と同じであるが，出力側においては，単語を出力する機構が追加されている．たとえば内部状態 $\boldsymbol{h}^5$ に行列 W が積算され $W\boldsymbol{h}^5$ となるが，ここで W は（語彙サイズ）×（内部状態の次元）の行列であるので，$W\boldsymbol{h}^5$ は語彙サイズのベクトルである．語彙とは，出力する可能性のある単語の集合であり，$W\boldsymbol{h}^5$ はその各単語にスコアを与える．このスコアが最大の単語が出力されることになる．実際は，スコアは softmax 関数により確率化される．出力された単語は，図で点線で表さ

[5] Clarke and Lapata は英語を対象にしていたので，実際の制約はここで紹介したものとは少し異なる．

[6] Rush らの論文 [18] がその先駆けであるが，彼らのモデルはリカレントニューラルネットワークに基づいた系列変換モデルとは異なるモデルである．

[7] $\boldsymbol{emb}$ は，ベクトル表現を指す表現としてよく使用される embedding（埋め込み）の略である．

れているように，次の時点での入力単語となる．語彙には，文の最後を表すEOSも含まれており，EOSが出力されたら，そこで文は終わりになる．

系列変換モデルの訓練時は，「すごく　空腹　です　EOS」と「ハラ　減った　EOS」のような原文と要約文の対を訓練データとして大量に準備し，それらがなるべく正しく要約されるようにパラメータ H, V, W, $\boldsymbol{emb}_{\text{すごく}}$, $\boldsymbol{emb}_{\text{空腹}}$, $\cdots$ の値を推定する．

以上が最も基本的な系列変換モデルであるが，実際はさまざまな変種や拡張が存在する．

5.2　工夫と課題

基本的な系列変換モデルだけではあまり高い性能は出ないことが多く，実際は入力側に双方向リカレントニューラルネットワークを用いる，リカレントニューラルネットワークの隠れ層を増やす，注視機構 (attention mechanism) を用いる [19] などの工夫がなされる．特に注視機構の有無は性能に大きな影響を与えることが知られている．注視機構とは，単語を出力する際に入力側の隠れ層の状態を利用する機構である．入力系列のどの部分を重要視して単語を出力するかを判定する仕組みを備えており，その意味で注視機構と呼ばれている．より具体的には，たとえば「減った」を出力する際に $\boldsymbol{h}^5$ だけを用いるのでなく，入力側の隠れ層の状態の重み付き和 $a_{1,5}\boldsymbol{h}^1 + a_{2,5}\boldsymbol{h}^2 + a_{3,5}\boldsymbol{h}^3$ を $\boldsymbol{h}^5$ に連結し，この長くなったベクトルを変換したものを $\boldsymbol{h}^5$ の代わりに利用して単語を生成する．このときの重み $a_{1,5}$, $a_{2,5}$, $a_{3,5}$ は，それぞれ $\boldsymbol{h}^1$ と $\boldsymbol{h}^5$, $\boldsymbol{h}^2$ と $\boldsymbol{h}^5$, $\boldsymbol{h}^3$ と $\boldsymbol{h}^5$ を用いて算出され，ここで入力系列の各部分の重要度が測られているといえる．自然言語処理におけるニューラルネットワークに基づく手法については，坪井らの書籍 [21] がよい解説書として挙げられる．

ニューラルネットワークによる文要約については，多くの研究がありいろいろなことがわかってきた．しかし，文書要約の場合は，系列が長過ぎて，リカレントニューラルネットワークがうまく働かないことが多い．ニューラルネットワークによる文書要約については，現在まさに多くの論文が発表されつつあり，今後多くのことが明らかになるだろう．

6.　おわりに

文書要約のための数理的な手法について説明した．特に，要約における冗長性の削減において，組合せ最適化が活躍することがわかる．最近は，ニューラルネットワークによる手法が急速に発展してきているが，長い文書の文書構造をうまく捉えることができるモデルは開発途中にあるといえる．また，組合せ最適化による手法で得られた知見とニューラルネットワークに基づく手法との融合も興味深い方向性だろう．

参考文献

[1] J. Kupiec, J. Pedersen and F. Chen, “A trainable document summarizer,” In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 68–73, 1995.

[2] C.-Y. Lin, “Training a selection function for extraction,” In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, pp. 55–62, 1999.

[3] M. Osborne, “Using maximum entropy for sentence extraction,” In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, Volume 4, pp. 1–8, 2002.

[4] T. Hirao, H. Isozaki, E. Maeda and Y. Matsumoto, “Extracting important sentences with support vector machines,” In *Proceedings of the 19th International Conference on Computational Linguistics*, Volume 1, pp. 1–7, 2002.

[5] A. Nenkova, “Automatic text summarization of newswire: Lessons learned from the document understanding conference,” In *Proceedings of the National Conference on American Association for Artificial Intelligence*, pp. 1436–1441, 2005.

[6] R. McDonald, “A study of global inference algorithms in multi-document summarization,” In *Proceedings of the 29th European Conference on Information Retrieval*, pp. 557–564, 2007.

[7] W.-T. Yih, J. Goodman, L. Vanderwende and H. Suzuki, “Multi-document summarization by maximizing informative content-words,” In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 1776–1782, 2007.

[8] D. Gillick and B. Favre, “A scalable global model for summarization,” In *Proceedings of the NAACL HLT Workshop on Integer Linear Programming for Natural Language Processing*, pp. 10–18, 2009.

[9] 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*, pp. 781–789, 2009.

[10] H. Takamura and M. Okumura, “Text summarization model based on the budgeted median problem,” In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 1589–1592, 2009.

[11] Z. Drezner and H. W. Hamacher (eds.), *Facility Location: Applications and Theory*, Springer, 2004.

[12] H. Nishikawa, T. Hasegawa, Y. Matsuo and G. Kikui, “Opinion summarization with integer linear programming formulation for sentence extraction and ordering,” In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 910–918, 2010.

[13] H. Lin and J. Bilmes, “A class of submodular functions for document summarization,” In *Proceedings of*

the Annual Conference of the Association for Computational Linguistics, pp. 510–520, 2010.
[14] S. Khuller, A. Moss and J. S. Naor, "The budgeted maximum coverage problem," *Information Processing Letters*, **70**, pp. 39–45, 1999.
[15] J. Clarke and M. Lapata, "Global inference for sentence compression: An integer linear programming approach," *Journal of Artificial Intelligence Research*, **31**, pp. 399–429, 2008.
[16] A. F. T. Martins and N. A. Smith, "Summarization with a joint model for sentence extraction and compression," In *NAACL-HLT Workshop on Integer Linear Programming for NLP*, pp. 1–9, 2009.
[17] T. Berg-Kirkpatrick, D. Gillick and D. Klein, "Jointly learning to extract and compress," In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 481–490, 2011.
[18] A. M. Rush, S. Chopra and J. Weston, "A neural attention model for abstractive sentence summarization," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, 2015.
[19] S. Chopra, M. Auli and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," In *Proceedings of Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 93–98, 2016.
[20] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gùlçhere and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, 2016.
[21] 坪井祐太，海野裕也，鈴木潤，『深層学習による自然言語処理（機械学習プロフェッショナルシリーズ）』，講談社，2017.

