4L－5

翻訳テンプレートの自動抽出
－緩やかに対応付けされたデータからの対訳抽出－

内野 一*　白井 諭*　池原 悟†
* ＮＴＴコミュニケーション科学研究所　† 鳥取大学

1　はじめに

機械翻訳において、専門分野に応じた高品質な訳文を得るためには、単語単位に翻訳を行うのではなく、その構文全体をとらえて、分野特有の言い回しや、表現方法に対応した翻訳を行うことが必要となる。このような翻訳を行うためには対象文を分析し、分野特有の表現をルールやデータとして蓄積していく作業が必要となる。品質を向上させるためには大量のデータに対する分析が必要となり、ルールを作成していく上での一つの障害となっている。

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、日本語のような言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方法 [3] は、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もので事前の処理を必要としないという特徴がある。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、記事ごとに対応付けされた日英の新聞記事コーパスに n-gram 統計処理を適用することによって、定型的な表現を抽出し、ルールを作成するための基データを自動的に収集する手法を提案する。

2　置換えを用いた n-gram 統計処理

我々は、固有名詞や、数詞の置換えを行なってから、n-gram 統計処理を適用することで、より長単位の定型パターンを抽出する手法を提案した [6]。この手法を用いることにより、一部の単語だけを置き換えるだけで、同じ文章となるような定型的な文を効率良く抽出することが出来る。本稿では、この手法をさらに英語文書にも用いて実験した結果を報告する。

2.1　複数の文への適用

理想的な対訳コーパスとしては、日本語の一文が英語の一文に翻訳され、各々の文の対応が付けられていることが望まれる。しかしながら、現実には簡単に収集することの出来る対訳データでは、日本語の一文が英語の一文に必ずしも対応するわけではない。例えば、新聞記事などでは日本語の記事データと英語の記事データを自動的に対応付けする [7] ことが可能となっているが、文単位での対応を見つけることは難しい。このような緩やかに対応付けられたデータから、対訳を抽出するためには複数の文にわたって出現する表現を考慮することが必要となる。

n-gram 統計処理を用いて、複数の文章からデータを抽出するには、n-gram の抽出範囲を文の区切りではなくパラグラフなどの区切りまで広げてデータを収集する、または文間の離散表現を抽出するという方法が考えられる。大きな範囲での離散表現の収集は、大量の記憶領域を必要とすることから、今回は前者の方法を採り、置換えを用いて複数文にわたる連鎖共起表現を抽出することとした。

2.2　各々の言語に対する置き換え処理

対象とした対訳データは、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（95.7-95.9、記事数2947）である。英語、日本語それぞれの記事データに対して、別個に統計処理を行なって共起表現を抽出した。日本語記事に対する前処理として、文献 [6] で行なわれている「数量詞、企業名の置換え」「引用部の置換え」「括弧内の削除処理」のすべてを行ない、さらに書き手による読点の使い方の影響をなくすため、読点をすべて削除した。

英語記事の n-gram 収集に関しては、単語の途中で表現を切らないように制御して収集を行なった。また、記事内で日付よりも 'last Monday' といった表現が使われているため、「曜日」を同一視するように置換え、スペルアウトされている「数詞、序数詞」、「月」、および「冠詞 (a,an)」の変換を行なった。

3　実験結果

日本語、英語記事、それぞれからの共起表現抽出結果（文字列長順）を表 1、 2に示す。文字列長順に整列した場合、数詞を多く含む文が上位に並び、ほぼ対応する文章を見つけ出すことができる。しかしながら、原文中では、多くの比率を占めていた企業名については連続した部分を圧縮してしまったため、上位に上がってこない。抽出された結果を分析すると、どちらの言語においても、７０～８０位程度に位置しており、対応する文章は同じ程度の頻度であることが分かった。これらの対応付けは、記録されている原文中の位置の記録から、対応する記事内から抽出された割合を調べることによって行なうことが出来る。

Automatic Extraction of Collocations for Machine Translation
* Hajime UCHINO, * Satoshi SHIRAI, † Satoru IKEHARA
* NTT Communication Science Laboratories, † Tottori University

4 考察

抽出された連鎖共起表現を基に、文章全体がパターンとなっている記事に関しては、ある程度自動的に対訳テンプレートを抽出をすることが可能であった。しかしながら、文章内で離れた位置にある2つの文が別の言語で1つの文になっているような表現はこの方法では抽出することができない。これらを抽出するには離散共起表現を収集することになる。今回、抽出した表現を分析したところ、置換えを行なっていない語に関してもまだまだ変数とできるものが多々あることが分かった。例えば、ピリオドと2文が連続して抽出された以下のような例がある。

度数	抽出表現 (企＝企業名, 連＝企業名の連続)
9	. AMONG DECLINERS WERE 連 AND 企. GAINERS INCLUDED 連 AND 企.
6	. AMONG GAINERS WERE 連 AND 企. DECLINERS INCLUDED 連 AND 企.
6	. AMONG GAINERS WERE 連 AND 企. LOSERS INCLUDED 連 AND 企.
5	. AMONG LOSERS WERE 連 AND 企. GAINERS INCLUDED 連 AND 企.

これらは、単語の順が違うだけでほぼ同じ文型であり、(DECLINERS, GAINERS, LOSERS) をグルーピングすることで一つのパターンとみなすことができる。特に英語においては、語数が一致するため、このように一部のみが異なった文型を容易に収集することが出来る。これを用いてさらにパターン化をしていくことにより、収集すべき表現の種類を減らすことが出来れば、これを用いて離散表現を収集していくことが出来る。

5 おわりに

本稿では、記事単位で対応付けられた日英対訳記事から、自動的に対となる表現を抽出する手法を提案し、その効果を示した。今後は、得られたデータを元にさらに自動化を進め、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していく。

表 1: 日本語記事からの共起表現抽出結果 (文字列長順)

度数	抽出表現 ([数]＝数字)
18	日銀は [数] 日午前の短期金融市場で短期国債を対象に買いオペ実施を通知した. 買い入れ額は [数] 円程度. 売り戻し条件付き現先方式で買い入れ期間は [数] 日から [数] 月 [数] 日.
13	日銀は [数] 日政府短期証券を [数] 円市中売却する. 買い戻し条件付きの現先方式で買い戻し日は [数] 月 [数] 日. 短資会社から銀行証券会社向けの実質転売レートは [数] %.
5	大蔵省は [数] 日短期金融市場で資金運用部による債券現先買いオペ実施を通知した. 買い入れ予定額は [数] 円程度. 期間は [数] 月 [数] 日から [数] 月 [数] 日まで.

表 2: 英語記事からの共起表現抽出結果 (文字列長順)

度数	抽出表現 (数＝数字, 曜＝曜日, 月＝月の名)
8	THE BANK OF JAPAN WILL SELL 数 YEN OF FINANCING BILLS UNDERREVERSE GENSAKI AGREEMENT 曜, THE CENTRAL BANK SAID 曜. THE BOJ WILL BUY BACK THE BILLS ON 月 数. MONEY BROKERS WILL RESELL THE BILLS TO BANKS AND SECURITIES HOUSES AT 数
7	THE BANK OF JAPAN NOTIFIED FINANCIAL INSTITUTIONS 曜 MORNING THAT IT WILL BUY 数 YEN OF BILLS VIA AUCTION LATER IN THE DAY, THE CENTRAL BANK SAID. PAYMENT FOR THE BILLS, WHICH MATURE ON 月 数,WILL BE MADE ON 曜.
5	THE BANK OF JAPAN WILL SELL 曜 数 YEN IN FINANCING BILLS UNDER GENSAKI AGREEMENTS, MONEY BROKERS SAID 曜. THE REPURCHASE DATE IS 月 数. BILL BROKERS WILL RESELL THE BILLS TO BANKS AND SECURITIES HOUSES AT 数

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, 110-11, pp. 71–78 (1995).

[6] 内野, 白井, 池原, 新田見: 置換えを用いた n-gram による言語表現の抽出, 情報処理学会自然言語処理研究会報告, 114-10 (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).