共起頻度と語順制約を利用した分野依存性の高い定型表現の自動抽出

下畑さより　　杉尾俊之　　永田淳次
Sayori SHIMOHATA Toshiyuki SUGIO Junji NAGATA
沖電気工業(株) 研究開発本部 関西総合研究所
Kansai Laboratory, Research and Development Group, Oki Electric Ind. Co., Ltd.

1 はじめに

定型表現とは、型にはまった慣用的な表現形式で、専門用語やイディオムのように連続的なものから、呼応表現や動詞と格要素の組合せのように非連続、かつ語順や挿入される語句の種類が自由なものまで、様々な種類がある。定型表現は、特殊な文型であったり、語句が一般的な用法とは全く違った使われ方をしていたりすることが多く、表層的な処理を行なうだけでは内容を正しく解釈することが困難である。しかし、読み手が表現形式に関する知識を有する場合には少ないコストで必要な情報を曖昧性なく伝達することができるという側面もあり、分野依存性の高い文章において頻繁に使用される。

本論文では、コーパスから分野依存性の高い定型表現を抽出する方法について述べる。また、英文コンピュータマニュアルを対象に、コーパスに頻出する分野固有の表現を抽出する実験についても報告する。

2 定型表現の自動抽出

2.1 基本的な考え

分野依存性の高い定型表現には、以下のような特徴がある。

- 特定分野の文章に繰り返し出現する。
- 複数の表現を構成要素とする。
- 表現の語順は固定的である。

この特徴を利用して、我々は共起頻度と語順制約に基づく定型表現の抽出方式を提案する。この方式は、コーパスから定型表現の構成要素となる表現（以下では、表現ユニットと呼ぶ）を抽出し、共起頻度の高い表現ユニットの組合せをコーパス中での語順に従って配置することにより、定型表現とするものである。

2.2 表現ユニットの抽出

表現ユニットの抽出には、文献[4]で提案したエントロピー基準に基づく文字列抽出方式を採用する。この方式は、隣接文字の分散の度合（エントロピー）を基準に、意味的なまとまりの強い文字列（英語などの場合は単語列）を抽出するものである。

隣接文字を $c_1, c_2, ...c_n$、w_j の出現頻度を $f(w_j)$、各々の隣接文字の出現確率 $p(c_i) = \frac{f(c_i)}{f(w_j)}$ とする時、文字列 w_j のエントロピー $H(w_j)$ は以下の式で求められる。

$$H(w_j) = \sum_{i=1}^{n} p(c_i) \log p(c_i) \qquad (1)$$

エントロピーの値は、文字列の長さや出現回数に依存しないため、句や節のようにある程度まとまった表現も、単語と同じように抽出することができる。また、形態素解析や辞書を必要としないためあらゆる言語に適用可能である[1]という利点もある。

[0]連絡先：下畑さより
沖電気工業(株) 研究開発本部関西総合研究所
〒540　大阪市中央区城見1-2-27 クリスタルタワー
Tel:(06)949-5101,Fax:(06)949-5108,
Email: *sayori@kansai.oki.co.jp*

[1]日本語のように単語境界が曖昧な言語では文字 n-gram を、英語のように単語境界が明確な言語では単語 n-gram を使えば良い。

本方式では、n-gram 統計によってコーパスから抽出した文字列のうち、エントロピーの値が閾値$T_{entropy}$を越えるものを表現ユニット$u_i(i = 1,2,...n)$として抽出する。

2.3 定型表現の抽出

定型表現の抽出処理は4段階に分かれる。

まず、任意の表現ユニットu_kに着目し、u_kを含む文をすべてコーパスから取り出す。表1は、u_kが"Refer to"の場合に抽出された文の例である。また、下線の文字列は各々表現ユニットに対応している。

次に、u_kとu_iの共起頻度を計数し、閾値T_{freq}を越えるu_iを抽出する。表2は、$T_{freq} = 2$の場合に表1から抽出されたユニットu_iのリストである。

次に、以下の処理を繰り返すことにより、表現ユニットの最適化を行なう。

- u_iとu_jが一部重複したり、隣接して出現する確率が閾値T_{ratio}より大きい場合、つまり、u_iとu_jが式(2)を満たす場合、2つのユニットを結合させて新たなユニットを生成する。

$$\frac{f(u_i,u_j)}{f(u_i)} \geq T_{ratio} \qquad (2)$$

- u_jがu_iを包含して出現する確率が、閾値T_{ratio}より大きい場合、つまり、u_iとu_jが式(3)を満たす場合、u_iを削除する。

$$\frac{f(u_j)}{f(u_i)} \geq T_{ratio} \qquad (3)$$

$T_{ratio} = 0.75$の場合、まず式(2)から"manual for specific instruction"が生成され、次に式(3)から"manual"と"for specific instruction"が削除される。この処理は、対象となるユニットがなくなるまで繰り返し行なわれ、最終的に適正な単位の表現ユニットだけが残る。表2の表現ユニットに対してこの処理を行なった結果を表3に示す。

最後に、u_kとの共起頻度が高いu_iから順に、コーパスの語順を保持した状態で生成する。表3の例では、

u_i	$f(u_k,u_i)$
the	4
manual	4
for specific instructions	3
on	2

表 2: 共起頻度が閾値を越える表現ユニット

u_i	$f(u_k,u_i)$
the	4
manual for specific instructions	3
on	2

表 3: 最適化された表現ユニット

まず"the"が対象となる。表1を参照すると、"the"はいつも"Refer to"の直後に出現するので、"the"の位置は"Refer to"の直後に決まる。次に、"manual for specific instruction"が対象となり、可変部をはさんで"Refer to the"の後ろに決まる。同様に"on"を対象に処理を行ない、以下の定型表現を抽出する。

Refer to the ... manual for specific instruction on ..

点線の部分は、語句の割り込みがあることを示している。例えば、"the"と"manual"の間には、"appropriate"や"install"などの単語が入る。

3 実験

定型表現の抽出実験を行なった。実験に用いたコーパスは英文コンピュータマニュアル(130万語,12万文)である。実験の条件を表4に示す。

[2]の手法によりコーパスから抽出した約16万件のn-gram文字列から、エントロピーの値が$T_{entropy}$を越える6,774件を表現ユニットとして抽出した。そのうち上位10件を表5に示す。

次に、抽出した表現ユニットをキーに定型表現の抽出を行なった。その結果、571件の表現が抽出された。

Refer to the appropriate manual for instructions on...
Refer to the manual for specific instructions.
Refer to the installation manual for specific instructions for ...
Refer to the manual for specific instructions on ...

表 1: "Refer to" を含む文の例

No.	定型表現
1	For more information on ..., refer to the ... manual.
2	You can use the ... to help you.
3	The syntax for .... is : ...
4	output from the execution of ... commands.
5	..., use the ... command with the ... option
6	... have a special meaning in this manual.
7	... to a (such as ..., and ...).
8	... if the system ...or a ...for a...,the..

表 6: 抽出された定型表現の例

$$T_{entropy} = 1$$

$$T_{ratio} = 0.8$$

$$T_{freq} = \frac{f(u_i)}{f(u_j)} \times 0.1$$

表 4: 実験条件

No.	表現ユニット
1	the current functional area
2	Before you install this device :
3	This could introduce data corruption .
4	All rights are reserved .
5	Note that the
6	, such as
7	Information on minor numbers is in
8	, for example ,
9	The default is
10	, you can use the

表 5: 抽出された表現ユニットの例

表6は、下線の表現ユニットをキーとして抽出された定型表現である。

3.1 評価

表現ユニットとして抽出された文字列は、コンピュータに関連する専門用語やマニュアルに特有の表現がほとんどであった。また、句読点を含む文字列や冠詞で終わる文字列も多く抽出された。これらの文字列は、文法的な単位ではないものの、テキスト中での語句の用法を決定する要素となるもので、定型表現を抽出する上で重要な手がかりとなる。その結果得られた定型表現にも、文型や構成要素数に関係なくコンピュータマニュアルで多用される表現が多数見られた。

No.7,8 は、不適切な抽出結果の例である。"to a" や ", the" のように不要な語句を定型表現の一部として抽出している。抽出に失敗した原因のほとんどは、このように不要な語句を抽出したものである。

過剰に抽出された語句は、句読点、前置詞、冠詞といった機能語で構成された比較的語数の少ない文字列であることが多い。こうした文字列は、表現ユニットを抽出する際に除去すべきである。

4 関連研究

近年、コーパスから定型表現を抽出する研究が盛んに行なわれている。特に最近では、非連続な定型表現の抽出に関する研究も行なわれるようになっている [1] [3] [5] [6] [7]。

Smadja[1] は、単語間の距離を考慮した共起関係から、非連続な表現を抽出する方式を提案した。また、日本語では、池原ら [3] が、n-gram 統計で抽出された文字列の 1 文中での共起頻度を計数し、共起頻度の高い組合せを定型表現として抽出する方式を提案している。

Smadja[1] の手法は、単語間の距離を単語数で計っている。しかし、日本語のように分かち書きされない言語では、単語区切りの概念が曖昧であるため、この方式は適さないと予想される。さらに、n-gram 統計による抽出文字列は未知語に頑健であること、コーパスによって最も適切なまとまりに分割可能であることから、英語のような分かち書き言語においても本手法の方が有効であると考えている。

池原ら [3] の手法では、我々と同様に、n-gram 抽出文字列を用いて共起頻度の高い文字列の組合せを抽出している。しかし、この研究では、定型表現の構成要素数が固定されてしまうという問題がある。これに対して、我々の手法は 2 項間の共起関係の組合せであるため、定型表現の構成要素数を予め意識する必要がない。

5 まとめ

本論文では、文字列の共起頻度と語順制約を利用した定型表現の抽出方式について述べた。この方式は、形態素解析や辞書を使用しないためあらゆる言語に適用が可能である。また、非連続な定型表現も連続する定型表現と同様に抽出できるという利点がある。

今後は、今回の実験で明らかになった問題点を解消し、定型表現抽出精度の改善を計る予定である。また、本研究では単言語を対象としているため、[8] [9] などの 2 言語を対象とする知識獲得の研究と融合を計り、具体的な応用技術について検討を進めたい。

参考文献

[1] Smadja,F.: Retrieving Collocations from Text: Xtract, Computational Linguistics, Vol.19, No.1, pp143-177(1993).

[2] 長尾, 森: 大規模日本語テキストの n グラム統計の作り方と語句の自動抽出, 情報処理学会自然言語処理研究会報告 96-1, pp1–8(1993).

[3] 池原, 白井, 河岡: 大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法, 情報処理学会論文誌 Vol.34, No.9, pp1937-1943(1995).

[4] 下畑, 杉尾, 永田: 隣接文字の分散値を用いた定型表現の自動抽出, 情報処理学会自然言語処理研究会 110-11 pp71-78(1995).

[5] 國吉, 中西: ギャップのある n-gram による言い回しの抽出, 情報処理学会自然言語処理研究会 117-6 pp37-44(1996).

[6] 尾本, 北: 距離反比例型スコアを導入したコロケーションの自動抽出法, 情報処理学会自然言語処理研究会 112-11 pp75-82(1996).

[7] Haruno,M., Ikehara,S., Yamazaki,T.: Learning Bilingual Collocations by Word-Level Sorting, COLING96, pp525-530(1996).

[8] Smadja,F., McKeown,K.R., Hatzivassiloglou,V: Translating Collocations for Bilingual Lexicons: A Statistical Approach, Computational Linguistics, Vol.22, No.1, pp1-38(1996).

[9] 北村, 松本: 対訳コーパス中の共起頻度に基づく対訳表現の自動抽出, 情報処理学会自然言語処理研究会 112-11 pp75-82(1996).