平成１５年度　研究開発成果報告書

「多言語標準文書処理システムの研究開発」

目　　次

参考資料、参考文献

（添付資料）

1　研究発表、講演、文献等一覧

１　研究開発課題の背景

ブロードバンドの普及、国際社会のグローバル化により、国際標準の文書や全世界で使われる機器のマニュアル、特許等を多言語へ翻訳するという必要性は増える一方である。このような文書は改版が付きまとい、その度に翻訳需要が発生するため、その翻訳作業は膨大になる。

機械翻訳システムが商用化されて久しいものの、多言語翻訳はもちろん、英日・日英においてもこれらの文書は通常、専門用語が多く表現も複雑で、複雑な表現を対処する文法が存在しない、専門用語が未登録などの理由により、機械翻訳することができない。その一方で、現在、翻訳文書の電子化やその公開が急速に進んでおり、翻訳者の仕事の形態が急変している。翻訳者は、過去に翻訳した結果や専門用語の対訳辞書をデータベース（トランスレーションメモリと呼ばれる）に蓄積しておき、そのデータベースを参照することにより、翻訳するという形態をとることにより、翻訳作業の効率化を図っている。さらに、最近ではインターネット上には多くの翻訳ボランティアが存在し、彼らは自国の技術水準を高めるために又は自国内での情報共有のために、Web 上の技術サイトを分担して自国語に翻訳する作業をおこなっている。

翻訳者の仕事の変化にみるように、機械翻訳においても過去の翻訳結果を利用して翻訳したり、翻訳結果から辞書を自動的に学習させたりすることができれば、機械翻訳が翻訳業務や多言語文書作成のシーンでも利用可能となるに違いない。また、インターネット上の翻訳ボランティアにおける協調作業にみるように、技術者や翻訳者などの多くの人間が協調して翻訳できるような翻訳支援環境が存在すれば翻訳作業は加速されるに違いない。

多種多様な分野で、多言語間にまたがった対訳文書は増大する一方である。そこで我々は、様々な知識を有する人々が既存の翻訳結果を利用して、協調的に翻訳作業を行なう多言語標準文書処理システムを提唱した。

２　研究開発分野の現状

グローバル化が進む中で、製品のマニュアル、特許文書等の翻訳に対する需要は非常に多い。また、多言語、特に中国語をはじめとするアジア言語への関心は、依然として高い。こうした状況の中で、多岐にわたる分野の文書に対して高品質な翻訳を得るための手段の模索が続いている。

本章では、我々と関連の深い研究領域の動向として、株式会社富士通研究所から発表された統合翻訳プラットフォーム Cliché と情報アクセス技術のための評価型ワークショップ NTCIR を紹介する。

(1) 「統合型翻訳支援システム Cliché」

Cliché は従来の機械翻訳システムと翻訳メモリシステム[1]を統合し、翻訳作業の効率化を目指した翻訳プラットフォームで、以下の特長を有する[文献 1]。

①機械翻訳技術と訳例検索(翻訳メモリ)技術の統合
②訳例検索機能の向上
③ネットワークによる訳例データベース、機械翻訳辞書の共有
④最新のアプリケーションの随時ダウンロード
⑤ユーザビリティテストによる GUI 設計
⑥産業翻訳の現場を想定した翻訳作業の効率化

Cliché では、これらの技術を導入することにより、人手翻訳の３～４倍の効率化を達成している。Cliché の設計思想は、翻訳者の振舞いを徹底的に分析し、作業効率をよくすることで翻訳の品質向上および翻訳時間の短縮を図るものであり、我々の提唱する多言語標準文書処理システムと共通する部分が多い。機械翻訳システムを真に使いやすいものにするためには、こうした観点からのアプローチが必須であろう。

(2) NTCIR(NII-NACSIS Test Collection for IR Systems)のワークショップ

NTCIR ワークショップは、情報検索、テキスト要約、情報抽出、質問応答、テキストマイニングなど、「情報アクセス」技術の研究を発展させることを目的とした評価ワークショップで 1999 年より始まった[文献 2]。評価タスクは言語横断検索や質問応答、テキスト要約などで、第 1 回ワークショップでは、6 カ国 28 研究グループがタスクを遂行したが、参加者は年々増加し、第 3 回ワークショップでは、9 つの国と地域から 65 グループが結果を提出している。NTCIR プロジェクトの特徴として、中国語、韓国語といったアジア言語文書の検索に重点を置いている点が挙げられる。参加者は提供されるテストデータを利用したり、研究上のアイディアや技術の交換・移転のための研究者のフォーラムに参加したりすることができるので、NTCIR に参加することにより研究開発を加速することができる。我々も本研究の成果を用いて、第 4 回ワークショップに参加した。

[1]機械翻訳システムは辞書や翻訳規則を用いて原文を対象言語の文に変換するシステムであり、翻訳メモリシステムはすでに翻訳した文書を蓄積して次回以降の翻訳に利用するシステムである。

3　研究開発の全体計画

3-1 研究開発課題の概要

数多くの人間が、現存する大量の国際標準の文書や特許等の翻訳文書を利用して、ネット上で協調的に翻訳作業を行なうことができる多言語標準文書処理システムを研究開発する。多言語標準文書処理システムの中核をなす技術は、既存の対訳文書や翻訳の用例を与えることによって、翻訳テンプレートを自動的に抽出する技術である。本技術を実現するための手法として、我々は、(1)構造照合技術を利用する手法、(2)統計的学習を利用する手法、の２つの方法について研究開発を行なう。

さらに、翻訳プロセスのシステム化という観点から、獲得した翻訳テンプレートを利用して翻訳する言語非依存型翻訳エンジンの技術、および、獲得した翻訳テンプレートを専門性や汎用性の高低によって、自動分類・自動階層化（以降、自己組織化と呼ぶ）する技術についても研究開発を行い、トータルな翻訳支援環境構築を目指す。多言語標準文書処理システムのシステム構成図及び本システムの利用の形態を図3-1に示す。

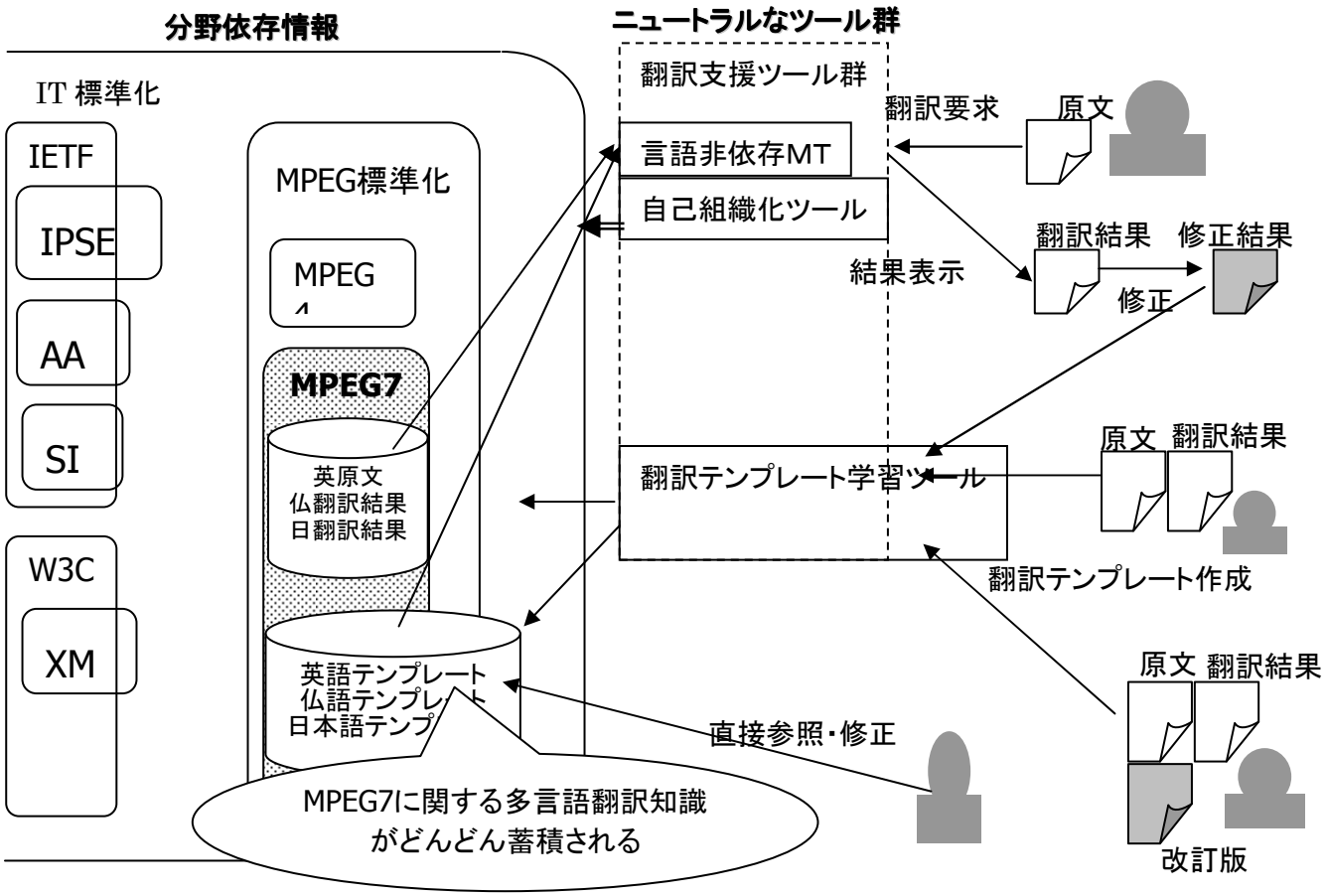


図3-1　多言語標準文書処理システムの構成図及びユーザによる利用形態

3-2　研究開発目標

3-2-1 最終目標

多言語標準文書処理システムの研究開発

(1) インターネット上のどこからも本システムが利用可能であること。
(2) 国際標準等、5分野以上の対訳文書DB､翻訳テンプレートDBを構築していること。
(3) 対訳文書DB、翻訳テンプレートDBを備えており、直接参照したり、修正したりすることができること。
(4) 以下の翻訳プロセスを実現するシステムであること。
  a. ユーザがインターネットを通じて原文を与えると日本語の翻訳結果が出力される。
  b. その翻訳結果に満足すれば対訳文書DBにその対訳文を格納する。満足しなければユーザが翻訳結果を修正する。修正した結果を対訳文書DBに格納し、修正した部分に関する翻訳テンプレートを自動的に作成し、翻訳テンプレートDBに格納する。
  c. 以降の翻訳では、1，2で格納された対訳文書DBと翻訳テンプレートを利用した翻訳結果となり、同じ翻訳間違いは２度としない。

ア．対訳文書及び改版の差分や後編集知識を利用した翻訳テンプレート作成に関する研究開発

(1) 対訳文書(英語以外の2つ以上の言語と日本語の対訳)を与えることにより、翻訳テンプレートを作成する。作成された翻訳テンプレートは簡単に修正でき、翻訳テンプレートDBに格納される。本ツールにより、翻訳テンプレート作成作業工数が50%以上削減されること。
(2) 構造照合利用型と統計的手法利用型の両方の技術を用いて翻訳テンプレートを作成できること。
(3) 文対応がついていない対訳文書についても専門用語の翻訳テンプレートDBが精度80%で抽出できること。

イ．多種多様な分野辞書の自己組織化に関する研究開発

(1) 5分野以上の翻訳テンプレートDBにおいて、自己組織化が行われること。自己組織化後は、翻訳結果の精度が向上すること。

ウ．言語非依存の翻訳エンジンの研究開発

(1) 多言語標準文書処理システムの研究開発の(4)において、英語以外の 2 言語以上を原文としても同様の翻訳プロセスが実現できること。
(2) 英語以外の 2 言語以上の翻訳文書 DB、翻訳テンプレート DB が存在すること。

3-2-2 中間目標

多言語標準文書処理システムの研究開発

(1) 多言語標準文書処理システムにおいて、翻訳エンジン部、改版文書を利用した翻訳テンプレート作成部、対訳文書 DB、翻訳テンプレート DB の試作システムが完成していること。
(2) 翻訳実験、翻訳テンプレート作成・DB 格納実験ができること。
(3) 国際標準等、2 分野の対訳文書 DB、翻訳テンプレート DB を構築していること。

ア. 対訳文書及び改版の差分や後編集知識を利用した翻訳テンプレート学習に関する研究開発

(1) 既存の対訳文書とその改版文書を与えることにより、改版文書に関する翻訳テンプレートを獲得できること。
(2) 構造照合技術を利用して、対訳の対応付けが精度 80%以上で実現されていること。
(3) 統計的手法を用いた翻訳テンプレートの汎化技術に関する手法を確立していること。
(4) 文対応がついていない対訳文書についても専門用語の対応付けが精度 80%以上で実現されていること。

イ. 多種多様な分野辞書の自己組織化に関する研究開発

予め人間の手で人によって分類・階層化されている翻訳テンプレート DB に対し、新しく獲得した翻訳テンプレートを最適な分類・階層の DB に格納できる技術が精度 80%で実現されていること。（精度の判定はここでは人手による客観評価とする。）

ウ. 言語非依存の翻訳エンジンの研究開発

言語に依存する部分は全て抽象化した翻訳エンジンの実装が終了していること。

3.3 研究開発の年度別計画

（金額は非公表）

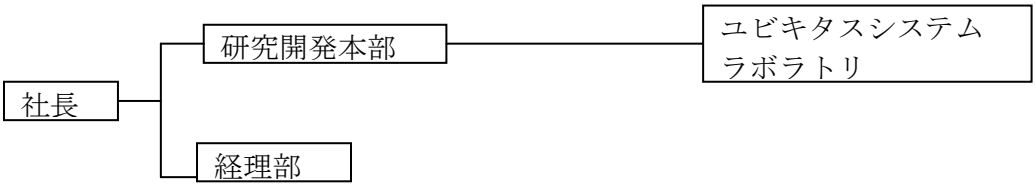
研究開発項目	１４年度	１５年度	１６年度	１７年度	年度	計	備　考
多言語標準文書処理システムの研究開発 ア．翻訳テンプレート自動学習の研究開発 ・構造照合型テンプレート自動学習システムの開発 ・統計的手法型テンプレート自動学習システムの開発 イ．翻訳テンプレートの自己組織化の研究開発 ・分類されたものへの選択手法の開発 ・自己組織化システムの開発 ウ．言語非依存型機械翻訳システムの研究開発 ・翻訳エンジンの開発 ・翻訳知識 DB の開発	→	→	→	→			
間接経費							
合　　計							

注）１　経費は研究開発項目毎に消費税を含めた額で計上。また、間接経費は直接経費の３０％を上限として計上（消費税を含む。）。
　　２　備考欄に再委託先機関名を記載。
　　３　年度の欄は研究開発期間の当初年度から記載。前年度（１４年度）までは、合計が当該年度の契約額の実績値となるよう記載。

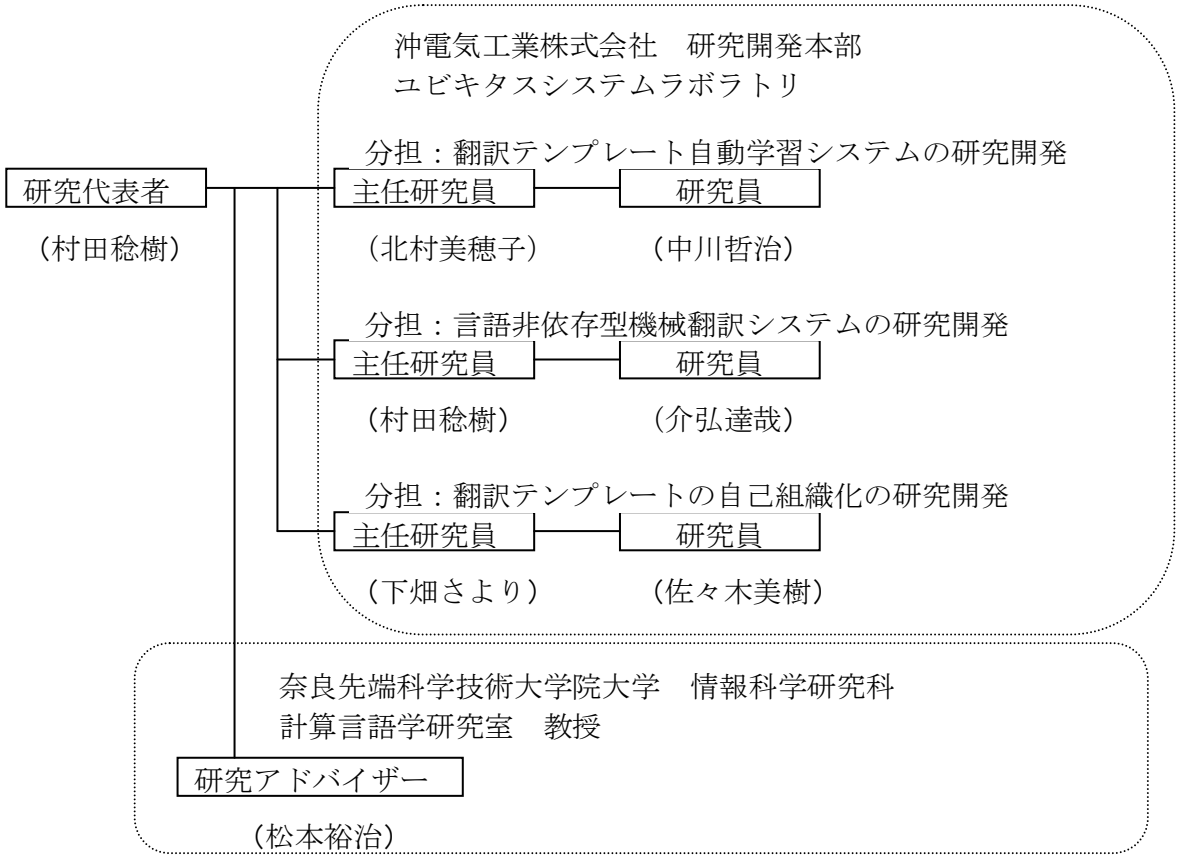
3-4 研究開発体制

3-4-1　研究開発管理体制

（注　受託者の経理部門の体制、経理責任者（所属、氏名、電話、ＦＡＸ、Ｅ－メールの連絡先）を含む。）



3-4-2　研究開発実施体制



４　研究開発の概要（平成 15 年度まで）

4-1　研究開発実施計画

4-1-1 研究開発の計画内容

多言語標準文書処理システムは、大きく、次の３つの技術から成り立っており、各技術を実現するために、次のア、イ、ウの研究開発が必要となる。

・既存の対訳文書や翻訳の用例を与えることによって、翻訳テンプレートを自動的に抽出する技術
⇒**ア．**対訳文書及び改版の差分や後編集知識を利用した翻訳テンプレート学習に関する研究開発

・獲得した翻訳テンプレートを専門性や汎用性の高低によって、自動分類・自動階層化（以降、自己組織化と呼ぶ）する技術
⇒**イ．**多種多様な分野辞書の自己組織化に関する研究開発

・翻訳プロセスのシステム化という観点から、獲得した翻訳テンプレートを利用して翻訳する言語非依存型翻訳エンジンの技術
⇒**ウ．**言語非依存の翻訳エンジンの研究開発

上記のサブテーマに対して、本年度の研究目標及び研究開発内容を記す。

ア．対訳文書及び改版の差分や後編集知識を利用した翻訳テンプレート学習に関する研究開発
(1) 既存の対訳文書とその改版文書を与えることにより、改版文書に関する翻訳テンプレートを獲得できること。
(2) 構造照合技術を利用して、対訳の対応付けが精度 80%以上で実現されていること。
(3) 統計的手法を用いた翻訳テンプレートの汎化技術に関する手法を確立していること。
(4) 文対応がついていない対訳文書についても専門用語の対応付けが精度 80%以上で実現されていること。

イ．多種多様な分野辞書の自己組織化に関する研究開発
(1) 予め人によって分類・階層化されている翻訳テンプレート DB に対し、新しく獲得した翻訳テンプレートを最適な分類･階層の DB に格納できる技術が精度 80%で実現されていること。（精度の判定はここでは人手による客観評価とする。）

ウ．言語非依存の翻訳エンジンの研究開発
(1) 言語に依存する部分は全て抽象化した翻訳エンジンの実装が終了していること。
(2) 国際標準等、2 分野の対訳文書 DB、翻訳テンプレート DB を構築していること。
(3) 多言語標準文書処理システムにおいて、翻訳エンジン部、改版文書を利用した

翻訳テンプレート作成部、対訳文書 DB、翻訳テンプレート DB の試作システムが完成していること。

4-1-2　研究開発課題実施計画

（金額は非公表）

研究開発項目	第１四半期	第２四半期	第３四半期	第４四半期	計	備　考
多言語標準文書処理システムの研究開発 ア．翻訳テンプレート自動学習の研究開発 ・構造照合型テンプレート自動学習システムの開発 ・統計的手法型テンプレート自動学習システムの開発 イ．翻訳テンプレートの自己組織化の研究開発 ・分類されたものへの選択手法の開発 ・自己組織化システムの開発 ウ．言語非依存型機械翻訳システムの研究開発 ・翻訳エンジンの開発 ・翻訳知識 DB の開発	→	→	→	→		
間接経費						
合　　計						

注）１　経費は研究開発項目毎に消費税を含めた額で計上。また、間接経費は直接経費の３０％を上限として計上（消費税を含む。）。
（合計の計は、「３－１の研究開発課題必要概算経費」の総額と一致）
２　備考欄に再委託先機関名を記載。

4-2　研究開発の実施内容

以下のア、イ、ウの研究開発を行い、目標を達成した。

ア．対訳文書及び改版の差分や後編集知識を利用した翻訳テンプレート学習に関する研究開発を行なった。

(1) 既存の対訳文書とその改版文書を与えることにより、既存文書と改版文書の差分を検知し、未変化部分と変化部分を出力する「既存文書と改版文書の差分検知システム」の実装を完了した。

(2) 統計的手法と構造照合の融合技術を考案した。システムの試作を行なうことにより、動画関連の国際標準文書において、1000 文中 992 文の構造照合結果を抽出することができ、人手による評価の結果、対訳の対応付けが 83%の精度で対応付けられることが確認できた。

(3) 統計的手法を用いた翻訳テンプレートの汎化技術に関する手法を確立した。構造照合結果を複数の部分木に分解し、語彙の一部を変数化した部分木をテンプレート形式に自動的に変換するツールの基本設計を完了し、その有効性を机上シミュレーションによって確かめた。

(4) 文対応がついていない対訳文書についても専門用語の対応付けが上位 5 位において精度 82%以上で実現された。実験の結果、

① 上位 5 位以内に正解の対訳表現が含まれている割合は 82%

② 従来手法に比べて、約 1.2 倍多くの対訳表現を抽出することができること

を確かめた。

イ．多種多様な分野辞書の自己組織化に関する研究開発を行なった。

(1) 予め人によって分類・階層化されている翻訳テンプレート DB に対し、新しく獲得した翻訳テンプレートを最適な分類・階層の DB に格納できる技術を精度 80%以上で実現した。また、あらかじめ人手によって分類・階層化されている翻訳テンプレート DB に対して、新しく獲得した翻訳テンプレートを最適な分類・改装の DB に格納するための手法を考案した。階層化された分野への対応を行なうことができた。さらに、訳語の情報を使うことにより、複数の意味を持つ翻訳テンプレートも分野判定することができるようになった。実験を行なった結果、精度 80%以上の結果が得られた。

ウ．言語非依存の翻訳エンジンの研究開発を行なった。

(1) 「形態素解析モジュール」の言語非依存版を実装し、日本語及び中国語における実験を行なった。また、言語非依存翻訳エンジンの検証のために、予定より早く中日翻訳システムの開発に着手した。

(2) 対訳文書 DB については、ISO 標準文書、動画関連の国際標準文書、FCC(米国通信委員会)の規則文書、という 3 分野の英日の対訳文書 DB を構築した。

また、翻訳テンプレート DB については、動画関連の国際標準文書に関する翻訳テンプレートを人手で作成し、その DB の構築を完了した。

(3) 多言語標準文書処理システムにおいて、翻訳エンジン部、改版文書を利用した翻訳テンプレート作成部、対訳文書 DB、翻訳テンプレート DB の仕様検討が終了し、試作システムの実装に入った。

５　研究開発実施状況（平成 15 年度）

5-1　翻訳テンプレート学習に関する研究開発

5-1-1　序論

既存の対訳文書や翻訳の用例から、翻訳知識(翻訳テンプレート)を自動的に抽出する技術は、翻訳品質の向上、および、翻訳プロセスの効率化において必須であり、多言語標準文書処理システムの中核をなす技術である。

本サブテーマでは、前年度に引き続き、以下に示す３種類の文書を用いて翻訳知識を獲得する研究を行なっている。

ア．　改版文書：改版文書とは、原本の一部に修正、加筆を行なった文書のことをいう。
イ．　文対応付き対訳文書(パラレルコーパス)：パラレルコーパスとは、第 1 言語と第 2 言語の文に 1 対 1 で対応がついた対訳文書のことをいう。
ウ．　文対応なし対訳文書(コンパラブルコーパス)：コンパラブルコーパスとは、同一分野、あるいは同一内容といったレベルの対応関係のある 2 言語の文書のことをいう。

前年度までは、各文書から翻訳知識を獲得するシステムの基本設計及びシステムの試作を行い、各々の有効性を確かめたが、本年度は、より実システムへの適用を意識した研究開発を行なっている。

5-1-2　改版文書を利用した翻訳テンプレート獲得に関する研究開発

(1) 研究の内容

改版文書を利用した翻訳テンプレート獲得のための要素技術として、対訳文書の文の対応付け技術、改訂前後の原文書間の文の対応付け技術、変更や追加された文に対する訳文付与技術等がある。なかでも、改訂前後の原文書間の文を対応付ける技術は、その対応付けの精度が翻訳の精度に直接影響するため、最も重要な技術となる。本年度は、文の類似度だけでなく文脈を考慮するという点に着目した改訂前後の原文書間での文対応付けの手法について検討した。以下では、文脈情報を利用した文対応付け手法について、詳しく説明する。

原文書と改版文書の文対応付けを行なう際、文レベルの類似度のみで対応付ける方法では文脈が考慮されず正しい対応が得られない。そこで、我々は、文書を章、段落といったブロックに区切り、まずブロック単位での対応関係を確定してから、この対応付けたブロック内で文を対応付ける手法を提案する。

・文対応付けの課題

改訂前の原文と改訂後の文書の文を対応付ける最も簡単な方法として改訂前の原文書の対訳をそれぞれ一文毎に区切り、各文を翻訳済みデータベースに格納しておき、改訂後の原文書の各文で、翻訳済みデータベースから最も類似する文を検索するという方法が挙げられる。しかし、この方法では図 5-1-1 に示すように、入力文①に最も類似する原文は③であるか⑤であるかの判断が難しい。但し、翻訳済みデータベースが

改訂前の文書中での順に文が格納されているならば、①②と③④、及び①②と⑤⑥を比較することができる。このように、その前後の文も含めて類似性を考慮することで、③ではなく⑤に対応付けることができる。

さらに、単純な類似文検索において、完全一致する文を対応付けることは容易であるが、完全一致する文の訳文も同一である保証はない。文の前後のつながりによって、訳や表現方法を変えたりすることは翻訳では常套手段である。したがって、単純な類似文検索だけでなく、改訂前後の文書内の文脈を考慮して文の対応付けをする必要がある。

また、訳文は直訳とは限らず、２文が１文に翻訳されていたり、章や段落でひとまとまりの意味を持ったりすることも多い。このような複数文は、別々に対応付けられるのではなく、できるだけまとまって対応付けられていると、その対応度の信頼性が増し、ユーザが翻訳結果を再チェックする場合もあまり時間をかけなくて済む。
このようなことを鑑みて、我々はその前後の文脈にどれだけ一致しているか等の情報をみながら、文の対応付けを判断することにした。

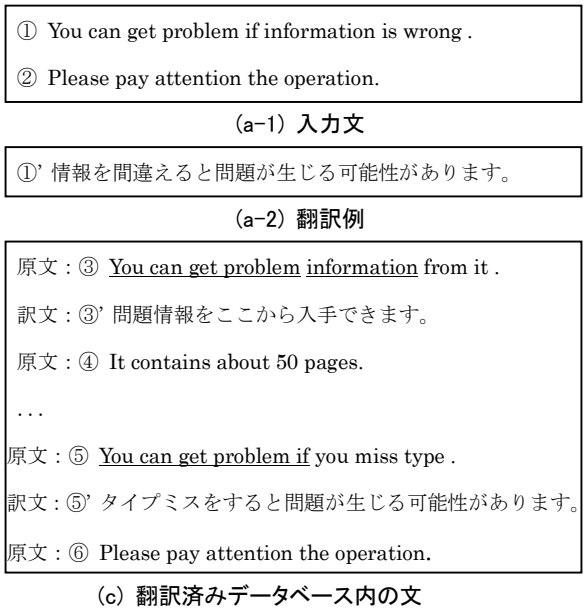


図 5-1-1　対応付け例

・対応付けアルゴリズム

対応付けアルゴリズムの流れを図 5-1-2 に示す。対応付けを行なう改訂前の原文書及びその改訂文書をそれぞれ、章、節、段落といった単位のブロックに区切り階層化する。対応付けはまずブロック同士で行い、文脈が考慮されるように、章から段落レベルと階層的に範囲を狭めていく。この際、連続性の考慮、出現位置の移動も考慮する。最も小さなブロック単位（段落レベル）の対応付けの終了後、そのブロック内での文単位の対応付けを行なう。以下に、各レベルでの処理について詳しく説明する。

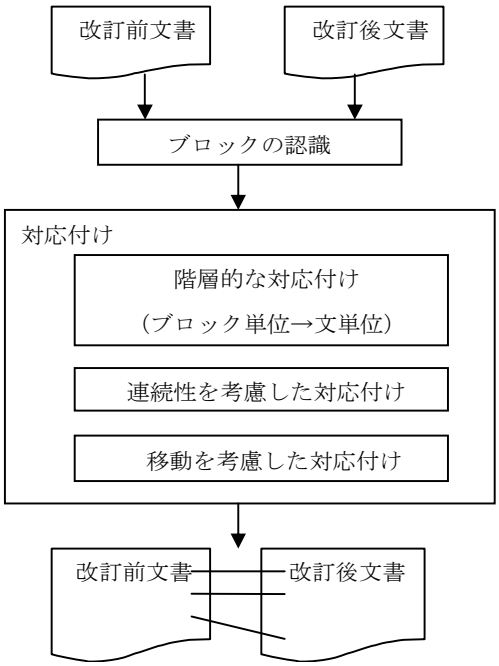


図 5-1-2　文対応付けアルゴリズム

・ブロックの認識

ここでは、ブロックの分割、階層化、及び文区切りを行なう。まず、章番号や字下げ等を基にして章や段落単位でブロックの分割を行い、次に、章とその章中の節、段落というように文書の階層構造を解析し、ブロックを章等の広範囲の第１階層から段落等の狭範囲の最下位階層まで階層化する。さらに文区切りも行なう。

・階層的な対応付け

図 5-1-3 を用いて、階層的な対応付けの方法を説明する。本手法では、まず第１階層のブロックについて改訂前後の文書の全ブロック間(図 5-1-3a)で類似度を計算する。類似度は式(1)で求める。類似度の計算単位は、文字や語単位でもできるが、計算量を抑えるために計算単位は文とした。ただし、数値、記号は改版されると変更されやすい。また、文中に段落番号を含んでいると改版時にずれ文単位で比較していると対応しないことが起こり得る。よって、数値、記号のみの語は除外して比較するようにしている。

$$\text{類似度}(\%) = \frac{2 \times NUMmatch}{NUMold + NUMnew} \times 100 \qquad (1)$$

NUMmatch: 完全一致する文数
NUMold:　 改訂前文書の１ブロック内の文数
NUMnew:　 改訂後文書の１ブロック内の文数

全ブロック間の類似度を計算した後、あらかじめ定めていた閾値以上の類似度を持つブロックのペアの対応を決定する。次に、対応決定したブロックの１ペアを取り出し、その中で次の階層のブロック間(図 5-1-3b)の対応付けを行なう。これを繰り返し最下位階層のブロックの対応付けまで行なう。ブロックの対応付けが完了したならば、次に対応付けしたブロックのペア内で文の対応付けを行なう。

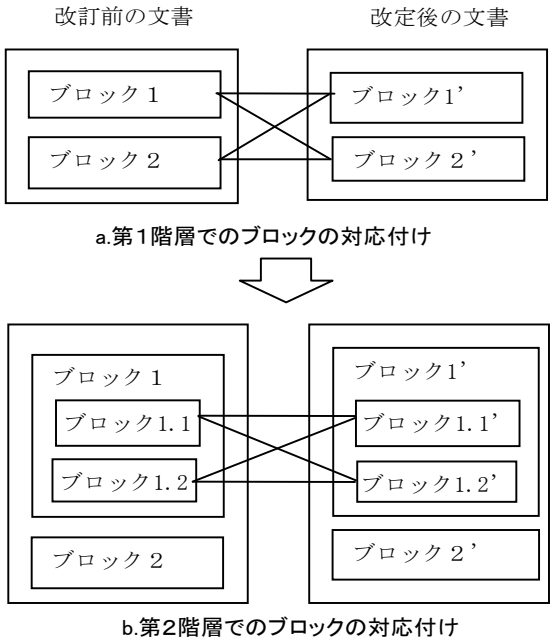


図 5-1-3　階層的な対応付け

・連続性を考慮した対応付け

類似度の計算は一致度だけでなく、対応関係の連続性によって類似度に重みを付ける。対応が決定しているペアの前後は対応付けられる可能性が高いと考え、対応を決定したペアの前同士、後ろ同士は対応決定が優先されるようにする。対応ペアを決定したら、そのペアの前同士または後ろ同士のペアは式(2)のように重み付けした類似度とする。そして、最も高い類似度のペアを選択するという処理を、類似度が閾値以上の間繰り返し対応を決定する。

$$\text{類似度}(\%) = w \times \frac{2 \times NUMmatch}{NUMold + NUMnew} \times 100 \qquad (2)$$

w:重み(>1)

例えば、図 5-1-4 では、まず類似度の高い順に 1 と 1’(①)、3 と 3’(②)のペアが対応付けられる。2 と 2’のペアについては、この前同士のペアが既に対応決定済みなので類似度に重み付けされており、一致度だけで見ると 50%で 4 と 2’のペアと同一であるが、２と 2’(③)のペアの方が対応付けられる。

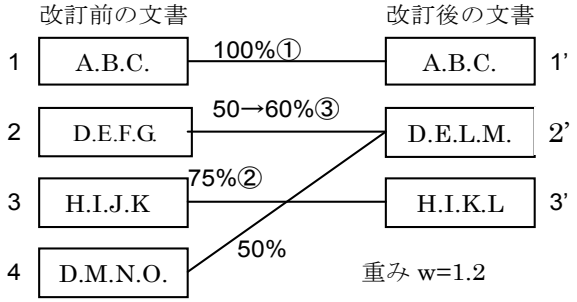


図 5-1-4　対応の決定順序

・移動を考慮した対応付け

上述の処理では、改訂前の文書から未変更、及び一部変更された場合、及び同一ブロック内での移動については対応付けができるが、別のブロックの下に移動した場合については対応付けできない。そこで、そのような場合でも対応付けができるように以下の処理を追加する。各階層での対応付け後、対応付けできなかったブロック全て類似度を再計算し対応付けする。この処理を行なっても対応付かなかったブロックは、削除または追加されたブロックとみなす。１ブロックが複数ブロックに分割された場合、または複数ブロックが１ブロックに結合された場合については、分割後・結合前のレベルの単位のブロックでの対応付けで対処できる。

・文単位の対応付け

文単位の対応付けは形態素の一致度をもとにした文類似度を利用する。まず、対応付けられた最下位階層のブロックの１ペアを取り出し、各文を形態素に分解する。そして、連続性を考慮した対応付けでの類似度の計算と同様にして全ての文間の類似度を計算し対応付けていく。

$$\text{文類似度}(\%) = w \times \frac{2 \times NUMmatch}{NUMold + NUMnew} \times 100 \quad (3)$$

w:重み
NUMmatch: 一致する形態素数
NUMold:　 改訂前文書の１文内の形態素数
NUMnew:　 改訂後文書の１文内の形態素数

・ブロック対応付け精度の評価

本対応付け方式の有効性を評価するために改訂前後の文書のブロック対応付け、文対応付けの実験を行なった。
実験は、以下の２方式について再現率、適合率を求めて比較した。

1. 本提案方式による文の対応付け手法。
2. 改訂前、改訂後の文書をそれぞれ文に区切り、改訂後の文書の各文について、改訂前の文書の全文の中で単語の一致度による類似度の最も高いものを選択していくことで文を対応付ける従来手法。

また、再現率、適合率は以下の式により定義した。

$$\text{適合率}(\%) = \frac{NUMa}{NUMb} \times 100 \quad (4)$$

$$\text{再現率}(\%) = \frac{NUMa}{NUMc} \times 100 \quad (5)$$

NUMa: A)またはB)の方式で対応付けたペアの中で正解した数
NUMb: A)またはB)の方式で対応付けたペアの総数
NUMc　人手により対応付けたペアの総数

適合率は対応付けされた文のペアの正解率、再現率は対応付けしなければならないペアのうち、どの程度対応付けられたかの割合である。さらに、改訂後に新規追加された文に対し、間違って対応付けてしまった割合を調べた。

評価用データには ITU/ISO の技術標準化文書の英文草案及びその改訂版[4]を利用した。類似度は式 (3) の重み w で変わるが今回は w=1.2 とした。また、対応決定するための類似度の閾値によっても結果が変わるが、ブロック単位の対応付けについては５%とし、文単位の対応付けの実験については 40%から 100%まで値を変化させ、それぞれの適合率、再現率を調べた。

表 5-1-1 は最下位階層のブロックの対応付け結果である。(a)は人が対応している、または追加とみなしたブロック数、(b)は提案方式での判定数、(c)は、提案方式での判定と人が判定したものとが一致した数である。人が対応しているとみなしたもののうち 84%を対応付けし、そのうち 99%が正しかった。また、人手により追加と判定した 30 ブロックに対し 29 ブロックを正しく認識しできた。対応付けしたペアは信頼性が高いので、ユーザが複数文まとまった単位で結果をチェックする場合あまり時間をかけなくて済み、さらに文単位の対応付けする際も信頼性が期待できる。人手による判定では対応するペアがあるとしたブロックを、追加と判定してしまったブロックが 27 あった。どのような場合に判定を間違えているか調べてみると、１ブロック中に完全一致する文がなかったり、割合が低かったりした場合であった。人手で対応付けられたブロックでも実際には文の対応ペアが少なく訳文を再利用できない場合が多い。したがって、対応付けられなかった 27 ブロックの対応付けへの影響は少ないと考える。

表 5-1-1 最下位階層のブロック対応付け結果

	(a) 人手判定数	(b) 提案方式による判定数	(c) 正解数	(d) 再現率 b/a(%)	(e) 適合率 c/b(%)
対応	170	143	141	84	99
追加	30	57	29	53	51

・文対応付け精度の評価

表 5-1-2 は、提案方式(A)、従来方式(B)それぞれ対応決定の類似度の閾値を変化させた時の再現率、適合率を調べた結果である。閾値が 50%の時、従来方式では誤対応付けの割合が 10%を超えるが、提案方式では 3.4%と低かった。ブロックによる対応付けを行い文の検索範囲を絞ったことにより、文脈を無視した文は対応付けない効果が得られた。

表 5-1-3 により再現率についても､従来方式と同レベルに維持できることがわかる。ブロック内に完全一致する文が全くない箇所や、ブロック内の全センテンス数に対し類似する文の割合が低い箇所で再現性が低い場合にブロックの対応付けができず、このため文の対応付けにも影響が出る可能性があった。しかし、本提案方式での検索範囲を絞り込んでいることによる対応付けもれの影響はそれほど見られなかった。

表 5-1-4 は、人が追加と判定した文について、追加文と判定した割合を示す。提案方式では、閾値を下げても類似しているが文脈に合わない文を対応付けしてしまう数が少ないことがわかる。

以上のように、ブロックの対応付けをすることで２文間の類似度が 50%前後と低い場合でも適合率、再現率ともに 90%以上の精度を有する。したがって、以下に述べる改版文書翻訳システムにおいて本提案手法を採用した場合、高い訳文再利用率を維持

することができる。

表 5-1-2 文対応付け結果（適合率）

閾値(%)	A)提案方式(%)	B)従来方式(%)
40	93.6	84.0
50	96.6	89.3
60	96.7	92.8
70	97.0	95.2
80	97.4	96.5
90	97.9	97.2
100	98.2	97.0

表 5-1-3 文対応付け結果（再現率）

閾値(%)	A)提案方式(%)	B)従来方式(%)
40	92.4	91.2
50	92.3	90.9
60	91.3	88.7
70	86.7	84.2
80	79.4	76.7
90	64.6	61.5
100	43.2	41.8

表 5-1-4 追加文の判定結果（再現率）

閾値(%)	A)提案方式(%)	B)従来方式(%)
40	80.4	45.6
50	82.0	74.5
60	85.6	78.7
70	89.6	86.7
80	92.8	91.3
90	94.1	94.6
100	96.8	96.0

（２）研究の効果

実改版文書に本手法を適用した実験を行なった結果、高い精度で原文書と改版文書の文対応付けを行えることが確認できた。

この成果は、既存の対訳文書とその改版文書を与えることにより、既存文書と改版文書の差分を検知し、未変化部分と変化部分を出力する「既存文書と改版文書の差分検知システム」として実装を完了している(詳細は「5-3-3 多言語翻訳データベースの研究」を参照のこと)。また、このシステムにより検知された差分から翻訳テンプレートを作成するモジュールを試作中である。

5-1-3　構造照合による訳語対応付けの研究

機械翻訳システムの品質向上のためには、専門用語や新語の辞書開発が欠かせない。対訳辞書を自動的に作成する方法として、対訳文書から対訳表現を自動的に抽出する多くの試みが存在するが、その中でも文対応のついた対訳文書から統計的な手法により対訳表現を自動抽出する試み[文献 3][文献 4]は、精度が高く、有効な手法の一つである。しかし、従来技術において、実用に耐える精度とカバレッジ(抽出率)を保証するためには 1 万文以上の規模の大きい文書を必要する。我々は、対訳文書の句分割の情報や既存の対訳辞書を利用することによって、また、人間が補助的に介入することによって、比較的入手可能な千文程度の対訳文書からでも対訳表現を抽出する手法について試みた。以下に本手法及び実験結果について述べる。

・貪欲的統計手法による対訳表現の半自動抽出

貪欲的な統計手法で対訳表現を抽出する手法[文献 4]とは、文対応のついた対訳文書に出現する原言語と目的言語の任意長の単語列を対応付けて単語列組を生成し、統計的確信度の高い単語列組から順番に対訳表現として抽出していくという方法である。統計的確信度は、単語列組の出現回数が多いほど、高くなるように設定する。この手法では、確信度を順々に下げる際、抽出された対訳表現を次回の抽出候補から除きながら抽出を繰り返すため、出現回数が少ない場合でも、抽出されなかったもの同士を対応付けることにより、対訳表現を獲得することができる。

この特長を活かして、文書に含まれる各単語の出現回数が少ない千文程度の小さな対訳文書において本手法を試みた。表 5-1-5 の a.の従来方法が、その結果である。この結果にみるように、統計的確信度が 0.5 の場合の精度は極めて悪い。ここで抽出された対訳表現を分析した結果、

・対応関係はあるが、不要な語が混じっている。
・一目瞭然に対応関係がない。

という 2 つが不正解の主な要因であることが確認された。

そこで、これらの課題を解決するために、図 5-1-5 に示す抽出方法を提案した。まず、前者の問題は、処理の最初に生成する単語列組が構文的に正しい区切りになっていないことに起因すると考えられる。したがって、図 5-1-5 の 2)における単語列組の生成時に、句範囲を超える単語列組は生成しない、という手法を導入した。

一方、後者に関しては、図 5-1-5 の 5),7)で示すように、対訳表現の抽出時に既存の辞書を参照し、辞書中で対訳関係のないものの抽出を遅延させることを試みた。図 5-1-5 の 5)で、まず、辞書に登録されている単語列対を優先して対訳表現が抽出され、8)で、その抽出された単語列対は次回の抽出候補から除かれる。次回の 7)の抽出では、対応付けの候補が減っているので、辞書に登録されていない単語列対でも、正しく抽出できると考える。

さらに、より積極的な方法として、図 5-1-5 の 6)に見るように、対訳表現抽出時に人間が抽出可否のチェックを行い「否」と判断された場合、候補外対訳表現として、その対訳表現を抽出対象から外す手法を試みた。人間の補助的な作業により、精度が格段に上がるならば、それも有用な手法であると考えられるためである。

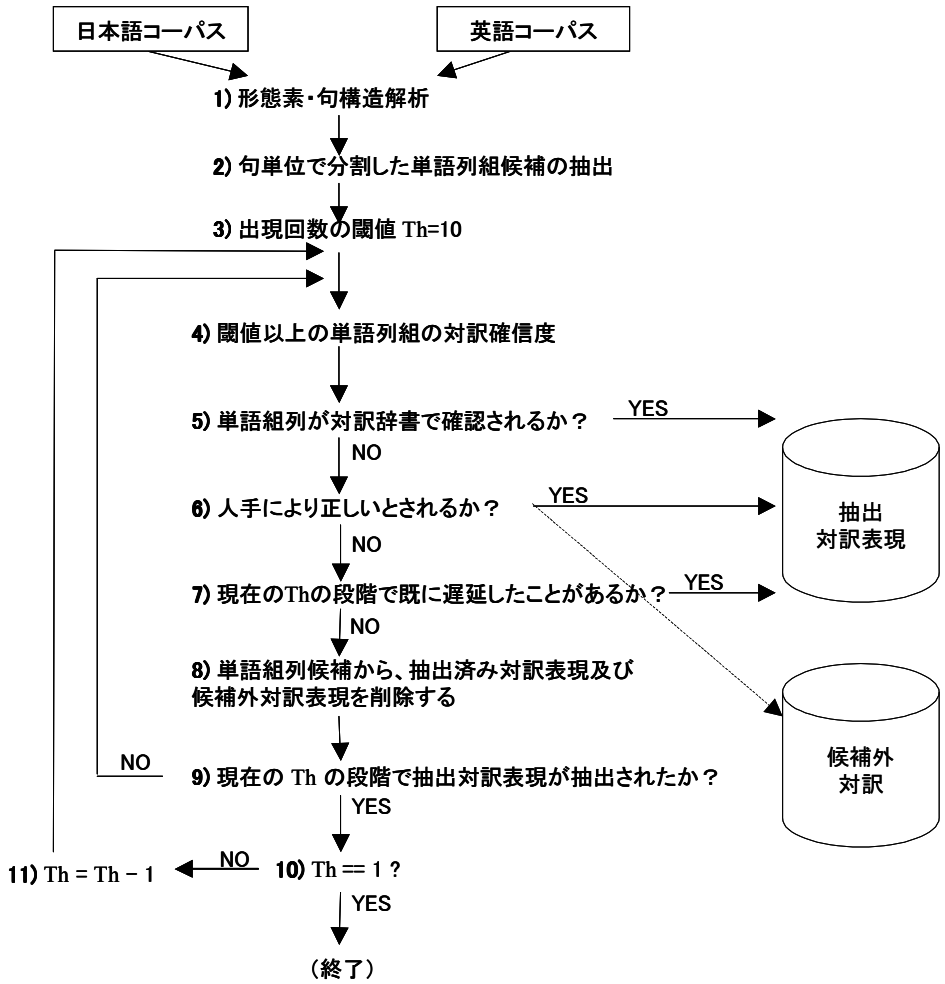


図 5-1-5 対訳表現の半自動抽出の流れ

・実験および評価

日英新聞記事対応付けデータ[文献5]から1,000文の対訳文を抽出し、a: 従来方法[文献4]、b: 従来方法の単語列抽出を句単位で区切った場合、c: bの手法に加え、辞書を参照した場合、d: cの手法に加え、人手による確認を行なった場合、の4つの手法について実験した。またベースラインとして、e: 既存の辞書を直接対応付けることによって抽出した場合のカバレッジについても調査した。

英語の形態素解析及び参照に用いる既存対訳辞書[2]は、自社製のものを用いた。英語の句単位分割は charniak[資料 6]パーザを用いた。日本語の形態素解析及び構文解析は ChaSen[資料 7]及び CaboCha[資料 8]を用いた。また、日本語単語列 w_J と、英語単語列 w_E の統計的確信度は、Dice 係数を応用した次の式を用いた。

$$sim(w_J, w_E) = (fje/2) \cdot (2fje / (fj + fe))$$

$$sim(w_J, w_E) \geqq Th/2$$

fje: w_J, w_Eの同時出現回数

fj: 日本語文書における w_Jの出現回数

[2]現在の総対訳表現数は、約 50 万語である。

fe: 英語文書における w_E の出現回数
Th: 出現回数の現段階の閾値

精度の評価方法は、人手により正解と思われる場合を 正解(○)、抽出した対訳表現の一方の訳がもう一方の表現の一部となっている場合(表 5-1-7“△”の例を参照)を半正解(△)、上記以外を不正解(×)として数え上げ、抽出総数に占める割合を各々求めた。[3]またカバレッジの計算は、以下の式により各言語の文書から求めた。

$$\text{coverage}_{(\%)} = \left(1 - \frac{\text{未抽出自立語数}^{4}}{\text{文書中の自立語の総数}}\right) \times 100$$

表 5-1-5 に、a～d における、統計的確信度(sim)別の精度及びカバレッジの集計結果を記す。上述した式の性質からわかるように sim>0.5 は、w_J,w_E の同時出現回数が 2 回以上の場合であり、sim=0.5 は同時出現回数が 1 回の場合である。精度に関してはレベル毎の値を示し、カバレッジに関しては、そのレベルまでの値を示す。なお、この表から、句単位の利用、辞書利用、人手利用を順に施すことによって、カバレッジを下げることなく、精度を向上させている、つまり、正解を減らすことなく不正解を除去できることがわかる。辞書のみを利用した場合では(e)、抽出できる対訳表現の数は本手法に比べて多いが、逆にカバレッジは低くなった。その理由は、本手法では複数の単語からなるイディオムや専門用語が多く抽出されるのに対し、辞書による対応付けでは、通常辞書に載っているような単語の対訳しか抽出されないためである。

表 5-1-6 は、手法 d において人手確認に要した時間を示す。sim=0.5 の場合は、sim>0.5 に比べて、確認すべき語数が急増するため、確認に要する時間も急増する結果となった。現在は、辞書で確認されなかった全ての候補に対し、特別な加工を施さず、対訳一語一語を確認しているが、より効率良く確認するためには候補を予め分類する等の処理が必要となる。

次に、表 5-1-7 に d の手法で抽出された対訳表現の例を示す。手法 d では、正解(○)の例に見られるように、対訳辞書では確認されない表現[5]を 335 語(完全正解の 22%)抽出することができた。一方、半正解 (△)となった原因は、“月面着陸:land”のように、英語と日本語の句分割単位の違いに起因するものと、“製薬会社:pharmaceutical”のように、文書中の訳語の揺れに起因するものに大別される。前者に関しては、単純に句単位で分割するのではなく句単位の情報を生かした分割手法について検討したい。

最後に、不正解(×)を分析すると、辞書確認の間違いが大きな原因であることがわかった。例えば“start:開始”及び“福祉:well being”の辞書が存在した場合、“hope to start:今秋開始”、“福祉:well”は、辞書確認の結果、正しいとみなされ、抽出されてしまう。今回の実験では辞書確認は単純な方法に留めたが、今後、工夫する必要がある。

[3] a の sim=0.5 は、抽出量が膨大であったため、任意の 1000 語に対して精度を求めた。カバレッジの計算はしていない。
[4]正解と半正解の対訳表現を取り除いた結果、残った自立語数
[5] 完全一致ではなく、英語、日本語のそれぞれにおいて少なくとも 1 語が含まれていれば確認できたとした。

表 5-1-5 対訳表現の抽出精度

手法		確信度 sim	総対訳抽出数	精度 ○	精度 △	カバレッジ 英	カバレッジ 日
a	従来手法	>1	550	96%	98%	43%	59%
		>0.5	582	75%	80%	54%	67%
		=0.5*	11,389	7%	12%	-	-
b	a+句単位	>1	600	96%	98%	47%	59%
		>0.5	619	81%	85%	59%	66%
		=0.5	7,905	13%	21%	77%	79%
c	b+辞書	>1	600	97%	99%	48%	59%
		>0.5	573	86%	90%	59%	66%
		=0.5	3,682	28%	42%	78%	78%
d	c+人手	>1	595	98%	99%	48%	59%
		>0.5	525	93%	97%	59%	66%
		=0.5	1,742	57%	84%	77%	78%
e	辞書のみ利用		2,134	—	—	53%	65%

表 5-1-6 表 5-1-5 の“d”において人手確認に要した時間

sim	時間 (hour.min.sec)	確認語数
>1	00.00.00.	3
>0.5	00.05.13.	166
=0.5	01.13.19.	2436

表 3 表 5-1-5 の“d”における対訳表現の抽出例

評価	英語	日本語	sim
○	ＰＫＯ	peacekeeping operations	0.5
○	下方修正する	downgrade	0.5
△	月面着陸	land (*on the moon*)*	0.5
△	製薬会社	pharmaceutical (*firm*|*company*)*	1.5
×	今秋開始	hope to s tart	0.5
×	福祉	well	0.7

*(*斜体*)は参考用に記したものであり、実際は抽出されていない。

(2) 研究の効果

統計的手法と構造照合の融合技術を考案し、システムの試作を行なうことにより、動画関連の国際標準文書において、1000 文中 992 文の構造照合結果を抽出することがで

き、人手による評価の結果、83%の精度で対応付けられることが確認できた。

5-1-4　テンプレートの汎化に関する研究

対訳文書から翻訳テンプレートを自動抽出するためには、原文のどの部分が翻訳結果のどの部分に対応しているかを見つける技術と、対応が見つかった段階で、その対応のうちどの部分を汎化(変数化)するかを決定する技術が必要となる。我々は以下の 2 つの方法を用いて、翻訳テンプレートの汎化を行なうこととする。

ア．構造照合を利用する方法
構造照合とは、対訳文書の対訳文をそれぞれの言語において構文解析し、構文解析の結果に対し、片言語の構文構造のどの部分がもう一方の言語のどの部分に対応するかを、文全体の構文情報及び単語と単語の対応度の情報から求めるという手法である。本手法を利用することにより、原文のどの部分が翻訳結果のどの部分に対応しているかを自動的に検出することができる。

イ．統計的手法を利用する方法
原文と訳文の構造が対応付けられると、次に、対応付けられた部分の中でどの部分を汎化(変数化)するかを決める処理を行なう。ここで、統計的手法を利用する。統計的手法の最も単純な手法は、その出現頻度により汎化する否かを決める手法であるが、我々は、出現頻度だけでなく、語の意味、構成品詞、前後の単語の関係、その用語の専門性、既存の翻訳テンプレートとの類似度などを総合的に分析することにより最適な汎化部分を決定する統計的学習モデルを採用する。

以上のプロセスのうち、今年度は、構造照合結果を複数の部分木に分解し、語彙の一部を変数化した部分木をテンプレート形式に自動的に変換するツールの基本設計を完了した。また、その有効性を机上シミュレーションによって確かめた。
今後は、構造照合と統計的手法の両方の技術を用いて翻訳テンプレートを作成するシステムを試作する。

5-1-5　コンパラブルコーパスからの専門用語自動抽出

本研究は、各言語で書かれた同一内容、同一分野の文書に出現する単語やイディオムを言語間で対応付ける技術に関する研究である。コンパラブルコーパスを用いた手法は、抽出対象となる文書が多数存在するので、実用的な手法であるが、従来手法[文献 9]（以下、Fung の手法と呼ぶ）では対応付けの精度が低く、実用化のためにはさらなる改善が望まれる。そこで、本年度は既存の文対応がない対訳文書を利用した単語の対訳対応付け手法を改良した新手法を考案し、新手法の有効性を確かめる実験を行なった。

・Fung の手法
ある言語で共起する単語のペアは別の言語でも共起するという考えを前提とする。既に対応がわかっている対訳単語リストを使って、各言語での単語リスト中の単語と対応付けの対象語(以下、候補語と呼ぶ)の間の共起パターン(Word Relation Matrix：WoRM)を抽出する。二言語間で WoRM が類似する候補語のペアが、ここで抽出すべき対訳表現となる。(ここで「共起」とは、ある単語とある単語が一定の範囲内(例えば、

文や段落)に同時に出現する状態を示している。)

・本手法の提案
上記の方法には

(a) 単語リストが固定(不変)のため、コーパスのサイズやコーパス中に含まれている単語の種類によっては、抽出できる対訳表現の数が少ない
(b) 単語リスト中のすべての単語を同等に扱うため、対象コーパスでの出現頻度に偏りがあると適切な類似度が計れず、精度低下の要因となる

という２つ課題がある。この課題を解決するため、我々は、Fung の手法に対して以下の改良を加えた。

(a) 単語リストの単語に重み付けを行なう
(b) 獲得した対訳対を対訳単語リストに加え、再帰的に処理を行なう

重みの学習は、対訳単語リスト中の 1 対の対訳を取り出し、対象コーパスにおけるその語とそれ以外の単語リストとの WoRM を抽出するという処理を、すべての対訳に対して同様に行い、単語リストの各語が類似度決定にどの程度影響があるかの情報(つまり、単語の重み)を学習した。
単語の重みを学習した後、獲得した重みつき単語リストを用いて、各候補語の WoRM を抽出し、類似度を計算し、ある閾値以上の類似度をもつ候補語のペアを対訳表現として抽出する。抽出した対訳表現を対訳単語リストに加え、重みの学習処理に戻り、抽出される対訳表現がなくなるまで、本処理を繰り返す。
上記の改良により、日本語、英語それぞれ 5000 文程度の対訳文書(経済白書 1 年分)を用いた小規模な実験を行い、従来手法を上回る結果を確認した。
また、実験の結果、「本手法では単語リストは対訳文書と別に用意しているため、コーパスによっては有効な WoRM を作成することができない」という問題が明らかになったため、コーパスに含まれる語から有効な単語リストを作成する方法について検討した。この手法は、コーパス中の単語同士の共起傾向を調べ、共起分布の偏った単語を抽出して単語リストに追加するもので、机上シミュレーションにより、実効性が確認できた。

5-1-6　結論と今後の課題

上述したとおり、本サブテーマは、当初予定した目標を達成することができた。以下に、項目別の結論と今後の課題を示す。

(1) 既存の対訳文書とその改版文書を与えることにより、既存文書と改版文書の差分を検知し、未変化部分と変化部分を出力する「既存文書と改版文書の差分検知システム」の実装を完了した。今後は、検知された差分から翻訳テンプレートを作成するモジュールを完成させる。

(2) 統計的手法と構造照合の融合技術を考案し、システムの試作を行なうことにより、動画関連の国際標準文書において、1000 文中 992 文の構造照合結果を抽出することができ、人手による評価の結果、対訳の対応付けが 83%の精度で対応付けられることが確認できた。今後は実際の翻訳作業の効率化にどの程度貢献

できているかを定量的に図り、精度やユーザビリティの向上に取り組みたい。

(3) 統計的手法を用いた翻訳テンプレートの汎化技術に関する手法を確立した。構造照合結果を複数の部分木に分解し、語彙の一部を変数化した部分木をテンプレート形式に自動的に変換するツールの基本設計を完了し、その有効性を机上シミュレーションによって確かめた。

(4) 文対応がついていない対訳文書についても専門用語の対応付けが上位 5 位において精度 82%以上で実現された。文対応がついていない対訳文書からの専門用語の対応付け手法において、Fung の手法を改良した新手法を考案し、試作システムを開発した。日本語、英語それぞれ約 5000 文の対訳文書（経済白書 1 年分）を用いた実験の結果、

① 上位 5 位以内に正解の対訳表現が含まれている割合は 82%

② 従来手法に比べて、約 1.2 倍多くの対訳表現を抽出することができる

ことが確かめられた。今後は、5-1-5 で述べた課題の解決に取り組み、よりいっそうの精度向上を図る予定である。

5-2　分野辞書の自己組織化に関する研究開発

5-2-1　序論

ユーザが、多種多様な分野辞書を利用することを想定した場合、ユーザは常に辞書の構成やエントリを熟知し、それを状況に応じて使い分ける必要がある。これは必ずしも現実的とはいえない。我々が本テーマでいう自己組織化とは、ユーザに代わって、システムが自動的に最適な辞書を選択するために、辞書の構築及び辞書の選択を自動化する技術を指す。以下に、以前に策定した研究方針の概要を記す。

(a) 人手で多種多様な分野をあらかじめ設定し、ある語がどの分野に分類されるかを自動判定する基本方式を研究開発する。(課題を単純化するために分類は単層とする) これにより、ある語を登録したい場合、どの分野に登録すべきかを自動的に判定することができる。

(b) 上記の方式を応用し、分類に階層性を持たせる。さらに、ある語の情報だけでなく、その語の訳語の情報を利用した分野判定方式を研究開発する。

(c) 上記の方式を語の分野判定だけでなく、文書の分野判定にも応用する。これにより、ある文書を翻訳したい場合、どの辞書を利用すべきかを自動的に判定することができる。

(d) 上記の方式を応用し、分野辞書の自動階層化・分類手法を研究開発する。具体的には、ある語群に対し、異種の語を発見したり、さらに下層に分類すべきサブ語群を発見したりする方式の開発である。また、未分類の語群を、既存の分野に分類し、もし、適切な分野が存在しなかった場合には、適切な階層位置に新たな分野を自動作成する方式も研究開発する。

上記の(a)について取り組み、その結果今後の研究の柱となる「コアワード」を利用した分野の自動判定手法を考案した。本年度は、上記の(b)について取り組み、その結果、コアワードを利用した分野の自動判定手法の応用として、階層化された分野における分野自動判定手法および訳語の情報を利用した分野自動判定手法を考案した。

5-2-2　コアワードを利用した分野の自動判定の研究

(1) 研究の内容

我々は、Web ベースのコミュニティ型機械翻訳サイト「訳してねっと」を開発している。「訳してねっと」の特徴は、多数のユーザがインターネットを通して協力して分野毎に辞書や文書を登録することによって翻訳品質を高めることである。しかし、語を登録する際、ユーザが数多くの分野から適切な分野を選択するのは負担である。また、ユーザによって選択する分野が不統一であると翻訳品質の精度にも影響するため、システムが適切な分野を自動的に選択することが望まれる。

そこで、分野に特徴的でかつ代表的な単語を「コアワード」と定義し、コアワードを用いて分野辞書に単語を自動分類することを試みた。

これは、ある語が、あるコアワードと共起関係にあれば、その語はそのコアワードと同じ分野に分類してよい、という考え方に基づいている。コアワードには、分野に属する度合いを示す分野関連度を付与する。分野関連度の値が大きいほど、分野に属

する度合いが強いとする。具体的には、コアワードを分野に特徴的な文書から前もって自動的に作成して、分野関連度を付与しておき、判定対象語と共起関係にあるコアワードを適当な文書から抽出して、コアワードと分野関連度を利用して、自動的に分野を判定する。

ある語の分野を自動的に判定する従来研究には、翻訳を行なう際に収集した未登録の名詞句を、分野別辞書に含まれるパターンに何回ヒットしたかという情報から、主分野､副分野と呼ぶ二段に階層化された分野別辞書へ分野を判定する方法がある[文献10]。まず、全副分野で計算を行なって、固有かつ頻度の高いものを副分野辞書に分類する。それらを取り除いた後に、全主分野で同じ計算を行なって、主分野辞書に分類する。計算は、分野を要素としたベクトル空間中のベクトルとみなして行なう。訳語は、訳語を知らないシステムに訳語を生成させるか、訳語を与える部分だけを人手で行なう。

この方法では､階層毎に計算を繰り返さなければならないため､階層が深くなると、計算回数が増えるし、分野別辞書の登録語に依存しているため、分野別辞書の登録語数が増えると、計算量が増加する。また、訳語を人手で与えたとしても、分野推定時には訳語の意味を考慮することができないため、訳語が違えば分類先を変えたい語句を区別して分野推定することはできないので､正しく分類できないことになる｡更に、分野推定は分野別辞書の登録語に依存しているため、正しく判断できなかった語句が登録されて分野別辞書の語彙が増えると、その後の分野推定が正しく判断できない場合が生じ、登録するほど分野別辞書の質が落ちる可能性がある。

本年度は、コアワードを利用した分野の自動判定の手法を応用し、階層構造の各分野にコアワードを作成することによって、階層の深さに関係なく、階層化された分野において分野を自動判定する手法を試みた。更に、原語のコアワードと訳語側から原語側への単語の訳を利用して訳語の分野を判定し、原語の分野を判定した結果と訳語の分野を判定した結果を統合することによって、原語と訳語の多義を解消し、訳語の情報を利用して分野を自動判定する手法を試みた。

以下に、最初に、分野が単層の場合のコアワード自動作成と分野自動判定について記し、次に、階層化された分野における分野自動判定手法と訳語の情報を利用した分野自動判定手法とに分けて、それぞれの処理の概要を説明し、最後に、実験とその結果について述べる。

・単層におけるコアワードの自動作成

各分野のコアワードを､分野毎に既に分類されている文書を利用して作成する｡まず、分野毎に既に分類されている文書を形態素解析する｡次に､形態素解析結果で､名詞、動詞、形容詞、形容動詞、未知語を各分野のコアワードとする。

次に、各コアワードの分野関連度を計算する。分野関連度とは、その分野にどれだけ関連しているかを示した値である。分野関連度の値は、tf*idf で計算した値を利用することにする。tf*idf は、文書の自動索引付けにおいて、索引語の重みを計算する手法である。

TF(Term Frequency)　tf(d,t)
　ある文書ｄにおける索引語ｔの生起頻度（文書毎の文書中の単語数)。
DF(Document frequency)　df(t)
　索引語ｔが一回以上生起する文書の数（ある単語を含む文書の数)。
IDF(Inverse Document frequency)　idf(t)＝log(N/df(t))
　文書の数Ｎと、DF の逆数をかけて、対数をとる。
w(t,d)＝tf(d,t)*idf(t)

索引語 t の文書 d における重み w(t,d)。

語がどのくらい特定性を持つかを idf によって反映させる。多くの文書中に現れる一般的な語の場合には idf は小さくなり、逆に、特定の文書にしか現れない語の場合には idf は大きくなる。tf を用いるのは、文書中で繰り返し生起する語はその文書において重要な概念であると考えるためである。

ある文書に多数出現するほど大きくなる値 tf と特定の文書に偏って出現するほど大きくなる値 idf をかけた tf*idf では、総単語数が多いほど大きい値を取り得るので、その分野との関連性を表すだけでなく、各分野のコアワード作成に利用した文書の量にも依存するという問題がある。その問題を解消するために、分野間での調整が必要である。そこで、分野毎に、tf*idf をコアワード総数で割った値を、分野関連度とする。

分野関連度(分野,コアワード) ＝ tf*idf／分野毎のコアワード総数

表 5-2-1 は作成したコアワードの例である。

・単層におけるコアワードを利用した分野の自動判定

分野を判定したい語を判定対象語と定義する。分野を自動判定するために、「コアワード検索用文書」を準備しておく。ここで、コアワード検索用文書とは、判定対象語と共起関係にあるコアワードを抽出するための文書で、特に分野に分類されている必要はない。コアワード作成に利用した分類済の文書を利用しても構わない。

判定対象語と共起関係にあるコアワードの出現回数が多ければ多いほど、判定対象語とそのコアワードの共起関係は強いといえるが、tf*idf のみではその強さが反映されない。この問題を解消するために、共起関係の強さを反映する重み付けとして、抽出した共起関係にある語の出現回数を、語が一致する分野関連度にかけることにする。分野関連度に重み付けをした値を、分野判定度とする。

分野判定度(分野,コアワード) ＝
　　分野関連度(分野,コアワード)×出現回数(コアワード)

分野を自動判定する手順は以下である。判定対象語を、コアワード検索用文書から検索し、一文内に同時に出現したコアワードを全て抽出する。次に、抽出された全てのコアワードに対して分野判定度を計算し、分野判定度が高い順に順位付けをする。最後に、最も分野判定度が高いコアワードが属する分野を、判定対象語の分野と判定する。図 5-2-1 はコアワードを利用した分野判定の図である。

・階層化された分野におけるコアワードの自動作成

階層化された分野とは、例えば、スポーツ分野の下に野球分野やサッカー分野があるように、ある分野の下にその分野に含まれる分野があるような、階層構造になっている分野のことである。図 5-2-2 は階層化された分野の例である。●が最下層の分野でそれ以外の○が中間層の分野である。ある分野の上にある分野がその分野の親分野であり、ある分野の下にある分野がその分野の子分野である。単層の分野に判定する場合と比べて、出来る限り狭い対象範囲の分野に、すなわち出来る限り下層の分野に判定するのが効果的であるため、適切な分野に判定するのは、単層の分野に判定する場合よりも困難である。例えば、サッカーの用語がサッカー分野より上のスポーツ分野にあっても間違いではないが、その他のスポーツには使用しない用語であればサッカ

一分野にあるのが適切である。ただし、サッカーで使用される用語であっても、スポーツ全般で使用される用語であれば、サッカー分野より上のスポーツ分野にある方がふさわしい。

本手法では、最下層以外の途中の階層も含む全ての分野に適切にコアワードを作成し、コアワードを利用して階層の深さに関係なく一度に分野を判定するという方針を採ることにする。

全ての分野にコアワードを作成するには、最下層の分野のコアワードは文書から作成し、親分野のコアワードは直下の子分野のコアワードから作成することにする。文書は、最下層の分野の単位に分類された文書だけを用意すればよい。文書からコアワードを作成する方法は、単層におけるコアワードの自動作成の場合と同じで、分野毎に既に分類されている文書を利用して作成する。最下層の分野と中間層の分野で処理が違うのは以下の理由による。

階層化された全ての分野に、分野に属する文書を与える方法では、子分野の内容を含まない「その他」にあたる文書の場合はその分野に属する文書として問題ないが、子分野の内容を含む「全般」にあたる文書の場合は親と子の区別が曖昧になるので、正しくかつ出来る限り狭い対象範囲に分類することができないことがある。途中の階層も含む全ての分野に適切に分類された文書を用意することは、階層が深くなるほど労力を要する。ゆえに、親分野のコアワードは文書からは作成しないこととし、最下層の分野別に分類された文書を用いて最下層の分野のコアワードを作成し、作成された最下層の分野のコアワードを用いて中間層の分野のコアワードを作成する。

親分野のコアワードを作成する方法は、作成した子分野のコアワードの偏り具合から判断するというものである。子分野に万遍なく存在する場合には親分野のコアワードにし（親分野にも作成する）、いずれかの子分野に突出している場合には親分野のコアワードにしない（子分野にのみ存在する）、という考え方である。ある親分野の直下の子分野全体で、コアワード毎に、分野関連度の平均値(mean)と標準偏差(sd)を、正規分布と仮定して、計算する。コアワードがない分野の分野関連度は 0 として計算する。すべての分野関連度がばらつきの範囲内であれば、そのコアワードを親分野のコアワードにする。例では、範囲を「平均値＋標準偏差×3」(mean+3sd)とする。「平均値＋標準偏差×3」(mean+3sd)の範囲内にデータが入る確率は 99.73%である。範囲の計算は正しくは平均値±プラスマイナスであるが、分野関連度に負の値はないため、マイナスの方は無視してよい。

分野関連度は、最下層の分野では、tf*idf をそのまま利用するとコアワード作成に利用した文書の量に左右されるので、分野毎にコアワードの総数で割った値にしているが、親分野のコアワードにはコアワード作成用の文書が存在しないので、親分野の分野関連度は、適当に設定しなければならない。ここでは、範囲の上限値「平均値＋標準偏差×3」(mean+3sd)とする。親分野のコアワードの分野関連度を、どの子分野の分野関連度よりも高い値にするためである。

・階層化された分野における分野の自動判定

階層化されていることを考慮して全ての分野にコアワードと分野関連度を作成しておくので、分野の判定は単層における場合と同様の手法を用いればよい。

分野を自動判定する手順は以下である。判定対象語を、コアワード検索用文書から検索し、一文内に同時に出現したコアワードを全て抽出する。次に、抽出された全てのコアワードに対して分野判定度を計算し、分野判定度が高い順に順位付けをする。最後に、最も分野判定度が高いコアワードが属する分野を、判定対象語の分野と判定する。

また、階層化されたことによって分野数が多くなることから起こり得る問題を解消

するために、分野判定度を分野毎に集計することも考えた。これは、作成したあるコアワードが、多数の分野に属する一般的な語で、判定対象語と共起関係にある語の出現回数が多くなることによって、分野判定度が高くなった場合を想定している。その場合、どこかの分野に特徴的で判定に有効なコアワードの方が、分野判定度が低くなり、順位が低くなることによって、判定に反映されないということが起こり得る。そこで、分野判定度を分野毎に集計することで、複数のコアワードから総合的に分野を判定することができる。

・訳語の情報を利用した分野の自動判定

訳語の情報を利用して分野を判定するとは、例えば、原語「ティー」だけではゴルフのティーのことなのかお茶のことなのか分野は曖昧であるが、訳語が"tee"ならばゴルフ分野で"tea"ならば料理分野であると、判定できるようになることとする。訳語"tee"がゴルフ分野であるとわかるのは共起関係にある語が"golf", "bunker"のようにゴルフに関連した語が多くなるからであり、訳語"tea"ならば"pot", "coffee"のように料理に関連した語が多くなるからであるとすると、コアワードを利用した分野判定手法と考え方は同じであり、訳語の分野もコアワードを利用して判定できるといえる。訳語と共起関係にある語の原語訳「ゴルフ」「バンカー」や「ポット」「コーヒー」をコアワードと考えれば、訳語側に新たにコアワードを作成する必要はなく、原語側のコアワードを利用すればよいことになる。

実際に訳語の情報を利用しようとすると、語の多義がいっそう多く含まれるので、多義性の解消を考慮しなければならない。例えば、原語が「選手」で訳語が"player"である語の分野を判定するために、"player"を訳語のコアワード検索用文書から検索する場合を考える。"baseball", "team"のようなスポーツ分野に含まれる語が共起関係にあるのならば問題ないが、「レコードプレーヤー」の意味の"player"が多く検索されて"disc", "recorder"のような音楽分野に含まれる語ばかりが共起関係にある語になれば判定を誤ることになる。更に、"recorder"の原語訳が「テープレコーダー」であれば原語訳は訳語と同じ音楽分野に含まれるが、原語訳が「記録係」であれば訳語と同じ音楽分野には含まれない。原語の多義を解消するために訳語の情報を利用するのにも、訳語の多義と訳語を原語訳にする際の多義が含まれるのである。

本手法では、原語を分野判定した結果と訳語を分野判定した結果をそれぞれ複数用意し、それらを統合することによって、多義性を解消して、分野を判定するという方針を採ることにする。

原語を分野判定した結果と訳語を分野判定した結果を統合して分野を自動判定するために、以下を定義する。分野判定したい訳語付の語を判定対象語、分野判定したい原語を判定対象原語、分野判定したい訳語を判定対象訳語と定義する。ここでは、原語を日本語、訳語を英語とする。判定対象語または判定対象原語または判定対象訳語に対して、分野に属する度合いを示すコアワードの値を、分野判定度と定義し、分野とコアワードと分野判定度の組を、分野判定度のリストと定義する。判定対象語に対して、分野に属する度合いを示す値を、統合分野判定度と定義し、分野と統合分野判定度の組を、統合分野判定度のリストと定義する。訳語の多義に対応するために、訳語に対しては複数の原語訳を与えることにする。訳語に原語訳を与えるために、訳語側から原語側への単語の訳を単語毎に複数保持する単語辞書を準備しておく。複数の原語訳があることが出現回数の重み付けに影響することを解消するために、原語訳の数の重み付けとして、語が一致する分野関連度を訳語に対する原語訳の数で割ることにする。分野関連度に重み付けをした値を、分野判定度とする。

本手法の考え方は以下である。原語を分野判定するには、原語と原語と共起関係にある語と原語側のコアワードを利用する。訳語を分野判定するには、訳語と共起関係

にある語と訳語側から原語側への単語辞書と原語側のコアワードを利用する。判定対象語に対して、判定対象原語の分野判定度を計算して判定対象原語の分野判定度のリストを作成し、判定対象訳語の分野判定度を計算して判定対象訳語の分野判定度のリストを作成し、判定対象原語の分野判定度のリストと判定対象訳語の分野判定度のリストから判定対象語の分野判定度のリストを作成して判定対象語の分野毎の統合分野判定度を計算し、分野判定した判定対象語が属する分野を判定する。

本手法の詳細な手順は以下である。判定対象原語を原語のコアワード検索用文書から検索し、判定対象原語と共起関係にある語を原語のコアワード検索用文書から抽出し、抽出した共起関係にある語をコアワードから検索し、検索したコアワードの分野関連度に対して、出現回数の重み付けをして、分野判定度を計算し、判定対象原語の分野判定度のリストを出力する。

判定対象原語の分野判定度(判定対象原語, 分野, コアワード)
＝　分野関連度(分野, コアワード)×出現回数(判定対象原語, コアワード)

判定対象訳語を訳語のコアワード検索用文書から検索し、判定対象訳語と共起関係にある語を訳語のコアワード検索用文書から抽出し、抽出した共起関係にある語の原語訳を単語辞書から検索し、原語訳を形態素解析した結果から不要語を除いた単語を抽出し、抽出した語をコアワードから検索し、検索したコアワードの分野関連度に対して、出現回数の重み付けをして、原語訳の数の重み付けをして、分野判定度を計算し、判定対象訳語の分野判定度のリストを出力する。原語訳を形態素解析して不要語を除くのは、原語訳をコアワードと一致させるためである。

判定対象訳語の分野判定度(判定対象訳語, 分野, コアワード)
＝　分野関連度(分野, コアワード)×出現回数(判定対象訳語, コアワード)／原語訳の数(判定対象訳語)

判定対象原語の分野判定度のリストと判定対象訳語の分野判定度のリストを統合して判定対象語の分野判定度を計算して判定対象語の分野判定度のリストを作成し、判定対象語の分野判定度のリストから判定対象語の分野毎の統合分野判定度を計算し、判定対象語が属する分野を判定する。分野は、統合分野判定度が最大の分野とする。判定対象語の分野判定度の値は、原語側と訳語側の両方の分野判定度が高いほど高くなるようにしたい。ただし、原語側と訳語側では、共起関係にある語を抽出するための文書の内容と量が違うことや、共起関係にある語が多義であることによって、分野判定度を単純に比較することができない。そこで、判定対象原語の分野判定度と判定対象訳語の分野判定度をそれぞれの分野判定度の最大値で割って正規化し、調和平均をとることにする。

判定対象語の分野判定度(判定対象語, 分野, コアワード)
＝　２×（Ｖ１／Ｍ１）×（Ｖ２／Ｍ２）／（(Ｖ１／Ｍ１）＋（Ｖ２／Ｍ２））
Ｖ１：判定対象原語の分野判定度(判定対象原語, 分野, コアワード)、
Ｍ１：判定対象原語の分野判定度(判定対象原語, 分野, コアワード)の最大値、
Ｖ２：判定対象訳語の分野判定度(判定対象訳語, 分野, コアワード)、
Ｍ２：判定対象訳語の分野判定度(判定対象訳語, 分野, コアワード)の最大値

判定対象語の統合分野判定度を計算する方法は、分野判定度がある閾値以上ならば分野毎に分野判定度を合計することにする。閾値で足切りするのは、分野判定度が低

いものが数で勝る悪影響を排除するためである。閾値は、分野を最適に決定することができる任意の値とする。

判定対象語の統合分野判定度(分野)
＝Σ判定対象語の分野判定度(分野, コアワード)

ただし、判定対象語の分野判定度≧α （αは閾値）

・単層における実験とその結果

我々が現在開発中の「訳してねっと」が所有する分野を用いて、本手法の有効性を検証した。まず、「訳してねっと」上に存在する分野から毎日新聞の記事の分類とほぼ一致するように 23 分野を選択し、毎日新聞(1995 年)の記事からコアワードを作成して分野関連度を計算した。テストデータは「訳してねっと」の各分野辞書に登録済のデータから毎日新聞(1995 年)の記事に存在するものをランダムに 100 個抽出したものとし、「訳してねっと」で登録されている分野を正解とした。コアワード検索用文書は、コアワード作成に用いた毎日新聞の 1995 年とコアワード作成とは別の 1996-1999 年の 2 種類を用いた。

その結果、上位 1 位、上位 5 位以内に正解が含まれた精度は、それぞれ、1995 年で 72%、88%、1996-1999 年で 69%、88%であった。これにより、本手法の有効性を確認した。また、コアワード作成に用いる記事と語の検索に用いる記事は別でよいことが確認できた。

表 5-2-2 は単層において分野を判定した結果の例である。コアワードを利用することによって、固有名詞や新語なども分野を判定することができることを示している。表 5-2-3 は単層において分野を判定した分野毎の正解率である。「政治」では、1 位で正解した率が 100%と高く、「野球」「水泳」などのスポーツ関連の分野もコアワードの量に関係なく正解率は高い。その理由は、新聞記事では政治面とスポーツ面は特に区別されており、そこに出現する語は、その分野に限定された専門的な語が多いためであると考えられる。これは、内容が限定された分野ではコアワードの量に関係なく正解率が高くなることを示している。逆に、「海外」や「家族・生活」や「趣味・娯楽」の正解率は他より低い。これは、内容が広く多岐にわたる分野ではコアワードの量に関係なく正解率が低くなることを示している。

・階層化された分野における実験とその結果

我々が現在開発中の「訳してねっと」が所有する分野を用いて、本手法の有効性を検証した。まず、単層における実験で「訳してねっと」上に存在する分野から毎日新聞の記事の分類とほぼ一致するように選択した 23 分野に対して、子分野や親分野を追加して、最下層 30 分野、中間層 4 分野にした。毎日新聞(1995 年)の記事を最下層の分野に再分類して、最下層の分野のコアワードを作成して分野関連度を計算し、最下層のコアワードと分野関連度から中間層のコアワードを作成して分野関連度を計算し、全ての分野にコアワードを作成して分野関連度を計算した。テストデータは単層における実験の場合と同じとした(「訳してねっと」の各分野辞書に登録済のデータから毎日新聞(1995 年)の記事に存在するものをランダムに 100 個抽出したものとし、「訳してねっと」で登録されている分野を正解とした)。コアワード検索用文書は、コアワード作成とは別の 1996-1999 年を用いた。

その結果、上位 1 位、上位 5 位以内に正解が含まれた精度は、それぞれ、62%、85%であった。単層における実験の場合は、それぞれ、69%、88%で、精度の低下は、分野が階層化されて判定が困難になったことを考慮すると、許容範囲である。これによ

り、本手法の有効性を確認した。

・訳語の情報を利用した実験とその結果

我々が現在開発中の「訳してねっと」が所有する分野を用いて、本手法の有効性を検証した。分野とコアワードと分野関連度は、階層化された分野における実験の場合と同じとした(「訳してねっと」上に存在する分野から毎日新聞の記事の分類とほぼ一致するように最下層 30 分野、中間層 4 分野を選択し、最下層の分野に分類した毎日新聞(1995 年)の記事から最下層の分野のコアワードを作成して分野関連度を計算し、最下層のコアワードと分野関連度から中間層のコアワードを作成して分野関連度を計算し、全ての分野にコアワードを作成して分野関連度を計算した)。テストデータは、分野が違う複数の訳語があって、原語の分野を判定した結果の 1 位が訳語を考慮すると正解ではないように訳語を選択した、4 語とした。原語のコアワード検索用文書は、コアワード作成とは別の 1996-1999 年を用いた。訳語のコアワード検索用文書は、JapanTimes1989 年を用いた。

その結果、4 語とも、原語の分野を判定した結果では 1 位ではないが訳語を考慮すると適している分野が、1 位になった。これにより、本手法の有効性を確認した。

(2) 研究の効果

1. 「コアワード」を利用した分野の自動判定の手法を確立した。これにより、ユーザが多種多様な分野に分類された辞書に対して、語を登録する際、ユーザが自ら分野を選定する必要がなく、システムが自動的に登録分野を選定することができる。

2. コアワードを利用して分野を判定することによって、分野別辞書の内容や量に依存することなく分野を判定することができる。また、判定対象語の分野を判定する際に使用するコアワード検索用文書は分野に分類されている必要はないため、未分類の文書を準備するだけで精度の向上が図れる。

3. 階層構造の各分野に前もってコアワードを作成しておくことによって、階層の深さに関係なく一度に語の分野を判定することができる。その際、コアワードを作成するために、全ての分野に対して分類済の文書を用意する必要はない。

4. 訳語の分野判定に原語のコアワードを利用することによって、原語と訳語のそれぞれに含まれる多義や原語と訳語の間の多義を解消して、訳語付の語の分野を判定することができる。その際、原語側のコアワードさえあれば、訳語側から原語側への単語辞書と訳語側の未分類の文書を変更するだけで、多言語に対応することができる。単語辞書には単語毎に複数の訳があればよく、分野の情報や構文的な情報などは不要であるので、特別に作成する必要はない。

5-2-3　結論と今後の課題

上述した通り、本サブテーマは、当初予定した目標を達成することができた。今後の課題は、まず、翻訳したい文書がどの分野辞書を利用すべきかを自動的に判定する手法を研究開発することである。次に、分野辞書の自動階層化・分類手法を研究開発することである。具体的には、ある語群に対し、異種の語を発見したり、さらに下層に分類すべきサブ語群を発見したりする方式の開発である。また、未分類の語群を、既

存の分野に分類し、もし、適切な分野が存在しなかった場合には、適切な階層位置に新たな分野を自動作成する方式も研究開発する。

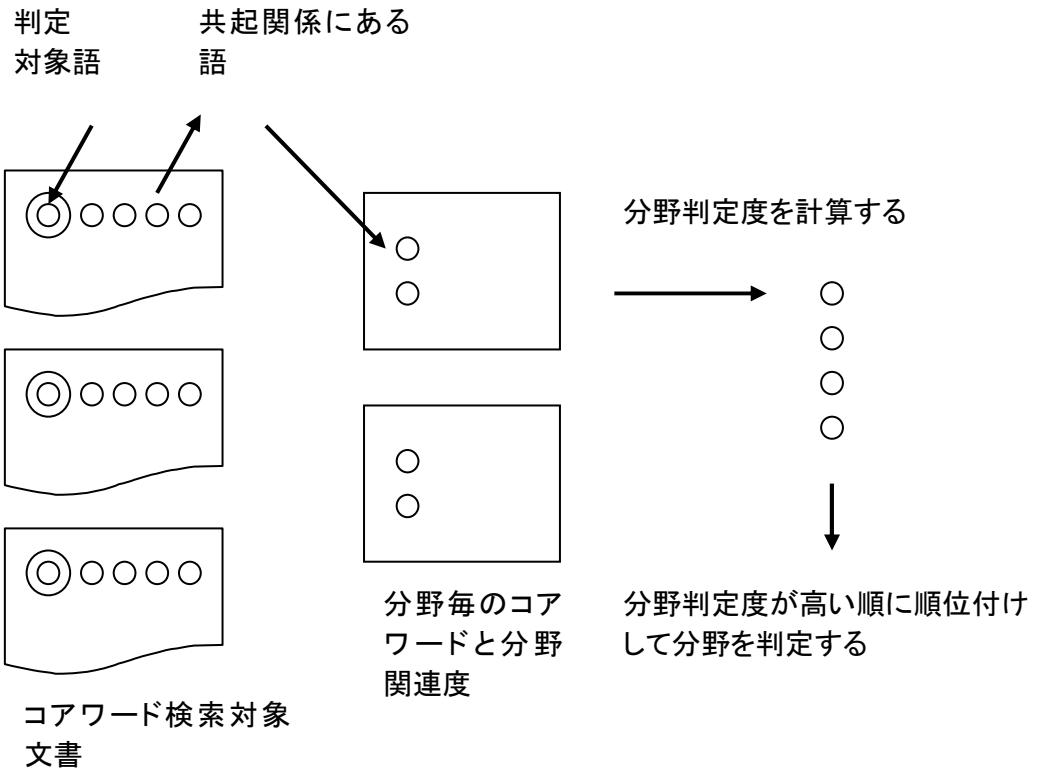


図 5-2-1　コアワードを利用した分野判定の図

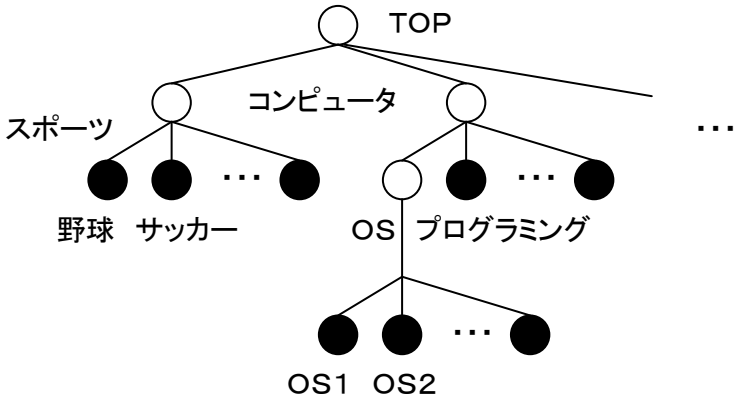


図 5-2-2　階層化された分野の例

表 5-2-1　単層において作成したコアワードの例

分野	コアワード（分野関連度が高い順に 10 個）
経済・ビジネス	金融、債権、円、金利、大蔵省、投資、ＡＰＥＣ、経済、市場、証券
料理・食事	料理、店、完、値段、刺し身、盛り合わせ、ワイン、ランチ、定食、同右
野球	本塁打、安打、登板、打線、投手、日本ハム、オリックス、ロッテ、野球、三塁

表 5-2-2　単層において分野を判定した結果の例

判定対象語	コアワード（分野関連度が最高）	分野（１位で正解）
行政改革	首相	政治
最優遇貸出金利	金利	経済・ビジネス
ドラクエ	ゲーム	趣味・娯楽
ハットトリック	サッカー	サッカー
高橋尚子	マラソン	陸上競技

表 5-2-3　単層において分野を判定した分野毎の正解率

分野（文書の多い順）	検索対象記事 1995 年		検索対象記事 199[6-9]年	
	１位	上位５位以内	１位	上位５位以内
経済・ビジネス	83%	100%	50%	100%
海外	0%	60%	0%	80%
政治	100%	100%	100%	100%
野球	100%	100%	100%	100%
：				
家族・生活	17%	50%	17%	17%
：				
生命科学	57%	100%	71%	100%
水泳	100%	100%	100%	100%
趣味・娯楽	50%	50%	50%	50%
料理・食事	60%	80%	40%	100%

表 5-2-4 階層化された分野において作成したコアワードの例

分野	コアワード	分野関連度（降順）
スポーツ	試合	0.00745
スポーツ	選手	0.00294
...	...	...
野球	投手	0.00929
野球	試合	0.00497
野球	選手	0.00145
...	...	...

表 5-2-5 階層化された分野における分野判定度の計算結果の例

判定対象語	分野	コアワード	分野判定度（降順）
公式戦	スポーツ	試合	4.30
公式戦	野球	野球	2.94
公式戦	野球	試合	2.87
...	...	...	...
防御率	野球	投手	4.46
防御率	スポーツ	試合	3.74
...	...	...	...

表 5-2-6 判定対象原語の分野判定度のリストの例

判定対象原語	分野	コアワード	分野判定度（降順）
審判	野球	野球	2.581
審判	政治	選挙	1.875
審判	政治	首相	1.184
審判	サッカー	サッカー	1.035
...	...	...	...

表 5-2-7 判定対象訳語の分野判定度のリストの例

判定対象訳語	分野	コアワード	分野判定度（降順）
judgement	政治	国会	0.01076
judgement	海外	議会	0.00654
judgement	政治	選挙	0.00481
judgement	科学	原子力	0.00387
...	...	...	...

表 5-2-8 判定対象語の分野判定度のリストの例

判定対象原語	判定対象訳語	分野	コアワード	分野判定度（降順）
審判	judgement	政治	選挙	0.553
審判	judgement	海外	選挙	0.292
審判	judgement	政治	党	0.170
審判	judgement	政治	国会	0.123
...	...	...	...	...

表 5-2-9 判定対象語の統合分野判定度の計算結果の例

判定対象原語	判定対象訳語	分野	分野判定度（降順）
審判	judgement	政治	1.138
審判	judgement	海外	0.572
...	...	...	...

5-3　言語非依存の翻訳エンジンの研究開発

5-3-1　序論

本サブテーマでは、「5-1 翻訳テンプレート学習に関する研究開発」および「5-2 分野辞書の自己組織化に関する研究開発」で得られた結果を用いて、実際に翻訳を行なうシステムの研究開発を行なう。研究開発の概要は以下のとおりである。

ア. 言語非依存翻訳エンジン：エンジン全体および個々のモジュールについての設計、形態素解析システムの中国語への適用。

イ. 多言語翻訳データベース：試作システムの開発、および、試作システムを使ったデータベースの構築。

ウ. 協調的翻訳支援環境：支援環境の開発、および、実証実験

翻訳エンジンは、さまざまな言語対から得られた翻訳テンプレートおよび対訳文書を利用して、指定された言語間の翻訳を行なう。本翻訳エンジンは多言語への展開を容易にするため、言語に依存する部分を最小限に抑えるよう設計されている。これまでは、日英間の翻訳のみ動作していたが、本年度は中国語への取り組みを開始し、本翻訳エンジンが多言語の翻訳に適応可能であることを示す。

多言語翻訳データベースの研究では、「5-1-2 改版文書を利用した翻訳テンプレート獲得に関する研究開発」を利用して得られた文対応のついた対訳文書や「5-1-3 構造照合による訳語対応付けの研究」を利用して得られた翻訳パターンを格納するデータベースを実装する。

協調的翻訳支援環境の研究では、翻訳エンジンおよび多言語翻訳データベースを用いて、翻訳を行なう環境を構築する。

前年度までは基本的なモジュールの設計が中心であったが、今年度は他の研究成果も取り入れ、実際にユーザに開放しての実証実験に着手するなど、実用化に向けた研究開発になっている。以下では、それぞれのテーマにおける本年度の取り組みについて、具体的に説明する。

5-3-2　中国語・日本語形態素解析システムの研究

(1) 研究の内容

中国語と日本語を高い精度で形態素解析するための手法を検討し、単語単位の情報と文字単位の情報を利用する形態素解析の手法を考案し、実装・評価を行なった。

形態素解析を行なう上で、大きく次のような 3 つの課題がある。1 つ目の課題は、曖昧性解消の問題である。入力された文に対する単語分割や品詞タグ付けの候補は多数存在するため、その曖昧性を解消する必要がある。曖昧性の解消に失敗して誤った解を出力すると、その結果を利用するアプリケーションも誤った処理をしてしまうため、形態素解析器はできるだけ高い精度で解析を行なう必要がある。2 つ目の課題は、未知語の問題である。未知語とは、形態素解析システムの辞書中に存在しない単語のことであるが、固有名詞や新語・造語などがしばしば未知語として出現する。このような未知語に対しては単語に関する情報が存在しないため、単語分割を決定して品詞

タグを推定するのは非常に難しい。3 つ目の課題は、言語への依存性の問題である。形態素解析システムは、言語に固有の現象を解析する必要があるが、特定の言語に特化して作り込んでしまうと、他の言語も扱いたい場合に拡張して利用するのが困難になる。以上のような課題に対処して形態素解析を行なうために、これまでに様々な方法が研究されている。特に近年、人手で作成した規則を用いて解析を行なうルールベースの手法に代わり、学習用のデータから形態素解析を行なうのに必要なパラメータを自動的に獲得する統計ベースの手法が広く用いられるようになっている。そのような統計ベースによる日本語や中国語の形態素解析手法として、コスト最小法と文字タグ付け法が知られている。

・コスト最小法と文字タグ付け法

コスト最小法は、精度や実行速度の面で優れた解析手法で、実用的な形態素解析システムで広く採用されている[文献 17]。この手法は、次の 2 つのステップにより形態素解析を行なう。

1. (解候補の生成) 文が与えられると、形態素辞書を使用して、その文中に含まれるあらゆる形態素を列挙し、ラティス構造で表現される解析結果の全候補を生成する (図 5-3-1)。

2. (解の決定) ラティスの最初のノードと最後のノードを結ぶパスの中から、正解であると思われるものを 1 つ選び出し、それを解として出力する。このパスを選び出す際の基準として、パス上のノードの単語と品詞列が生成される確率を使用する。この確率の逆数の対数をとったものをそのパスのコストとし、このコストを最小にするパスが選択される。

この手法は、既知語(未知語とは逆に、システムの辞書中に存在する単語) に対して高い精度で解析を行なうことができ、計算時間も速いという特徴がある。しかしながら、そのままでは未知語を処理できないため、ヒューリスティックな方法で未知語の候補を作り出し、あらかじめラティス上に追加する方法が一般的に用いられる。例えば日本語の場合、「連続するカタカナやアルファベットは一単語としてまとめ、漢字やひらがなは一文字を一単語にする」というルールが使われる。このような方法は未知語に対する精度が比較的低く、また言語に大きく依存してしまう問題がある。

文字タグ付け法では、単語に関する情報は用いずに文字単位のタグ付けにより形態素解析を行なう[文献 15] [文献 16]。この手法では、入力された文中の全ての文字に対して、その文字が単語のどの位置にあるか、またその文字が構成する単語の品詞は何であるかを決定することにより、形態素解析を行なう。図 5-3-2 に例を示す。この例では、文中の各文字に対して、B、 I、 E、 S で表される単語中の位置を示す文字と、単語の品詞名から構成されるタグを付与している。B は単語の先頭の文字を、I は単語の中間にある文字を、E は単語の末尾にある文字を、S は 1 文字だけで単語を構成する文字を表している。このような文字に対するタグ付けを考えることにより、形態素解析の問題は文中の各文字がどのタグを持つかという分類問題として考えられる。このような分類問題は、サポートベクターマシーン等の機械学習アルゴリズムを用いて解くことができる[文献 16]。この手法の利点として、文字の情報だけを使い形態素解析を行なうため未知語に強いという点がある。一方で、単語の情報を使わないために既知語に対しては精度がやや低いという問題がある。

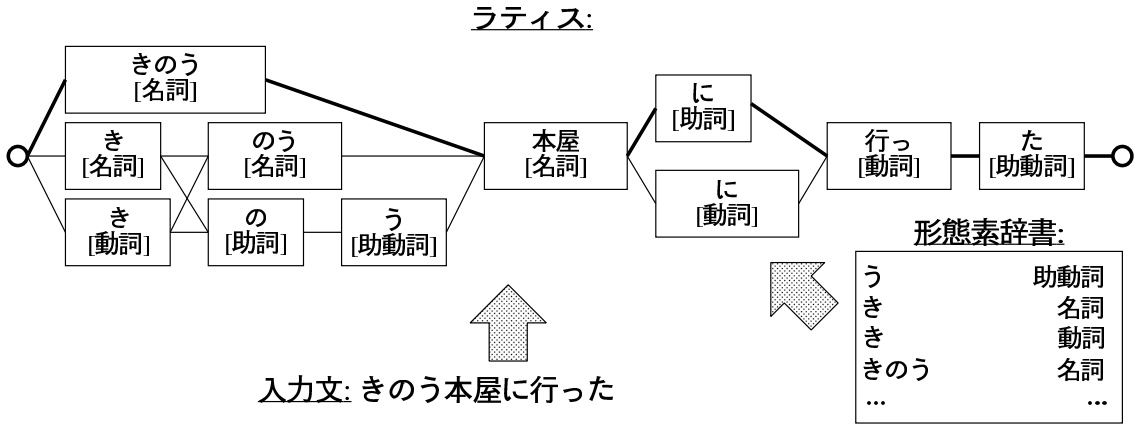


図 5-3-1: コスト最小法のラティス

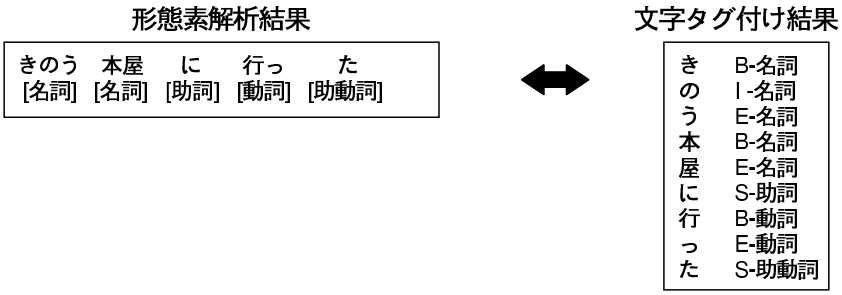


図 5-3-2: 文字タグ付け法による形態素解析

・提案手法

上記の 2 つの従来手法を踏まえ、言語に比較的依存せずに中国語と日本語の形態素解析を行えるシステムを検討し、実装した。このシステムは、コスト最小法をベースにして、未知語の処理には文字タグ付けの手法を応用する。つまり、品詞タグと文字タグを同じコスト最小法の枠組みの中で使用する。これにより、単語に関する情報があらかじめ分かっている既知語に対しては単語単位で高精度な解析を行い、単語に関する情報が無い未知語に対しては文字単位で頑健な解析を行なうことが期待できる。また、未知語の処理を特定の言語に依存すること無く行なうことができ、中国語と日本語を同一のシステムで解析することが可能になる。

図 5-3-3 に提案手法における解析の例を示す。これは、「細川護熙首相が訪米」という文が入力され、そのとき「護熙」という単語が未知語である場合の例である。まず解析対象の文が入力されると、通常のコスト最小法と同様に既知語に対するラティスのノードが作成される。そして次に、文中の全ての文字に対して、B、 I、 E、 S の文字タグを持ったノードが作成される。そして、このように既知語を処理するための品詞タグを持った単語のノードと、未知語を処理するための文字タグを持った文字ノードが混在したラティスの中から、最適なパスを選択し、形態素解析を行なう。なお、この際に I タグから B タグ、あるいは任意の品詞タグから E タグへの遷移など、不適切な状態遷移は無視する。

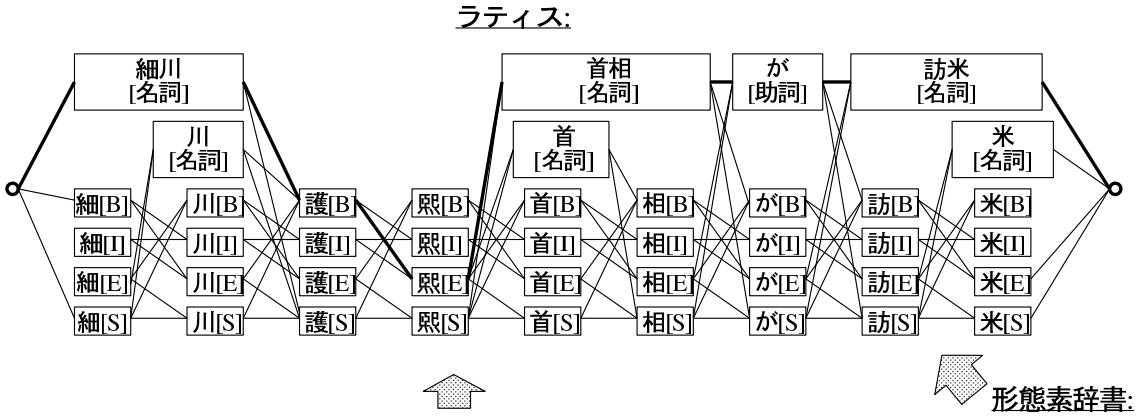


図 5-3-3 提案手法による形態素解析

通常のコスト最小法で使用される品詞 bigram モデルは、直前の 1 つの品詞だけを考慮して次に出現する品詞を予測するため、十分な精度が得られない。そこで、次のようなモデルによりラティス中のパスの生成確率を計算することにする:

$$
\begin{aligned}
&P(w_1t_1\cdots w_nt_n)\\
&=\prod_{i=1}^{n} P(w_it_i \mid w_1t_1\cdots w_{i-1}t_{i-1})\\
&\approx\prod_{i=1}\left\{\lambda_1P(w_it_i)+\lambda_2P(w_i \mid t_i)P(t_i \mid t_{i-1})+\lambda_3P(w_i \mid t_i)P(t_i \mid t_{i-2}t_{i-1})+\lambda_4P(w_it_i \mid w_{i-1}t_{i-1})\right\}
\end{aligned}
$$

ここで、w_i と t_i はそれぞれ文頭から i 番目の単語と文字である。上の式中の確率は、品詞タグの付与された学習データから最尤推定により計算される。また、$\lambda_1,\lambda_2,\lambda_3,\lambda_4$ の値は削除補間法によって計算される。文字レベルの豊富な情報をモデルに取り入れるため、文字タグに関する単語の出現確率はベイズの定理を用いて以下のように計算する:

$$P(w_i \mid t_i \in \{\mathbf{B},\mathbf{I},\mathbf{E},\mathbf{S}\}) = \frac{P(t_i \mid w_i)P(w_i)}{P(t_i)}$$

そして、この式の右辺中の $P(t_i \mid w_i)$ の値を、最大エントロピー法[文献 11]によって推定する。w_i が文頭から i' 番目の文字である時、c_x は文頭から x 番目の文字を、y_x は文字 c_x の文字種をそれぞれ表すとして、最大エントロピー法の素性として以下のものを使用した:

(1) 文字($c_{i'-2}$, $c_{i'-1}$, $c_{i'}$, $c_{i'+1}$, $c_{i'+2}$)

(2) 文字 bigram($c_{i'-2}c_{i'-1}$, $c_{i'-1}c_{i'}$, $c_{i'-1}c_{i'+1}$, $c_{i'}c_{i'+1}$, $c_{i'+1}c_{i'+2}$)

(3) 文字種($y_{i'-2}$, $y_{i'-1}$, $y_{i'}$, $y_{i'+1}$, $y_{i'+2}$)

(4) 文字種 bigram($y_{i'-2}y_{i'-1}$, $y_{i'-1}y_{i'}$, $y_{i'-1}y_{i'+1}$, $y_{i'}y_{i'+1}$, $y_{i'+1}y_{i'+2}$)

この最大エントロピーモデルのパラメータは、学習データから GIS アルゴリズム[文献 12]等に
より計算することができる。

なお、この提案手法では未知語の品詞推定までは行わないため、最大エントロピー

単語分割の精度					
システム	再現率	精度	F 値	未知語再現率	既知語再現率
提案手法	**0．993**	**0．994**	**0．993**	**0．586**	**0．993**
茶筌	0．991	0．992	0．991	0．243	0．991
単語分割と品詞タグ付けの精度					
システム	再現率	精度	F 値	未知語再現率	既知語再現率
提案手法	**0．982**	**0．983**	**0．982**	**0．388**	**0．983**
茶筌	0．977	0．978	0．977	0．042	0．978

表 5-3-1 日本語形態素解析の精度の比較

法を用いた後処理によって、前後の品詞や単語情報から未知語の品詞を推定する。

(2) 研究の効果

上記提案手法に基づいた中国語・日本語形態素解析システムを試作し、日本語と中国語のコーパスを用いて評価を行なった。

・日本語形態素解析の評価

日本語の形態素解析精度を評価するために、毎日新聞の記事 37、589 文からなる RWCP コーパスを使用した。このコーパスの 33、831 文を学習用データとして中国語・日本語形態素解析器のパラメータの学習を行い、残りの 3、758 文に対して形態素解析を行い解析精度を求めた。解析精度の指標には、再現率、精度、F 値を使用し、単語分割のみを行なった場合の精度と、単語分割と品詞タグ付けの両方を行なった場合の精度を求めた。再現率は正解の中のどれだけを正しく求めたかの割合を、精度はシステムの出力中のどれだけが正しかったかの割合を、F 値は再現率と精度の調和平均を表し、次のように定義される:

・再現率 = 解析結果の正解形態素数／正解データ中の形態素数
・精度 = 解析結果の正解形態素数／解析結果の形態素数
・F 値 = 2×再現率×精度／(再現率 + 精度)

再現率に関しては、未知語に対する再現率と既知語に対する再現率も求めた。
表 5-3-1 に結果を示す。比較のため、日本語形態素解析システム「茶筌」による精度も示してある。茶筌は、コスト最小法を改良したアルゴリズムとヒューリスティックなルールによる未知語処理を使用した、高精度な形態素解析システムである。この結果を見ると、単語分割のみの場合も単語分割と品詞タグ付けの場合も、提案手法は茶筌と比較して特に未知語に対して高い再現率を得ていることが分かる。

・ 中国語形態素解析の評価

中国語の形態素解析精度を評価するために、Academia Sinica コーパス(AS)、Hong Kong City University コーパス(HK)、そして Beijing University コーパス(PK)の 3 つを使用した。これら 3 つのコーパスは、SIGHAN Chinese Word Segmentation Bakeoff 2003 (中国語単語分割の評価型ワークショップ)[文献 14]で使用されたものである。これら

のコーパスは単語分割はされているが品詞は付与されていないため、そのままではコスト最小法で利用することはできない。そこで、教師無し学習アルゴリズムである Baum-Welch 法[文献 13]を利用して、コーパス中の各単語に対して品詞の代わりとなる状態を付与した。解析精度の評価は日本語の場合と同様に行い、単語分割の精度のみを評価した。

表 5-3-2 に結果を示す。比較のため、SIGHAN Bakeoff に参加した上位 3 つの単語分割システム(Bakeoff-1、 2、 3)の精度も示してある。作成したシステムは未知語と既知語の再現率のバランスがとれており、全体的な性能の指標である F 値は最も高い値を得た。

表 5-3-2 中国語形態素解析の精度の比較

コーパス	システム	再現率	精度	F 値	未知語再現率	既知語再現率
AS	**提案手法**	**0．973**	**0．971**	**0．972**	**0．717**	**0．979**
	Bakeoff-1	0．966	0．956	0．961	0．364	0．980
	Bakeoff-2	0．961	0．958	0．959	0．729	0．966
	Bakeoff-3	0．944	0．945	0．945	0．574	0．952
HK	**提案手法**	**0．951**	**0．948**	**0．950**	**0．715**	**0．969**
	Bakeoff-1	0．947	0．934	0．940	0．625	0．972
	Bakeoff-2	0．940	0．908	0．924	0．415	0．980
	Bakeoff-3	0．917	0．915	0．916	0．670	0．936
PK	**提案手法**	**0．957**	**0．952**	**0．954**	**0．774**	**0．970**
	Bakeoff-1	0．962	0．940	0．951	0．724	0．979
	Bakeoff-2	0．955	0．938	0．947	0．680	0．976
	Bakeoff-3	0．955	0．938	0．946	0．647	0．977

(3) まとめ、今後の課題

本研究では中国語と日本語を解析するための形態素解析器について検討した。特に、文字単位のタグ付けをコスト最小法と組み合わせることで、言語に依存せずに未知語を処理する方法を検討し、中国語・日本語形態素解析システムを開発した。日本語と中国語のコーパスを使い評価を行なった結果、作成した形態素解析器は既存の手法以上の精度が得られることを確認した。今後の課題としては、この形態素解析器の高速化や英語への対応が挙げられる。

5-3-3 多言語翻訳データベースの研究

(1) 研究の内容

本研究は、文対応のついた対訳文書や翻訳テンプレートをデータベースに登録し、翻訳に利用するための技術に関する研究である。本研究は前年度までに基本仕様の検討が終わっているため、今年度は実際に対訳文書 DB ならびに翻訳テンプレート DB を作成した。対訳文書 DB については、ISO 標準文書、動画関連の国際標準文書、FCC(米国通信委員会)の規則文書、という 3 分野の英日の対訳文書 DB を構築した。また、翻訳テンプレート DB については、動画関連の国際標準文書に関する翻訳テンプレートを人手で作成し、その DB の構築を完了した。

また、作成したデータベースを利用した翻訳の枠組みを検討し、試作システムの実装に着手した。これは、既存の対訳文書とその改版文書から得られた翻訳テンプレー

トをデータベース化し、それを次回以降の翻訳に利用する改版文書翻訳システムで、「5-1-2 改版文書を利用した翻訳テンプレート獲得に関する研究開発」の成果を利用している。図 5-3-4 に示すように、この翻訳システムは、翻訳済みデータベース、改訂前後の原文の対応付け機能、翻訳機能、文書生成機能で構成される。

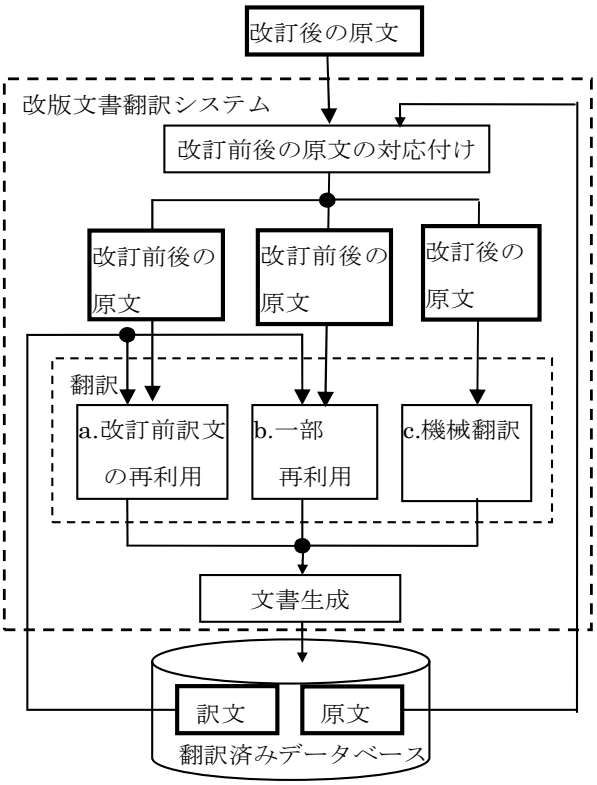


図 5-3-4 改版文書翻訳システムの図

翻訳済みデータベースには、対訳が構造解析された状態で登録されている。翻訳対象文を入力すると、まず改訂前後の原文の対応付け処理を行なう。改訂後の原文に対応している改訂前の原文が存在するならば、その訳文及び新旧原文の類似度等の情報が翻訳機能に渡される。翻訳は全て人手による翻訳でもよいが、改訂前後の原文の対応付け結果により、以下の３方式から選んで訳文を出力するという方式をとり、できるだけ旧版文書の訳文を再利用しながら機械翻訳を利用する方式を検討している。

1. 変更されなかった文は改訂前の訳文をそのまま再利用し出力する。
2. 部分変更されたと判定した文については、構文構造を比較し、変更されなかった部分を再利用し残りを機械翻訳する。
3. 追加と判定された文については１文全体を機械翻訳する。

最後に、生成した改訂後の原文と訳文の対訳データファイルが翻訳済みデータベースに格納され再改訂時に利用される。

(2) 研究の成果

多言語翻訳データベースの試作システムの構築を完了した。対訳文書 DB については、ISO 標準文書、動画関連の国際標準文書、FCC(米国通信委員会)の規則文書、という 3 分野の英日の対訳文書 DB を構築した。また、翻訳テンプレート DB については、動画

関連の国際標準文書に関する翻訳テンプレートを人手で作成し、その DB の構築を完了した。
また、改版文書翻訳システムの実装に着手した。現段階では単体で動作する状態であるが、最終的には、多言語標準文書処理システムの機能として組み込む予定である。

5-3-4 協調的翻訳支援環境の研究

（1） 研究の内容

多言語標準文書処理システムの重要な要素である、翻訳エンジン部、改版文書を利用した翻訳テンプレート作成部、対訳文書 DB、翻訳テンプレート DB は、それぞれ仕様検討が終了した。次の段階として、我々は、これらの機能を統合的に運用するための翻訳支援環境の構築に着手した。

協調型翻訳支援環境とは、多数のユーザが持つ翻訳知識を相互に利用可能にし、効率良く翻訳が行なえるようなしくみである。我々は、このような環境を構築する際の様々な課題を検討し、「訳してねっと」(図 5-3-5)上に実装した。

図 5-3-5 訳してねっとのトップページ

・ コミュニティ型機械翻訳サイト「訳してねっと」

「訳してねっと」では、様々な分野の辞書を作成することができる。これら一つ一つの分野はコミュニティと呼ばれている。コミュニティは図 5-3-6 のようにツリー構造をなしていて、下の層に行くほど細かいジャンルになっている。ユーザは、必要があれば新しいコミュニティを自由に作成することができる。それぞれのコミュニティに辞書がひとつずつ存在し、そのコミュニティのメンバーは翻訳テンプレートの追加、修正、削除、検索などを行なうことができる。

各コミュニティの管理は、そのコミュニティに参加しているユーザ主導で行われる。

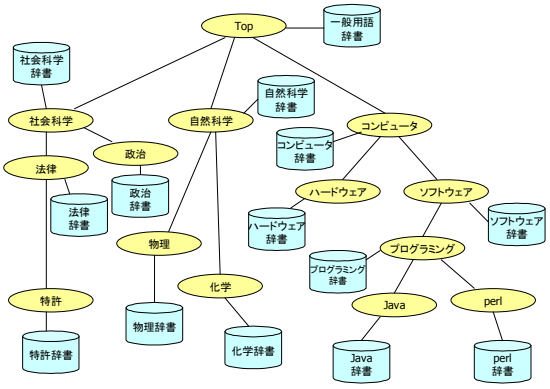


図 5-3-6 「訳してねっと」のコミュニティ構成

また､コミュニティリーダーと呼ばれるコミュニティのまとめ役をおくこともできる。コミュニティリーダーは他のメンバーが登録した辞書の内容の修正や、掲示板などに投稿された不適切な記事の削除などができる。

コミュニティ上で翻訳処理を行なうと、そのコミュニティの辞書、及びそのコミュニティの上位の辞書を使って翻訳される｡例えば Java コミュニティで翻訳する場合は、Java 辞書、プログラミング辞書、ソフトウェア辞書、コンピュータ辞書、一般用語辞書を使って翻訳するようになっている。

また、翻訳に使用する辞書を自由に選択することもでき、例えばコンピュータ関係の特許の翻訳を行なう場合は、特許特有の言いまわしなどが登録された特許辞書とその特許の分野であるコンピュータ辞書を使うといった指定もできる。

翻訳は英語から日本語と日本語から英語の２通りが可能で、テキストや Web ページや Microsoft Office のファイルなどが翻訳できる。

翻訳テンプレートの辞書への登録が行われると、すぐに翻訳用辞書の再構築が行われるので、登録した辞書データが翻訳結果に正しく反映されているかどうか、その場でチェックすることができる。よく翻訳するサイトや文書はあらかじめ登録しておくことができ、一度登録すると１クリックで翻訳できる。

・「訳してねっと」を使った協調型翻訳支援環境の研究

「訳してねっと」をベースに、協調的翻訳支援環境を実現するためのフレームワークの設計を行い、以下の８つの機能の基本開発を完了した。

1. ユーザ間の情報伝達手段としてのメッセージ機能
2. コミュニティ専用の辞書の登録・検索・更新機能（辞書機能）
3. 他のコミュニティの辞書参照機能
4. 文書や URL を共有知識として管理する文書管理機能
5. 各種文書の翻訳および翻訳結果の修正機能（翻訳機能、ポストエディット機能）
6. ユーザを限定したり、アクセス権を設定したりするユーザ管理機能
7. コミュニティに関する情報の多言語表示機能
8. サブコミュニティ作成機能

これらの機能を、従来から実験していたコミュニティ型機械翻訳サイト「訳してね

っと」に取り込み、上記のフレームワークに基づく新しい「訳してねっと」として、2003 年 10 月 31 日より、限定ユーザによる実験を開始した。また、実際のユーザからのフィードバックを得ることを目的に、2004 年 3 月 1 日より一般公開し、インターネット上で辞書構築の本格実証実験を開始している。

（2） 研究の成果

コミュニティ型機械翻訳サイト「訳してねっと」に研究成果を取り込み、2004 年 3 月 1 日に一般公開し、辞書構築の本格実証実験を開始した。これは今後の事業展開のコアとなるものである。

さらに、本研究開発の成果の一部分を利用した「会社で訳してねっと」という製品を開発し、販売開始のプレスリリースを 10 月 16 日に発表した。本製品は、上記フレームワークを企業用にカスタマイズし製品化したものである。本製品は、本研究開発のビジネス化において、テストマーケティングとしての役割を担っており、研究開発完了後の事業化を確実なものにするために大変重要なものである。本製品を販売することで市場からの反応を本研究開発に対してフィードバックさせ、より成果が大きくなるように研究開発を行なっていく予定である。

また、中国語の形態素解析を試作し、言語横断検索システムを試作した。2003 年 10 月～11 月に開催された第 4 回 NTCIR ワークショップ（評価型ワークショップ）において、CLIR（言語横断検索）の英語・日本語・中国語におけるすべてのサブタスクを実行し、特に英日検索において好成績を獲得した。さらに、中日翻訳のための基本辞書・基本文法を構築し、言語非依存翻訳エンジンが実際に多言語翻訳が行なえることを示すデモシステムの構築を行なった。

5-3-5　結論と今後の課題

今後は、「訳してねっと」公開により、運用上の、あるいは、ユーザビリティの問題点および改良点を明らかにし、システムの完成度を高める。また、中日翻訳システムの開発を通して、多言語拡張のための開発モデルを作成したい。

5-4　総括

上記のように、我々は、当初設定した３つのサブテーマ

ア．翻訳テンプレート学習に関する研究開発
イ．分野辞書の自己組織化に関する研究開発
ウ．言語非依存の翻訳エンジンの研究開発

に対して、さらに個別の研究課題に分け、各課題に対して取り組んだ。本年度の成果をまとめると以下のようになる。

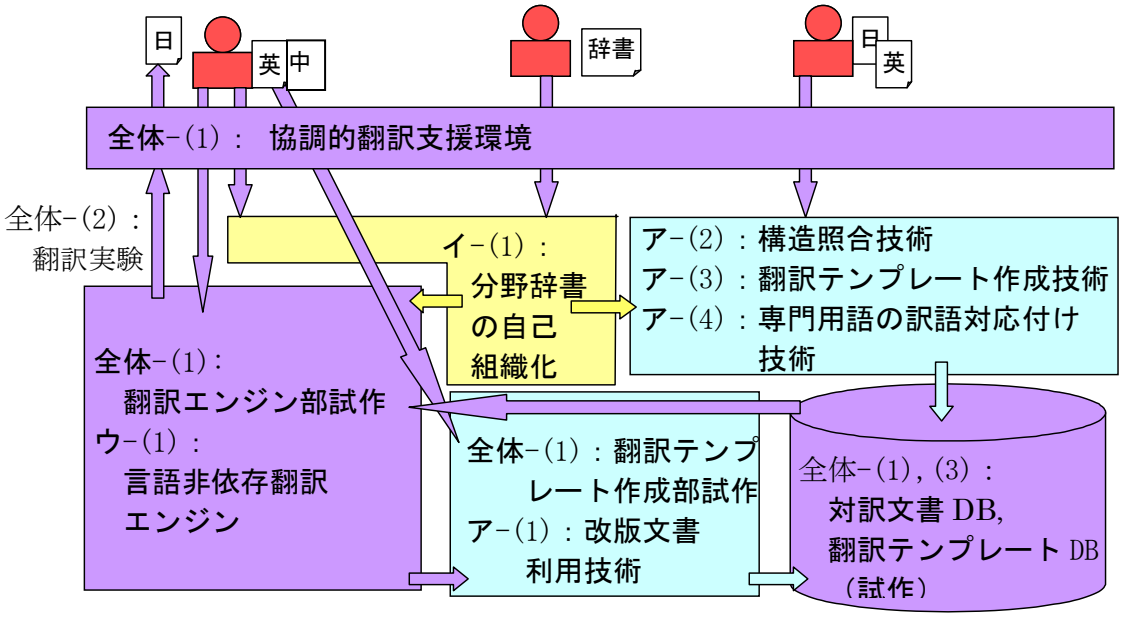
ア．翻訳テンプレート学習に関しては、改版文書、パラレルコーパス、コンパラブルコーパスという３つタイプの翻訳文書を利用した翻訳テンプレート学習の手法を検討し、実験により、それぞれ高い精度で翻訳テンプレートを抽出できることを確認した。

イ．分野辞書の自己組織化に関しては、「コアワード」を利用した分野の自動判定の手法を確立した。

ウ．言語非依存翻訳エンジンの研究では、コミュニティ型機械翻訳サイト「訳してねっと」に、研究成果を取り込み、2004 年 3 月 1 日に一般公開して辞書構築の本格実証実験を開始した。また、予定より早く、中日翻訳システムの実装に着手した。

以上のように、本年度は、それぞれの研究テーマが本格的に動き出し、具体的な成果が得られた。と同時に、あるテーマの研究成果を別のテーマの研究で利用するなど、各サブテーマの連携が図られ、最終目標を見据えた多言語標準文書処理システムの全体像が見えてきた（図 5-4-1)。

本年度獲得した成果物及び知見をもとに、今後は、翻訳品質および翻訳作業効率化の定量的評価手法を検討し、より実用的な機械翻訳システムの実現を目指して研究開発を進めていく予定である。



全体-(2)：翻訳テンプレート作成・DB 格納実験

図 5-4-1：各目標の関連および多言語標準文書処理システム全体像

参考資料、参考文献

[文献1] 潮田他: 「自動翻訳から翻訳支援へ、そして…」, 情報処理, 44巻, 9号, pp.931-939, 2003

[文献2] 「特集: NTCIR:情報アクセスにかかわるテキスト処理技術の評価ワークショップ」, 人工知能学会誌, Vol. 17, No. 3, pp.295-319, 2002

[文献3] 北村, 松本: 「対訳コーパスを利用した対訳表現の自動抽出」, 情報処理学会論文誌, Vol.38, No.4, pp.727-736, 1997

[文献4] 山本, 松本: 「統計的係り受け結果を用いた対訳表現抽出」, 情報処理学会論文誌, Vol.42, No.9, pp.2239-2247, 2001

[文献5] 内山, 井佐原: 「日英新聞記事対応付けと精度評価」, 情報処理学会研究報告, 2002-NL-151, pp.15-22, 2002

[資料6] Charniak, E.: "A Maximum-Entropy-Inspired Parser", Proceedings of NAACL-2000
その他 http://ftp.cs.brown.edu/pub/nlparser/ 参照

[資料7] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原 正幸: "日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書", Dec, 2000
その他 http://chasen.aist-nara.ac.jp/ 参照

[資料8] Kudo, T. and Matsumoto, Y.: "Fast Methods for Kernel-Based Text Analysis", Proceedings on ACL, 2003
その他 http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/ 参照

[文献9] Fung, P. and McKeown, K.: "Finding Terminology Translations from Non-parallel Corpora", Proceedings of 5th International Workshop of Very Large Corpora (WVLC-5), pp.192－202, 1997

[文献10] 神山, 伊藤: 「自律的語彙拡充を行なう機械翻訳システム」, 情報処理学会第65回全国大会, pp.2-5～2-6, 2003

[文献11] Berger, A. L., Pietra, S. A. D. and Pietra, V. J. D.: "A Maximum Entropy Approach to Natural Language Processing", *Computational*

Linguistics, Vol. 22, No. 1, pp. 39-71, 1996

[文献12] Darroch, J. and Ratcliff, D.: “Generalized iterative scaling for log-linear models”, *The annuals of Mathematical Statistics*, Vol. 43, No. 5, pp. 1470-1480, 1972

[文献13] Rabiner, Lawrence R. and Juang, Biing-Hwang: *Fundamentals of Speech Recognition*, PTR Prentice-Hall, 1993

[文献14] Sproat, R. and Emerson, T.: “The First International Chinese Word Segmentation Bakeoff”, *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pp. 133-143, 2003

[文献15] Xue, N.: “Chinese Word Segmentation as Character Tagging”, *Computational Linguistics and Chinese Language Processing*, Vol. 8, No. 1, pp. 29-48, 2003

[文献16] 吉田辰巳, 大竹清敬, 山本和英: 「サポートベクトルマシンを用いた中国語解析実験」, 自然言語処理, Vol. 10, No. 1, pp. 109-131, 2003

[文献17] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸: 「形態素解析システム『茶筌』version 2.2.9 使用説明書」, 奈良先端科学技術大学院大学松本研究室, 2002

（添付資料）

1　研究発表、講演、文献等一覧

ACL(Association for Computational Linguistics) 2003 において
“Collaborative Translation Environment `Yakushite.Net’” デモ展示

松永聡彦、北村美穂子、村田稔樹:
「改版文書翻訳システムにおける文脈を考慮した文対応付け手法」,
電子情報通信学会技術研究報告言語理解とコミュニケーション(NLC),
NLC2003-22, pp.43-48, 2003

北村美穂子:
「小さな対訳文書からの対訳表現の半自動抽出」,
FIT(情報科学技術フォーラム)2003 論文集, pp.87-88, 2003

佐々木美樹、北村美穂子、下畑さより、中川哲治:
「コアワードを利用した単語の分野自動判定」,
FIT(情報科学技術フォーラム)2003 論文集, pp.171-172, 2003

Toshiki Murata, Mihoko Kitamura, Tsuyoshi Fukui, and Tatsuya Sukehiro:
“Implementation of Collaborative Translation Environment `Yakushite Net’”,
MT(Machine Translation) Summit IX, pp.479-482, 2003

Mihoko Kitamura, Toshiki Murata:
“Practical Machine Translation System allowing Complex Patterns”,
MT(Machine Translation) Summit IX, pp.232-239, 2003

Mihoko Kitamura and Yuji Matsumoto:

“Practical Translation Pattern Acquisition from Combined Language Resources”,
First International Joint Conference on Natural Language Processing (IJCNLP-04),
pp.652-659, 2003

Mihoko Kitamura, Tetsuji Nakagawa, Seika Kim, Toshiki Murata:
“Development of Chinese-Japanese MT System based on Language-Independent Translation Engine”,
Asian Symposium on Natural Language Processing to Overcome Language Barriers,
pp.39-45, 2003