自然言語処理 114－10
（1996. 7. 18）

置換えを用いた n-gram による言語表現の抽出

内野 一* 白井 諭* 池原 悟† 新田見 緑‡

* NTTコミュニケーション科学研究所
† 鳥取大学
‡ 北海道大学

uchino@nttkb.ntt.jp

あらまし 機械翻訳システムにおいてユーザにあった正しいスタイルで翻訳を行なうためあらかじめ用意したテンプレートを用いて翻訳を行なう方法が知られている。このようなシステムに用いる定型的パターンを人手によって抽出することは容易でなく、これらを機械的に抽出することが必要となる。本稿では、原データに対して変換を行った後に、n-gram 統計処理による共起表現の抽出をすることで、長単位の定型パターンを抽出する手法について提案する。

キーワード 自動抽出, 定型表現, コーパス, n-gram

Automatic Extraction of Template Patterns Using n-gram with Tokens

Hajime UCHINO* Satoshi SHIRAI* Satoru IKEHARA† Midori SHINTAMI‡

* NTT Communication Science Laboratories
† Tottori University
‡ Hokkaido University

uchino@nttkb.ntt.jp

Abstract

In machine translation a useful method for translating repetitive text is template translation, that is using fixed templates with only a few variables. However, it is hard to compile these templates by hand. This paper describes a method of automatic extraction of template patterns from Japanese corpora using n-gram after replacing potential variable elements with tokens.

key words automatic extraction, template pattern, corpus, n-gram

1 はじめに

近年、パーソナルコンピュータおよびインターネットの普及に伴い、機械翻訳システムへのニーズが高まりつつあり、それに合わせて、安価で操作性の良いシステムが供給されてきている。このような形態で個人が補助として機械翻訳システムを使用する場合、訳文の意味が正しければ、実用上大きな問題はない。しかしながら、翻訳を業務として行なっている部所においては、文書も専門的な内容で、訳すべきスタイルや言い回しが定められていることが多く、単に翻訳が正しいのみならず、その訳文がスタイルに沿っているか、十分にこなれた文章であるかどうかが問題とされる。たとえ機械翻訳が正しい訳を出したとしても、その表現が通常のスタイルと大きくかけ離れていれば、結局その訳文は使用されない結果となる。

通常、日英機械翻訳を行なう場合、多くの種類の文章に対応するため、形態素解析、構文解析と順に、辞書情報、文法情報を用いて、英文に変換する方法をとる。この方法をとった場合、最終的な翻訳のスタイルを定めることは難しく、また、専門分野における独特の言い回しに対処することも困難であり、有用な翻訳文をなかなか得ることが出来ない。これらに対応するには、その分野における定型的な表現パターンをとらえて翻訳を行なう仕組みを合わせ持つことが必要となる。しかし、このような定型的パターンを人手によって抽出するのは用意ではない。そのため定型パターンをコーパスから機械的に抽出する手法が各所で研究されている [1] [2] が、これらの方法は対象となるテキストが単語単位に分割されていることが前提とされており、日本語のような原語では事前に形態素解析などにより分割をしておく必要があった。

これに対して長尾らによって提案された、テキストに対して n-gram 統計処理を行ない、テキストデータ内の文字列をその文字長の順および出現頻度の順に抽出する手法 [3] は事前の処理を必要としないものであった。この手法においては、断片的な文字列がかなりの割合で混在すると言う問題があるが、これを解決する手法として相互に重複する文字列を除去する方法 [4]、エントロピー基準を用いる方法 [5] などが提案されている。本稿では、この n-gram 統計処理を用いて日英機械翻訳で定型パターンとして扱うべき表現を自動的に抽出する手法を提案する。

2 n-gram による共起表現抽出時の問題点

我々は日経新聞社のオンラインサービスで提供されている市況速報記事に対して n-gram 統計処理を適用し、定型パターンを抽出する方式についての検討を行なった。

この市況速報データには市況情報特有の言い回しや、繰り返し現れる定型的パターンが頻出するため、定型的パターンを利用しての機械翻訳を行なうことが適していると考えられる。文献 [4] における n-gram 統計処理を市況速報記事９カ月分（95.6-96.2、記事数 6,315、文字数 1,460,112）に適用した結果、表 1のような結果が得られた。

表 1: 原データからの抽出結果

文字列長順	頻度順
先週末のニューヨーク市場では ... %に急低下した（９６文字、頻度２） ８時５０分に発表時間が変更に ... ドル買いに動いた（７７文字、頻度２） ７月のドイツの通貨供給量M３ ... 買い・マルク売りが（６９文字、頻度２）	ただ、(187) 市場では (126) という (81) 円相場はもみあい (63) 小動き (59) 一方、(58) 円相場は小動き (53)

文字列長から見た場合、１０文字以上の長さをもつ文字列はほとんどが頻度５以下であり、定型パターンとみなせるだけの意味のあるデータとはならなかった。また、頻度順に見た場合、意味のある単語が抽出されているため、離散型の共起表現の抽出を試みたが、要素数２の場合でも、ほとんどが頻度２しかなく、意味のある文型を抽出することは出来なかった。また、弱抑制型連鎖共起表現抽出方式 [6] による抽出も大きな差異は得られなかった。

これらの連鎖および離散共起表現の抽出結果を分析した結果、以下のことが原因であることが分かった。

- 数字の場合、一つのまとまった数値がその部分文字列で抽出される
  例えば、"100" と "300" のような数字では "00" の部分文字列が抽出される。
  また、数が違うだけでほぼ同じ形式を持った表現が、別の表現として分解されて集計されるため頻度が高くならない。

例 いずれも頻度６

２７万株と　２９万株と　３００株　３０円で引けた
３枚の売り越し　４５７　４６万株と　５０円でこの日の取引を終えた
５０円で取引を終えた　５２１５３７　５円に

- データ内の固有名詞が抽出される
  その固有名詞が同属性であっても、文字の違いにより別々に抽出されるため、低頻度の共起表現が出てくる。

例 いずれも頻度７

、セーレンも安い　、タカラブネ、　、ドレスナー、
、ルシアン、セーレン、　、ルシアンも安い
、住友シチックも安い　、神東塗も安い

3　データ変換を用いた抽出処理

3.1　データ変換手法

前記問題を解決するため、あらかじめ問題となるデータを別の文字列で置き換えることにより、抽出効果をあげる手法を考案し、実験を行なった。

- 数字の置き換え（数字列をひとまとまりにして文字列に置き換える）
- 企業名の置き換え（固有名詞である企業名を文字列に置き換える）
  また、企業名が連続している部分をひとまとまりにしての変換
- 括弧内の省略
- 引用文全体を置き換え

問題と考えられた数量、固有名詞表現に加え、文型を乱してしまう引用文や、括弧なども同時に置き換えの効果を検討した。置き換えにあたっては単純な字面レベルでの置き換えを行なうこととした。具体的には、数字列、括弧、引用文に関しては正規表現処理によって置き換えた。固有名詞に関しては、人名や、役職名などを認定して置き換えることも可能であるが、大量のデータに対して形態素解析を行なうのは、時間的、労力的に問題があるため、字面レベルでの置き換えを行なった。また、企業名に関しては、会社四季報や、新聞の株式面などから簡単にリストを作成することが出来るため、それを用いて変換を行なった。実際の変換リストにはデータの一部から抽出した企業名も加えて置き換えを行なった。このような方式で単純に置換えを行なったため、置き換え後のデータの中にも企業名が一部そのまま残ってしまったり、企業名以外の部分も置き換えてしまっているが、十分体量のデータを対象とすれば統計誤差とみなすことができる。

3.2　連鎖共起表現の抽出

前節でのデータ加工を行い、連鎖共起表現を求めた統計結果を表 2に、実際に抽出されたデータ例を表 3に示す。なお、表３における文字列長順、頻度順の抽出結果の例示にあたっては、その変換の効果を如実に表すデータをピックアップしている。たとえば、実際には、最長の抽出文字列はすべての変換方法において、原データからのものと一致しているが、例示ではそれは省略している。

- 数字変換
  表 2によると、総度数、種類数、平均文字長のすべてにおいて、最も効果があることがわかる。

表 2: 置き換えによる統計データの変化

加工法	種類数	総度数	平均文字列長
原データ	101,566	236,817	8.5
数字変換データ	81,843	196,352	9.0
企業名変換データ	94,981	221,130	8.8
企業名重複変換データ	94,544	221,040	8.6
括弧内省略データ	97,827	229,757	8.3
引用文変換データ	94,817	220,150	8.6
全変換適用データ	66,019	160,414	9.2

また、実際に抽出されたデータにおいても、ほぼ完全に１つの文がパターン化されて抽出されている。また、局所的にも効果が現れていることが、頻度順データから読みとれる。

- 企業名変換
  文全体としてみると企業名の連続数が１つでも違ってしまうと別のパターンとなるため、これだけでは文全体をパターン化するだけの効果がないことが文字列長順で抽出したデータから判断することが出来る。一方、局所的にはパターン化の効果があり、頻度順データにおいてそれを確かめることができる。

- 企業名重複変換
  文字列長順で見る限り、大きな効果は見られないが、頻度順データを見ると文全体をパターン化したものが抽出されており、また、部分的なパターン化も行なっていることが読みとれる。

- 括弧内省略 頻度順のデータに一部効果が現れているが、目立った効果は現れていない。

- 引用文変換 括弧内省略と同様に頻度順のデータに一部効果が現れているが、目立った効果はない。

- 全加工法適用 上記の置換えの効果がすべて働き、文字列長順、頻度順双方で効果を確かめることが出来る。しかし、一つの文に対して、省略以外の複数の変換が適用された例はほとんどない。これは、この市況情報の一つの性質と推測される。

3.3 離散共起表現の抽出

全データ変換適用後のデータに対しては、強抑制型連鎖共起表現抽出とともに、弱抑制型連鎖共起表現抽出を合わせて行なった。その結果から離散共起表現の抽出を行なった結果の一部を表 4に示す。

本実験においてはなるべく、長い表現を収集するような設定で行なったため収集された多くの表現は文の一部だけが欠けた形式となっている。たとえば、最後の行の、「ドル台後半」と「推移している」の間には、「で」「で軟調に」などが入っている。これらの抽出結果を逆に使用することにより、変数となり得る表現の組を求め、定型パターンの作成に利用することが可能である。

例えば、次のような３つの文があった時、

前場はＡＡＡが売られた。
前場はＢＢＢが売られた。
前場はＣＣＣが売られた。

離散共起表現を求めると

前場は　が売られた。

となる。この結果から逆に「ＡＡＡ」「ＢＢＢ」「ＣＣＣ」が同じ属性を持つ単語ではないかと推測でき、これらを定型的パターンの中の変数部分として扱うといったことが可能になる。

表 3: 置き換えによる抽出例

加工法	文字列長順	頻度順
原データ	先週末のニューヨーク市場では ... %に急低下した（９６文字）(2) ８時５０分に発表時間が変更に ... ドル買いに動いた（７７文字）(2) ７月のドイツの通貨供給量M３ ... 買い・マルク売りが（６９文字）(2)	ただ、(187) 市場では (126) という (81) 円相場はもみあい (63) 小動き (59) 一方、(58) 円相場は小動き (53)
数字変換	ＴＯＰＩＸ先物数月物は同数ポイント安の数ポイント、日経数先物数月物は同数ポイント安の数ポイントで前場の取引を終えた (11) 売買が成立したのは数銘柄（値付き率数%）で、このうち値上がり数、値下がり数、変わらず数、比較できず数だった (16) 日銀は数日数時数分、短期金融市場で手形買いオペ数円（期日数月数日、レート数%）を通知した (21)	ただ、(187) 売買高は数枚 (182) 前日比数銭円安・ドル高の数ドル＝数円数－数銭で取引されている (145) 売買高は概算数株 (111) 数日続伸 (106) という (81) 売買の成立した数銘柄（値付き率数%）のうち、値上がり数、値下がり数、変わらず数、比較できず数だった (73)
企業名変換	企、企、企、企、企、企が売られ、企、企、企、企、企、企、企、企、企、企(2)	（企）という (268) ただ、(186) 市場では (126)（企）(91) という (81) 円相場はもみ合い (63)（企）との声が聞かれた (59) 一方、(56) 円相場は小動き (53) 商いは低調 (53) 企、企が軟調 (43)
企業名重複変換	先週末のニューヨーク市場では ... %に急低下した（９６文字）(2) ８時５０分に発表時間が変更に ... ドル買いに動いた（７７文字）(2) ７月のドイツの通貨供給量M３ ... 買い・マルク売りが（６９文字）(2)	（企）という (268) ただ、(186) 市場では (126) 連企も安い (108) 連企も高い (101)（企）(91) という (81) 連企が軟調 (77) 円相場はもみ合い (63) 連企が売られ、連企も安い (30)
括弧内省略	先週末のニューヨーク市場では ... %に急低下した（９６文字）(2) ８時５０分に発表時間が変更に ... ドル買いに動いた（７７文字）(2) ７月のドイツの通貨供給量M３ ... 買い・マルク売りが（６９文字）(2)	という (556) ただ、(187) 市場では (126) との声が聞かれた (95) との声も聞かれた (76) との見方が多い (74)
引用文変換	先週末のニューヨーク市場では ... %に急低下した（９６文字）(2) ８時５０分に発表時間が変更に ... ドル買いに動いた（７７文字）(2) ７月のドイツの通貨供給量M３ ... 買い・マルク売りが（６９文字）(2)	引 (498) 円相場はもみ合い (62) という (60) 小動き (59) 市場では引 (57) ただ、引 (56) 円相場は小動き (53) 商いは低調 (53)
全加工法適用	ＴＯＰＩＸ先物数月物は同数ポイント安の数ポイント、日経数先物数月物は同数ポイント安の数ポイントで前場の取引を終えた (11) 数時数分現在の概算売買高は金が数枚、銀は数枚、白金、パラジウムはそれぞれ、数枚、数枚となった (10)	引 (525) という (308) 売買高は数枚 (182) 前日比数銭円安・ドル高の数ドル＝数円数－数銭で取引されている (145) 数日続伸 (138) 引という (123) 前日比数銭円高・ドル安の数ドル＝数円数－数銭で取引されている (112) 連企も安い (110) 連企も高い (110)

表 4: 離散共起表現抽出結果

文字列長順	
アジアのディーラー間ドル建て相場は数字トロイオンス数字ドル台後半	下落している
アジアのディーラー間ドル建て相場は数字トロイオンス数字ドル台後半	強含んでいる
アジアのディーラー間ドル建て相場は数字トロイオンス数字ドル台後半	弱含んでいる
アジアのディーラー間ドル建て相場は数字トロイオンス数字ドル台後半	上昇している
アジアのディーラー間ドル建て相場は数字トロイオンス数字ドル台後半	推移している

4 おわりに

本稿では、原データに対して変換を行った後に、n-gram 統計処理による共起表現の抽出をすることで、長単位の定型パターンを抽出する手法について提案し、その効果を示した。また、離散共起表現のデータから、同じ属性を持つ単語を推測し、定型的パターン内の変数として扱う手法を提案した。今後は、得られたデータを元に文の範囲を越えた定型パターンの抽出方法について検討を進めていく。また、日本文記事に対する英文記事もサービス上で提供されており、それら記事間を自動的に対応付けする方式 [7] についても研究されているため、今後、自動的対訳テンプレート作成についても検討していく予定である。

参考文献

[1] 浦谷, 加藤, 相沢: ＡＰ電経済ニュースからの定型パターンの抽出, 情報処理学会第 42 回全国大会, 6E-4 (1991).

[2] 北, 小倉, 森元, 矢野: 仕事量基準を用いたコーパスからの定型表現の自動抽出, 情報処理学会論文誌, Vol. 34, No. 9, pp. 1937–1943 (1993).

[3] Nagao, M., Mori, S.: New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, *COLING ' 94*, pp. 611-615 (1994)

[4] 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. 36, No. 11, pp. 2584–2596 (1995).

[5] 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報告, 100-11, pp. 71–78 (1995).

[6] 内野, 池原, 白井: 弱抑制による連鎖共起表現の抽出とそれに基づく離散共起表現の抽出, 言語処理学会第 2 回年次大会, pp. 257–260 (1996).

[7] 高橋, 白井, 藤波, 池原: ＤＢから抽出した日英新文記事の自動対応付け, 言語処理学会第 2 回年次大会, pp. 201–204 (1996).