軽量な文短縮手法

平尾 努　　鈴木 潤　　磯崎 秀樹
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
{hirao,jun,isozaki}@cslab.kecl.ntt.co.jp

1　はじめに

原文の大意を保持したままそれを短縮するためには，主語，述語の関係を崩すことなく短縮文に保存しなければならない．これを実現するため，多くの文短縮手法は文を構文木として表現し，その枝を刈ることによって短縮文を得ている [4, 12, 11, 10]．しかし，構文解析処理に要する時間は無視できるほど短いものではないので，リアルタイムに短縮文を得たい場合には問題がある．さらに，人間が文を短縮する場合には，木の枝を刈るだけでは不可能な短縮を行う場合もあるため，人間のような柔軟な文の短縮が困難であるという問題もある．

一方，文を木とみなすのでなく，単語列とみなし，各単語を短縮文に採用するか否かを決定する問題としてみなす手法も提案されている [8, 9, 1, 2]．これらの手法には木の枝刈りという強い制約がないため，より柔軟な文の短縮が可能である．しかし，単語をあらわす素性を工夫しなければ性能が劣化するため，構文情報を素性として埋め込んでいる．よって，構文木の枝刈りに基づく手法と同様，処理時間に問題がある．また，短縮文も構文構造をなるべく保存する傾向にあるため，人間に近い文の短縮が難しいという点も同様である．

そこで，本稿では，構文解析器を必要としない文短縮手法を提案する．文を構文木ではなく単語列としてとらえる点では既存手法と同様であるが，素性として，新たに単語の出現位置に基づく重要度 (intra-sentence positional term weighting:IPTW) と文短縮のための言語モデル (patched language model:PLM) を提案する．素性の重みパラメータは誤り最小化 (MCE) [3] 学習を用いて最適化する．

提案手法を従来手法，依存構造木の枝刈りによるベースラインと比較した結果，統計的に有意な差で優れていることを確認した．また，IPTW，PLM の双方の有効性を確認し，依存構造情報が必ずしも有効でないことを確認した．

2　文短縮モデル

2.1　組合せ最適化問題としての文短縮

文短縮は，N 個の単語列から M 個の部分単語列を選択する問題としてとらえることができる．原文を $\boldsymbol{x}$= x_1 ,x_2, …, x_j, …, x_N，短縮文を $\boldsymbol{y}$ =y_1, y_2, …, y_i, …, y_M とする．ここで，Hori らの定式化 [2] に基づき，原文 $\boldsymbol{x}$ から得た短縮文 $\boldsymbol{y}$ の評価関数を以下の式で定義する．$\boldsymbol{\lambda}$ はパラメータベクトルである．

$$f(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\lambda}) = \sum_{i=0}^{M} g(y_i;\boldsymbol{\lambda}_g) + h(y_{i+1},y_i;\boldsymbol{\lambda}_h) \qquad (1)$$

式 (1) の第 1 項は短縮文に含まれる単語の重要度を評価し，第 2 項は単語間の繋がりを評価する．ここで，式 (1) を最大とする $\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, f(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\lambda})$ は，動的計画法を用いて得ることができる．そのアルゴリズムを図 1 に示す．なお，$\phi[i,x_k]$ は，y_0 から y_i(原文における x_k) までの部分単語列のスコアの最大値をあらわす．

$x_0 = \hat{y}_0 =$ <s>
$x_{N+1} = y_{\hat{M+1}} =$ </s>
for $(k=1\ ;\ k \le N-M+1\ ;\ k{+}{+})$ **do**
  $\phi' \leftarrow g(x_k;\lambda_{\mathbf{g}}) + h(x_k,x_0;\lambda_{\mathbf{h}})$
  if $\phi' > \phi[1,x_k]$ **then**
    $\phi[1,x_k] \leftarrow \phi'$
    $\hat{y_1} = x_k$
  end
end
for $(i=1\ ;\ i \le M-1\ ;\ i{+}{+})$ **do**
  for $(j=i\ ;\ j \le N-M+i\ ;\ j{+}{+})$ **do**
    for$(k=i-1\ ;\ k \le j-1\ ;\ k{+}{+})$ **do**
      $\phi' \leftarrow \phi[i,x_k] + g(x_k;\lambda_{\mathbf{g}}) + h(x_k,x_j;\lambda_{\mathbf{h}})$
      if $\phi[i+1,x_k] \leftarrow \phi'$
        $y_{\hat{i+1}} = x_k$
      endif
    end
  end
end
return $\hat{\boldsymbol{y}}$

図 1　動的計画法を用いた短縮文の取得アルゴリズム

2.2　素性

単語重要度を評価する $g(y_i;\boldsymbol{\lambda_g})$，単語間の繋がりを評価する $h(y_{i+1},y_i;\boldsymbol{\lambda}_h)$ をそれぞれ以下の IPTW，PLM に基づき定義する．

2.2.1　IPTW

本稿では，文内の単語の重要度は出現位置に依存して決まるという仮定に基づきその重要度を以下の混合正規分布を用いて定義する

$$G(k(y_i,\boldsymbol{x});\boldsymbol{\lambda_G}) = \lambda_{m1}\frac{1}{\sqrt{2\pi}\lambda_{\sigma_1}}\exp\left(-\frac{1}{2}\left(\frac{k(y_i,\boldsymbol{x})-\lambda_{\mu_1}}{\lambda_{\sigma_1}}\right)^2\right) + \lambda_{m2}\frac{1}{\sqrt{2\pi}\lambda_{\sigma_2}}\exp\left(-\frac{1}{2}\left(\frac{k(y_i,\boldsymbol{x})-\lambda_{\mu_2}}{\lambda_{\sigma_2}}\right)^2\right) \qquad (2)$$

正規分布を 2 つ混合した理由は文の主語，述語が頻出する位置の重要度を他の位置よりも相対的に高くするためである．ここで，$k(y_i,\boldsymbol{x})$ は，単語 y_i の原文 $\boldsymbol{x}$ における出現位置を返す関数である．$k(y_i,\boldsymbol{x}) = 0$ であれば，y_i=BOS であり，$k(y_i,\boldsymbol{x}) = 1$ であれば，y_i=EOS である．λ_μ,λ_σ

は正規分布の平均，分散をあらわし，λ_m は混合比をあらわす．

式 (2) を用いて，$g(y_i;\boldsymbol{\lambda_g})$ を以下の式で定義する．

$$g(y_i;\boldsymbol{\lambda_g}) = \begin{cases} \lambda_w \mathrm{IDF}(y_i) \times N(k(y_i,\boldsymbol{x});\boldsymbol{\lambda_N}) \\ \qquad\qquad y_i \text{ が内容語の場合} \\ \mathrm{Constant} \times N(k(y_i,\boldsymbol{x});\boldsymbol{\lambda_N}) \\ \qquad\qquad \text{それ以外} \end{cases} \tag{3}$$

2.2.2 PLM

従来の文短縮手法では短縮文の言語尤度を評価する際に N グラム言語モデルが用いられている．しかし，一般的な N グラム言語モデルは様々な長さの文を含む大量のコーパスを用いて計算される．このため，文短縮というタスクに適用する際，必ずしも我々の直観に合致した結果とならない．そこで，本稿では，原文から得られる情報と一般的な N グラム言語モデルを組合せた以下の言語モデルを提案する．

$$P_{\mathrm{plm}}(y_{i+1}|y_i) = \begin{cases} 1 & y_i \text{ と } y_{i+1} \text{ が原文において} \\ & \text{隣り合っている} \\ P_{\mathrm{ng}}(y_{i+1}|y_i) & \text{それ以外} \end{cases} \tag{4}$$

なお，$P_{\mathrm{ng}}(y_{i+1}|y_i)$ は，通常のバイグラム確率である．

2.2.3 品詞バイグラム

明らかな非文は，品詞 N グラムを評価することで除外することができる場合が多い．そこで，以下の品詞バイグラムも素性として用いた．

$$P_{\mathrm{pos}}(y_{i+1}|y_i) = P(\mathrm{pos}(y_{i+1})|\mathrm{pos}(y_i)). \tag{5}$$

PLM と品詞バイグラムを用いて $h(y_{i+1},y_i;\boldsymbol{\lambda_h})$ を以下の式で定義する．

$$\begin{aligned} h(y_{i+1},y_i;\boldsymbol{\lambda_h}) = \lambda_{\mathrm{lm}} & P_{\mathrm{plm}}(y_{i+1}|y_i) \\ & + \lambda_{(\mathrm{pos}(y_{i+1})|\mathrm{pos}(y_i))} P_{\mathrm{pos}}(y_{i+1}|y_i)\} \end{aligned}$$

2.3 パラメータの最適化

文短縮は単語を短縮文に含めるか否かを決定する 2 値分類問題としてみなすことができる．しかし，ある単語を短縮文に採用するかどうかは他の単語に依存して決定する場合が多い．よって，独立に単語を 2 値に分類するのではなく，単語列としての良さを学習することのできる MCE 学習 [3] を採用し，パラメータの最適化を行った．

$\boldsymbol{x}^{\boldsymbol{\ell}}$ を訓練データ中の原文とし，$S(\boldsymbol{x}^{\boldsymbol{\ell}},M)$ を $\boldsymbol{x}^{\boldsymbol{\ell}}$ から得らることのできるすべての短縮文 (その単語数は M) とする．$\boldsymbol{y}^{*\boldsymbol{\ell}}$ を $\boldsymbol{x}^{\boldsymbol{\ell}}$ に対する正解 (人間が作成した短縮文) とし，$\boldsymbol{y}$ をシステムが推定した短縮文とする．ここで，誤り推定関数は，正解短縮文のスコアとシステムが推定した短縮文のスコアの差として以下の式で定義できる．

$$\begin{aligned} d(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\lambda}) = & \sum_t -f(\boldsymbol{y}^{*\boldsymbol{\ell}},\boldsymbol{x}^{\boldsymbol{\ell}};\boldsymbol{\lambda}) \\ & + \operatorname*{argmax}_{\boldsymbol{y}\in S(\boldsymbol{x}^{\boldsymbol{\ell}},M)\setminus\boldsymbol{y}^{*\boldsymbol{\ell}}} f(\boldsymbol{y},\boldsymbol{x}^{\boldsymbol{\ell}};\boldsymbol{\lambda}) \end{aligned} \tag{6}$$

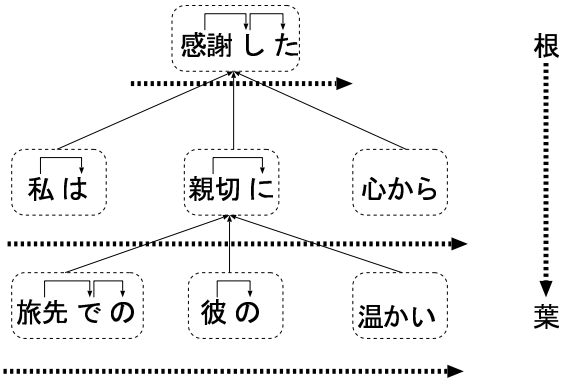

図 2　依存構造木の例

ここで，平滑化関数としてシグモイド関数を用いると，以下の式を最小化する問題に帰着する．

$$L(d(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\lambda})) = \frac{1}{1+\exp(-c\times d(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\lambda}))} \tag{7}$$

式 (7) を最小化する $\boldsymbol{\lambda}$ は，勾配法などを用いることで求めることができる．

3 評価実験

3.1 コーパス

毎日新聞 1994 年から 2002 年までの 9 年分の記事より，ヘッドラインを除き，30 単語以上で構成される 1 文目をランダムに 1000 文取り出し，それらを原文として評価用コーパスを作成した．各原文に対し，5 名による異なる参照短縮文を用意した．なお，参照短縮文は，原文の単語数に対して約 6 割の単語で構成されるように被験者に対して制約を与えた．つまり，要約率は 0.6 である．原文の平均単語数は 42 単語，短縮文の平均単語数は 24 単語である．

3.2 比較した文短縮手法

提案手法の有効性と IPTW，PLM の有効性を確認するため，以下の手法との比較を行った．

3.2.1 依存構造木の枝刈りによる手法

日本語の場合，依存構造木の枝を刈ることで簡単に短縮文を得ることができる．本稿ではこれをベースラインとする．

原文を依存構造木へと変換し，根から葉の方向に向けて，指定の単語数を満たすまで文節から単語を抽出し，短縮文を得る．ただし，同一階層の文節に関しては左から右への順に単語を取り出す．

図 2 を用いて説明する．根は「感謝した」なので，この文節に含まれる単語が最優先で短縮文に採用される．次に，その 1 つ下の階層には 3 つの文節があるが，これは左側が優先されるので，「私は」，「親切に」，「心から」の順で文節から単語を取り出し短縮文に採用する．なお，文節内の単語に関しては，右側の単語を優先して取り出す．実際の係り受け解析には工藤らの CaboCha [5] を用いた．

2 の例において，7 単語，11 単語からなる短縮文はそれ

表 1　ROUGE による評価結果 (5 分割交差検定)

Evaluation Measure		提案手法	w/o IPTW	w/ DEP	w/o PLM	Hori	Tree trimming
ROUGE-1	max	**.789**	.766	.781	.776	.739	.769
	avg	**.690**	.677	.687	.686	.653	.667
	min	.589	.585	**.591**	**.591**	.562	.563
ROUGE-2	max	**.670**	.630	.638	.625	.596	.635
	avg	**.540**	.511	.510	.500	.483	.498
	min	**.413**	.393	.383	.374	.373	.363
ROUGE-3	max	**.587**	.536	.546	.529	.506	.544
	avg	**.435**	.398	.397	.386	.376	.390
	min	**.292**	.268	.256	.247	.255	.244
ROUGE-4	max	**.516**	.455	.475	.456	.427	.466
	avg	**.347**	.306	.312	.301	.285	.303
	min	**.197**	.173	.166	.158	.161	.156

ぞれ以下の文となる．
- 私　は　親切　に　感謝　し た
- 私 は　旅先 で の 親切　に　心から 感謝　し た

3.2.2　Hori らの手法 [2]

提案手法と同じく，文を単語列としてとらえ，その部分単語列を短縮文とする手法である．$g(y_i;\boldsymbol{\lambda}_g)$，$h(y_{i+1},y_i;\boldsymbol{\lambda}_h)$ はそれぞれ以下の式で定義される．

$$g(y_i;\boldsymbol{\lambda_g}) = \begin{cases} \lambda_w \mathrm{IDF}(y_i) & y_i \text{ が内容語の場合} \\ \mathrm{Constant} & \text{それ以外} \end{cases} \tag{8}$$

$$h(y_{i+1},y_i,y_{i-1};\boldsymbol{\lambda_h}) = \lambda_{\mathrm{lm}} P_{\mathrm{tri}}(y_{i+1},y_i|y_{i-1}) + \lambda_{\mathrm{dep}} P_{\mathrm{dep}}(y_{i+1}|y_i) \tag{9}$$

なお，単語間の係り受け確率を得るためには，工藤らの手法を用いた [6]．また，パラメータは提案手法と同じく MCE 学習を用いて最適化を行った．

3.2.3　提案手法のバリエーション

素性の有効性を調べるため，下記の設定で提案手法を評価した．

w/o IPTW　単語重要度に IPTW を用いずに IDF のみを用いた場合．$g(y_i;\boldsymbol{\lambda}_g)$ として以下の式を採用．

w/o PLM　PLM を通常のバイグラム言語モデルに置き換えた場合．$h(y_{i+1},y_i;\boldsymbol{\lambda_h})$ として以下の式を採用．

$$h(y_{i+1},y_i;\boldsymbol{\lambda_h}) = \lambda_{\mathrm{lm}} P_{\mathrm{ng}}(y_{i+1}|y_i) + \lambda_{(\mathrm{pos}(y_{i+1})|\mathrm{pos}(y_i))} P_{\mathrm{pos}}(y_{i+1}|y_i)$$

w/ DEP　単語間のかかり受け確率を素性として採用した場合．$h(y_{i+1},y_i;\boldsymbol{\lambda_h})$ として以下の式を採用．

$$h(y_{i+1},y_i;\boldsymbol{\lambda_h}) = \lambda_{\mathrm{lm}} P_{\mathrm{plm}}(y_{i+1}|y_i) + \lambda_{\mathrm{dep}} P_{\mathrm{dep}}(y_{i+1}|y_i) + \lambda_{(\mathrm{pos}(y_{i+1})|\mathrm{pos}(y_i))} P_{\mathrm{pos}}(y_{i+1}|y_i) \tag{10}$$

なお，評価指標には ROUGE-N [7] を用いた．5 つの参照短縮文があるため，最大値，平均値，最小値をそれぞれ記録した．MCE 学習を用いた手法は 5 分割の交差検定で評価を行った．

4　実験結果と考察

表 1 に実験結果を示す．提案手法とそれ以外の手法との間の差を t 検定で比較した結果，ROUGE-1 以外のすべての組合わせで有意差があった．

表 1 より，IPTW を用いないことで成績は大きく低下していることからその有効性がよくわかる．訓練時に学習されたパラメータをにおける混合正規分布を図 3 に示す．図より，文の先頭付近と末尾付近の重みが大きくなっていることがわかる．これは，文の主部，述部がそれぞれ出現する位置であると考える．

また，PLM を一般的なバイグラム言語モデルに置き換えることでも成績は大きく低下している．これは，長短さまざまな長さの文を含むコーパスに基づくバイグラム言語モデルが短縮文の言語尤度を評価するのにふさわしくないことを示している．文短縮タスクにおいては，原文でのバイグラムに高い重みを与えることが有効であることがわかる．

さらに，単語間の係り受け確率を素性として追加したことでも成績が低下している．係り受け確率を素性とした Hori らの手法，依存構造木の枝刈りによる手法とも ROUGE スコアは提案手法と比較して低い．原文の係り受け構造をできるだけ保持するような短縮手法では，人間のような短縮が難しいことを示唆している．また，IPTW，PLM を用いた場合には必ずしも係り受け情報を追加することが有効でないこともわかる．

以上より，IPTW，PLM 双方の組合せが文短縮に有効であることがわかった．さらに，文の係り受け構造に強く依存する文短縮手法では人間のような文の短縮が難しいこともわかった．

5　まとめ

本稿では，構文解析器を必要としない文短縮手法を提案した．構文情報に依存せず，高性能な文短縮を実現するため，IPTW，PLM を新たな素性として提案した．また，素性の重みパラメータは MCE 学習を用いて最適化を行った．評価実験の結果より，提案手法が従来手法より統計的に有意な差で高性能であることを確認した．また，文短縮タスクに IPTW，PLM が素性として有効であることを確認し，係り受け情報が必ずしも有効でないことがわかった．

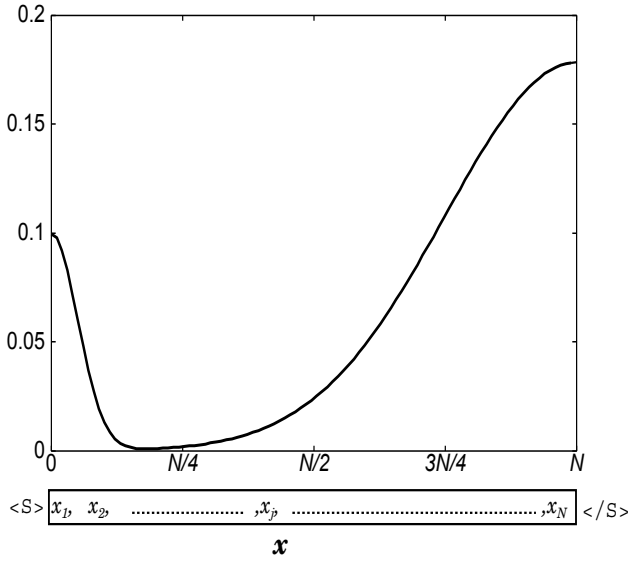

図 3　訓練データによって学習された混合正規分布の例

参考文献

[1] J. Clarke and M. Lapata. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proc. of the 21st COLING and 44th ACL*, pages 377–384, 2006.

[2] C. Hori and S. Furui. A New Approach to Automatic Speech Summarization. *IEEE trans. on Multimedia*, 5(3):368–378, 2003.

[3] B. H. Juang and S. Katagiri. Discriminative Learning for Minimum Error Classification. *IEEE Trans. on Signal Processing*, 40(12):3043–3053, 1992.

[4] K. Knight and D. Marcu. Summarization Byond Sentence Eextraction. *Artificial Intelligence*, 139(1):91–107, 2002.

[5] T. Kudo and Y. Matsumoto. Japanese Dependency Analysis using Cascaded Chunking. In *Proc. of the CoNLL*, pages 63–69, 2002.

[6] T. Kudo and Y. Matsumoto. Japanese Dependency Parsing Using Relative Preference of Dependency (in japanese). *IPSJ Journal*, 46(4):1082–1092, 2005.

[7] C-Y. Lin and E. Hovy. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proc. of the HLT and NAACL*, pages 150–157, 2003.

[8] R. McDonald. Discriminative Sentence Compression with Soft Syntactic Evidence. In *Proc. of the 11th EACL*, pages 297–304, 2006.

[9] T. Nomoto. Discriminative sentence compression with conditional random fields. *Information Processing and Management*, 43(6):1571–1587, 2007.

[10] K. Takeuchi and Y. Matsumoto. Acquisition of Sentence Reduction Rules for Improving Quality of Text Summaries. In *Proc. of the 6th NLPRS*, pages 447–452, 2001.

[11] J. Turner and E. Charniak. Supervised and Unsupervised Learning for Sentence Compression. In *Proc. of the 43rd ACL*, pages 290–297, 2005.

[12] Y. Unno, T. Ninomiya, Y. Miyao, and J. Tsujii. Trimming CFG Parse Trees for Sentence Compression Using Machine Learning Approach. In *Proc. of the 21st COLING and 44th ACL*, pages 850–857, 2006.