4P−6

形態素解析情報に基づく長い日本語ニュース文の分割

金　淵培　江原暉将　相沢輝昭
NHK 放送技術研究所

1　はじめに

我々は日本語テレビニュース文を英語に機械翻訳する研究を行っている。日本語ニュース文は約８０％が６０文字（約３０単語）以上で書かれている。このニュース文は、ほとんど引用、並列構造を持っている。このような長文を構文分析する場合、その成功率を高めるには、長文をいくつかの単文に分割することが有効である。人手による分割実験によると分割される前の状態での構文分析の成功率は約６０％で、文分割後の構文分析の成功率は８２％に増加する。

ここでは、形態素解析情報、特に、格助詞「が」、係助詞「は」と用言の活用情報によって接続構造と分割点を把握できることに注目して、適切な分割点の検出と主語の補完を含む分割文の生成を行う方法を提案し、その有効性を述べる。

2　長文の分割アルゴリズム

分割作業は図１のように３つの段階に別れて形態素列上で実行する。

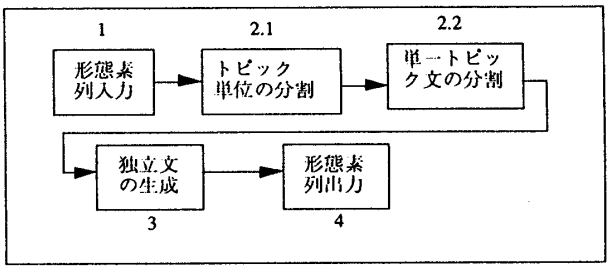

図１：分割処理の流れ

2．1　トピック単位の接続構造の検出

ニュース文のトピック単位の分割は次の現象を前提にする。１）ニュース文での係助詞「は、も」はトピックマーカとしての役割を明確する。２）係助詞はおおよそ文末に係り、特殊な場合を除いて連体埋め込み文には係らない。

トピック単位の分割は次の３段階（マージ、チェイン、レンジ）で構成されている。

以下、図2の例に即して説明する。

・マージ(Merge)

ここでは、文節と文節をマージして一つのブロックにすることと共にトピックや主語を検出する。

a) (スペインの, バルセロナで) --> bc (Case Block)
b) (bc, 行われた) --> brt (Rentai Block)
c) (１５か国が) --> * (Subject)
d) (日本は) --> bt (Topic)

・チェイン(Chain)

各ブロック間の依存関係を明確にして従属文とトピックを抽出する。

a) (brt, bt) --> # (Topic Chain)
b) (*, bry) --> Sry (Renyou Chain)

・レンジ(Range)

レンジでは、係助詞の性質を利用して考案された以下のようなトピックパターンを使って、文内の各トピックの範囲を決定してトピック分割点を提示する。

S1.....Sk (分割点) # Sk+1 ..... Sn
S1.....Sn (分割点無し)

Snは従属文に相当するものである.

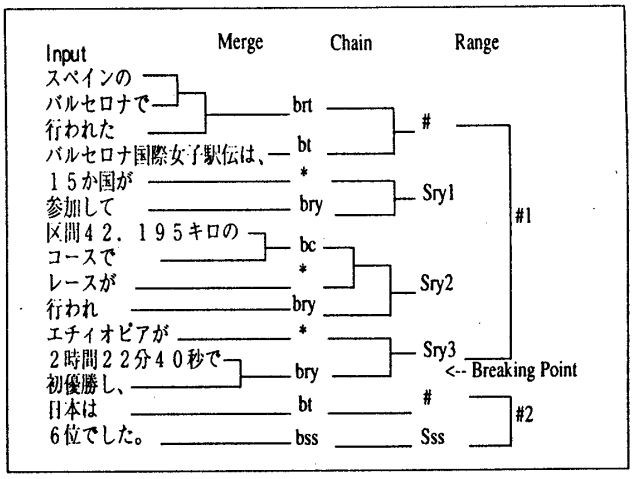

図２：トピック単位の分割の例

2．2　単一トピック文内の接続構造

トピック分割から得られた１つの文を単一トピック文と呼ぶ。しかし、この文が複数の従属文で構成されている場合、再分割を行う必要がある。ここでは、チェインによって抽出された従属文を接続表現を核とするパターンマッチング方法で分割の可能性を調査し、文の距離(長さ)による分割点の最終選択を行う。

Breaking long Japanese news sentences based on morphological information. Yeun Bae KIM, Terumasa EHARA, Teruaki AIZAWA
NHK Science and Technical Research Laboratories.

2.2.1 分割点候補の決定

上の例文(図2)は次のように２つの単一トピック文になる。 (a) # Sry1 Sry2 Sry3 (b) # Sss
(a)文は３つの従属文で構成されている。この場合はSry1とSry2が分割可能な連用文であるかどうかをパターンで調べれば良い。パターンは次のフォーマットを持つ。

[左パターン ＜中心パターン＞ 右パターン：アクションコード]

中心パターンは分割点になる部分を記述する。中心パターンとして連用中止、接続詞、またはニュース文の独特な表現が上げられる。左と右パターンは中心パターンが分割点になるために必要な条件が記述されている。複数のパターンがマッチングした場合は、左と右パターンの長さの長いものが優先される。アクションコードについては分割文の生成で述べる。

2.2.2 最終分割点の選択

いまの例文(a)では、Sry1とSry2が分割可能である。分割後の結果として３つの可能性が上げられる。この３つの可能性の中で最もバランスとれている分割点を選択する一つの方法として、分割文の長さを利用することが出来る。

これを実現するためには長さの単位と分割文としての最低長さを決めるべきである。ここでは、文節を長さの単位として採用し、分割文の最低長さは４にする。これによって、例文は次のように同一のトッピクを持つ文(a1)、(a2)とトピック分割からの(b1)に分割される。

a1) スペインのバルセロナで行われたバルセロナ国際女子駅伝は１５か国が参加して
a2) (バルセロナ国際女子駅伝は)区間４２．１９５キロのコースでレースが行われエチオピアが２時間２２分４０秒で初優勝し
b1) 日本は６位でした。

3 独立分割文の生成

通常、機械翻訳機への入力文は独立文であるため、分割された文を独立分割文に変える必要がある。

3．1 分割点部位の仕上げ

(2.2.1)で述べたパターンはアクションコードによって分類されている。例えアクションコードがRY0100の場合は分割パターンが連用タイプ１であるため分割点(連用形)を文末の述語の活用形に合わせて直し、句点を追加した後、仕上げを終了することを意味する。もちろん、分割点の接続表現を十分反映できる接続句を分割点の右側の文の文頭に投入させることも重要な作業である。これは各パターンと共に適切な接続句をあらかじめアクションコードに登録する必要がある。

3．2 主語の補完

元来主語がない日本語文を英語に機械翻訳する方法はある。しかし、分割されて主語をなくした文の翻訳結果は思わしくない場合が多い。

トピックの場合はトピックが主語の役割（述語に直接係る）を演ずる以外は補完されてもされなくてもよい場合が大部分である。補完の原理は簡単である。主語が支配している範囲内に分割点が存在する場合、その分割点の右側の文の主語は左側の文の主語と同一とする。

4 評価

分割の方法を評価するために、ＮＨＫの放送データベースから日本語ニュース文を約４００文をランダムに選定し、実験を行った。最初は、ニュース全文を人手で分割し、その分割文を実際に機械翻訳をしながら最も正し分割点（理想分割結果）を決定して基準として使用した。その次は、４００文から１００文を分割パターンの学習データとして利用した後、この分割方法を活用して４００文全体を対象に分割を行った。評価は実験結果と理想分割結果をトピック単位の分割成功率(tbr)と単一トピック文の分割成功率(sbr)に分けて比較することにした。

実験結果：

ｔｂｒ： ８６．５％
ｓｂｒ： ８８．９％

トピック単位分割の失敗原因は、1) 係り先の認定の誤り、2) 係助詞の前に来る格助詞の影響であった。

単一トピック文の分割の失敗原因は、1) 埋め込み引用文の認定失敗、2) 形態素情報の不足であった。

5 おわりに

以上、長い日本語ニュース文の分割方法について述べ、高い分割成功率を得た。分割成功率をもっと上げる方法として、1) 意味情報を利用する、2) ニュース文コーパスの統計処理によるニュース文係り先の定型パターンの抽出がある。

今後の課題としては、長い連体修飾文の分割処理や埋め込み引用文の抽出方法、より自然な主語の補完などが上げられる。

参考文献

1）池田尚志：助詞「は」の動きについての認知的なレベルからの考察、電総研彙報第５４巻第８号
2）児玉徳実：依存文法の研究、研究出版社(1987)
3）YOKO M.McCLAIN:Handbook of Modern Japanese Grammar, The Hokuseido Press (1981)