自然言語処理 93－3
（1993. 1. 13）

日英機械翻訳のための日本語ニュース文自動短文分割と主語の補完

金 淵培　　江原 暉将
NHK 放送技術研究所
email: kimyb@strl.nhk.or.jp　eharate@strl.nhk.or.jp

あらまし　日本語ニュース文等を英語に翻訳する作業を支援することを目的として、日英機械翻訳の研究を進めている。文の長さが長くなると、係り受け構造が複雑となり、翻訳精度が悪くなるが、話し言葉の一種であるテレビニュース文の約８０％は３０単語以上と非常に長くなっている。そのため、長いままのニュース文の機械翻訳は難しい。そこで、以下のことを行ない、良好な結果を得た。

・日本語ニュース文の短文分割を自動的に実行する。

・短文分割を行なった結果、主語が無くなった文に対して、主語を補完する。

約４００文のニュース文を対象に分割と主語補完の実験を行なった。分割点の認定には、分割点が記述されているパターン約１００個を用いてパターン・マッチングを行い、８８．９％の分割点認定率を得た。主語補完には、学習データを用いて、主語になる名詞の特徴ベクトルの確率分布を推定した後、各主語候補に対して主語になれる確率値を算出して主語補完を行なう統計的方法を使って７６％の補完率を得た。
この報告では、短文分割の有効性と方法及び主語補完について述べる。

An Automatic Sentence Breaking Method for Japanese-to-English Machine Translation

Yeun-Bae Kim　　Terumasa Ehara
NHK Science & Technical Research Laboratories, Japan
email: kimyb@strl.nhk.or.jp　eharate@strl.nhk.or.jp

Abstract　One ubiquitous problem in machine translation of spoken Japanese news into English is the translation of long news sentences. In general, they cause multiple analyses increasing the complexity of the translation process, and eventually lead to the wrong results or frequent failures. About 80% of the news sentences are composed of more than 30 words.

We propose an efficient method of analysis to break a long news sentence into short ones. The method consists of the following two distinct measures; First one concentrates on the recognition of break points (BP) via local patterns describing conjunctive structures. The second one supplements the proper subjects for those sentences generated without subjects.

An experiment was conducted to test efficiency of the method using 400 Japanese news sentences. We obtained 88.9% of proper BP recognition, and 76% of proper subject supplement.

1　はじめに

我々は日本語ニュース文等を英語に機械翻訳する研究を進めている。文が長くなると係り受け構造が複雑となり、翻訳精度が悪くなる場合が多い。その精度を高めるため長文を複数の短文に分割し、主語が無くなった文に対して主語補完を行った。

従来の文分割の研究としては、推敲支援を目的とした研究[1]や曖昧性を排除するための研究[2]がある。これらの手法は連用中止を含む接続表現の分類に基づいて解析ルールを設定し、分割を行っている。しかし、この手法ではルールの獲得や連体節、引用節の様な接続形式の分割への拡張が容易ではない。また、ニュース文の様な長文を機械翻訳する目的の場合、分割文の数が十分でない。

一方、主語補完の技術としては、待遇表現や発語内行為の制約を用いた補完手法[3,4]が提案されているが、これらの手法は、ニュース文の様な単独に発話される文には適用し難い。

本稿では、まず、短文分割の有効性について述べ、形態素列、品詞列、短文列の情報を用いた多層パターンを使用した分割の方法を述べる。これは連体節と引用節の分割にも適用できる。最後に、主語と述語間の関係を数量化し、統計的手法によって主語補完を行う方法について述べる。これは、ニュース文のような広範囲の文に適用できる。

2　短文分割の有効性

我々がニュース文の長さを調査した結果、日本語ニュース文の約８０％が３０単語／文以上で書かれていることが明らかになった。このニュース文を長いまま機械翻訳すると大部分は失敗してしまう。長文の構文解析の失敗原因のひとつは長さであると指摘した研究報告[5]もある。

約５００個のニュース文に対して人手による分割実験を行ない、その長さの分布を調べた（表１）。その分布によると、３０単語未満の文は、分割前の２２．５％から分割後の７８．５％に増加した。更に分割後では、７０単語以上の文はほとんどなくなった。

表１　長さによるニュース文の分布

文の長さ（単語）	分割前	分割後
０１～３０	２２．５％	７８．５％
３１～７０	６８．８％	２１．０％
７１以上	８．７％	０．５％

分割の効果を実際に測定するために、ニュース原文（３７８文・引用節が含まれてない文）と人手による分割文を機械翻訳し、比較した（表２）。原文では、３７８文の内、１５０文が構文解析に失敗した（訳文生成にも失敗）。

しかし、失敗文に対応する分割文側は、その１５０個の失敗文の内、８０文が構文解析に成功している。この様に短文分割は構文解析の成功率を高めるのに有効である。

表２　構文解析成功率の比較

	成功文	失敗文	構文解析成功率
原文	２２８	１５０	６０．３％
分割文	２９５	８３	７８．０％

3　短文分割と主語補完

分割処理のフローは次の５ステップで行われる。

1）形態素列の入力
2）情報素列の抽出
3）分割点の認定
4）主語の無い文の主語補完
5）形態素列の出力

長いニュース文は、表３の様な接続方法を組み合わせて作られている。分割はこれらの接続点（以下分割点と呼ぶ）で行われる。分割を行うと主語が無くなる場合があるので、その場合は主語の補完を行う。以下、順次これを説明する。

表３　ニュース文に現われる長文の種類

接続の種類	出現頻度	分割の対象
(ア) 連用中止法	（４０％）	対象
(イ) 引用法		
・直接引用	（１８％）	対象
・間接引用	（０２％）	対象外
(ウ) 連体法		
・形式名詞を修飾する	（１０％）	対象
・形式名詞を修飾しない	（２０％）	対象外
(エ) 他の接続法	（２０％）	対象

4　短文分割

4．1　分割 方策

（ア）連用中止の分割：表３から、連用中止法による復文の形成はニュース文でも最も一般的で、分割の第一対象であることが分かる（ここでは連用中止と

連用形＋「て」両方とも連用中止とする）。

一方、連用中止表現の中では、単なる並列接続ではないものがあるため、分割点文節を終止形に変える方法では有効な分割文の生成が難しい場合がある[6]。文の非並列接続については６．１章で詳しく述べる。例文１の様に連用中止は基本的に全て分割する。

「例文１」海部総理大臣はきょうの閣議のあと、吹田自治大臣と会談し、今後の政治改革への取り組みについて協議しました。

「分割文」海部総理大臣はきょうの閣議のあと、吹田自治大臣と会談しました。海部総理大臣は今後の政治改革への取り組みについて（吹田自治大臣と）協議しました。

しかし、次のような場合は分割しない。

1）副詞的用法を持つ表現：を初め、対し、これに関し、引き続き、に伴い、に加え、必要に応じ、によって、など約４０個

2）連用＋接尾辞：起きて以来、降りしだい、当たり券

3）連用＋連体：利用して出かけた人

4）連用の前に主語が見つからない場合：事故で大やけどをして病院で治療を受けているセルゲイ君．．．

イ）引用の分割：ニュース文は人の発言やその発言の内容を引用の形式を取って長く表現する場合が多い（ここでは引用節を含む文を一つの文として考える）。引用の形式には直接引用と間接引用がある[7]。直接引用は人の発言をそのまま引用し、間接引用では主に発言内容が引用される。ニュースに現われる直接引用節は「」によってはっきりマークされるので引用節の認定と抽出が簡単である（例文２）。

一方、間接引用では記号によって引用節がマークされてないので、その認定は簡単ではない。そのため、今回の報告では、間接引用節の分割は対象外とする。

「例文２」ミッテラン大統領は中東の戦後処理の問題について「われわれは国連の枠の中ですべての人にとって公正な形の平和の基礎作りを目指さなければならない」と述べました。

「分割文」ミッテラン大統領は中東の戦後処理の問題について次のように述べました。「われわれは国連の枠の中ですべての人にとって公正な形の平和の基礎作りを目指さなければならない。」

ウ）連体節の分割：また、ニュース文では、名詞を修飾する長い連体節を補足節として取る傾向が強い。この連体節の分割は分割文の再編成の必要な場合が多い。更に文中の連体節の認定も容易ではないので、ここでは分割の対象外とする。しかし、連体節が「こと、の、ところ」様な形式名詞、又は「結果、場合、際など」の様な名詞を修飾している際は分割を行なう（例文３）。

「例文３」会議は日程を１日延長して、連日明け方まで続けられた結果、全文で２６条からなる原案が本会議で採択されました。

「分割文」会議は日程を１日延長して、連日明け方まで続けられました。その結果、全文で２６条からなる原案が本会議で採択されました。

エ）他の接続節の分割：接続助詞（「て」以外）や接続表現による復文の分割には、まず接続点部位を調査し、分割の不可を確認した上、可能の場合は分割文間の接続関係を意味的に考慮して、適切な書き換えを実行する（例文４）。但し、日英機械翻訳には、接続表現による書き換えは人工的な英語を生成する場合があるので接続表現投入には注意が必要である。

「例文４」前回２位の日本は２区の寺沢選手が健闘しましたが、終盤、外国勢に抜かれ６位に終わりました。

「分割文」前回２位の日本は２区の寺沢選手が健闘しました。しかし、終盤、日本は外国勢に抜かれ６位に終わりました。

４．２　分割点の認定

基本的な考え方：分割対象文の入力形式は形態素列とする。分割点のパターン・マッチングを効率よく、かつ効果的にするために形態素情報を工夫して４種類の情報素列を得る（表４）。

「表４」情報素列の種類

表面素列(Surface Layer)：通信/所/で/は/、/郵政/省/の/免許/が/おり/しだい/、/インテルサット/の/予備/衛星/を/使って/、/埼玉県/に/ある/ＫＤＤ/上福岡研究所/と/の/間/で/電話/や/ＦＡＸ/通信/を/中心/に/およそ/２/年/間/送/受信/実験/を/行い/、/実用/化/に/こぎつけ/たい/と/し/て/い/ます/。

標準素列(Standard Layer)：通信/所/で/は/、/郵政/省/の/免許/が/降りる/次第/、/インテルサット/の/予備/衛星/を/使う/て/、/埼玉県/に/在る/ＫＤＤ/上福岡研究所/と/の/間/で/電話/や/ＦＡＸ/通信/を/中心/に/凡そ/２/年/間/送/受信/実験/を/行う/、/実用/化/に/漕着ける/たい/と/する/て/いる/ます/。

記号素列(Symbol Layer)：ncm sfx csp t S ncm ncm csp ncm * v1 sfx S npp csp ncm ncm csp v1 S npp csp v3 npp npp csp csp ncm csp ncm coo ncm ncm csp ncm csp ncm num sfx sfx ncm ncm ncm csp v1 S ncm sfx csp v2 csp v2 S

短文素列(Sentence Layer)：T , S br br br , bs .

同一の分割パターンの中で、違う情報素列が使えるので、多層・パターン・マッチング(Multi-Layered Pattern Matching)を行う。マッチングはパターンの長さ優先で行うので、あるパターンが合致したらその時点でマッチングは終了し、その点は分割点として認定される。ここでは、情報素列の抽出とパターン・マッチングの方法について述べる。

4．2．1　情報素列の抽出

表面素列と標準素列はそのまま形態素情報から得られる。表面素は原文の出現表現に当たる。標準素は出現表現の標準表現（用言の場合は原形、他は標準表記）からなる。記号素列は、約３０個の記号（表５）を使用して形態素情報から変換される。

表５　情報素列で使用されている記号の種類

記号素列の場合：	
ncm ：普通名詞	t：係り助詞
npp ：固有名詞	*：格助詞「が」
sfx ：接尾辞	v1：用言の連用形
csp ：格助詞	v2：用言の終止形
coo ：並列助詞	v3：用言の連体形
S：記号	など
短文素列の場合：	
T：主題	
S：主語	
br：連用接	
bs：終止接	

一方、短文素列はグルーピングと呼ばれる過程（図１）によって作られる。

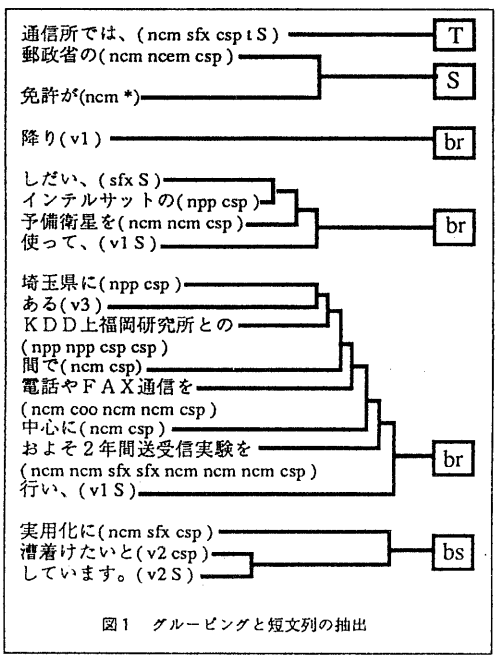

図１　グルーピングと短文列の抽出

このグルーピングは記号素列上で左から右へ前進しながら短文素を作っていく。しかし、途中で用言や提題助詞又は「が」が見つかると、グルーピングは一段落し、新たなグルーピングを開始する。これを文末が見つかるまで繰り替えす。分割パターンは大部分が用言を中心にして記述されるため、短文素と短文素の間は分割点候補になる。また、グルーピングは各文節の範囲を示すことができる。

4．2．2　パターン・マッチング

分割点のパターンを図２に示す。パターンの構成成分の内、重要な要素について説明する。

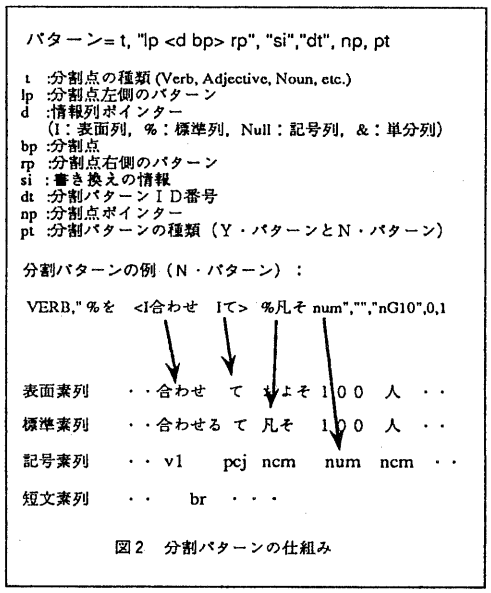

図２　分割パターンの仕組み

ｔ：分割点の種類を示す。これはパターンを種類別に管理し、マッチング処理のスピードを早くするためである。

ｌｐとｒｐ：分割点を中心にした左右のパターンを記述する。

ｂｐ：分割点の成分を記述する。

ｓｉ：分割後の分割文と分割文を接続する補助表現を示す。

パターンには「Nパターン」と「Yパターン」２種類が設定されている。「Nパターン」は分割不可点を記述し、「Yパターン」は分割可能点を記述する。即ち、「Nパターン」は「Yパターン」に対する制限(Constraint)である。もちろん「Nパターン」のマッチングは「Yパターン」より前に行われる。もし「Nパターン」が合致すればその点は分割点として失格である。また「Yパターン」が合致すればその点は分割点として認定される。表４の入力

文の場合、分割可能な点は次の２箇所である：

分割点１：免許がおり、（分割不可）

分割点２：衛星を使って、（分割可）

分割点３：実験を行ない、（分割可）

5　主語の補完

分割によって主語が無くなる場合がある。その時、主語がない日本語文を英語に機械翻訳する一方法として受動形化（Passivization）がある。英語には受動形より能動形が選好される傾向が強いので、できれば能動形の方が良い。そのためには、主語補完が必要である。また、さらに分割文の構文分析の失敗原因として主語喪失によるものが多いため、主語補完は分割文の翻訳には必要な作業である。ここでは、主語補完の方法について述べる。但し、次の３点を主語補完の前提とする：

・主語は補完対象述語の左側にある。

・主語は分割対象文内にある。

・主語は「は、では、が、には、を、の、も、に、で」のいずれかを持つ名詞である。

５．１　主語と述語の関係

主語候補と述語間の関係を把握することは主語補完のアプローチに決定的な影響を与えるため、最も重要な作業である。現在までの経験的観察によれば、その関係は次の特徴によって部分的記述できることが可能である。

ア）主語候補に付属する格助詞の種別

助詞は補足語と述語の関係を表したり主題を提示するため[7]、助詞の種別は正確な主語認定の一つの手掛かりになる。例えば、提題助詞「は」と格助詞「が」が付属している名詞は、他の助詞が付属している名詞より対象述語に対して、主語を表す可能性が高い。

イ）連体節との関係

連体節に対して主語となる名詞の係り受け範囲はその連体節に制限される場合が多いので、主語候補が連体節を修飾しているかいないかは主語認定の手掛かりになる（例文５）。

「例文５」鶴見町で父親と高校２年生の長女が乗った乗用車がガケから４０メートル下の海岸に転落しました。

ウ）主語候補と補完対象述語の意味的整合度

人間が主語の無い分割文に対して的確な主語を指摘できるのは、主語と述語の意味的整合がわかるためと考えられる。例文６について分割を行なう。

「例文６」政府は湾岸危機に対する中東貢献策のひとつとして国連平和協力法案を去年の秋の臨時国会に提出しましたが、野党側が強く反発し、結局、廃案となりました。

「分割文１」政府は湾岸危機に対する中東貢献策のひとつとして国連平和協力法案を去年の秋の臨時国会に提出しました。

「分割文２」しかし、野党側が強く反発しました。

「分割文３」（主語：法案）結局、廃案となりました。

分割文３の主語はもちろん「政府」や「野党側」ではなく、「法案」である。このような主語・述語間の意味的整合を計るため、「が」を中心とした語と語の係り受け解析資料[8]を利用した（約４万件）。まず、補完対象の述語に対して主語になれる名詞が属する意味マーカのリストを抽出する。その際、各主語候補（名詞）に対して付与されている意味マーカのリストも作って、その意味マーカのリストのマッチングを取る。意味マーカは分類語彙表[9]に基づいている。

図３の中の述語「作動する」は車やシステムような操作できる名詞とマッチし、「政府」のような名詞はマッチしない。この様に主語と動詞の意味的整合性も主語認定の一つの手掛かりである。

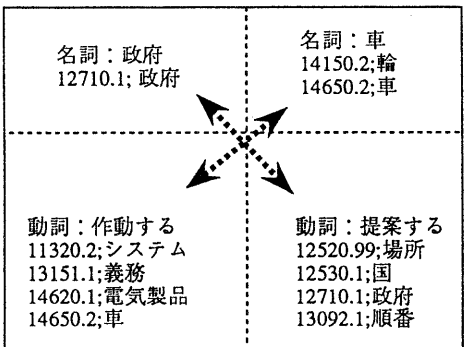

図３　主語と動詞の意味的整合

エ）主語候補と補完対象述語間の距離

主語が補完対象述語からどの程度離れているのかは、主語認定のもう一つの手掛かりになる。一般的に遠く離れているほど主語になれる可能性は低くなる。ここではその離れている度合、即ち距離を計る基準について述べる。

・主語候補と述語との間にある「は」格要素の数

提題助詞である「は」の動きは複雑であり[10]、「は」について全ての係り受けパターンを決定するのはあまり容易ではない。　一方、「は」はある主題を表すので、同一文内で他の「は」によって主題の切り替えを行なった際、前者が後者を越えて係る場合はあまり多くない（例文７）。即ち、同一文内の提題助詞は相互に影響を受けるため、提題助詞の先にある他の提題助詞の数は重要である。

「例文７」バルセロナで行われた国際女子駅伝は、１５か国が参加して区間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、日本は６位でした。

・主語候補と述語との間にある「が」格要素の数

格助詞「が」も提題助詞のように相互に影響を受けるため、他の「が」の存在は重要である（例文７）。

・主語候補と述語との間にある「は」と「が」以外の格要素の数

「は」と「が」以外の格要素の数は、主語と述語の間に存在する文節（又は格パターン）の数を表す。通常、主語の候補の内、述語により近いものが主語として認定される可能性が高いのでこの格要素の数が少ないほど主語として認定されやすい。

・主語候補と述語との間にある動詞の数：

述語のヘッドが動詞であるので、その数は主語候補と補完対象述語の間の距離を表していると考えられる。この情報は特に、「が」による補足語の主語認定に効果的である。ニュース文では、主語と述語の間にある他の述語の数が多いほど、その文に対する理解度は落ちると考えられる。そのため、述語が少ない方が主語になれる可能性が高い。

５．２　主語候補・述語関係の数量化と統計的接近方法

前項で述べた主語候補・述語の関係を明確に把握し、規則の形式で記述するのは簡単な作業ではない。そして、ここでは、この関係を数量化して統計的分析を行なう。この数量化方法（表６）を学習データに適用し、そこから主語になる名詞の特徴ベクトルの確率分布（Ｐ）と、ならない名詞の特徴ベクトルの確率分布（Ｑ）を推定する。

なお、特徴ベクトルの確率分布が多次元正規分布であると仮定して、確率密度関数のパラメータである平均値ベクトル（μ）と分散共分散行列（Λ）を学習

表６　主語特徴の数量化

１）主語に付属する格助詞	３）主語と述語の意味的整合度
「は」　：0	
「では」：1	合致する　：1
「には」：2	合致しない　：0
「も」　：3	わからない　：0．1
「が」　：4	
「の」　：5	４）主語と述語間の「は」の数
「を」　：6	
「で」　：7	５）主語と述語間の「が」の数
「に」　：8	
	６）「は」と「が」以外の格助詞の数
２）連体接との関係	
連体節の一部分　：1	７）主語と述語間の用言の数
連体節と無関係　：0	

データを用いて推定した。

「μ」と「Λ」を用いて式１[11]からある候補が主語になれる確率（ｐ）が算出できる。同様に主語になれない確率（ｑ）も算出できる。

式１　確率密度関数

$$p(x)=\left(\frac{1}{\sqrt{2\pi}}\right)^{k}\frac{1}{\sqrt{|\Lambda|}}\exp\left[\frac{-1}{2}(x-\mu)^{t}\Lambda^{-1}(x-\mu)\right]$$

５．３　主語の認定

表４の例文では、助詞「は、が、を、で」が付属する名詞を主語の候補とする。その際、複数の主語候補（S1, S2, ...,Sn）について、ｐ(Si)とｑ(Si)を求め、「ｐ(Si)／ｑ(Si)」を評価関数として利用し、その値が最大となる候補を主語として判定する。

$$\text{主語} = \text{Argmax}\{\, p(Si) / q(Si) \,\}_{(i=1,2,.....,n)}$$

表４の例文の中、「行う」に対して主語補完を実行した結果（表７）、通信所が主語として最も適当であった。

表７　述語「行う」に対する主語認定の結果

候補	p(i)	q(i)	p(i)/q(i)
候補１）通信所	0.043663	0.000411	106.326981
候補２）免許	0.015053	0.000443	33.963880
候補３）衛星	0.007791	0.001197	6.508704
候補４）通信	0.006579	0.006436	1.022223

6　分割と主語補完の精度評価実験

6．1　短文分割実験

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文選定し、評価実験を行った。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正しい分割点を決定して評価基準として使用した。評価はシステムによる分割結果と人手による分割結果を比較することにして次のような分割成功率を得た。

分割成功率：８８．９％

考察：少ないローカル・パターンの数（約１００個）で良好な分割結果を得たのは、ニュース文には定型的な文の接続パターンが存在し、更に、高頻度のパターンは少数であるからであると考えられる。

分割失敗の主な原因は連体節による非並列接続構造の分割と分離された複合動詞の分割である。例文８の場合、連用節「スキーや海水浴などで日焼けし」は連体節の一部分である。このような連用節の分割を避けるために、現在、我々は連体節の範囲認定の研究に取り組んでいる。一方、表４の例文の場合では、「衛星を使って、」と「実験を行い、」は副詞節によって分離されている。この場合、「使って」で分割すると、分割によって情報の流失が発生する。「連用１＋名詞＋格助詞＋連用２」又は「連用１＋副詞節＋連用２」ような接続パターンは分割パターンで記述できるが、「連用１」の分割可能性は連用節を作る動詞１と２の性質によるもので、全てのケースを調査してパターンとして登録する作業はコストが高くなる可能性があるので他の方法も講ずる必要がある。一つの方策として、構文解析情報を用いて分割点の認定を行う手法がある。しかし、コスト・パフォーマンス面を十分考慮して実行するべきであろう。

「例文８」到着ロービはスキーや海水浴などで日焼けし、お土産をいっぱい抱えた家庭連れなどで、ごったがえし、宅急便の窓口や都心に向かうバス乗り場には一日中、長い列ができていました。

6．2　主語補完の実験

分割点認定の実験で使用した４００文のうち、主語の無い約１００個の分割文に対して主語補完手法の精度評価の実験を行なった。補完対象文の数があまり多くないため、対象文のうち、７５％を学習データとして使用し、残りの２５％を試験データとして用いた。これを４回繰り返し、結果の平均値を精度評価の対象にした（表８：Jack Knife Test)。

試験文に対して、正解の主語が第１位であった場合が７６％で、正解が１位、又は２位の場合が８６％であった。

表８　主語補完実験の結果

候補	学習データ	試験データ
１位	81.9%	76.0%
１位と２位	89.0%	86.0%

考察：本手法の補完率（７６％）は、予備実験で実施した１）ルールによる補完率（６０％）、又は２）意味的整合性のみを用いた補完率（４０％）と比較して高い。これは、主語特徴の数量化分析にルールと意味的整合の効果がうまく反映されたからと考えられる。

今度の実験では、例文６のように「を」格を持つ主語が１位として認定されないケースがあった。その原因の一つは、学習文の中で「を」格を持つ主語の例が不足するからである。もう一つ考えられる原因は、主語と述語間の意味的整合性の強度が考慮されなかったからであろう。この整合の強度は、マッチングの頻度や意味マーカのマッチング・レンジ（今回は分類語彙表マーカの３～７桁を使用した）などを利用して表現可能であると思われる。また、「なる」、「する」のように汎用的で補助的な役割（ほとんどの名詞とマッチング可能）を持つ動詞に対して、意味的整合の精度はまだ低い。その精度を上げるために「混雑になる」のように名詞を含む述語パターンを多数登録するか、又は、述語の主動詞やヘッドを抽出する方法が考えられる。

現在の主語補完率は主語特性に全面的に依存しているため、主語特性の選定は非常に重要な作業である。

表６で提案している７個の主語特徴パラメータ以外にも主語の特徴はある。例えば、文末の文節は比較的多くの文節の修飾を受ける性質[12]を持っている。特に「は、が」のような助詞は文末によく係る。そのため、補完対象述語が文末要素であるかないかは主語認定の重要な手掛かりになる可能性がある。しかし、パラメータの数が多すぎるとオーバ・チューニングが起こる危険性があるので、特徴パラメータの最適化を行う必要がある。これは主語特徴ベクトル分布の固有ベクトル値を分析して特徴分布に対する各パラメータの

影響力を調べることで最適化が可能であろう。

高精度の主語補完にはまだ難しい問題がある。例えば、例文９の内、動詞「減少し」に対するの主語認定は現手法ではできない。即ち、主語の候補「輸出」と「内需」に対して「内需が増加すれば輸出は減少する」ような知識が正しい認定に要求されるからである。

「例文９」昨年度の鉄鋼輸出は中国・ソ連向けの輸出の大幅な減少に加え好調な内需が影響して６年連続して減少し、２１年ぶりの低い水準になる見通しです

このような知識の活用は現段階では無理である。しかしながら、例文９のような典型的な構文パターン、または文型パターンがニュース・コーパスから見つかれば、解決方法はあるので、興味深い研究テーマである。

7　今後の課題

短文分割：　現在、計算機による平均分割率（入力文の数／分割文の数）は２．８程度である。しかし、まだ長い分割文が頻繁に生成されている。これらの文はほとんど連体節による長文の場合が大部分である。

我々は、長い連体節を分割するための第１ステップとして連体節の範囲認定の研究を始めている。その成果を利用して分割率を４以上まで上げる予定である。

計算機による平均分割率が人手による平均値（２．５）と比べて高いのは分割点認定に失敗しているからである。その失敗率をもっと下げるため、分割点パターンの改良化とパターンの数の増強を行うべきである。現在、我々はＮＨＫの放送データベースを元に日本語ニュース・コーパスを構築している。今後、これを用いてニュース文の定型パターンを調査し、分割パターンの作成に利用する予定である。

この分割パターンの形式を用いて翻訳に邪魔になる要素の認定もできるので、自動前編集への応用も興味深いことである。

主語補完：　統計的アプローチ手法を用いて良好な補完結果を得た。この方法の長所は、複雑な動きを持つ助詞の係り受けのようにルールの形式では捕捉しにくい関係をつかむには最適である。しかし、メンテナンスが面倒であるのが指摘される。この手法の効果を最大にするためには、最適な学習と正確な主語特徴の把握が必要である。

今後の課題として、入力原文の外にある主語の補完がある。ニュース・テキストは記事単位で区切っているし、ある程度固定されたスタイルを持って書かれているので、一般テキストと比較して処理しやすいと思われる。

また、現手法を目的語の補完や代名詞の指示対象の同定など機械翻訳に必要な他の情報の補完に応用するのも興味深いことであろう。

8　まとめ

本稿では日英機械翻訳システムのための長いニュース文の分割点の認定手法と主語の無い分割文の主語補完手法、そしてその問題点について述べた。分割システムを翻訳システムのフロント・エンドとして使用するためには、分割点部位の用言を活用させて完全文を生成しなければならない。用言の活用はテンス、アスペクト、モーダル、スタイルなどを考慮する必要がある。この点につては別途報告したい。

参考文献

1）武石，林：接続構造に基づく日本語復文の分割，情報処理学会論文誌　第３３巻　第５号，(1992)

2）阿部，奥西，三吉：接続助詞に注目した文分割の一方式，情報処理学会第４２回全国大会，１Ｃ－７，(1991)

3）堂坂，小暮：対話参加者に関するゼロ代名詞の同定，情報処理学会第３９回全国大会，５Ｆ－５，(1989)

4）鈴木：対話翻訳における領域知識による補完手法の検討，情報処理学会第４５回全国大会，２Ｅ－１，(1992)

5）Nagao： Are the grammars so far developed appropriate to recognize the real structure of a sentence?, TMI-92, (1992)

6）武石，林：文分割における連用中止表現の扱い，情報処理学会第４２回全国大会，５Ｑ－９，(1990)

7）益岡，田窪：基礎日本語文法，くろしお出版，(1989)

8）田中：語と語の関係解析用資料（朝日新聞記事データー一年分）" が" を中心とした，(1989)

9）国立国語研究所：分類語彙表，秀英出版，(1964)

１０）池田：助詞「は」の動きについて認知的なレベルからの考察，電子技術総合研究所彙報第５４巻第８号，(1990)

１１）丸山：荻野日本語における文節間係り受けの統計的性質，情報処理学会第４５回全国大会，４Ｆ－７，(1992)

１２）小川：池田近代統計入門，森北出版，(1963)