日英機械翻訳のための日本語長文自動短文分割と主語の補完

金 淵培† 江原暉将†

日英機械翻訳の精度を低下させる要因の一つとして，文が長すぎるということがある．文が長くなると係り受け構造が複雑となり，構文解析ができず，翻訳に失敗することが多くなる．この問題を解決するため，われわれは日本語の長文を複数の短文に自動的に分割する研究を行った．われわれの手法は，形態素，品詞，文節カテゴリのようなさまざまな情報をフレキシブルに組み合わせて分割点の認定が行えるという特徴をもつ．さらに，分割を行うと，分割後の文に主語がなくなることがあり，この現象も機械翻訳の精度を悪くする．そこで，主語のなくなった文に対して，自動的に主語を補完する研究を行った．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う統計的方法を用いている．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，約88%の分割点認定率を得た．また，主語補完の補完率は76%であった．本論文では，短文分割の有効性と方法および主語補完について述べる．

An Automatic Sentence Breaking and Subject Supplement Method for J/E Machine Translation

YEUN-BAE KIM† and TERUMASA EHARA†

One ubiquitous problem in machine translation of Japanese into English is the translation of long sentences. In general, they cause multiple analyses increasing the complexity of the translation process, and eventually lead to the wrong results or frequent failures. As an example, about 80% of the Japanese TV broadcasted news sentences are composed of more than 30 words. To overcome the problem, we propose an efficient method of analysis to break (or partition) a long Japanese sentence into short ones. The method consists of the following two distinct measures; 1) Recognition of break points (BPs) using multi-layered patterns describing conjunctive structures around BPs. 2) Recognition of the proper subjects for those short sentences generated without subjects after breaking from the multiple candidates using a statistical approach. A series of experiments has been conducted to test efficiency of the method using 400 Japanese news sentences. We obtained approximately 88% of accuracy for proper BPs recognition, and 76% for proper subject supplement.

1. はじめに

機械翻訳の精度を低下させる要因の一つとして，文が長すぎるということがある．文が長くなると係り受け構造が複雑となり，構文解析に失敗することが多い．最もよく用いられるトランスファー方式の機械翻訳では解析・変換・生成の3段階を経て翻訳しているので，構文解析の失敗はそのまま機械翻訳の失敗となる．例えば，日本語のニュース文は長い文が多く，そのまま日英翻訳すると生成段階までも行かず構文解析の段階で失敗し，結局，翻訳結果が悪くなる場合が多い．これを解決するため，われわれは日本語の長文を複数の短文に自動的に分割する研究（以後，短文分割の研究と呼ぶ）を行い，良好な結果を得た．さらに，短文分割を行うと，分割後の文に主語がなくなることがあり，この現象も機械翻訳の精度を悪くする．主語のなくなった文に対しては，自動的に主語を補完する方法の提案をする．

従来の日本語文における短文分割の研究としては，推敲支援を目的とした研究[1]や自然言語における曖昧性を排除するための研究[2]等がある．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接続構造の解析ルールを設定し，分割を行っている．しかし，この接続分類（カテゴリ）の数は多くないため，コンテ

† 日本放送協会放送技術研究所先端制作技術研究部
NHK Science and Technical Research Laboratories, Program Production Technology Research Division

クストによって接続の構造が変わる場合には解析ルールの適用がしにくくなる．さらに，従来の方法では最適な分割点を1個所だけ求めており，十分な短文化が行えないことがある．複数の分割点を得るために，このルールを再帰的に適用した場合，分割対象文（前段階の分割文）が短くなる．それに従ってコンテクストへの依存度が高くなり，接続構造の解析がいっそう難しくなる．また，並列構造の推定によって長い文を正しく構文解析する研究[3]もあるが，並列構造と異なる連体節や引用節による接続形式の分割を取り扱っておらず，これらを含む長文への対処が不十分である．

提案する手法は，形態素，品詞，文節カテゴリのようなさまざまな情報が記述できるフレキシブルな多層パターンを使用して分割点の認定を行う点が特徴である．これによってコンテクストに敏感な接続詞や接続表現による主節と従属節の構造の記述が可能である．また，すべての分割点の候補に対して，分割の可能性を一挙に認定するため，各候補に対して文全体の情報を考慮することができるし，多数の分割点を同時に認定できる．さらに，この手法は連体節や引用節の分割にも部分的ではあるが適用可能である．

一方，主語補完の技術としては，従来，待遇表現や発語内行為の制約を用いた補完手法[4),5]が提案されているが，これらの手法は対話文を対象にしており，ニュース文のような単独に発話される文には適用し難い．また，従属節の統語的な情報を手掛かりに主語の同定を行う研究[6]もある．しかし，この手法は取り立ての「は」と主格助詞「が」でマークされた体言文節のみを主語補完の候補対象にしているため，例えば「を」でマークされた体言を主語として補完しなければならない場合には適用できない．

われわれは，主語と述語間の統語的関係と意味的関係を数量化し，統計的手法によって主語補完を行う方法を提案する．これは，ほとんど全種類の助詞でマークされた体言文節に対して主語になれる確率を計算し，最も適当な主語を選択するのでニュース文を含む広範囲の文に適用できる．

本論文では，まず，2章で短文分割の有効性と分割処理のフローについて述べる．3章では，各種の接続種類とそれに対する分割方策について述べる．また，分割点認定に必要な情報素列の抽出とパターン・マッチングの方法についても述べる．4章では，主語補完について説明する．ここでは，主語・述語間の特徴の数量化と統計的なアプローチについて説明する．5章では，短文分割と主語補完の精度実験の結果と問題点について述べる．

2. 短文分割の有効性と分割処理フローの概略

日本文の構文解析の失敗原因のひとつはその文の長さであり，50文字以上の文の解析は非常に困難で，80文字以上の文はほとんど構文解析に失敗すると指摘した研究報告[3),7)]がある．筆者らの調査でも，日本語のテレビニュース文の長さは，その約80%が30単語/文（単語の平均文字数を2とすると，60文字に相当する）以上で構成されており，このニュース文を長いまま機械翻訳すると大部分は構文解析段階で失敗してしまうことがわかった．これは文献3)と7)の記述とほぼ一致している．

構文解析に対する短文分割の効果を確かめるため，約500のニュース文に対して人手による分割実験を行い，分割前の原文と分割後の分割文を機械翻訳してその成功率を比較した．原文と分割文の長さの分布は**図1**に示すとおりである．その両者によると，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（図1の累積率 Acc:Before と Acc:After を参照）．さらに分割後では，70単語以上の文はほとんどなくなった．

われわれの機械翻訳システムで原文と分割文の翻訳実験をした結果を**表1**に示す．原文では，378文のうち，228文が構文解析に成功し，さらに96文が翻訳

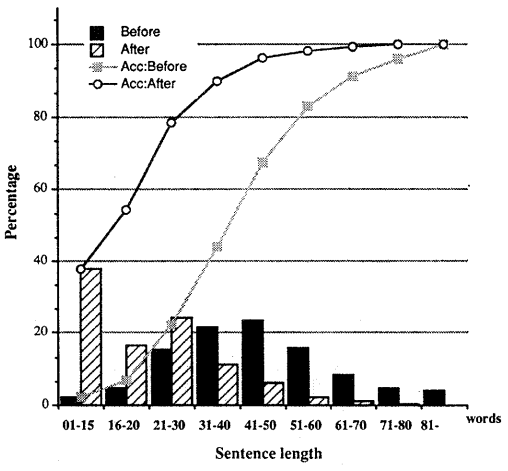


図1 日本語テレビニュース文の長さの分布
Fig. 1 Distribution of length of Japanese TV broadcasted news sentence.

に成功した．一方，分割文側は，構文解析に成功したのは295文で，翻訳に成功したのは，118文でありともに増加している．なお分割によって原文の1文が複数の文になりうる．そこで，統一をとるために表1では，原文も分割文も，原文の1文の範囲を単位としてカウントした．分割文側の成功の意味はそれら複数の文全てが成功したことを意味する．また，機械翻訳の成功は文意が誤りなく伝わるかどうかで判断した．このように短文分割は構文解析の成功率を高め，その結果機械翻訳の成功率を高めるのに有効である．

次に分割処理の概略フローを説明する．分割処理のフローは次の5ステップで行われる．

ア）形態素解析

イ）情報素列の抽出

ウ）分割点の認定

エ）主語のない文の主語認定

オ）形態素列の出力

ア）入力日本語を一文単位*で通常の方法で形態素解析する．

イ）形態素解析結果およびいくつかの形態素をグルーピングした結果から必要な情報素列（後述）を抽出し，分割点の候補を認定する．

ウ）その分割点候補に対してパターンマッチングを最長一致法によって行い，分割点を認定する．

エ）主語がない分割文に対して主語候補から主語を認定する．

オ）分割点の仕上げと主語補完の結果を形態素列として出力する．分割点の用言の活用では，原文の文末の活用状態（テンス，アスペクト，モーダル）と同一とする．主語補完では，選択された主語の後に「は」を挿入する．

3. 短文への分割

長いニュース文は，**表2**のような接続法を組み合わせて作られている．3.1節では，これらの接続法の説明と本手法での対象となる接続点の説明を行う．分割はこれらの接続点（以下分割点と呼ぶ）で行われる．3.2節では，分割点の認定手法について順次に説明する．

3.1 接続法の分類と分割の基本的方策

（ア） 連用中止の分割：表2から，連用中止法による複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わかる（ここでは連用中止と「連用形＋て」の両方とも連用中止とする）．連用中止の分割の実例を例文1で示す．

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協議しました．

ただし，次の1～4に示す連用中止の接続構造が検出される場合は，分割点の候補として認定しない．

1)コンテクストによって副詞的性質を表す連用形である場合：「をはじめ，に関し，に引き続き，に伴い，に加え，必要に応じ，によって」など約40個．

2)連用形に接尾辞や接尾的な述語成分が接続する場合：「起きて以来，降りしだい，当たり券」など．

3)連用形にすぐ連体形が接続する場合：「利用して出かけた人」など，連体節の一部である可能性が強い．

4)主語のない連用節（下線部分）で始まる場合（この場合は，連用節が連体節の一部であるケースが多い）：「カムチャッカで大やけどをして札幌医大病院で治療を受けているセルゲイ君の母親のタマラさんが，今日船で…」など．

（イ） 引用の分割：ニュース文は人の発言やその発

表 1 構文解析と翻訳成功率の比較
Table 1 Comparison of successful parsing and translation rates between before and after breaking.

	構文解析に成功	機械翻訳に成功	総文数
原文	228	96	378
分割文	295	118	378

* 直接引用を従属節として含む文は，それの主文の範囲を1文とする．

表 2 ニュース文に現れる接続法の種類
Table 2 Conjunctive types used in Japanese news.

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

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[8]．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．直接引用節は“「”と“」”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比較的簡単である．直接引用節の認定は，“「”と“」”の中に用言が存在するかしないかによって判別する．直接引用節の分割は，引用節を主文の後に移動することで行う（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いないので，その認定は簡単ではない．そのため，現在は，間接引用節の分割は対象外としている．

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，日本語文では，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が必要な場合が多い．さらに文中の連体節の範囲（特に連用節が連体節の一部である場合）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．ただし，連体節が「こと，の，ところ」のような形式名詞，または「結果，場合，際」などのような関係名詞を修飾している際は分割を行う（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：「て」以外の接続助詞や接続表現による複文の分割には，まず接続点部位に対してパターン・マッチングをすることによって分割の可能性を調べ，可能の場合は分割する．この場合，分割文間の接続関係を意味的に考慮して，分割点付近の適切な書き換えを実行する必要がある．例えば，例文4では分割文に「しかし」を補っている．

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．

3.2 分割点の認定手順

分割対象文を形態素解析し，形態素情報を得る．次に，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 3**）．

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）が可能である．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．

3.2.1 情報素列の抽出

表面素列と標準素列はそのまま形態素解析結果から得られる．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（用言の終止形や標準表記に寄せた表現）からなる．記号素は，ほぼ品詞に相当する約30個の記号である．記号素群の一部を**表 4** に示す．記号素も形態素解析から得られる．

一方，短文素列はグルーピングと呼ばれる過程（**図**

表 3 情報素列の種類
Table 3 Types of information used in pattern matching.

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, ncm ncm csp ncm * v1 sfx, npp csp ncm ncm csp v1, npp csp v3 npp npp csp csp ncm csp ncm coo ncm ncm csp ncm csp ncm ncm sfx sfx ncm ncm ncm csp v1, ncm sfx csp v2 csp v2.

短文素列 (Sentence Layer): Td, S Br Brt, Br, Bs.

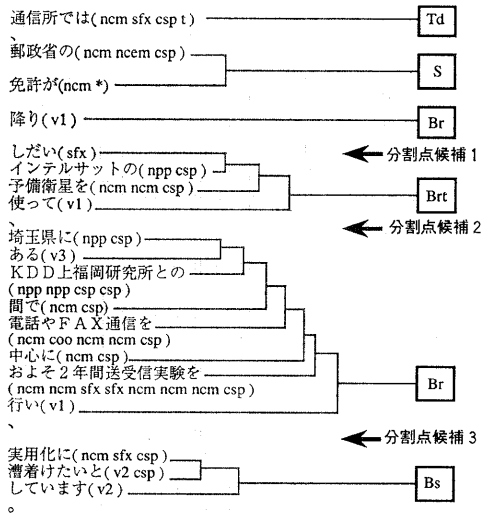
* この情報素列の種類は分割処理の必要に応じて何個でも追加できる．例えば，将来，意味情報がわかるような意味素列の追加が可能である．

2) によって作られるものであり，約20個の記号からなる．短文素群の一部を表4に示す．短文素列は体言グループと用言グループに大きく分けられる．前者はグループの主辞（ヘッド）が名詞であり，後者はヘッドが用言である．グルーピングは記号素列上で文頭から文末へ前進しながらグループを作っていくことで行われる．しかし，途中で用言，係助詞（「は」，「も」など），格助詞「が」，または連体に係る特殊要素，例えば，関係名詞，形式名詞などが見つかると，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．

表 4 使用されている記号素列と短文素列の記号の例
Table 4 Examples of symbols used in "Symbol and Sentence Layer".

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

短文素列の場合
Td：「では」でマークされた場合
S：「が」でマークされた場合
Br：連用文節
Brt：連用＋「て」文節
Bs：終止文節



グルーピングによる短文素列の抽出結果：
Td, S Br Brt, Br, Bs.

図 2 グルーピングと短文素列の抽出
Fig. 2 Example of Grouping with extraction of "Sentence Layer".

これを文末が見つかるまで繰り返す．図2の例文の中のように「降り＋しだい」のような分割点候補にはなれないグルーピングも生じうるが，後述する分割パターンを用いて分割点としては認定されない．グルーピングを行うことによって体言グループ（Td, S 等）と用言グループ（Br, Brt, Bt, Bs 等）が把握できる．用言グループの場合は用言の活用形と用言に接続されている接続詞または接続表現によって細かく分割される．これらは分割パターンの記述要素として利用できる．一方，現在のグルーピングでは，各グループ内の構成要素としてヘッド（体言グループの場合は名詞，用言グループの場合は用言）しか把握してないが，もっと詳しい構成要素の把握へ拡張することも可能である．しかし，グルーピングではグループ間の係り受け関係の分析は行われない．その関係は分割パターンに記述される．分割はほとんどが用言を中心に行うため，二つの用言グループ（Bで始まるグループ）の境界を分割点の候補とする．

3.2.2 パターン・マッチング

1) パターンの仕組み：分割点の認定は入力文に対応する情報素列と分割パターンとのマッチングによって行われる．ここでは，**図3**を用いて分割パターンの構成成分とその役割について説明する．

t：分割点の種類（動詞☆，形容詞，形容動詞）を示

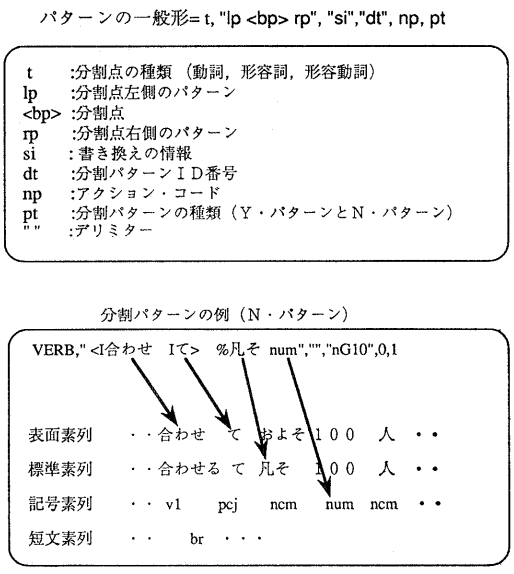


図 3 分割パターンの構造
Fig. 3 Structure of breaking pattern.

☆ 述語名詞を含む．

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．

bp, lp, rp：bp は分割点自身のパターンの要素を記述し，lp と rp では，分割点を中心としたそれぞれの左と右パターンの要素を記述する．マッチングは bp, lp, rp の順で行い，この三つのパターンすべてが合致すればマッチングは成功する．bp は "〈" と "〉" によって lp と rp から識別される．これらのパターンの各要素は情報素列へのポインタ（I：表面素列，%：標準素列，Null：記号素列，&：短文素列）と共に記述できるため，マッチングのとき，ポインタによる各情報素列の検索が可能である．例えば，図3の分割パターンの例では，「I合わせ」は表面素列上の「合わせ」を検索し，「%凡そ」は，標準素列上で，「およそ」に当たる標準表現「凡そ」を検索し，「num」は記号素列上で数詞に当たる記号 num を検索する．また lp, rp, bp に短文素を記述する際，短文素とともにそのヘッドを記述することができる．例えば，連用節で，ヘッドが「使って」を記述する場合は，"Brt：(I使っ I て)" とすればよい．

si：分割後の分割文と分割文を接続する補助表現を示す．si が指定されていない場合は，接続詞の種類をもとに次のような表現を自動的に生成する：

- 逆接：「しかし」
- 原因・理由：「そのため」，「したがって」
- 結果の場合は「その結果」

しかし，これらの接続表現による書き換えは例文5のように機械翻訳に悪い影響を与える場合がある（分割文A，機械翻訳文A）．この場合，si を活用して最適な接続表現を差し込むことができる．例えば，例文5で「—」を指定することによって接続表現を生成しないこともできる（分割文 B，機械翻訳文 B）．また，例文4のように接続助詞「が」が用いられる場合にもこの手法は適用できる．

例文 5　私は映画を見るのに対し，あなたは小説を読む．

分割文 A　私は映画を見る．それに対し，あなたは小説を読む．

機械翻訳文 A　I watch a movie. You read a novel to it.

分割文 B　私は映画を見る．あなたは小説を読む．

機械翻訳文 B　I watch a movie. You read a novel.

dt：各分割パターンの識別番号を示す．これは，分割パターンを種類別に管理するために設定されている．

np：アクション・コード：これは，分割文を生成する際の仕上げに必要な情報を記述する．例えば，np の値が4の意味は，「分割点から右側に存在する4個の形態素は分割文の仕上げのために不必要な情報である」を意味する．

pt：分割パターンの種類を示す．パターンには，pt が1の場合の「Nパターン」と pt が0の場合の「Yパターン」2種類が設定されている．「Nパターン」は分割不可能な点を記述し，「Yパターン」は分割可能な点を記述する．すなわち，「Nパターン」は「Yパターン」に対する制限（Constraint）である．そのため「Nパターン」は一般的に「Yパターン」より細かく記述されている．

2) パターンマッチング：パターン・マッチングはパターンの長さ優先で行い，「Nパターン」を用いて，まず分割不可能な点の処理を行った後，「Yパターン」で分割可能点の処理を行う．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．そのため，「Nパターン」と「Yパターン」が同時に合致する場合はない．一方，「Nパターン」が合致しなかった場合には，「Yパターン」のマッチングが行われる．しかし，「Yパターン」も合致しない場合は分割点として認定されない．このように二重でマッチングを取ることにより間違った分割を避けることができる．**表5** に示すパターンを表3の例文に適用し，分割点認定について説明する．

グルーピングから次の3箇所（BP1, BP2, BP3）の分割点候補がわかる．

分割点候補 1：免許が降り（BP1）しだい
分割点候補 2：衛星を使って，（BP2）
分割点候補 3：実験を行い，（BP3）

まず，表5の「Nパターン」と「Yパターン」が示

表 5　分割パターンの実例
Table 5　Examples of the actual breaking patterns.

「Nパターン」の例
1) VERB, "〈I 合わせ I て〉 %凡そ num", "", "nG10", 0, 1
2) VERB, "〈v1〉 %以来", "", "nG8", 0, 1
3) VERB, "〈v1〉 adv", "", "nG9", 0, 1
4) VERB, "〈v1〉 sfx", "", "nG4", 0, 1
「Yパターン」の例
5) VERB, "〈v3〉 %際%に", "その際に、", "N01", 2, 0
6) VERB, "〈v2〉 %と%共%に%、", "", "N02", 4, 0
7) VERB, "〈v1〉 %、", "—", "RY1", 0, 0
8) VERB, "〈v1〉", "—", "RY2", 0, 0

す意味は以下のとおりである．

「Nパターン」

パターン 1:「合わせて＋凡そ＋数詞」

パターン 2:「連用形＋以来」

パターン 3:「連用形＋副詞」

パターン 4:「連用形＋接尾辞」

「Yパターン」

パターン 5:「連体形＋際＋に」

パターン 6:「終止形＋と＋共＋に＋“，”」

パターン 7:「連用形＋“，”」

パターン 8:「連用形」

「Nパターン」を用いて，BP1 とマッチングを取るとパターン4と合致する．これで BP1 は分割不可能な点として認定される．BP2 と BP3 は「Yパターン」であるパターン7と合致するので，分割可能な点として認定される．

もし，BP2 で分割したくない場合は，例えば，「〈I 使っ I て I，〉& Br I,」を「Nパターン」として登録すれば「使って＋“，”＋連用節＋“，”」のような所では分割しない．

4. 主語の補完

表3の例文を図2の分割点候補2と3で分割すると，分割文「埼玉県にある…受信実験を行い」と分割文「実用化に…しています」に対して主語がなくなる．その際，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．しかし，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，できれば能動形のほうがよい．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省略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．ただし，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．

本主語補完手法では，まず「主語・述語」と「非主語・述語」の構文的と意味的特徴を数量化によって特徴ベクトル化する．これらを用いて主語になれる名詞と主語になれない名詞の特徴ベクトルの確率分布を各々推定した後，各主語候補に対して主語になれる確率値を主語になれない確率値で割った値を主語補完の判断基準として扱う統計的手法である．以下で，この手法について説明する．

4.1 主語・述語間の特徴と数量化

まず，「主語・述語」と「非主語・述語」の関係を次のような七つの特徴 ($x1$〜$x7$) に分けてとらえ，それぞれ数値を対応させることで数量化を行う．

ア）主語候補に付属する格助詞の種別 ($x1$)

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，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，係助詞「は」と主格助詞「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，主語となる可能性が高い．そこで，これらの格助詞「は，では，には，も，が，の，を，で，に」をそれぞれ数値「0, 1, 2, 3, 4, 5, 6, 7, 8」に経験的に数量化した．ここで，「は」，「では」，「には」などの文法的な差についてはそれらの数値が異なることで統計的に考慮されている．

イ）連体節との関係 ($x2$)

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．例えば，例文6では「株が」は連体節「売却された当時」を修飾しているため，「申告しています」の主語になれる可能性は低い．連体節を修飾する場合は「1」，しない場合は「0」に数量化した．

例文 6 平安閣では株が売却された当時，都内に住む 76 歳の男性が3日間だけ社長を勤め，株の売却益とみられる 30 億円に上る所得を申告していますが，所得税のほとんどは…

ウ）主語候補と補完対象述語の意味的整合度 ($x3$)

われわれが，主語のな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．例えば，例文7で，「廃案となりました」の主語は「政府」や「野党側」ではなく，「法案」である．

例文 7 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．

このような主語・述語間の意味的結合の整合度も主語認定の手掛かりである．この整合度を計るため，「が」を中心とした語と語の係り受け解析コーパス[9]を

---

* 「の＋用言」の場合のみを対象とする．例えば，「安全装置のない車」など．

利用した．このコーパスは，新聞から抽出された約4万個の「名詞（意味マーカ付き）＋が＋述語」形式の単文でできている．このデータを用いて意味的整合度のテストを行う．例えば，ある補完対象の述語「作動する」に対して，「政府」と「車」のように2個の主語候補が存在する場合に，意味的整合度のテストの手法を**図4**を用いて説明する．まず，コーパス上で，補完対象の述語1（作動する）に対して主語になれる名詞が持つべき意味マーカのリストAを作成する（意味マーカは分類語彙表[10]に基づいている）．リストAから，「作動する」は「システム，義務，電気製品，車」のような名詞を主語として取るのがわかる．次に，各主語候補1「政府」と主語候補2「車」に付与されている意味マーカのリストBとCを作り，リストAとB，そしてAとCのマッチングを取る．この例では，「車」は「作動する」に対して意味的に整合し，「政府」は整合しない．すなわち，「車」の方が述語「作動する」に対して主語になる可能性が「政府」より高い．意味的整合度の数量化を次のようにした：

整合する場合：1．0

整合しない場合：0．0

対象述語や主語候補がコーパスから見つからない場合：0．1

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，すなわち距離を計る基準を1～4に分けて述べる．

1）主語候補と述語との間にある「は」格要素の数（$x4$）：一般的に「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った際，前者が後者を越えて係る場合はあまり見られない（例文8）．すなわち，同一文内の係助詞「は」は相互に影響を受けるため，「は」格要素の先にある他の「は」格要素の数は重要である．

例文 8 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．

2）主語候補と述語との間にある「が」格要素の数（$x5$）：主格助詞「が」も係助詞「は」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8）．

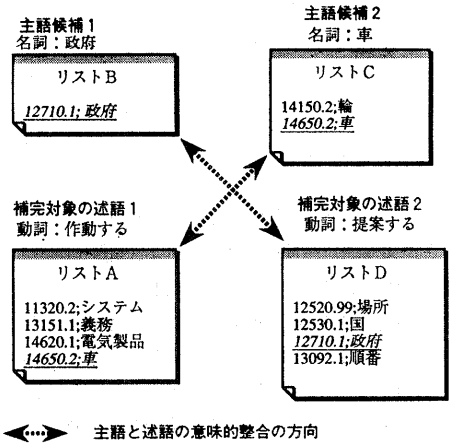


図 4 主語と述語の意味的整合
Fig. 4 Semantic agreement between subject and predicate.

3）主語候補と述語との間にある「は」と「が」以外の格要素の数（$x6$）：「は」と「が」以外の格要素の数は，主語と補完対象述語の間に存在する体言文節の数を表す．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．ただし，補完対象述語の格要素は数えない．

4）主語候補と述語との間にある動詞の数（$x7$）：主語候補と述語の間に存在する動詞の数を示す．一般的に，文末に係る主語を除いて，この数値が大きいほど主語になれる可能性は低い．ただし，連体形の動詞は数えない．

表6は，特徴（$x1$）～（$x7$）の数量化を表している．

表 6 主語・述語特徴の数量化
Table 6 Quantitative analysis of Subject-Predicate agreement.

$x1$) 主語に付属する格助詞 「は」：0 「では」：1 「には」：2 「も」：3 「が」：4 「の」：5 「を」：6 「で」：7 「に」：8	$x3$) 主語と述語の意味的整合度 合致する：1．0 合致しない：0．0 データがない：0．1 $x4$) 主語と述語間の「は」の数 $x5$) 主語と述語間の「が」の数 $x6$)「は」と「が」以外の格要素の数 $x7$) 主語と述語間の動詞の数
$x2$) 連体節との関係 連体節の一部分：1 連体節と無関係：0	

4.2 統計処理による主語の認定手法

4.1 節で述べた数量化された種々の特徴（$x1$〜$x7$）をパラメータとして用いて以下のように主語補完に必要な分布の推定を行う．まず，処理対象文から標本文を摘出し，各標本文内に存在するすべての述語に対する主語候補の中から人間が主語と非主語を分離して正確に認定する．この標本文群を学習データ☆とする．

人間によって「主語・述語」と「非主語・述語」とに分離された学習データに表6の数量化方法を別々に適用し，特徴ベクトル化を行う．これらを用いて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．

実際には，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する．$x1$ から $x7$ の7個の特徴パラメータを用いているので7次元となる．これに対して，確率密度関数のパラメータである平均値ベクトル（μ_P, μ_Q）と分散共分散行列（Λ_P, Λ_Q）を学習データを用いて推定する．

μ_P と Λ_P が求まったのでこれを用いて，式(1)（$k=7$）からある特定の候補が主語になれる確率密度 p が算出できる．同様に μ_Q と Λ_Q から主語になれない確率密度 q も算出できる．

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

複数の主語候補（$S_1, S_2, \cdots, S_n$）について，$p(S_i)$ と $q(S_i)$ を求め，$p(S_i)/q(S_i)$ 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補完する．例えば，後述する実験の結果によれば表3の例文中，「行う」に対する主語候補として「通信所」，「免許」，「衛星」，「通信」があり，これらに対して p/q を計算した結果（**表 7**），「通信所」の p/q 値（106.326981）が最も大きいので主語として補完される．

表 7 述語「行う」に対する主語認定の結果
Table 7 Subject recognition for verb "Okonau".

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

☆ 実験に用いた学習データについては，5.2 節で述べる．

5. 分割と主語補完の精度評価実験

5.1 短文分割実験

分割の方法を評価するために，1991年1月1日から2月10日までの日本語テレビニュース文から 381 文をランダムで選定し実験文とした．その中，23文は短文で，分割点がないため，実験の対象外とした．まず，ニュース全文を人手で分割し，正しい分割点を求めた．次に，約100個の分割パターンを用いて，機械によって分割を行った．両者を比較し，すべての分割点が完全一致する文を成功とした．その結果，358 文の中，315 文が成功した．

分割成功率：87.9%

このように分割点を中心としたさまざまの情報を利用すれば，良好な分割結果を得ることができる．さらに，分割パターンの数を増加することによって，現在より高い分割成功率が期待できる．

一方，分割失敗の原因として次のようなものがある．1)「連用形＋て」が副詞的に使用される場合で，例えば「2年連続して全国で最も高い…」ような非並列接続である．2)連用節が連体節の一部分になる場合である．例えば，例文9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．

例文 9 到着ロビーは，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．

1)の場合は，「連続して＋連用節」のような「Nパターン」を登録して解決できる．2)の場合は，連体節の範囲の認定が必要であり，その解決は今後の課題である．ただし，このような文の数は381文の内7文と少ないため，成功率には大きい影響はない．

5.2 主語補完の実験

分割点認定の実験で使用した文のうち，分割によって主語のなくなった108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（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，または2位の場合が86%であった（**表 8**）．また4回の実験の標準偏差は第1位に対して3.6%であった．

本手法の補完率（76%）は，予備的な実験で実施した「文法ルールのみによる補完率（60%）」，または「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相補的に働いたからと考えられる．今回の実験では，例文7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．失敗の原因のもう一つは，「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，意味的整合精度が低いためであった．

表8 主語認定の実験結果
Table 8 Result of subject recognition experiment.

	学習データに対する正解率 (%)			試験データに対する正解率 (%)		
	文番号	1位	1位と2位	文番号	1位	1位と2位
1回目	01～81	80.7	87.1	82～108	76.6	86.6
2回目	28～108	80.9	90.4	01～27	79.1	87.5
3回目	01～27 55～108	79.2	87.0	28～54	70.9	74.1
4回目	01～54 82～108	80.7	91.5	55～81	80.0	96.0
平均値		80.4	89.0	平均値	76.7	86.1
標準偏差		0.7	1.7	標準偏差	3.6	7.9

表3の例文に対する短文分割と主語補完の結果を原文と分割文の英語翻訳結果とともに付録に示した．

6. おわりに

本論文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ない分割文の主語補完手法，およびそれらの評価実験結果と問題点について述べた．ここで述べた手法は比較的簡単なものであるにもかかわらず，かなり高い成功率を得ており，ニュース文をはじめとする，広範囲なテキストに適用可能であると思われる．

短文への分割については，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が長い連体節を含む文である．われわれは，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．

また，現在の手法では，「Nパターン」の数が多くなる可能性があるため，特に非並列接続構造の分割パターンの記述をさらに効率的に行う必要がある．

主語補完については，統計的アプローチ手法を用いて良好な補完結果を得た．しかし，より的確な主語認定結果を得るため，次のような点を検討する必要がある．

1) 提案した主語・述語の特徴で十分か？

2) 提案した数量化方式より効果的な方式は？例えば格助詞の数量化の最適性は？

3) 提案した学習モデルより優れたモデルは？

また現在は，原文の外にある名詞は主語の候補から除かれているが，この制限をなくすことも課題である．さらに，認定された主語候補を代名詞化してより適切な英文主語（it, they など）を出力することも可能である．ただし，主語候補の意味特徴を正確にとらえる必要がある．

謝辞 本研究のテーマを与えて下さった当所の相沢輝昭主幹研究員に心から深く感謝します．日頃熱心に議論していただいた当所自動翻訳グループの諸氏に感謝します．

参考文献

1) 武石，林：接続構造解析に基づく日本語複文の分割，情報処理学会論文誌，Vol. 33, No. 5, pp. 652-663 (1992).

2) 阿部，奥西，三吉：接続助詞に注目した文分割の一方式，第42回情報処理学会全国大会論文集，1C-7, pp. 13-14 (1991).

3) 黒橋，長尾：長い日本語文における並列構造の推定，情報処理学会論文誌，Vol. 33, No. 8, pp. 1022-1031 (1992).

4) 堂坂，小暮：対話参加者に関するゼロ代名詞の同定，第39回情報処理学会全国大会論文集，5F-5, pp. 644-645 (1989).

5) 鈴木：対話翻訳における領域知識による補完手法の検討，第45回情報処理学会全国大会論文集，2E-1, pp. 89-90 (1992).

6) 福本ほか：係り受けの強度に基づく依存文法—制限依存文法—，情報処理学会論文誌，Vol. 33, No. 10, pp. 1211-1223 (1992).

7) Nagao, M.: Are the Grammars so far Developed Appropriate to Recognize the Real Structure of a Sentence?, *Proceedings of 4th International Conference on Theoretical & Methodological Issues in MT*, Montreal, pp. 127-137 (1992).

8) 益岡，田窪：基礎日本語文法，くろしお出版 (1989).

9) 田中，吉田：自然言語の知識獲得（その1）—語

と語の関係について，朝日新聞記事データの分析（“が”について）—，情報処理学会自然言語処理研究会資料，69-3 (1988).
10) 国立国語研究所：分類語彙表，秀英出版 (1964).

付録 原文と分割文の機械翻訳結果
Appendix: Result of machine translation of before and after breaking

Input Sentence:

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．

As soon as license of the Ministry of Post and Telecom gets down, the transmission-and-reception experiment during about 2 years is performed focusing on a telephone or FAX communication between KDD Kamifukuoka research lab in the Saitama prefecture using the reserve satellite of an Intelsat, and it is considering as wanting to reach utilization in the communication station.

Short Sentence 1:

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いたいとしています．

As soon as license of the Ministry of Post and Telecom gets down, it is considering as wanting to use the reserve satellite of an Intelsat in the communication station.

Short Sentence 2:

通信所は埼玉県にある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たいとしています．

The communication station is being taken as wanting to perform the transmission-and-reception experiment during about 2 years focusing on a telephone or FAX communication between KDD Kamifukuoka research lab in the Saitama prefecture.

Short Sentence 3:

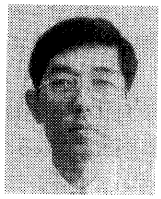
通信所は実用化にこぎつけたいとしています．

The communication station is being taken as wanting to reach utilization.

(平成5年4月7日受付)
(平成6年2月17日採録)

金 淵培（正会員）

1958年生．1983年University of São Paulo大学工学部化学工学科卒業．1984年(株)ブラビス・インターナショナル入社．1991年NHK入局．現在，放送技術研究所・先端制作技術研究部においてニュース用日英機械翻訳，自然言語処理等の研究に従事．



江原 暉将（正会員）

1967年，早稲田大学第1理工学部電気通信工学科卒業．同年，日本放送協会に入局．1970年より，放送技術研究所に勤務．現在，先端制作技術研究部主任研究員．その間，1989年から1991年までATR自動翻訳電話研究所に出向．かな漢字変換，機械翻訳，言語データベースなどの研究に従事．電子情報通信学会，言語処理学会，Association for Computational Linguistics各会員．