日本語話し言葉を要約するための文短縮の試行

池田 英明　　鈴木 寿
中央大学大学院 理工学研究科 情報工学専攻

{hikeda@suzuki-lab., suzuki@}ise.chuo-u.ac.jp

概 要

構文情報に依存せずバイグラムの統計情報を最大限に活用する先行研究, および, 品詞に基づく素性を豊富に勘案したヒューリスティックを活用する先行研究の中核的アイデアを組み合わせることにより, 構文情報を利用せずかつ短縮率を指定しない部類で, バイグラムの統計情報となるべく少数のヒューリスティックとを用いて有効に機能しうる文短縮の原理を考え, 要約の妥当性を試行的に評価したところ, ある程度良好な結果が得られた旨を報告する.

1 はじめに

いわゆる話し言葉は, 言い直し, 言い淀み, 発話計画の随時変更などの現象に伴い, 冗長表現および非流暢性を含みやすい. 特に日本語話し言葉には, 副詞, 前置き, 挿入句, 文末の表現などにおいて冗長表現が多いという傾向が見られる. これらの冗長表現は文脈に依存することなく現れるので, いわゆる “文短縮” によって冗長表現を削除することにより, 日本語話し言葉を要約することが可能となる.

削除対象にすべき冗長表現を明らかにするべく, 要約筆記の観点から日本語話し言葉の要約方法をまとめた文献 [1] を参考にして, 表 1 のように冗長表現の分類を試みた. このような冗長表現を含んだ文短縮の例を次に示す.

[原文 1]
短調 独特 の 旋律 の え 旋律 を 形作っ て いる と いう 風 に なっ て い ます.

[短縮文 1]
短調 独特 の 旋律 を 形作っ て いる.

[原文 2]
三 つ の 群 に 対し て えー 先程 言っ た ように 実験 を 行ない まし た.

[短縮文 2]
三 つ の 群 に 対し て 実験 を 行ない まし た.

表 1: 冗長表現の例

種類	例
フィラー (filler)	“えーっと” “あー”
副詞	“本当に” “例えば”
前置き	先ほど申し上げたように... 皆さんご存知の通り...
挿入句	..., 私が思うに, ...
間接的な物言い	... というものに つきましては
可能表現	... することができる (... できる)
本来の意味では用いられない接続助詞	“けど” “が” “ので”
言い淀み	こ こちら の 被験者 は
言い直し (言い換え)	ランプ えー 赤い ランプを この 中央 中心 に
文末の表現	... と思います ... という風に言えます

[原文 3]
五 か月 児 で は えー やはり 特に 差 は ない ん です けれど も.

[短縮文 3]
五 か月 児 で は 差 は ない.

従来さまざまな文短縮方法が考案されており, 代表的な一つとして, 平尾らは構文情報を利用しない方法を提案している [2]. 文短縮を組合せ最適化問題として捉え, 識別学習によりパラメーターを最適化するうえで, 原文に含まれるバイグラムが候補短縮文にも含まれるとき上限得点 1 を付与し, それ以外のときはバイグラム確率の反映された得点を付与する Patched Language Model (PLM) により, 構文情報の要らない文短縮を実現した. 教師あり訓練データの量を抑制しつつ原文およびコーパスの統計情報が活用できる. また, 短縮率を指定できる.

短縮率を指定しない部類では, 平尾らに先立ち McDonald が, 同じく識別学習による組合せ最適化問題の

枠組下で効果的な英文短縮の方法を提案している [3]. 英文に対し依存構造解析および句構造解析をおこなったうえで品詞に基づく素性を豊富に勘案することにより, 構文木から部分木を得る方法などに比べ構文情報にはさほど依存しない文短縮方法を実現した. その際, Margin Infused Relaxed Algorithm (MIRA) でパラメーターを最適化するうえで, 辞書的な素性を用いないことにより, 訓練データが少量で済むよう工夫している. 単語に得点を付与し組合せ最適化問題として文短縮をする原理上, 短縮文に含まれる単語数が多いほど得点は高くなるので, より長い短縮文が選好される. このことへの対策として, 削除される単語に対してもそれらの品詞や動詞に基づく得点を加算して均衡を図ることにより, 人手で得られた短縮文の短縮率とが同程度となる.

本稿は, 組合せ最適化問題を解く上述の 2 方法に基づき, 日本語話し言葉を対象に

- 構文情報を利用しない
- 短縮率を指定しない

なる部類における文短縮方法の一つを提案する.

書き言葉に比べて話し言葉の場合は, 言い直し表現や発話計画の随時変更などが含まれるという事情に加え, 文法的制約も緩いことから, 的確な構文解析が期待できず, 本稿では構文情報を利用しない立場をとる. また, 冗長表現のみを削除するなどの実際的な用途を想定し, 冗長表現の量に応じて文短縮するべく短縮率を指定しない立場をとる.

過去には, 話し言葉を対象とする文短縮方法の代表的研究として, 堀らが, 音声認識による書き起こしテキストを対象とした文短縮方法を提案している [4]. 短縮率を指定し, トライグラム確率や係り受け確率, 音声認識結果の信頼度などを素性とした組合せ最適化問題としての文短縮方法を, 日本語のニュース放送の音声に適用し良好な結果が得られている. 一方, 本研究は, 原稿準備やリハーサルなどのない自発的な日本語話し言葉を対象とする.

また Wang らは, 自発的な英語話し言葉を対象に, 構文木から部分木を得るという方法や組合せ最適化問題の枠組外で, 構文情報や辞書的な素性を扱う Conditional Random Fields (CRF) により個々の単語を短縮文に含むべきか否か硬判定する文短縮方法を提案している [5]. CRF により得られた N-best 候補を再順位化すれば, 大域的な素性も考慮した文短縮が可能となる. 一方, 本研究は, 辞書的な素性を用いず, 教師あり訓練データが少量で済むような文短縮方法を探求している.

2 日本語話し言葉の文短縮方法

2.1 文短縮方法の原理

2.1.1 短縮文の得点

日本語話し言葉の一文 (原文とよぶ) を単語の並置として $\mathbf{x} = x_1 x_2 \cdots x_{|\mathbf{x}|}$ ($|\mathbf{x}|$: 原文を構成する単語数) を表す. 同様に, 短縮文を単語の並置として $\mathbf{y} = y_1 y_2 \cdots y_{|\mathbf{y}|}$ ($|\mathbf{y}|$: 短縮文を構成する単語数) を表す.

また, 短縮文を構成する単語の位置 $j = 1,2,\cdots,|\mathbf{y}|$ を, 原文を構成する単語の位置 $i \in \{1,2,\cdots,|\mathbf{x}|\}$ に対応させる関数を I で表す. 各 $j = 1,2,\cdots,|\mathbf{x}|$ に対し, 明らかに $I(j-1) < I(j)$ である.

さらに, 原文 $\mathbf{x}$ に対する短縮文 $\mathbf{y}$ の得点を, 次のように定める.

$$s(\mathbf{x},\mathbf{w};\mathbf{y}) = \sum_{j=2}^{|\mathbf{y}|} f(\mathbf{x},\mathbf{w};j)$$

ここに $f(\cdot)$ は, 2.2 節に述べる 4 種類の尤度 $f_{\mathrm{PLM}}(\cdot)$, $f_{\mathrm{POS}}(\cdot)$, $f_{\mathrm{DPOS}}(\cdot)$, $f_{\mathrm{VERB}}(\cdot)$ の単純和である.

$$\begin{aligned} &f(\mathbf{x},\mathbf{w};j) \\ &\quad = f_{\mathrm{PLM}}(\mathbf{x},\mathbf{w}_{\mathrm{PLM}};j) + f_{\mathrm{POS}}(\mathbf{x},\mathbf{w}_{\mathrm{POS}};j) \\ &\qquad + f_{\mathrm{DPOS}}(\mathbf{x},\mathbf{w}_{\mathrm{DPOS}};j) + f_{\mathrm{VERB}}(\mathbf{x},\mathbf{w}_{\mathrm{VERB}};j) \end{aligned}$$

また, $\mathbf{w}_{\mathrm{PLM}}$, $\mathbf{w}_{\mathrm{POS}}$, $\mathbf{w}_{\mathrm{DPOS}}$, $\mathbf{w}_{\mathrm{VERB}}$ の各々は, 2.3 節のもとで教師あり訓練データを用いて学習される大量の重みパラメーターの組であり, $\mathbf{w}$ はさらにそれらをまとめた四つ組 $(\mathbf{w}_{\mathrm{PLM}},\mathbf{w}_{\mathrm{POS}},\mathbf{w}_{\mathrm{DPOS}},\mathbf{w}_{\mathrm{VERB}})$ であるとする.

2.1.2 動的計画法

原文 $\mathbf{x}$ の長さ $i \in \{1,2,\cdots,|\mathbf{x}|\}$ の接頭辞 $\mathbf{x}^{(i)} = x_1 x_2 \cdots x_i$ に対する短縮文を $\mathbf{y}^{(i)}$ で表すとき, $s(\mathbf{x}^{(i)},\mathbf{w};\mathbf{y}^{(i)})$ を最大化するような $\mathbf{y}^{(i)}$ を, $i = 1,2,\cdots,|\mathbf{x}|$ を順次増やしつつ動的計画法により求める.

2.2 バイグラム素性

2.2.1 単語バイグラム素性

$j = 2,3,\cdots,|\mathbf{y}|$ に対し, 単語 $\mathbf{x}_{I(j-1)}$ と単語 $\mathbf{x}_{I(j)}$ の共起確率を $\phi_{\mathrm{PRM}}(\mathbf{x};j)$ で表す. 一方, 2.3 節のもとで教師あり訓練データに現れる品詞バイグラムの全体集合を $\mathscr{P}$, 各品詞バイグラム $\mathbf{p} \in \mathscr{P}$ に対して学習された実数値の重みパラメーターを $w_{\mathrm{PLM}}(\mathbf{p})$, またすべての $\mathbf{p} \in \mathscr{P}$ に対する $w_{\mathrm{PLM}}(\mathbf{p})$ からなる組を $\mathbf{w}_{\mathrm{PLM}}$ で表す.

日本語版し言葉の係り受けには, “ある単語が短縮文に含まれるとき, その直後の単語も短縮文に含まれや

すい” という傾向が見られる. この傾向を反映した尤度として $j=2,3,\cdots,|\mathbf{y}|$ に対し

$$f_{\mathrm{PLM}}(\mathbf{x},\mathbf{w}_{\mathrm{PLM}};j)=$$
$$\begin{cases}1 & (I(j)=I(j-1)+1 \text{ のとき})\\ w_{\mathrm{PLM}}(\mathbf{p}(I(j-1),I(j)))\phi_{\mathrm{PLM}}(\mathbf{x};j) & (\text{その他のとき})\end{cases}$$

(ここに, 任意の $m,n\in\{1,2,\cdots,|\mathbf{x}|\}$ に対し $\mathbf{p}(m,n)\in\mathscr{P}$ は単語 x_m の品詞および単語 x_n の品詞からなる品詞バイグラムを表すとする) を定め, これを単語バイグラム素性とよぶ.

2.2.2 品詞バイグラム素性

$j=2,3,\cdots,|\mathbf{y}|$ に対し, 単語 $x_{I(j-1)}$ の品詞と単語 $x_{I(j)}$ の品詞との共起確率を $\phi_{\mathrm{POS}}(x;j)$ で表す. さらに, 単語バイグラム素性と同様な尤度として

$$f_{\mathrm{POS}}(\mathbf{x},\mathbf{w}_{\mathrm{POS}};j)=w_{\mathrm{POS}}(\mathbf{p}(I(j-1),I(j)))\phi_{\mathrm{POS}}(\mathbf{x};j)$$

を定め, これを品詞バイグラム素性とよぶ.

ここに, 各品詞バイグラム $\mathbf{p}\in\mathscr{P}$ に対し $w_{\mathrm{POS}}(\mathbf{p})$ は 2.3 節の学習下で実数値をとる重みパラメーター, また $\mathbf{w}_{\mathrm{POS}}$ はすべての $\mathbf{p}\in\mathscr{P}$ に対する $w_{\mathrm{POS}}(\mathbf{p})$ からなる組である.

2.2.3 削除品詞バイグラム素性

日本語話し言葉の係り受けには “ある単語が短縮文に含まれないとき, その直後の単語も短縮文に含まれにくい” という傾向が見られる. この傾向を反映したヒューリスティックな尤度として $j=2,3,\cdots,|\mathbf{y}|$ に対し

$$f_{\mathrm{DPOS}}(\mathbf{x},\mathbf{w}_{\mathrm{DPOS}};j)=$$
$$\begin{cases}w_{\mathrm{DPOS}}(\mathbf{p}(I(j-1),I(j-1)+1))\\ +w_{\mathrm{DPOS}}(\mathbf{p}(I(j)-1,I(j)))\\ +(I(j)-I(j-1)-2) & (I(j-1)+1<I(j) \text{ のとき})\\ 0 & (\text{その他のとき})\end{cases}$$

を定め, これを削除品詞バイグラム素性とよぶ.

ここに, 各品詞バイグラム $\mathbf{p}\in\mathscr{P}$ に対し $w_{\mathrm{DPOS}}(\mathbf{p})$ は 2.3 節の学習下で実数値をとる重みパラメーター, また $\mathbf{w}_{\mathrm{DPOS}}$ はすべての $\mathbf{p}\in\mathscr{P}$ に対する $w_{\mathrm{DPOS}}(\mathbf{p})$ からなる組である.

2.2.4 削除動詞バイグラム素性

日本語話し言葉において, 一つの動詞を含む文節 A の直前の文節 B は A に係りやすいので, A は B と共に削除するのが自然である.

また, 日本語版し言葉においては断定口調を避ける動詞を含む文節 (例: ... と思います) で文末を閉じる傾向が見られ, そのような文節を削除しても不自然にはなりにくい.

以上のヒューリスティックを簡素なルールとして記述した 0 または 1 をとりうる関数を $j=2,3,\cdots,|\mathbf{y}|$ に対し $\phi_{\mathrm{VERB}}(v,\mathbf{x};j)$ で表す. ここに v は, 2.3 節のもとで教師あり訓練データに現れる動詞の全体集合 $\mathscr{V}$ における各動詞 $v\in\mathscr{V}$ とする. 各動詞 $v\in\mathscr{V}$ に対して 2.3 節の学習下で実数値をとる重みパラメーターを $w_{\mathrm{VERB}}(v)$, またすべての $v\in\mathscr{V}$ に対する $w_{\mathrm{VERB}}(v)$ からなる組を $\mathbf{w}_{\mathrm{VERB}}$ で表す.

$x_{I(j-1)+1},\cdots,x_{I(j)-1}$ に含まれる動詞の全体集合を $\mathscr{V}(j)\subset\mathscr{V}$ で表すとき

$$f_{\mathrm{VERB}}(\mathbf{x},\mathbf{w}_{\mathrm{VERB}};j)=\sum_{v\in\mathscr{V}(j)}w_{\mathrm{VERB}}(v)\phi_{\mathrm{VERB}}(v,\mathbf{x};j)$$

と定め, これを削除動詞バイグラム素性とよぶ.

2.3 重みパラメーターの学習

McDonald[3] に倣い, 教師あり訓練データに基づく重みパラメーター $\mathbf{w}$ の学習を, MIRA を用いた $\mathbf{w}$ の最適化問題として解く. その概要を以下に示す.

損失関数 L を, 暫定の $\mathbf{w}$ を用いた原文各単語の削除可否判断が訓練データのそれと一致しないような単語数として定める. MIRA は k-best 候補に対して k 個の制約を満たす最小化問題を効率よく解く手法の一つであり, 特に k=1 に対しては, 次のアルゴリズムにより解析的に最適化問題が解ける.

Algorithm: 1-best MIRA
訓練データ: $\mathbf{D}=\{(\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2),\cdots,(\mathbf{x}_T,\mathbf{y}_T)\}$
$\mathbf{w}^{(0)}=0$; $\mathbf{w}_{\mathrm{sum}}=0$; $i=0$
for $n=1$ to N do
  for $t=1$ to T do
    $\min \frac{1}{2}||\mathbf{w}^{(i+1)}-\mathbf{w}^{(i)}||^2$
    s.t. $s(\mathbf{x}_t,\mathbf{w}^{(i+1)};\mathbf{y}_t)-s(\mathbf{x}_t,\mathbf{w}^{(i+1)};\mathbf{y}')\geq L(\mathbf{y}_t,\mathbf{y}')$
    where $\mathbf{y}'=\underset{\mathbf{y}}{\mathrm{argmax}}\, s(\mathbf{x}_t,\mathbf{w}^{(i)};\mathbf{y})$
    $\mathbf{w}_{\mathrm{sum}}=\mathbf{w}_{\mathrm{sum}}+\mathbf{w}^{(i+1)}$
    $i=i+1$
  end for
end for
$\mathbf{w}=\mathbf{w}_{\mathrm{sum}}/(N*T)$

アルゴリズム中の最小化問題は次のように解ける.

$$\mathbf{w}^{(i+1)}=\mathbf{w}^{(i)}+\lambda\mathbf{v}$$
$$\text{where } \lambda=\frac{L(\mathbf{y_t},\mathbf{y}')-d}{||\mathbf{v}||^2}$$

ここに, $d = s(\mathbf{x_t}, \mathbf{w}^{(i)}; \mathbf{y_t}) - s(\mathbf{x_t}, \mathbf{w}^{(i)}; \mathbf{y}')$, $\mathbf{v} = \frac{d}{d\mathbf{w}}\{s(\mathbf{x_t}, \mathbf{w}; \mathbf{y_t}) - s(\mathbf{x_t}, \mathbf{w}; \mathbf{y}')\}$ である. このとき, λ に任意の上限 α を定める, すなわち

$$\lambda = \min\left\{\frac{L(\mathbf{y_t}, \mathbf{y}') - d}{||\mathbf{v}||^2}, \alpha\right\}$$

により, 過学習を抑える効果が期待できる [1].

3 評価の試行

日本語話し言葉コーパス (CSJ)[7] を用いて, 要約を目的とすることを意識しつつ, 自然言語処理に詳しくない被験者に協力を募り, 文短縮方法の効果を試行的に評価した.

具体的には, 準備として, CSJ に含まれる情報に基づきフィラーと言い淀みを除いておく. CSJ において, 8 講演中の 727 文からなる原文の各々に対し人手による短縮文 (参照文とよぶ) が付されている. 一つの講演 X を除く 7 講演において原文と参照文とのペアを訓練データとし, 講演 X については評価用のデータとすることにより, いわゆる 8 分割の交差検定を計画した.

日本語を母語とする 6 人の被験者が, 5 単語以上からなる原文から無作為に選んだ 50 文の各々を対象として, 原文の提示および短縮文の提示 (人手による参照文, 本方法による短縮文, 対照方法による短縮文を無作為の順番で提示) を見て, 文法性および内容性に関し主観的に 5 段階 (1, 2, 3, 4, 5) で評価した. ここに対照方法とは, CSJ に含まれる構文情報を用いた単語バイグラム確率に基づく素性において, 着目する 2 単語が係り受け関係にあるとき上限得点 1, それ以外のときはバイグラム確率を反映した得点を付与した方法である.

評価結果の平均および短縮率を表 2 に示す. なお, 短縮率は 727 文すべてにおける平均である.

表 2: 評価の結果

文短縮方法	短縮率	文法性	内容性
本方法	0.819	4.14	4.06
対照方法	0.767	3.80	3.49
原文	1.00	4.35	—
参照文	0.700	4.63	4.38

おもな特徴を述べる. 対照方法では係り受け関係にあるバイグラムを選好されるので, 述語に係る目的語が削除される, “こと”, “もの” のように単体では意味をなさない単語が独立して残存する, などの現象により文法性に対する評価は低くなりやすい. 一方, 本方法では文節や係り受け関係によらず一つの単語とその直後の単語とが一緒に選好されるので, 不自然な単語の欠落は生じにくく文法性に対する評価は高くなりやすい.

反面, 本方法の短縮率は対照方法の短縮率よりも悪くなる.

問題として, 被験者は前後の文脈を見ないで評価するので, 原文から情報が失われると内容性を低く評価する傾向があり, より短い短縮文の内容性を低めに評価する傾向がある. さらに, 文法性の低さは内容性の低さの原因となっているように見える. 評価に伴うこれらの問題を解消するため, ある程度連続した文のまとまりを対象として内容性を評価する実験計画をおこなう必要がある.

4 おわりに

構文情報に依存せずバイグラムの統計情報を最大限に活用する平尾らの方法 [2], および, 品詞に基づく素性を豊富に勘案したヒューリスティックを活用する McDonald の方法 [3] の中核的アイデアを融合することにより, 構文情報を利用せずかつ短縮率を指定しない部類で, バイグラムの統計情報となるべく少数のヒューリスティックとを用いて有効に機能しうる文短縮の原理を考え, 要約の妥当性を試行的に評価したところ, ある程度良好な結果の得られることがわかった.

統計情報とヒューリスティックとを統一的に記述したうえで性能をシステマティックに改善するための開発方法論は未検討であり, 今後の課題である.

参考文献

[1] 三宅初穂, 話しことばの要約 要約筆記の探求から, 全国要約筆記問題研究会, 愛知県, 2012.

[2] 平尾努, 鈴木潤, 磯崎秀樹, “構文情報に依存しない文短縮手法,” 情報処理学会論文誌, Vol. 2, No. 1, 1–9, 2009.

[3] Ryan McDonald, “Discriminative Sentence Compression with Soft Syntactic Evidence,” *In Proc. EACL*, pp. 297–304, 2006.

[4] Chiori Hori, Sadaoki Furui, “A New Approach to Automatic Speech Summarization,” *IEEE Transactions on multimedia*, Vol. 5, No. 3, pp. 368–378, 2003.

[5] Dong Wang, Xian Qian, Yang Liu, “A Two-step Approach to Sentence Compression of Spoken Utterances,” *In Proc. ACL*, pp. 166–170, 2012.

[6] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer, “Online Passive-Aggressive Algorithms,” *Journal of Machine Learning Research 7*, pp. 551–585, 2006.

[7] 前川喜久雄, “『日本語話し言葉コーパス』の概要,” 日本語科学, 15, pp. 111–133, 2004.

[1] スラック変数を導入した最小化問題を解くことと同義 [6].