文生成のための機能語の補完

池田 諭史 , 沢井 康孝 , 山本 和英
長岡技術科学大学 電気系
E-mail:{ikeda,sawai,ykaz}@nlp.nagaokaut.ac.jp

1 はじめに

文意を決める主要な単語が与えられた状況で、それら単語を用いて不足する語を補いながら文を作る処理を考え、これを本稿では文生成と呼ぶ。文生成は要約、機械翻訳など自然言語処理の様々な分野に適用できる重要な要素技術の 1 つである。

要約においては、原文から要約に必要な単語群を抽出し、その単語群から文を生成することにより要約文とすることが可能である。我々はこのような要約を行うための第 1 段階として、ある要約文に使われた単語を用いて文生成を行い元の要約文を作成することを試みた。

内元ら [1]、肥塚ら [3] は { 国, 政策, 発足 } のような 3 つの単語からの文生成を行っている。これらの研究は、単語から文を生成することを目的としており、本稿で必要となる元の文との意味的な一致は考えていない。また、今回は 3 単語より多い単語での生成が必要となるが、これらの研究では単語が多い場合においても文を正しく生成できるかについては言及されていない。そのうえ、係り受け候補を全て用いているので単語数の増加と共に処理時間が指数的に増加することが予想される。

廣嶋ら [4] は文章から必要な内容語及び機能語を抜き出し、それらを連ねることで文章のヘッドラインを作成している。他に文生成が利用可能な状況としては失語症患者との意志の疎通や第 2 言語での作文支援等が考えられる。これらは両者とも単語は出てくるが、その単語を用いて文の作成を行えない状況である。またテキストマイニングで取り出された単語より日本語の再生成を行うことも検討されている [2, 5]。このように単語から文の生成が可能であれば様々なことに利用可能である。

2 問題設定

本研究では文単位の要約を次の 2 つの手順で行うことを想定し、今回は手順 2 を行うことを目的とする。

手順 1 要約したい文より要約に必要な単語群を抽出する。
手順 2 抽出した単語群より文生成を行う。

ここで手順 2 のみを行うためには手順 1 の処理が必要となる。本稿では、要約文の単語を使うことが要約に必要な単語を用いることと等価であると考え、要約文の単語を用いて要約文と同じ内容の文を生成することを目指す。ここで生成された文が要約文と同じ文になっていれば、要約したい文から要約に必要な単語を抽出し文生成することで要約可能となる。本稿では手順 1 で抽出した単語のことをキーワードと定義する。

ここで文生成をするために必要な単語について考える。人間が文を読むときに最低限必要な情報は内容語である。これは、人間が日本語を速読しようと試みると機能語部分を読み飛ばすことからもわかる。このことより内容語があれば文を補完して意味を把握することが可能であると考える。また、形容詞や副詞は機能語を間に挟むことがなく名詞や動詞の直前に出現するため、原文から抽出しておけば補完は容易に行うことができると考える。そこで今回はキーワードとしての使用及び補完の双方で形容詞や副詞は使用しないことにした。以上により手順 2 で文生成を行う際に必要となる単語は名詞と動詞であるとして、要約文の全ての名詞と動詞を用いて文生成を行うこととする。このとき入力する単語数に制限はない。また、今回は語順が与えられているものとした。

手順 1 に相当する処理として以下を行う。要約文を形態素解析する。その後、名詞、動詞、句点を語順を保持した状態で抽出する。このとき複合名詞のように名詞が連続している場合であっても形態素単位で取り出す。例 1 にキーワードの抽出例を示す。ここで抽出されたキーワードを用いて文生成を行う。

例 1) 旅客の安全検査を一部簡素化する方向で検討する。
→{ 旅客, 安全, 検査, 一部, 簡素, 化, する, 方向, 検討, する,。 }

3 提案手法

本稿では以下に示す手順で文生成を行った。

1. 機能語の補完対象箇所の同定
2. 助詞ノの補完
3. 機能語の補完

3.1 機能語の補完箇所の同定

抽出されたキーワードには機能語の補完を必要としない部分も存在する。例 1 における「安全」と「検査」はこの例の 1 つである。このため最初に機能語の補完が必要か否かを判断する。この判断を SVM(線形カーネル) を用いて行い、このときの素性は単語、品詞[1]、キーワードの出現位置を用いた。ここで補完の必要がないと判断された箇所は機能語の補完を行わず、前後のキーワードをまとめて 1 つのキーワードとして扱う。また、ここで補完の必要があると判断された箇所については、3.3.1 節で抽出する補完候補が存在しない場合以外は必ず何らかの機能語を補完する。

例 1 において補完不要箇所がすべて正しく同定されると例 2 のようになる。

例 2){ 旅客, 安全検査, 一部簡素化する方向, 検討する。 }

3.2 助詞ノの補完

所有を表す助詞ノを含む文節は形容詞的な働きをするので、3.3 節以降の文全体の機能語の並び (n-gram) をみている手法では、形容詞的な情報は上手く補完できないと考える。そこで最初にノが補完されるか否かという判断を行う。この判断は SVM を用いて行い、3.1 節と同一のカーネル及び素性を用いた。ここでノが補完されると判断された箇所にはノを補完し、ノが補完されないと判断された箇所については 3.3 節以降でノを補完することはない。

例 2 から完全にノの補完箇所が判定されて例 3 のようになる。

例 3){ 旅客の安全検査, 一部簡素化する方向, 検討する。 }

[1]できるだけ多くのタスクに対応できることを考え、形態素解析器で得られるような品詞の詳細情報は用いずに名詞や動詞といった品詞の第一階層のみを使用した。

3.3 機能語の補完

機能語の補完は以下の手順で行う。

3.3.1 補完候補の抽出

ここまでの処理で機能語の補完が必要でかつ助詞ノの補完が行われなかった箇所について補完する機能語の候補をコーパスより抽出する。ここでは機能語の連続はまとめて 1 つの機能語として扱う。例えば、例 4 の「からの」という部分は「から」と「の」という 2 つの助詞の連続であるのでこれを 1 つの機能語として扱う。以下の Step は補完候補が存在しなかったときに次の Step の処理を行い、補完候補を抽出した時点でこの節の処理を終了する。

例 4) 政府からの要請を受ける。

[Step 1] 補完箇所の前後のキーワードにより補完候補を抽出する。例えば補完箇所の前後のキーワードが { 安全検査, 一部簡素化する方向 } の場合は「安全検査+(機能語)+一部簡素化する方向」という形で出現する全ての機能語をコーパスから抽出し、それを補完候補とする。

[Step 2] 補完箇所の前のキーワードにより補完候補を抽出する。例えば補完箇所の前後のキーワードが { 安全検査, 一部簡素化する方向 } の場合は「安全検査+(機能語)」という形で出現する全ての機能語をコーパスから抽出し、それを補完候補とする。

[Step 3] 補完箇所の前のキーワードの最終形態素と補完箇所の後のキーワードの第 1 形態素により補完候補を抽出する。例えば補完候補の前後のキーワードが { 安全検査, 一部簡素化する方向 } の場合は、「検査+(機能語)+一部」という形で出現する全ての機能語をコーパスから抽出し、それを補完候補とする。

[Step 4] 補完候補の前のキーワードの最終形態素により補完候補を出力する。例えば補完候補の前後のキーワードが { 安全検査, 一部簡素化する方向 } の場合は、「検査+(機能語)」という形で出現する機能語を全てコーパスから抽出し、それを補完候補とする。

[Step 5] 本手法で補完候補を抽出できないため、補完を行わない。

3.3.2 機能語の決定

3.3.1 節で抽出した補完候補の中から 1 つの機能語を補完する。補完する機能語は式 (1) のスコアが最も高い機能語である。また補完は文頭から行うこととした。

$$Score(A,B,Z) = \log L(A,B,Z) + \lambda \log G(Z) \qquad (1)$$

ここで $Score(A,B,Z)$ はキーワード {A,B} の間に機能語 Z が補完されるときのスコア、λ は重み係数であり、L は局所的なスコア、G は大局的なスコアである。

局所的なスコア $L(A,B,Z)$ は形態素 3-gram 確率 $P(w_m|w_{m-2}w_{m-1})$ を用いて算出する。キーワード A,B と機能語 Z はそれぞれ複数の形態素が複合している場合もあるのでそれぞれ形態素単位にする。例えば A が $\{a_1,a_2\}$ の 2 形態素、Z が $\{z_1,z_2\}$ の 2 形態素、B が $\{b_1,b_2\}$ の 2 形態素からなっていたとする。そのときの $L(A,B,Z)$ は式 (2) のようになる。

$$\begin{aligned} L(A,B,Z) = & P(z_1|a_1a_2) \times P(z_2|a_2z_1) \\ & \times P(b_1|z_1z_2) \end{aligned} \qquad (2)$$

また、A が 1 形態素で形成される場合は、$P(z_1|a_1a_2)$ の計算ができない。この場合は、A の前に補完された機能語の最終形態素を用いて同様の計算を行う。またこのとき A が文頭のキーワードの場合も同様に $P(z_1|a_1a_2)$ の計算ができないが、これは文の最初に架空の文字 (以下文頭文字と呼ぶ) を仮定することで計算を行う。

大局的なスコア $G(Z)$ は機能語の 5-gram 確率を用いた。機能語の n-gram 確率はコーパス中の連続する機能語を 1 つの機能語として扱い、機能語以外の形態素を全て削除したコーパスを作成し、n-gram 確率を求めたものである。例 5 でこのコーパスの文の例を原文、作成したコーパスの文の順に示す。これを用いて $G(Z)$ を決定する。また局所的なスコアと同様に文頭文字を用いることにより、前方に機能語が 4 箇所補完されていない状況でも 5-gram を計算可能にしている。

例 5) 当時の社会状況では違法とは言えない。
→ の では とは ない

スコア $L(A,B,Z)$ 及び $G(Z)$ はそれぞれ n-gram 確率より求めているが、n-gram 確率が 0 になることがある。このときそれぞれの n-gram 確率の最低値の 1/100 の値を与えることにより対応している。

4 評価実験

本稿の提案手法の妥当性をはかるために、実際に実験を行った。実験には新幹線の電光掲示板でのニュースを用いた。これは日経ニュースメール (1) という NIKKEI-goo が行っているメールサービスで配信されるメールに含まれているニュース記事[2]である。我々はこれを 48618 文収集した。このうち 1000 文をテストデータ、1000 文を重み係数 λ を決めるための学習用データ、残り 46618 文を機能語の n-gram 確率、形態素 n-gram 確率、3.1 節と 3.2 節での SVM の学習データとして用いた。また、3.3.1 節の処理では日本経済新聞 2000 年度版を用いて補完候補を抽出した。形態素解析には ChaSen(3)、SVM 学習には TinySVM(4) をそれぞれ用いた。式 (1) の重み係数 λ の値は学習用データを用いて求めた。λ の値を種々に変化させて文生成を行い原文を正解例とした BLEU スコアを計算し、BLEU スコアの最も高かった重み係数を用いた。本実験では $\lambda = 10$ となった。

4.1 人手による正解評価

テストデータから文生成した 1000 文から無作為に 100 文抽出し、3 人の被験者が正しい文になっているかを独立に評価して、3 人の評価の多数決によって正解の評価を行なった。本実験では学習データに体言止めや助詞止めの文が多く含まれているため、体言止めや助詞止めになっている文も正解としている。評価は 2 つの指標で行った。1 つは生成した文の可読性の評価 (評価 1) で、もう 1 つはその意味が正しく生成されているかという評価 (評価 2) である。またそれぞれの人による正解についてのゆれをみるために、一人以上が正解としたときと 3 人が正解としたときの正解率をそれぞれ求めた。それを表 1 に示す。表 1 より人により正しく文生成が行われたという評価が大きく異なることがわかる。また評価 2 は意味の判断を行うので評価 1 が正解の場合にしか正解と判断されない。そこで評価 1 が正解のときの評価 2 の正解率を調べたところ 85%であった。

正解と不正解の例を示す。例 6、例 7 はそれぞれ正解例、不正解例を表す。またそれぞれの例は上から原文、キーワー

[2]我々はこのような文を新幹線要約と呼び、その特徴について観察を行っている [6]。

表 1: 正解の人数を変えたときの正解率

	≥1	≥2	=3
可読性の評価 (評価 1)	77%	53%	33%
意味同一性の評価 (評価 2)	46%	23%	15%

ド、生成された文となっている。

例 6) 米国防長官は２６日、来年早々にも海兵隊３大隊をイラクに派遣する運用計画を承認。
→{ 米, 国防, 長官, ２, ６, 日, 来年, 早々, 海, 兵隊, ３, 大隊, イラク, 派遣, する, 運用, 計画, 承認,。}
→ 米国防長官が２６日来年早々に海兵隊を３大隊イラクに派遣する運用計画を承認。

例 7) 地方機関の３割が暴力団や右翼から物品購入など不当要求を経験。
→{ 地方, 機関, ３, 割, 暴力団, 右翼, 物品, 購入, 不当, 要求, 経験 }
→ 地方機関は３割を暴力団右翼が物品購入で不当に要求経験。

4.2 先行研究との比較

先行研究 [1] との比較を行った。先行研究では 3 つのキーワードから文生成を行って評価 1 とほぼ同じ基準で正解を判断している。先行研究では 30 文生成している。そこで同じ 30 文の 3 つのキーワードから本手法で文生成を行い、3 人の被験者が独立に評価 1 の基準で評価を行った。結果を表 2 に示す。

表 2: 先行研究 [1] との正解率の比較

	先行研究	提案手法		
		≥1	≥2	=3
正解率	63.3%	80.0%	70.0%	46.7%

独立に評価をしているので、正解を多数決で判断をする。そこで先行研究の正解率と提案手法で 2 人以上が正解と判断したときの正解率を比較すると、提案手法の方が良い結果を得られた。

5 考察

5.1 補完部分の同定について

3.1 節で提案した機能語補完箇所の同定について考える。まず 3.1 節だけの精度を求め、グラフ化すると図 1 となる。学習の１単位をエントリと呼び、2 つのキーワードから抽出した素性と正例負例のラベルをセットにしたものを１エントリとした。正例と負例は、補完の必要がない箇所を正例、補完の必要な箇所を負例とした。ここで、精度は全体の正解率とし、適合率、再現率は一般的な指標を用いた。この図から補完部分は約 90%の精度で同定されていることがわかる。また曲線が飽和状態へと向かっているためにエントリ数を増やすことで多少の精度の向上は望めるが、これ以上は大きく上がらないだろうということも読み取れる。

5.2 助詞ノの補完について

3.2 節で提案した助詞ノの補完について考える。3.2 節だけの精度を求め、グラフ化すると図 2 となる。このとき正例と負例は、ノを補完する箇所を正例、ノを補完しない箇所を負例とした。この図からノの補完部分は約 90%の精度で同定されている。しかし再現率が著しく低いことがわかる。これはエントリを作る際に 2 キーワードの間にノが補完されるときに正例、ノ以外の機能語が補完されるときに、負例となるようにデータを作成したため、テストデータのほとんどが負例となる。そのため適合率や再現率が低い値

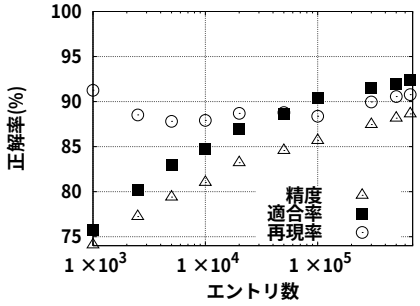

図 1: 学習数を変えたときの補完部分の同定性能の推移

となっていても、ノを補完する必要がない部分の大半を正解することができるので精度のみ高くなっている。つまり、ノを補完すべきところの多くを補完していない。また再現率は飽和傾向にみられないので、エントリ数を増やすことで精度が向上する可能性がある。

また、3.1 節と 3.2 節の処理を両方行ったときの精度は 85.0%であった。

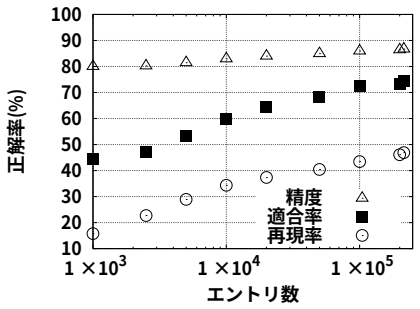

図 2: 学習数を変えたときの助詞ノの補完性能の推移

5.3 不正解の文について

4.1 節で被験者全員が不正解とした 23 文について考察する。これらの文を観察すると、文の大部分は正しく生成されているが一部分が文法的に致命的な間違いをしているため文全体として間違ったという傾向がみられた。またこの誤りには補完部分の同定やノの補完で間違いになっている箇所もみられた。そこでこれらの誤りが存在しないという条件でこれらの 23 文について提案手法を用いて文の生成を行った。この結果を表 3 に示す。

表 3: 被験者 3 人が不正解とした文において補完箇所の同定及びノの補完に誤りが存在しないときの正解率

	≥1	≥2	=3
可読性の評価 (評価 1)	65.2%	39.1%	21.7%
意味同一性の評価 (評価 2)	43.7%	30.4%	4.3%

表 3 より、本手法で補完箇所の同定及びノの補完を行ったときに被験者全員に不正解とされた文の約 65%が、少なくとも 1 人の被験者が正解であるとしたことがわかる。また多数決で正解を判断すると被験者全員に不正解とされた文の約 40% が正解であると判断された。このことより、提案手法の前処理部ともいえる 3.1 節や 3.2 節について精度の向上が不可欠である。

次に 23 文に対し抽出したキーワードから人手で文の生成

を行った。このときのキーワードは形態素区切りのものを使用し、補完する単語は本手法と同様に機能語のみとして生成を行った。このとき被験者が生成出来ない文も存在した。その例を例 8 に、原文、キーワード、本手法の出力の順に示す。これは「今後」というキーワードがあるときに次の時制は未来もしくは現在になる。しかし、「今後」の後のキーワードが「採用+し」という過去形の「した」に補完されやすい表現となっているために補完できなくなったと考える。また、提案手法でも「採用+し」の後は「た」が補完されている。これは「採用し」のあとは「た」がコーパスに多く出現したことを示している。このことより、「する」は機能を表す単語であるとして、キーワードとして用いるのではなく補完対象とすることも考えられる。また、この例は「今後」という自立語で時制が変わる例である。このことより、このような自立語での判断も必要になると考える。

例 8) 企業が今後採用したい年金制度は確定拠出年金が５割超す。
→{ 企業, 今後, 採用, し, 年金, 制度, 確定, 拠出, 年金, ５, 割, 超す,。}
→ 企業は今後採用した年金制度を確定拠出年金で５割を超す。

また、補完する上で読点が必要になる文があった。例 9 に原文、キーワード、人で手作成した文の順で例を示す。原文では「一元化」の後は読点になっている。しかし、提案手法では読点はキーワードではなく、さらに補完する語としても含んでいない。そのため、出力結果に表れることはない。提案手法で読点を補完候補から除いた理由は、読点は文の区切りを示す際に用いられるため、文を生成する際には不要だと考えたからである。しかし、実際に生成を行うと例 9 の場合は読点が必要であることがわかる。生成文で丸括弧で囲った読点は、実際には「する」の未然形「し」を補完することも可能だが、「する」は ChaSen の解析結果が動詞なので、今回は補完対象ではない。つまり、この文は本手法では正しい補完が不可能である。

例 9) 日産自動車は来年から、グループ各社の支払業務を本社に一元化、手形支払いは全廃する。
→{ 日産自動車, 来年, グループ, 各社, 支払, 業務, 本社, 一元化, 手形支払い, 全廃, する,。}
→ 日産自動車は来年にグループ各社の支払業務を本社で一元化 (、) 手形支払いを全廃する。

3.3.1 節で出力した補完候補の中に、正解の補完候補がない文も存在した。これは 3.3 節で補完候補を抽出する際に、Step1 や Step2 のように早い段階で出力された補完候補はかなり強い局所的な制約となっている。そのためにコーパスの量が十分でないとコーパスの過疎性が問題になってくる。そのため、今後より大きなコーパスを用い同様の補完候補の抽出を行ってみる必要がある。ある程度の量を用意しても変わらなければ制約を緩めて補完候補を抽出する必要がある。

人手で補完しても原文と意味が変わってしまう文も存在した。これは、１通りではなく複数の文の候補があるため、実際に元の文に戻すことを考えたときに名詞と動詞だけではなく他にも何らかの要素が必要であることを示している。どんな要素が必要であるかという調査も今後の課題である。また今回は新幹線の電光掲示板でみられるような特殊な要約文なので、他の一般的な要約文でも同様に名詞と動詞では意味の補完が可能かを調査する必要もある。

5.4 大局的なスコアについて

本稿では、大局的なスコアとして機能語の n-gram 確率を使用した。しかし、4.1 節で実際の要約文を用いるときより 4.2 節 3 つのキーワードからの生成の方が約 20%も正解率が良かった。これは、キーワードが少ない方が文生成を正しく行っているということになる。このことにより、大局的なスコアが文を生成する上であまり機能していない。本稿で機能語の n-gram 確率を用いたのは n-gram 確率という簡単なスコアでどこまで補完可能かということに加え、先行研究で最も問題になるであろう点、すなわちキーワードの増加に伴い処理速度が指数的に増加していく点に対処するためである。しかし、n-gram 確率だけではあまり良い結果は得られていないので、他手法も含めさらに検討する必要がある。

6 まとめ

要約したい文から要約に必要な単語を抽出し文生成を行うことで要約は可能であるという仮説に基づき、要約文の単語を用いて文生成を行い元の要約文に復元する手法を提案した。その結果、人手での評価で 53%の正解率を得た。また先行研究と比較した結果、先行研究の正解率を上回った。さらに、実際に文を生成することにより、生成を行う際には 名詞と動詞だけではなく、何らかの要素が必要であろうという知見を得た。

謝辞

本研究の一部は、科学研究費補助金 若手 (B) 「高密度表現を利用したまとめ型要約に必要な言語変換技術」 課題番号 16700134、及び科学研究費補助金 基盤 (A) 「円滑な情報伝達を支援する言語規格と言語変換技術」 課題番号 16200009 によって実施した。

使用した言語資源及びツール

(1) 日経ニュースメール,NIKKEI-goo, http://nikkeimail.goo.ne.jp/

(2) 日本経済新聞全記事データベース 2000 年度版, 日本経済新聞社.

(3) 形態素解析器 ‘‘ChaSen’’,Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/hiki/ChaSen/

(4) SVM 学習ツール ‘‘TinySVM’’,Ver.0.09, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/˜taku/software/TinySVM/

参考文献

[1] 内元 清貴, 関根 聡, 井佐原 均: キーワードからのテキスト生成: 言語処理学会第 8 回年次大会発表論文集,pp.375–378,2002.

[2] 坂尾 要祐, 池田 崇博, 佐藤 研治, 赤峯 享: 特徴的な意味内容を抽出する木構造マイニングのための日本語処理手法: 言語処理学会第 10 回年次大会発表論文集,pp.73–76,2005.

[3] 肥塚 真輔, 岡本 紘幸, 斎藤博昭: サポートベクタマシンを用いたキーワードからのテキスト生成: 言語処理学会第 10 回年次大会発表論文集,pp.409–412,2004.

[4] 廣嶋 伸章, 長谷川 隆明, 奥 雅博: Web ページのヘッドライン生成のための統計的要約: 言語処理学会論文誌「自然言語処理」,Vol.12,No.6,pp.113–128,2005.

[5] 森永 聡: – テキストマイニング技術の動向 – Key semantics マイニング、動的トピック分析による Knowledge Organization: 日本行動計量学会第 33 回大会発表論文抄録集,pp.370–373,2005

[6] 山本 和英, 池田 諭史, 大橋 一輝: 新幹線要約のための文末の整形: 言語処理学会論文誌「自然言語処理」,Vol.12,No.6,pp.85–112,2005.