

## 論理展開の情報を付与した英日対訳コーパスの開発 —英文アブストラクト作成支援ツールの構築に向けて—

成田真澄

(株) リコー 情報通信研究所

### 1. はじめに

インターネットを介した海外からの情報収集や海外への情報発信が日常茶飯事になりつつある現在、我々は今まで以上に英語でのコミュニケーション能力を高める必要に迫られている。特に、最新の技術情報や研究成果を英語で海外に発信する機会が多くなると、内容を正確に伝えられる英語文書をいかに効率よく作成するかが重要になってくる。

日本人が英語文書を書く際に最も留意しなければならないことは、読者が正確かつ容易に内容を把握することができるように直線的な論理展開を心がけることである。日本語文書に見られるような、状況や基本的な考え方といった周辺の事象の説明が続いて結論になかなかたどりつけない文書構成では受け入れてもらえない[1][5][7]。にもかかわらず、このような英語文書が要求する論理展開構造を考慮に入れた英文作成支援手法はこれまで提案されていない。

そこで我々は、英語の論理展開構造がひとつのパラグラフに凝縮された「英文アブストラクト」の作成支援ツールをコンピュータ上に構築する研究に取り組むことにした。具体的には、科学技術分野の英語論文や英文レポート用のアブストラクト作成を支援するツールを目指しており、ユーザが英語の論理展開の骨組みと文例を参照・利用することによって所望のアブストラクトを作り上げていくものである。本稿では、この英文アブストラクト作成支援ツールを支える重要な言語資源として開発した「論理展開の情報を付与した英日対訳コーパス」について説明する。

### 2. 論理展開情報つき英日対訳コーパスの作成

#### 2. 1 アブストラクトの英日対訳データの収集

リコーのホームページでは、『英語技術文献の日本語要約』を提供しており、米国の学会 IEEE (Institute

of Electrical and Electronics Engineers) 発行の論文誌に収録されている英文アブストラクトの日本語訳を同学会の許可を得て作成し、公開している。我々は、本コーパスの作成を開始した 1998 年 7 月の時点でホームページ上に公開されていたデータの中から、以下の論文誌を対象としてコーパスを開発することにした。

#### (1) *IEEE MultiMedia* (季刊誌)

対象範囲 : Spring, 1995 ~ Fall, 1996

英文アブストラクト数 : 36

#### (2) *IEEE Transactions on Pattern Analysis and Machine Intelligence* (月刊誌)

対象範囲 : July, 1995 ~ July, 1997

英文アブストラクト数 : 285

ホームページ上では日本語訳だけがまとめて提供されているので、我々は、原文である英文アブストラクトを IEEE のホームページから入手して、日本語訳との対応づけを行なった。さらに、各アブストラクトにおいて、英文と日本語訳文との文単位での対応づけを人手で行なった。この結果、英文約 2000 文に対して日本語訳文との対応づけがなされた、321 例のアブストラクト英日対訳用例を得た。

#### 2. 2 論理展開情報の付与

英文アブストラクトはひとつのパラグラフで書かれることが多いが、パラグラフの基本的な考え方は単一概念が論理的構成をとって過不足なく述べられることにある[2][3]。つまり、ひとつのパラグラフにはパラグラフの中心概念を要約した主題文 (トピックセンテンス) があり、パラグラフ内の他の文はこの主題文と論理的な関連づけがなされていなければならない。

我々は、収集したアブストラクト英日対訳用例に対して、主題文のパラグラフ内での生起位置（論理展開タイプ）と主題文に対して他の文が論理展開上どのような役割を果たしているのか（文役割）に着目して、以下のような情報を付与した。

#### <論理展開タイプ情報>

パラグラフ内の主題文（トピックセンテンス）の位置に応じて、各アブストラクトを頭括型、中括型、尾括型の3種類に分類した。但し、2つ以上のパラグラフで構成されているアブストラクトには一律にマルチパラグラフ型を付与した。

- ・頭括型 --- 主題文で始まる
- ・中括型 --- 主題文の前に導入文がある
- ・尾括型 --- 主題文で終わる
- ・マルチパラグラフ型 --- 複数のパラグラフで構成されている

#### <文役割情報>

アブストラクトを構成している各英文とその対訳文に主題文との論理的関係を表す情報を付与した。論理的関係は、以下のように分類した。各アブストラクトは、主題文を中心として以下の役割を担う複数の文がまとまりをなすように配置されている。

- ・背景（導入）文  
研究の動機や背景、従来手法の説明や問題点などを記述した文。
- ・主題文（トピックセンテンス）  
論文の主題や目的を記述した文。
- ・説明文  
主題を具体的に展開した文で、例えば新しい手法の概要を説明した文。
- ・実証文  
有効性検証などのために行なわれる実験の手順や実験結果を記述した文。
- ・補足／追加工  
主題に対して補足的な情報を提供する文。
- ・結論文  
実験結果から導出される結論を記述した文。

・縮文

今後の展開／応用可能性を記述した文。

## 2. 3 その他の情報付与

前節で述べた論理展開情報に加えて、我々は以下の情報をコーパスに付与した。

### (1) 書誌的信息

各アブストラクトの出典・タイトル・著者・キーワード情報。

### (2) アブストラクトタイプ

アブストラクトの作成目的による分類（以下の5タイプに分類）。

- ・新規モデル／アルゴリズムの提案
- ・既存技術／システムのサーベイ報告
- ・既存技術の発展または有効性の検証
- ・論評
- ・既存の技術／システム／理論等の解説

### (3) 文構造パターン情報

コーパス中の英文の各動詞にニューヨーク大学で開発された英語統語辞書『Comlex Syntax』[4]に準拠した文構造パターン情報を付与（図1参照）。

Recent approaches have [aimed@@NP1 aim to \_INF-SC@@] to [resolve@@NP1 resolve NP2 @@] these problems but can [compromise@@NP1 compromise NP2@@] other performance aspects.

図1 文構造パターン情報の付与例

各動詞に文構造パターン情報を付与したのは、文単位での英文作成を行なう際に、ユーザが動詞を指定するとその動詞が取りうる文構造パターンの候補を検索することができ、さらに所望の文構造パターンを選択すれば当該の文構造パターンを有する文例を参照することができるようにするためである[6]。

## 3. 「英文アブストラクト支援ツール」における論理展開情報の利用

我々が目指す「英文アブストラクト支援ツール」

の処理系の開発はまだ着手したばかりであるが、コーパスに付与した論理展開情報が具体的にどのように活用されるのかを示すために、暫定的に開発した検索・表示機能を紹介する。図2と図3は、それぞれ、論理展開タイプ指定によるアブストラクト用例検索結果と文役割指定による用例文検索結果を示している。

ユーザは、これから作成しようとしているアブストラクトと同じタイプに属するアブストラクト用例で、しかも所望の論理展開構造を持つ用例を参照するために、画面に表示されたリストの中から所望のタイプを指定する。図2の用例は、科学技術分野の論理展開タイプとして最も代表的、かつ効果的と言われる「頭括型」を選択した場合の表示結果である。この場合、「頭

括型」の用例は12例あり、スペースキーを押下することにより次々と用例を表示させ、自分が「借りる」のに最適な用例を探すことができる。次に、所望の論理展開構造を持つ用例の骨組みを借りて個々の文を作成する。この段階では、文役割に応じた用例文を参照する必要があり、図3の用例は主題文としての役割を持つ用例文を検索・表示させたものである。この場合も参照可能な用例文は25文あるので次々と用例文を表示させて「借文」できるものを見つけ出すことができる。また、この文レベルの作成をさらに支援するために、コーパス中の英文に付与した動詞の文構造パターン情報による用例文検索機能を開発中である。

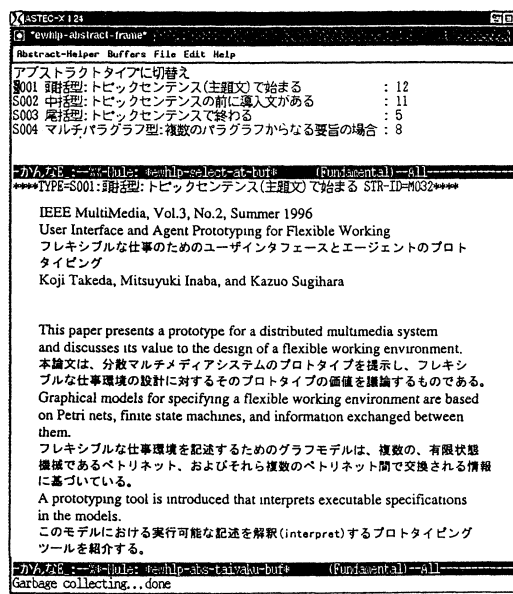


図2 論理展開タイプ指定によるアブストラクト用例検索結果

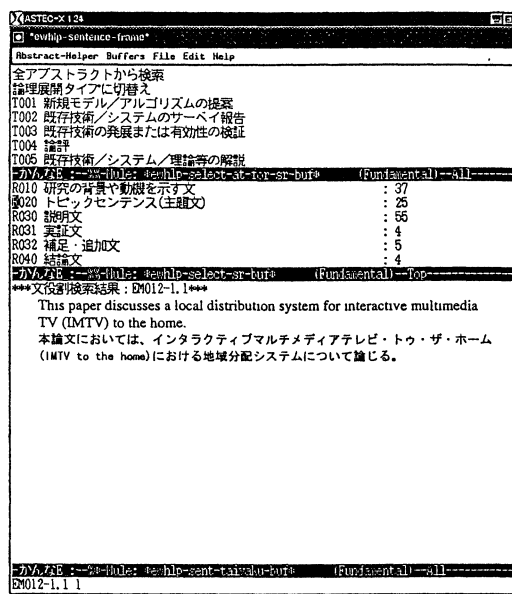


図3 文役割指定による用例文検索結果

#### 4. おわりに

本稿では、科学技術分野の英文アブストラクトの作成を支援するツールの構築に向けて我々が開発した英日対訳コーパスの概要を述べた。論理展開情報の分類には細かい点で問題があるが、我々の支援ツールのプロトタイプが出来上がった時点でユーザの声を聞きながら分類体系を再検討する予定である。また、英文作成においてより多くの支援を必要とするユーザのために、開発したコーパスに基づいて論理展開タイプ別に英文アブストラクトのテンプレートを用意する計画である。

さらに、コーパスの収録分野を拡大するために、IEEE の 2 つの論文誌に加えて計算言語学会 (Association for Computational Linguistics) 発行の論文誌 *Computational Linguistics* (1994 年～1997 年) と年次大会予稿集 *Proceedings of Annual Meeting of the ACL* (1990 年～1997 年) のアブストラクトについても同学会の許可を得て既に言語データの開発を進めている。

#### 謝辞

本研究は、郵政省平成 10 年度情報通信ブレークスルー基礎研究 21 の助成を受けている。本研究の実施にあたりご指導をいただいた郵政省通信総合研究所関西先端研究センターの井佐原均氏に感謝の意を表する。

#### 参考文献

- [1] 木下是雄. 『理科系の作文技術』中央公論社. 1981.
- [2] 篠田義明. 『テクニカル・イングリッシュ論理と展開』南雲堂. 1981.
- [3] 天満美智子. 『新しい英文作成法』岩波書店. 1998.
- [4] Grishman, R., Macleod, C. and Mayers, A. Complex Syntax: Building a Computational Lexicon. In *Proc. of COLING-94*, pp.268-272, 1994.
- [5] Leggett, A. J. Notes on the Writing of Scientific English for Japanese Physicists. 『日本物理学会誌』, pp. 790-805, 1966.
- [6] Narita, M. Language Resources for Writer's Helper. In *Proc. of First International Conference on Language Resources and*

*Evaluation*, pp.269-273, 1998.

- [7] Sageev, P. *Helping Researchers Write ... So Managers Can Understand*. Battelle Press. 1994.