係り受け構造の刈り込みと CRF による文の要約

野本 忠司
国文学研究資料館
nomoto@acm.org

1　はじめに

文要約は、社会の急速な情報化を背景に RSS フィード、メール・ニュース、街頭の電光掲示板、タクシー、バス、新幹線車内でのニュース速報等様々な場面で活用され、日常的に接する機会が多くなってきた。本稿では、このような文レベルの要約を自動的に構成する手法について検討する。

文要約は、従来から様々な手法が提案されてきたが、大きく二つの流れに分けることができる。一つは、要約・原文間の確率的なマッピングをベースにしたアプローチ [1] と、統語構造に基づく生成を特徴とするアプローチである [2]。前者は、要約と原文の対応付けさえすれば、(a) 煩雑なルールを考えなくともよい（データからすべて学習できる）、(b) 適用範囲が広いというメリットがある。しかし、その一方で要約率を含む、出力に対する柔軟な制御がしにくい。対して、後者はルールがすべて明示的に表現されるため、出力に対して微妙な調整ができる反面、ルールが煩雑になり、一般性、頑健性について不安が残る。

ところで、文要約では出力文が言葉として自然でなければならない、という強い要請がある。実際、読めなければ、使い物にならない。確率ベースのアプローチでは、一般にこの要件を十分満たすことができない。文が不自然なところで切れたり、読解に必要な文法要素が欠落することがよくある。そこで、本稿では、係り受け構造の刈り込みによりあらかじめ妥当な要約候補を生成し、CRF（条件付き確率場）を用いてそれらを選別する、新しい文生成型の要約手法を紹介する。本アプローチの有効性を検証するため、実データ（日経ネットニュース）を使い、従来手法との比較を行った [3] 。

2　アプローチ

本アプローチでは、文要約問題を以下のように定式化する。

$$\mathbf{y}^\star = \arg\max_{\mathbf{y}\in G(S)} p(\mathbf{y}|\mathbf{x}). \tag{1}$$

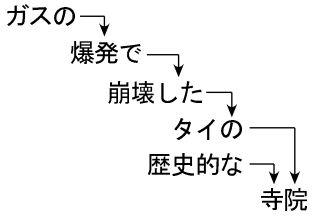

図 1: 係り受け構造

ここで、S は入力文、$G(S)$ は S から生成される文法性を満たす要約候補文の集合、$\mathbf{x}$ は入力文の属性列とする。さらに、p のモデルとして標準的な CRF を導入し、

$$\begin{aligned} &p(\mathbf{y}|\mathbf{x}) \propto \\ &\exp\left(\textstyle\sum_{k,j}\lambda_j f_j(y_k, y_{k-1}, \mathbf{x}) + \sum_i \mu_i g_i(x_k, y_k, \mathbf{x})\right) \\ &= \exp[\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})] \end{aligned} \tag{2}$$

と定義する。$\mathbf{x}, \mathbf{y}$ は、入力文の文節列に対応し、特に $\mathbf{y}$ は 0/1 のバイナリ列とする。（0 は「消去」、1 は「維持」の意。） $\mathbf{w}, \mathbf{f}$ はそれぞれ重み、属性の列とする。

従って、本アプローチは候補文の生成と CRF による候補の順位付けという２つの処理より成る。以下、それぞれの処理について説明する。

3　候補文の生成と選別

要約候補文は、入力文の係り受け構造（依存構造）を刈り込むことで生成する。刈り込みは末端の終端ノードからルートに向かって行う。例えば、「ガスの爆発で崩壊したタイの歴史的な寺院」の係り受け構造（図 1）に対する刈り込みは、表 1 のようになる。

以上のことを一般的に考えるために、*Terminating Dependency Path*（以下、TDP）という概念を導入する。TDP とは、係り受け構造においてどこからも係り受けを受けないノード（起点ノードと呼ぶ）からルートへのパスを

表 1: 刈り込み例。括弧内は省略可能。

a. ガスの爆発で崩壊したタイの (歴史的な) 寺院
b. 爆発で崩壊したタイの (歴史的な) 寺院
c. 崩壊したタイの (歴史的な) 寺院
d. タイの (歴史的な) 寺院
e. 歴史的な寺院
f. 寺院

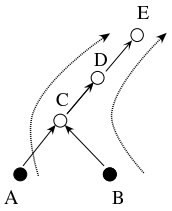

図 2: 係り受け構造と TDP

表す。図 1 では、「ガスの」と「歴史的な」が起点ノードになる。従って、TDP は < ガスの, 爆発で, 崩壊した, タイの, 歴史的な, 寺院 > と < 歴史的, 寺院 > となる。

次に、得られた各 TDP について、ルートまでのサフィックス（接尾）をすべて求めることにする。例えば、図 2 の 2 つの TDP <A C D E> と <B C D E> では、サフィックスは以下のようになる。(E をルートとする。< > は空サフィックス。)

$$\mathcal{T}(p_1) = \{\text{<A C D E>}, \text{<C D E>}, \text{<D E>}, \text{<E>}, \text{<>}\}$$
$$\mathcal{T}(p_2) = \{\text{<B C D E>}, \text{<C D E>}, \text{<D E>}, \text{<E>}, \text{<>}\}$$

ここで各サフィックスの組み合わせをすべて考える

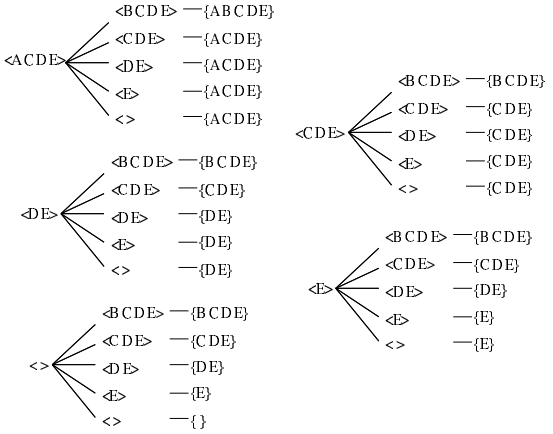

図 3: TDP サフィックスの結合

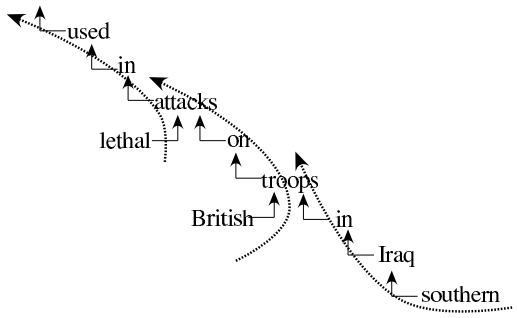

図 4: 英語の係り受け構造と TDP

(図 3)。このそれぞれの組み合わせが要約候補文を表す。(組み合わせとは各サフィックスの集合和と考える。)

このアプローチの興味深い点は、英語についても同等な議論ができるところにある。いま、

> *"A senior British official was quoted yesterday as accusing Iran of supplying explosive technology used in lethal attacks on British troops in southern Iraq."*

という文を考える。簡単のため、*technology used in lethal* 以降について係り受け構造を考えると図 4 のようになる。起点ノードは、*lethal*, *British*, *southern* となり、3 つ TDP を得る。日本語と同様、TDP サフィックスの組み合わせを考えると、表 2 のような刈り込みが得られる。

一般に、TDP の刈り込みと結合によって生成された短縮文は文法的である場合が多いが、個別言語の特殊性を考慮しないと文法性が確保できないケースがある。例えば、日本語における、「こと」や「の」などの形式名詞や引用の「と」の直前での刈り込み、英語における前置詞の直後での刈り込みは、不自然な文を生む。前者の例としては、「首相は、と述べた。」、後者の例としては、*"A senior British official was quoted yesterday as accusing Iran of supplying explosive technology used in."* などが挙げられる。このため本稿の実装では、KNP（黒橋・長尾パーサ）[1] における「隣受絶対」「形副名詞」「〜ため」「スルナル」での刈り込みを禁止した。一方、読み易さの観点から、提題表現「〜は」「〜も」は保持することにした。

ところで、刈り込みによって切り落とされた文節に 0 のラベル、残った文節に 1 のラベルを付与すれば、それぞれの要約候補文を 0/1 の列として表現できる。このバイナリ列と原文の属性情報を CRF に与え、最適候補文の選別を行う。

[1] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/top-e.html

表 2: 英語の刈り込み例

[...] supplying explosive technology used in lethal attacks on British troops in southern Iraq
[...] supplying explosive technology used in lethal attacks on British troops in Iraq
[...] supplying explosive technology used in lethal attacks on British troops
[...] supplying explosive technology used in lethal attacks on troops
[...] supplying explosive technology used in lethal attacks
[...] supplying explosive technology used in attacks

属性情報としては、JUMAN/KNP が出力する様々な文節レベル、形態素レベルの情報をもとに構成にした。文節レベルでは、例えば、係り受けタイプ、外の関係か否か、同格の別、文頭・文末の別、形態レベルでは、文節内の最初と最後の自立語（主辞）の情報（見出し情報、品詞情報）と、それらがタイトルに出現するか否かの情報を利用した。さらに、文節前後のコンテキスト (語彙、品詞など）の情報、地名、人名、組織名、時間といった分類も利用した。また、自立語については、その TFIDF も属性として導入した。今回の使用したデータでは、属性の総数は 80,000 余に上った。なお、CRF は、GRMM と呼ばれるツールキットを利用して構築した[2]。以降では、簡便のため、本アプローチを GST (Generic Sentence Trimmer) と呼ぶ。

4 係り受けパスを用いた確率的文要約

本節では、追試が容易なこと、要約モデルとしても十分な妥当性があることから、Yamagata ら [3] によって提案された係り受けパスを利用した文要約手法をベースライン・モデルとして考える。この手法は、ある意味でよく知られた Knight ら [1] の Noisy-Channel 型要約モデルの裏返しと言える。つまり、Knight らのモデルは、要約から原文を出力するモデルであるの対して、Yamagata らのモデルは、原文から要約を出力するモデルである。簡単のため、以下では Yamagata らのモデルを *Dependency Path Model* (DPM) と呼ぶことにする。DPM では次の式を考える。但し、$\mathbf{y} = \beta_0 \dots \beta_{n-1}$ は、任意の文を表す文節列とする。

$$h(\mathbf{y}) = \alpha f(\mathbf{y}) + (1-\alpha) g(\mathbf{y}) \tag{3}$$

f は文節重要度、g は係り受けの強度を示す。ここで、

$$f(\mathbf{y}) = \sum_{i=0}^{n-1} q(\beta_i), \tag{4}$$

また、

$$g(\mathbf{y}) = \max_s \sum_{i=0}^{n-2} p(\beta_i, \beta_{s(i)}). \tag{5}$$

とする。$p(\beta_i, \beta_{s(i)})$ はマッピング s のもとでの文節 $\beta_i, \beta_{s(i)}$ の間の係り受けの強度を表し、以下で定義する。

$$p(\beta_i, \beta_j) = \begin{cases} \log S(t) & \text{if DL}(\beta_i, \beta_j) \neq \infty \\ -\infty & \text{otherwise} \end{cases} \tag{6}$$

DL(β_i, β_j) は文節 β_i, β_j の係り受けパス上の距離を表す。例えば、図 1 では以下のようになる。

$$\text{DL(ガスの, 爆発で)} = 1$$
$$\text{DL(ガスの, 崩壊した)} = 2$$
$$\text{DL(ガスの, 歴史的な)} = \infty$$

文節「ガスの」は「爆発で」に一回の係り受けで到達できるため、その DL 距離は 1 となる。同様にして、「ガスの」と「崩壊した」の DL 距離は 2、しかし、「ガスの」から係り受けを順方向にたどって、「歴史的な」に到達することはできない。この時の DL 距離を ∞ とする。

次に、三つ組 $t =< C_s(\beta_i), C_e(\beta_j), \text{DL}(\beta_i, \beta_j) >$ を定義する。いま、β_i を係り文節、β_j を受け文節とする。$C_s(\beta_i)$ とは、係り文節のクラス（主辞の品詞、活用形等で定義）を、$C_e(\beta_j)$ は受け文節のクラス（主辞の品詞、文末か否か、判定詞を含むか否か等の観点で分類）を表す。（詳しくは、福冨 [5] を参照。）

$S(t)$ は次の式で与えられる。

$$S(t) = \frac{\text{要約文で出現した } t \text{ の総数}}{\text{原文データ中の三つ組の総数}} \tag{7}$$

従って、文節 β_i と β_j の係り受け強度は、関連文節のクラスと DL 距離によって定まる。

一方、文節重要度は以下の式で定義する。

$$q(\beta) = \log p_c(\beta) + \text{tfidf}(\beta) \tag{8}$$

$p_c(\beta)$ は、文節 β のクラス（主辞の品詞、その付属語の有無また助詞等により分類）が要約に出現する確率を表す [5]。tfidf は通常の情報検索におけるそれである。

最終的に、要約文は $\arg\max_{\mathbf{y}} h(\mathbf{y})$ を満たす $\mathbf{y}$ となる。本稿の実験では、α は、諸岡 [4] を参考に 0.1 とした。

[2]http://mallet.cs.umass.edu

表 3: 文の自然さ（平均）

モデル/要約率	50%	60%	70%
DPM	2.222	2.372	2.660
GST	**3.430**	**3.820**	**3.810**
人間	−	4.45	−

表 4: 意味的な近さ（平均）

モデル/要約率	50%	60%	70%
DPM	2.210	2.548	2.890
GST	**2.720**	**3.181**	**3.405**

5 実験と評価

実験コーパスは、日経ニュースメールから採取した要約文をもとに、日経ネットから対応する原文を人手で収集し構築した。最終的に要約・原文 1,401 対を得た。10 回交差検定で、50%, 60%, 70%と要約率と変動させながら、原文すべてについて、DPM, GST による要約を作成した。評価は、各要約率のもとでの DPM, GST のサンプルを交差検定の出力結果からそれぞれ 200 文を無作為抽出し、1（低）から 5（高）までのスケール で、日本語母語話者（大学院生、ポスドク）6 名に、出力要約の (a) 日本語としての自然さ（滑らかさ）と (b) 日経ニュースが実際に配信した要約との意味的な近さについて、直感的に評価してもらった。比較のため、配信要約の自然さ（200 文）も評価した。その結果が、表 3 と表 4 である。自然さ、意味的な近さ、いずれのカテゴリーでも GST が DPM を大きく上回った。結果は、読み易さについては、刈り込みによる候補生成、意味的な近さにおいては、CRF に依るところが大きいと考えられる。なお、配信要約の要約率は、60%程度であった。

6 おわりに

以上、機械要約文の可読性の向上に向け、係り受け構造の刈り込みにより、あらかじめ妥当な要約候補を生成し、CRF を用いてそれらを選別する新しい文要約手法を提案した。

実データをもとにコーパスを構築し、可読性と配信要約との内容的近さの観点から、人間による評価実験を行い、係り受けパスを用いた従来手法を上回る性能を確認した。本アプローチは一部を除いて言語に依存しない。そのため、英語へ適用も比較的容易であると想像される。今後はこの点を検証して行く予定である。

参考文献

[1] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 2002.

[2] Stefan Riezler, et al. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical functional grammar. *Proc. of HLT-NAACL*, 2003.

[3] Kiwamu Yamagata, et al. Sentence compression using statistical information about dependency path length. *Proc. of TSD*, 2006.

[4] 諸岡裕平, 江嵜誠, 高木一幸, 尾関和彦. 重要文抽出と文簡約を併用した新聞記事の自動要約. 言語処理学会 10 回大会論文集, 2004.

[5] 福冨 諭, 高木 一幸, 尾関 和彦. 確率的な手法による日本語文簡約. 言語処理学会 13 回大会論文集, 2007.

付録

DPM と GST の出力サンプル。要約率は 60%。

DPM　続く外国人の退避が相次ぎ、１９日までに数千人が近隣国に脱出した。

GST　イスラエルの攻撃が続くレバノンから外国人の退避が相次ぎ、１９日までに数千人が脱出した。

−−

DPM　オリンパスはデジタルカメラなどにソフトウエアの開発業務を、委託する。

GST　オリンパスはソフトウエアの開発業務を、全面的に日本ＩＢＭに委託する。

−−

DPM　書籍の年間ベストセラーが発表され、２２７万部の「国家の１位になった。

GST　２００６年の書籍の年間ベストセラーが発表され、「国家の品格」が１位になった。

−−

DPM　政府は中心市街地活性化本部の市町村の国が支援対象として認定する手順などを基本方針を

GST　政府は中心市街地活性化本部の初会合を開き、認定する手順などを示した基本方針をまとめた。

−−

DPM　名古屋南貨物駅で、専用列車の運転開始出発式を行った。

GST　日本貨物鉄道はトヨタ自動車向け専用列車の運転開始出発式を行った。

−−

DPM　防衛大手の所有する欧州航空機メーカーのエアバス株を欧州航空・防衛大手の方針を発表した。

GST　英ＢＡＥシステムズは６日、エアバス株を欧州航空・防衛大手のＥＡＤＳに売却する方針を決めたと発表した。