機械翻訳における文型パタンの部分的利用

松田聡史　山本和英
長岡技術科学大学 電気系
E-mail: {matsuda,ykaz}@nlp.nagaokaut.ac.jp

1 はじめに

近年、機械翻訳において、非線形な表現の部分に着目して大量の文型パタンを作成し、翻訳を行う手法が注目を浴びている。この目的のため、戦略的創造研究推進事業(CREST)の一環として「意味レベル文型パタン」(以下、CRESTパタン)が作成されている。

CRESTパタンでは、従来の結合価文型パタンの方法では扱うことのできなかった重文と複文の非線形な表現形式に対して、「線形要素と非線形要素を識別する方法を考案すること」との目標のもと、線形要素、非線形要素それぞれについて以下のような定義付けを行っている。

【言語表現の線形性と非線形性の定義】[1]

①一つ以上の代替要素が存在し、その要素に置き換えても「表現全体の意味」が変化しないような要素を「線形要素」と定義する。

② また、「線形要素」のみから構成される言語表現を「線形な表現」、1つ以上の非線形要素を有する言語表現は「非線形な表現」と定義する。

③　　但し、「表現全体の意味」とは、表現構造の表す意味、すなわち「抽象化された複合概念」のことであり、当該言語とは異なる任意の言語の意味的に対応する表現構造を用いて定義される。

CRESTではこれらの定義をもとに線形表現と非線形表現からなる言語モデルを構築し、非線形表現に着目した文型パタンの作成を行っている。

パタンを用いる手法では、適合するパタンが存在した場合に品質の良い翻訳が期待できるだけでなく、重文や複文といった他の手法では対応が難しい文についても対応がしやすいという利点がある。

しかし、目的とする文章が比較的長文であることや、非線形部分という特殊な文型に着目した手法であるため、入力文の微妙な違いに対する柔軟な対応が取りづらいといった問題点が存在する。

これらの問題に対して、短いパタンを部分的に適応させ、複数のパタンを組み合わせて翻訳を行うという手法が考えられる。しかし、変数として一般化できない非線形部分の存在する文型パタンでは、そのままの形での部分的な利用は困難である。そこで本稿では非線形部分および対応のとれている線形部分をもとにして部分的文型パタンを作成し、作成された部分的文型パタンによって入力文に対してどれくらいのパタンがマッチできるかについて実験および考察を行う。

2 関連研究

パタン翻訳は機械翻訳の中でも古くから行われている手法であるため、パタン作成については様々な手法が提案されており、入力文に対する部分的な網羅を目的とした短いパタンを作成する試みは複数行われている。神野ら[2]　　の名詞句の翻訳を目的とした手法では、名詞句コーパスをもとにして、和英辞書を用いた名詞句パタンの自動生成を行っている。また、白井ら[3]の意味解析を用いる方法では動詞と名詞の意味的な共起関係である結合価を用いた結合価パタンによる構文辞書の作成を行っている。

また、道祖尾ら[4]によって統計処理を用いたコーパスからのパタン自動抽出についても試みが行われているが、人手による判断が必要なものとなっている。

本稿ではコーパスからではなく、既存のパターンからのパターン作成を試みることで、新しい部分的なパターンの自動獲得を行う。

3 提案手法

図1-iに示すのはCRESTによって作成された文型パタンの一例である。文型パタンは日本語と英語のパタンが対になっており、それぞれの線形部分(N1[noun:名詞]、AJ4[adjective:形容詞]など)が対応している。本手法ではこのような文型パタンを元に、部分的に対応する箇所を探し、新たなパタンとして作成する。図1-iの中で下線部分が線形部分であり、それ以外が非線形部分である。

次に、日本語の助詞を除く非線形部分を元にパタンを2個に分割し、それぞれの非線形部分からパタンの先頭および末尾までを仮のパタンとして作成する。(図1-ii)

仮のパタンの中で、2つ以上の同じ線形部分を含むもの同士を取り出し、新しいパタンとして登録する。(図1-iii)

パタンによっては、英日どちらか一方の非線形部分が圧倒的に多いパタンも作成されるため、一方のパタンにおける非線形部分と、もう一方のパタンでの非線形部分の比が一定の閾値(文字数で4倍)以上のものについて削除を行った。(図1-iv)

表1 パタン数

	パタンの種類	パタン数
	CRESTパタン	122,692
分割位置	非線形部分	97,813
	前置詞	87,126
	関係代名詞	37,916

今回は、非線形部分を元に分割したパタンの他に、英

i) { WJ000168-00:<N1は>たったADV2する食事(ほど¦程)<N3は>AJ4^rentaiものはない。
WE000167-00:N3 never feel as AJ4 as when N1 have to eat ADV2.

ii) <N1は>**たった** **たった**ADV2する食事(ほど¦程)<N3は>AJ4^rentaiものはない。
<N1は>たったADV2**する食事(ほど¦程)** **する食事(ほど¦程)**<N3は>AJ4^rentaiものはない。
N3 **never feel as** **never feel as** AJ4 as when N1 have to eat ADV2.
N3 never feel as AJ4 **as when** **as when** N1 have to eat ADV2.
N3 never feel as AJ4 as when N1 **have to eat** **have to eat** ADV2.
※非線形部分(太字部分)をもとにパタンを2分割する

iii) { WJ:N1はたったADV2する食事(ほど¦程) { WJ:する食事(ほど¦程)N3はAJ4^rentaiものはない。
WE:as when N1 have to eat ADV2 WE:N3 never feel as AJ4 as when
※同じ線形部分(下線部)を含むもの同士をパタンとする

iv) { WJ:N1はN2には
WE:shoud not clothe N1.pron.poss N2 too heavily
{ WJ:^N1はN3をV3(ており¦でおり)ますので１７歳以下のN4の入場はお断り(いたし¦致し)ます
WE:N1 V3 only N2. N4 under
※一方のパタンの非線形部分(下線部)に対して、もう一方の非線形部分の割合が多いパタンを削除

v) (前置詞) { WJ:N1はN2にV3^rentai(まで¦迄)の間、 (関係代名詞) { WJ:N1はN2で出かけているあいだ
WE:until N1 V3.past N2 WE:While N1 be on N2

図1 提案手法によるパタン分割

文形態素解析器MXPOSTの形態素判定辞書(以下tagdict)から取得した前置詞および関係代名詞を元に上記と同様の手法でそれぞれ分割したパタンも作成した。(図1-v)

表1に作成したパタン数の一覧を示す。

4 実験

4.1 実験に使用する文

提案手法によって作成したパタンの有効性を評価するため、パタンを作成するに当たって使用された原文を用いて実験を行った。全体的な手法としては、英文に対してCRESTパタンを含む4種類のパタンをマッチさせ、何個のパタンが当てはまるかを調べた。

実験に使用する文の抽出には、tagdictを用いた。パタン作成に用いたのと同じく、tagdict中で前置詞、関係代名詞として判定される可能性のある単語を抜き出し、CRESTパタンの原文となった英文の中からそれらの単語(前置詞、関係代名詞)で始まる文を100文ずつランダムに選んだ。これと完全にランダムで選んだ100文のあわせて300文を実験用の文とした。

4.2 実験方法

実験には単純マッチを採用した。すなわち英文に対して形態素解析を行い、パタン中の単語及び変数(線形要素)が、実験文中でパタンと同じ順番で出てきた場合、マッチしたと見なした。ただし、パタンの作成手法を考慮し、パタンと例文それぞれの先頭の形態素が一致しない場合、マッチしていないものと見なした。また、いずれのパタンもパタン作成の元となった文に対してはマッチ数から除外した。図2にマッチする場合、しない場合それぞれの例を示す。

実験文	What shall I go in?
形態素解析	what/WP shall/MD I/N go/V in/RB
マッチする	what N V
パタンの例	what shall V
マッチしない	what ADV N
パタンの例	N V in

図2 パタンマッチングの例

5 結果および考察

5.1 パタン作成について

本稿では3種類のパタン分割(非線形部分、前置詞、関係代名詞-パタンの先頭および末尾)について作成および実験を行ったが、文中の非線形部分-文中の非線形部分といったパタンの先頭や末尾を含まない箇所についても作成を行った。これらの作成手法を応用することで、パタン中の特定の連語表現(in order to など。およびそれに対応する表現「...するために」)のみを抜き出してパタン化することも可能であると予想される。

また、線形部分-パタンの先頭および末尾、線形部分-線形部分についてもパタン作成を行ったが、意味の通らないパタンやあまりに短いパタンが大量に作成されたため、実験を行うまでには至らなかった。線形部分をもとにパタンを分割する手法については、分割した後に不必要なパタンを削除する手法の考案が必要となる。

5.2 ランダムに抽出した文に対する実験結果

図2はCRESTパターンを作成する際に原文となった文からランダムに100文を抽出し、4種類のパタンのうちいくつが当てはまるかを調べ、CRESTパタンのマッチ数で

ソートしたものである。各実験文にはソートした順に実験文IDをふってある。特にマッチ数が多い(1000以上)結果となっているのは"To"や"That"で始まる文であり、これらが考察に対するノイズとなりうる点やグラフの可読性を考え、以降の実験ではこれらの文は排除して行っている。"To"や"That"で始まる文のマッチ数が多い理由としては、もともとこれらの単語で始まるパタンが多く、分割されたパタンが多数作成されること、形態素解析によって先頭の単語に一般的なタグが付与されるため、多数のパタンがマッチするようになることが挙げられる。

グラフからCRESTパタンではマッチ数が少ない文(グラフ右側)で、作成したパタンでのマッチ数が多いことがわかる。CRESTパタンのマッチ数が0となった実験文に対するほかのパタンでのマッチ数を表2に示す。これら4文はすべて前置詞や関係代名詞で始まる文であったため、前置詞、関係代名詞で始まる文はCRESTパタンでは対応が難しく、かつ前置詞、関係代名詞から作成したパタンがこれらの文に対してうまく対応できるという仮定のもと、前置詞、関係代名詞で始まる文に対する実験を行った。

例1にCRESTパタンでは対応するパタンが発見できなかった実験文の例と、同様の文を入力した際に発見できた分割したパタンの例を示す。

例1) To give her room a new look, she decided to paper it in green.

接続詞 { WE:to V3 N1.pron.poss N2
WJ:N1はN2をV3^rentaiため

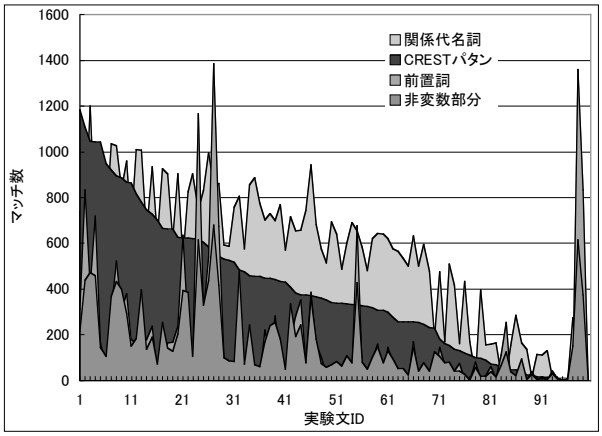

図2 パタンごとのマッチ数(ランダム100文)

表2 CRESTパタン

実験文ID	97	98	99	100
	各パタンのマッチ数			
非線形部分	150	614	374	1
前置詞	0	1,360	834	0
関係代名詞	274	3	3	0

5.3 接続詞から抽出した文に対する実験結果

図3は前置詞で始まる例文100文に対し、4種類のパタンをマッチさせたものである。強い相関関係はみられないものの、前置詞、非線形部分それぞれから作成したパタンで多くの場合CRESTパタンより多くのマッチ件数を得ることができている。tagdictでは多数の単語が前置詞として判断されるため、作成されるパタンや抽出された例文が多岐にわたることで、マッチ数の多い文と少ない文ができ、強い特徴が出なかったものと予想する。

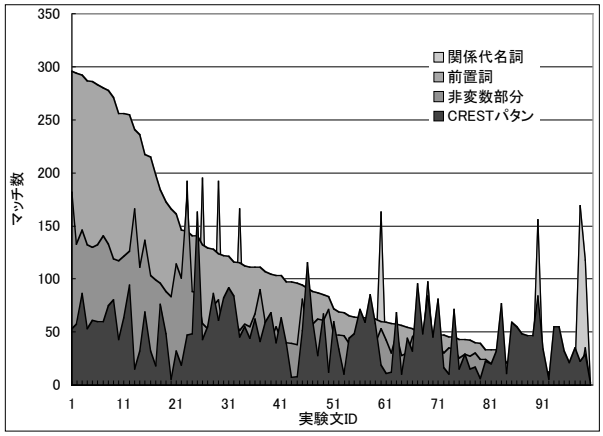

図3 パタンごとのマッチ数(前置詞100文)

5.4 関係代名詞から抽出した文に対する実験結果

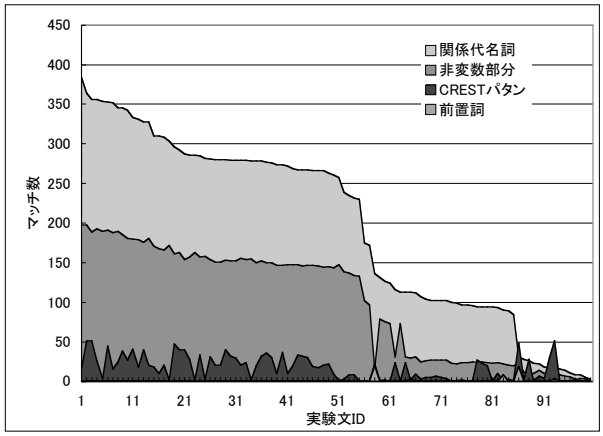

図4 パタンごとのマッチ数(関係代名詞100文)

図4は関係代名詞で始まる例文100文に対し、4種類のパタンをマッチさせたものである。図4から、非線形部分、関係代名詞それぞれをもとに分割したパタンについては、CRESTパタンからみてマッチ数が大幅に増加していることがわかる。CRESTパタンでは実験文IDが(55,56,57,61,63,73,76,81,89,95,98,99,100)の13点で入力された実験文に対して対応するパタンを発見できなかったが、非線形部分をもとに分割したパタンでは1つの実験文(ID:100)以外でパタンを発見でき、関係代名詞をもとに分割したパタンではすべての実験文に対して何らかのパタンを発見することができた。また、前置詞で分割したパタンからはほとんどマッチが得られなかった。ただし、これらのパタンは単純マッチで発見したものであるため、意味的な適合をとるためにはマッチングの条件を厳しくするなどの対策が必要である。

例2,3にCRESTパタンでは対応するパタンが発見できなかった実験文と、実験文を入力した際に発見できた分割したパタンの例を示す。

例2) When spring comes, flowers come out.

関係代名詞 { WE:when N1.pron V2
WJ:N1はV2^rentaiとき

関係代名詞 { WE:when N1 V2
WJ:N1のV2^rentai頃に

例3) What he is asserting is correct.

非線形部分 { WE:what N1 be V2.ing
WJ:N1のV2^rentai事はもっともだ

関係代名詞 { WE:what N1 be V2.ing
WJ:N1は何をV2.teiru^rentaiのか

また、図4のグラフ全体が横軸のID:55,85付近で大きく変化しているが、これはCRESTパタンの存在する数に大きく左右されている。グラフ中の最もマッチ数が多い部分(ID:1〜55)では、すべての実験文が"When"で始まる文章であった。また次にマッチ数の多い部分(ID:56〜85)では、ほぼすべての実験文が"What"で始まる文であった。パタン作成に使用したCRESTパタン約12万の中に、"When"および"What"で始まるパタンはそれぞれ1411、532パタン存在するため、それぞれのパタンからほぼ一定の割合でマッチするパタンが作成されているといえる。それ以降のマッチ数の少ない部分(ID:86〜100)では"Who"や"How"などで始まる文章が多く、これらの単語で始まるパタンが比較的少数なため、作成された部分パタン、マッチ数ともに少なくなったと見ることが可能である。

5.5 パタンの意味的対応

機械翻訳を行う際、パタンにおける原言語-対象言語間の意味的な対応がとれていることが不可欠である。そこで作成した3種類のパタンからそれぞれ100パタンをランダムに抜き出し、言語間の意味的な対応がとれているか否かを以下の3種類の基準で評価した。表3はその結果である。

○: 意味的な対応がとれている

△: 英→日、日→英いずれか片方向で意味的に対応の取れない非線形要素が存在するもの

×: 両方向で意味的に対応の取れない非線形要素が存在するもの。またはパタンが短いため、意味が一意に定まらないもの

表3 パタンの意味的対応

	○	△	×
非線形部分	43	47	10
前置詞	30	43	27
関係代名詞	35	58	7

例4に△と判断されたパタンの例を示す。

例4) { WE:N3 never feel as AJ4 as when
する食事(ほど¦程)N3はAJ4^rentaiものはない

※「する食事」に対応する部分が英語パタンに存在しない

前置詞で分割したパタンはほかよりも×と判定されたパタンが多い結果となったが、これは短いパタンが多くできたため、作成されたパタン部分だけでは意味の取れない構成となったことが要因として挙げられる。

片方向で対応の取れない非線形要素が存在するパタンがいずれの作成手法でも多く作成されたが、これは一方の言語では分かれて存在する非線形部分が、他方の言語ではまとまって現れるという言語間の特性によるものと推測できる。この点の解消のためには、パタン作成の際に非線形部分をさらに意味的要素を用いて分割したり、複数のパタンの意味的情報を見て共通部分を新たなパタンとするなどの工夫が必要になると考えられる。

6 おわりに

本研究では、CRESTによって作成された日英文型パタンをもとに提案手法を用いて部分的なパタンを作成し、CRESTパタンの原文に対して当てはめた際の各文に対するマッチ数について検証した。その結果、特定の文章に対してはマッチするパタン数が大幅に増加することを確かめた。一方、作成したパタンの意味的な対応についてはパタン作成の手法に工夫を加えるなど考慮の余地が残る結果となった。

今後の課題としては、対応する場所が文型パタン中にない非線形部分をどのようにして作成したパタン中から排除するかといった点が挙げられる。この点については、意味的な情報を考えてパタンの分割を行うことや、意味的な情報を考えて複数のパタンを統合することで解決が図れるものと予想できる。

謝辞

本研究の一部は、科学技術振興機構(JST) 戦略的創造研究推進事業(CREST)「セマンティック・タイポロジーによる言語の等価変換と生成技術」の支援によって実施した。

使用した言語資源およびツール

1) 改良型意味レベル文型パターンファイル
http://unicorn.ike.tottori-u.ac.jp/crest/

2) Adwait Ratnaparkhi, MXPOST: Maximum Entropy POS Tagger
ftp://ftp.cis.upenn.edu/pub/adwait/jmx/jmx.tar.gz

参考文献

[1] 池原悟,徳久雅人,村上仁一,佐良木昌,池田尚志,宮崎正弘: 非線形な重文複文の表現に対する文型パターン辞書の開発 情報処理学会 研究報告, NL170-11 (2005)

[2] 神野絵理,徳久雅人,村上仁一,池原悟: 文型パターンによる日英翻訳のための名詞句パターン辞書の構築 言語処理学会年次大会, pp.376-379 (2005)

[3] 白井論,横尾昭男,中岩浩巳,池原悟,宮崎正弘: 日英機械翻訳のための構文辞書 情報処理学会 研究報告, NL120-7 (1997)

[4] 道祖尾太祐,村上仁一,徳久雅人,池原悟: 日英対訳パターンの自動抽出に向けて 情報処理学会 研究報告, NL153-15 (2003)