平成１４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

１　研究計画の背景

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、専門辞書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学習し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、また、多言語間にまたがった対訳文書は増大する一方である。インターネットにアクセスする様々な知識を有する人々が既存の翻訳結果を利用して、機械翻訳の翻訳品質を高める技術は、今まででは不可能とされていた多種多様な分野における機械翻訳の適用に役立ち､将来的には､言語を超えた翻訳知識の抽出かつ多言語機械翻訳にまで発展すると期待される。

２　研究開発分野の現状

昨今のグローバル化の波を受けて、多言語・翻訳関係のプロジェクトがさまざまな領域で行なわれ始めた。特にわれわれの研究に関連があるものとして、機械翻訳をコミュニケーション手段として捉えて実際に実験を行っている研究と、多言語のリソースを収集して将来に活かすことを目的とした研究があるので、その２つを以下に記す。

(1) 異文化コラボレーション Intercultural Collaboration Experiment(ICE)2002 実証実験

京都大学、NTT、科学技術振興事業団戦略的基礎研究推進事業(CREST)が主催して行っている実証実験。2002 年度は、京都大学, 上海交通大学(中国), ソウル国立大学およびハンドン大学(韓国), そしてマラヤ大学(マレーシア)がその教員・学生総勢 40 名が参加した。本実験では、ソフトウェアを各国の学生が協力して作成することをゴールとし、そのコミュニケーションツールとして機械翻訳を用いてその有用性を確かめた。その結果、多少の問題があるにせよ機械翻訳は有用だったという結果が出ている。今後の課題としては学習機能を挙げており、我々の研究が大変有益であること

を示している。

(2) 多言語同時処理研究

大阪外国語大学、大阪府立産業技術総合研究所、松下電器産業株式会社先端技術研究所が共同で進めている科研費プロジェクト。大阪外国語大学に所属する各言語の教員リソースを利用して、アジア諸言語の文字・音声データの蓄積および応用研究を進めている。データの蓄積のためのプラットフォーム作りやそれを利用した音声翻訳システム、携帯端末による辞書検索システムを構築している。また、多言語データベース検索も進めており、我々の文書データベースと関連が深い。

３　研究開発の全体計画

3-1　研究開発課題の概要

3-1-1　研究開発全体の概要

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１に示す。

3-1-2　各サブテーマの概要

(1) 翻訳テンプレート自動抽出システムの研究開発

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。以下、2つの手法を利用してこれらの技術を研究開発する。

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、

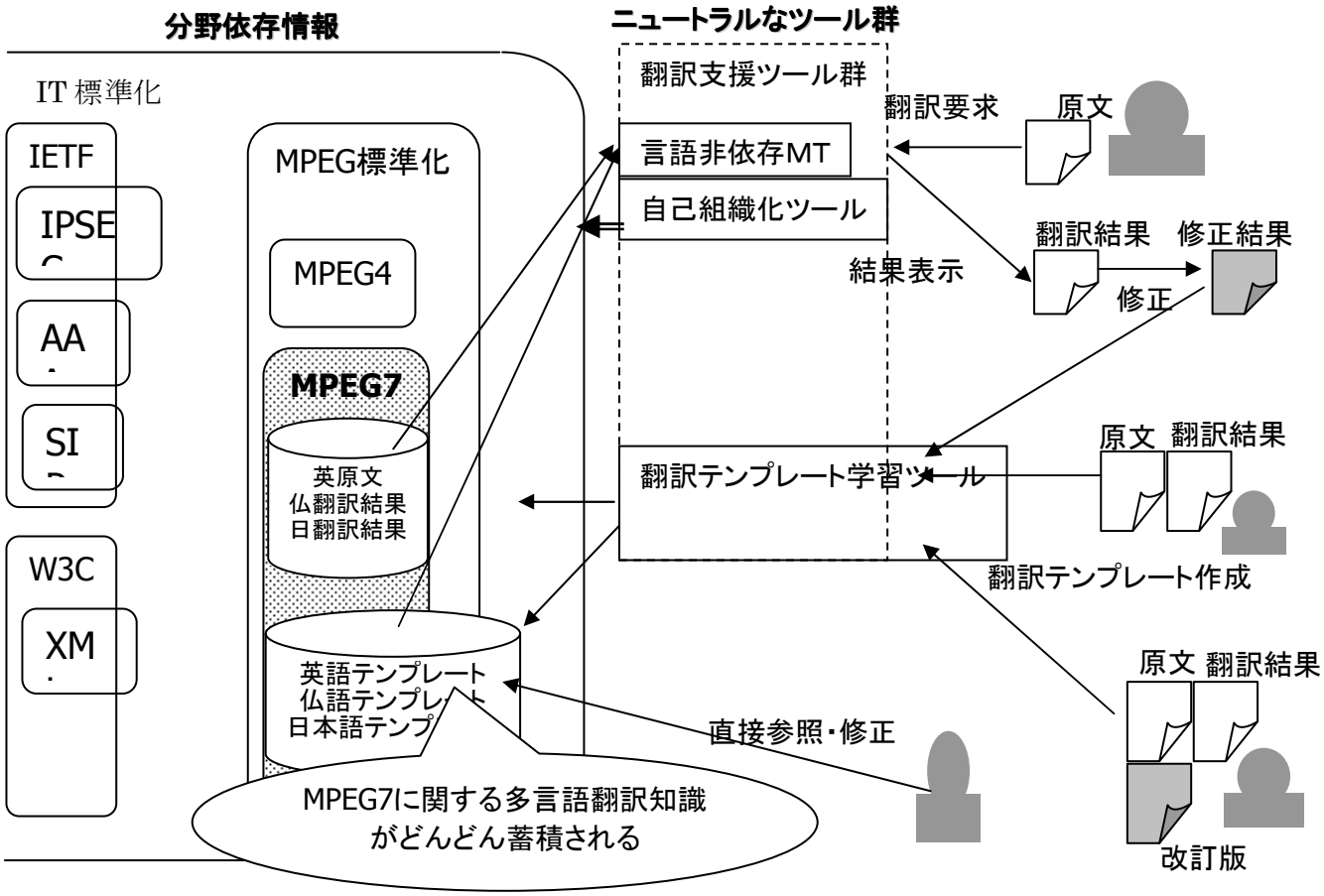


図1　多言語標準文書処理システムの構成図及びユーザによる利用形態

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。

しかし、本手法は、対訳文の両方の構文解析結果が必要となるため、構文解析ツールが存在しない言語に関しては、本手法をそのまま利用することはできない。構文解析ツールが存在しない言語に関しては、既存の機械翻訳で解析された翻訳テンプレート群を構文解析結果と見立てて構造照合する方法や、以下に述べる統計的手法と本手法を融合した方法(例えば、統計的学習によって簡易的な係り受け解析結果を求め、その結果を構造照合する等の方法)が必要となるが、その点が本提案の研究課題となる。

また、国際標準化に関する文書や機器のマニュアル等は、版を改訂していくことが頻繁に行われる。改訂前と改訂後の差分を構造照合によって検出することにより、改訂された部分に関する翻訳テンプレートを自動的に発見する技術が応用として考えられる。さらに、ユーザが機械翻訳した結果が気にいらない場合翻訳結果を修正するが、その機械翻訳結果と修正結果の差分を構造照合によって検出することにより、修正された部分に関する翻訳テンプレートを自動的に発見する技術にも応用することができると考える。

イ．統計的手法を利用する方法

構造照合を利用する方法によって原文のどの部分が翻訳結果のどの部分に対応付けられるかが検出される。次に、対応付けられた部分の中でどの部分を汎化(変数化)するかを決めるために統計

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か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利用した方法を利用することを考える。

世の中には、共通の内容を有するが原文と翻訳結果の文の対応がとれていない対訳文書も数多く存在する。そのような対訳文書に対しては、統計的手法を利用して、原文と翻訳結果の各々から専門用語を抽出し、専門用語同士を対応付けることにより、専門用語の翻訳テンプレートを作成するという手法についても研究開発する。

(2) 翻訳テンプレートの自己組織化の研究開発

(1)の手法によって獲得できる翻訳テンプレートは、獲得元の文書の改訂文書や獲得元文書に関連する文書の翻訳に利用できるのは当然であるが、その翻訳テンプレートがより一般的な表現に関するテンプレートであれば、別の文書の翻訳にも利用できる可能性がある。その一方で、獲得した翻訳テンプレートが全ての文書で適用できるわけではなく、翻訳結果に悪影響を及ぼす翻訳テンプレートも存在する。獲得した翻訳テンプレートに対して汎用的に利用できるものは汎用的に利用し、限定的に利用すべきものは、ある文書のみに適用すべき翻訳テンプレートとして利用を制限するようなしくみが存在することが望ましい。

この目的を達成するために、獲得した翻訳テンプレートを文書の種類、用途、専門性に応じて、自動分類、自動階層化する技術（ここでは自己組織化技術と呼ぶ）を研究開発する。それにより、目的文書の翻訳に最適な翻訳テンプレートの選択が可能となる。

方針として、翻訳テンプレートの文書中の出現頻度、出現の多様度(ごく限られた文書に集中的に出現するか多岐に渡る文書に出現するか)、および他の翻訳テンプレートの適用に与える影響度、テンプレートの構成品詞などの情報を統計的に分析し、予め人間の手で分類・階層化されている辞書に登録する方法をまず考案し、次に、分類の大きさや階層化のレベルも自動推定し、自己組織化する技術を研究開発する。

(3) 言語非依存型機械翻訳システムの研究開発

上記(1)の方法によって、獲得した翻訳テンプレートを辞書として利用する機械翻訳システムを研究開発する。本システムの構成図を図２に示す。本システムは、言語に非依存の翻訳エンジン部と、翻訳知識を格納する言語依存の翻訳辞書DBから構成される。後者は対訳文書DBと翻訳テンプレートDBから構成する。翻訳文書DBは、対訳文書そのものが蓄積されているDBである。翻訳テンプレートDBは、(1)で獲得した翻訳テンプレートが格納されているDBである。

翻訳のプロセスは、まず、対訳文書DBのみで翻訳処理を試し、解析が成功しなければ翻訳テンプレートDBを利用して翻訳処理を実行する。それでも解析が成功しなければ、汎用辞書や文法を用いて翻訳するか、汎用辞書や文法が存在しない場合は、翻訳テンプレートが存在しない原文の部分をユーザに表示して、ユーザにテンプレートの作成を要求する。

一方、翻訳結果が間違っている場合は、ユーザが修正し、その後、修正結果から(1)の方法を利用して翻訳テンプレートを再学習することもできるようにする。

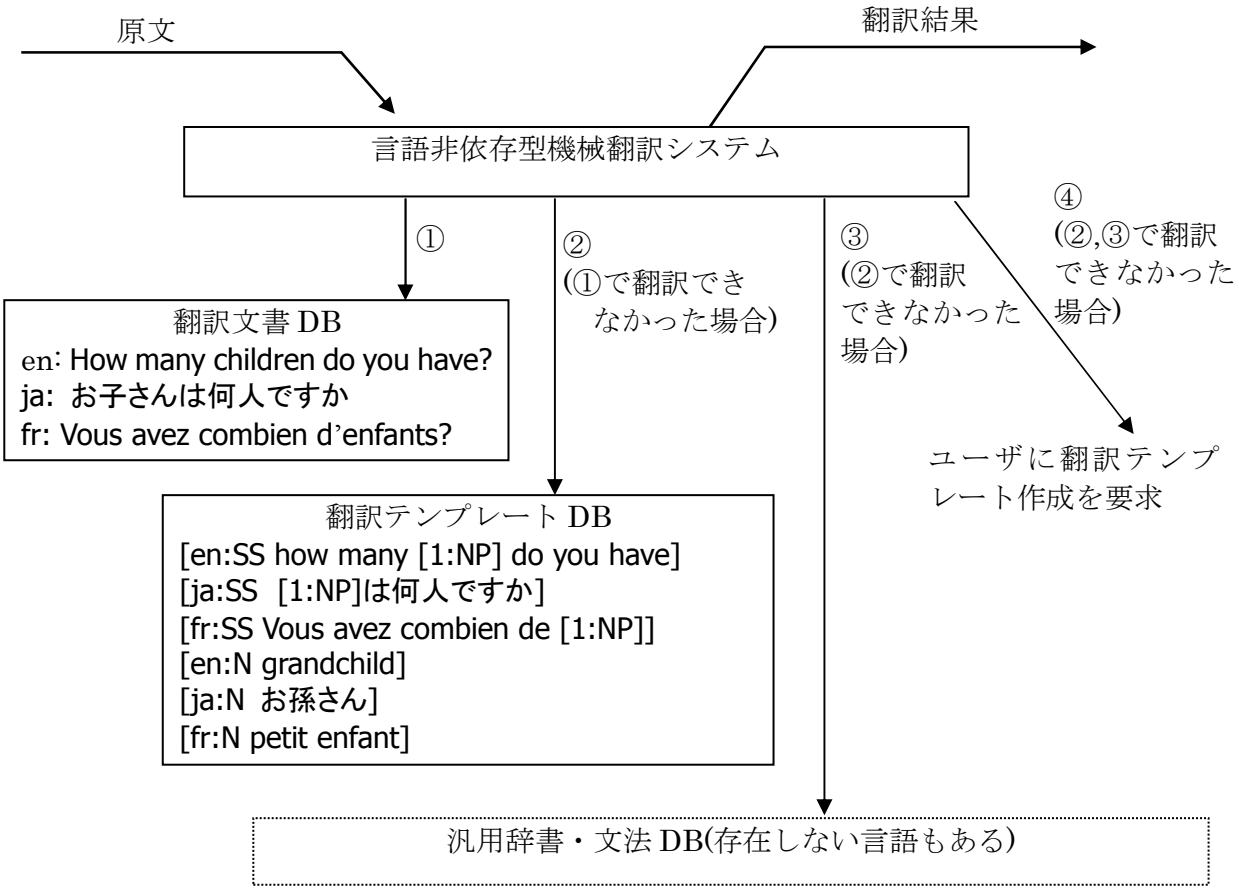


図2　言語非依存型機械翻訳システムの構成図

3-2　研究開発目標

3-2-1　最終目標（平成 18 年 3 月末）

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で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。
   1. 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。
   2. 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。
   3. 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（平成 16 年 3 月末）

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
(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。

3-3　研究開発の年度別計画

（金額は非公表）

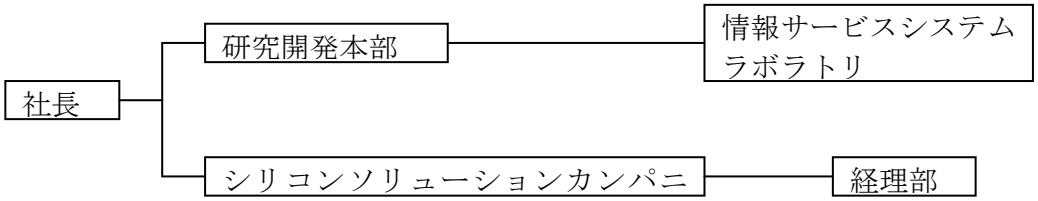
研究開発項目	14 年度	15 年度	16 年度	17 年度	年度	計	備　考
多言語標準文書処理システムの研究開発 ア．翻訳テンプレート自動学習の研究開発 ・構造照合型テンプレート自動学習システムの開発 ・統計的手法型テンプレート自動学習システムの開発 イ．翻訳テンプレートの自己組織化の研究開発 ・分類されたものへの選択手法の開発 ・自己組織化システムの開発 ウ．言語非依存型機械翻訳システムの研究開発 ・翻訳エンジンの開発 ・翻訳知識 DB の開発							
間接経費							
合　　計							

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。

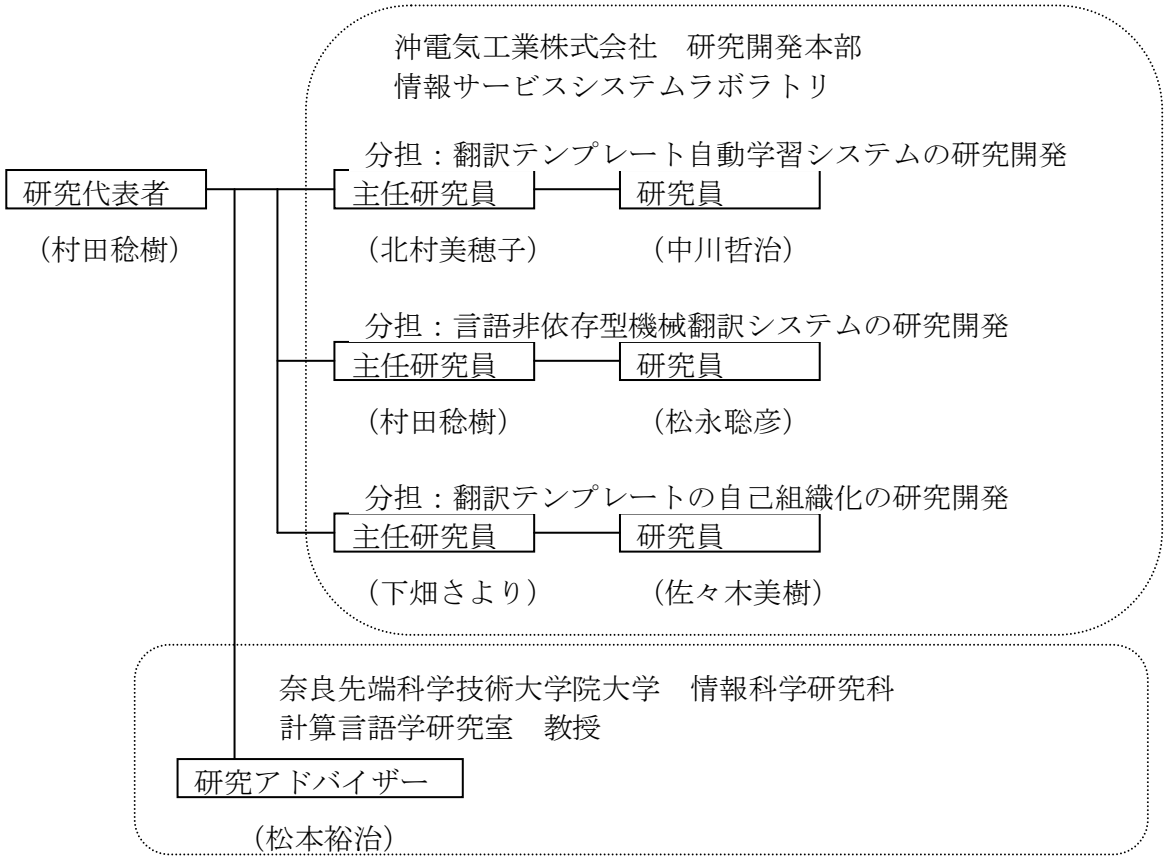
3-4　研究開発体制（平成１４年度）

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１４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、及び、その改版文書を入力手段とする翻訳テンプレート作成システムの仕様が確立していること。
既存の対訳文書と改版文書の差分を検出しその差分情報と構造照合技術から翻訳テンプレートを作成する手法を考案する。多くの実例を理論的にシミュレーションすることにより、この手法を確立し、翻訳テンプレート作成のシステム仕様を構築する。

(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。
構造照合ツールを試作し、試作したツールについて実例による実験及び検証を行い、試作したツールの精度を高める。

(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が確立していること。
(2)の対訳の対応付け結果の汎化手法を考案する。さらに、汎化された対応付けの結果を翻訳テンプレートとして利用するための方法を定式化する。対訳例

文数文に対して、翻訳テンプレート作成までの手順をシミュレーションし、本手法の有効性を実証する。

(4) 文対応がついていない対訳文書について、専門用語を対応付ける手法を確立し、専門用語対応付けの精度上位5位以内80%以上で実現されていること。

専門用語対応付け手法を考案・試作し、試作したツールについて実例による実験及び検証を行い、試作したツールの精度を高める。

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) 新しく作成した翻訳テンプレートを最適な分類･階層のDBに格納できる手法の確立。

自己組織化技術のための分類・階層化学習用データ(翻訳テンプレートDB)を人手で作成し、新しく作成した翻訳テンプレートを最適な分類・階層のDBへ格納できる手法を考案する。実際の例を想定したシミュレーションを行い、精度が上位5位以内80%以上で実現できることを実証する。

ウ．言語非依存の翻訳エンジンの研究開発

(1) 言語非依存の翻訳エンジンの設計を完了すること。

翻訳エンジンにおいて言語非依存のモジュール(形態素解析処理部及び形態素生成処理部)を設計し、仕様書を作成する。

(2) 国際標準等、1 分野の対訳文書 DB 及び翻訳テンプレート DB の構築に着手していること。

国際標準等の翻訳文書及びその多言語化について調査し、調査結果及び上記ア.イ.の仕様検討に基づいて、対訳文書 DB 及び翻訳テンプレート DB を設計し、構築に着手する。

(3) 国際標準等の翻訳プロセスを効率化するための協調的翻訳支援環境の基本開発を完了し、簡単な翻訳プロセスのデモができること。

国際標準等の翻訳作業を調査し、翻訳プロセスをシステム化するために必要なツール、モジュール群を開発する。さらに、そのツールやモジュールを用いて翻訳プロセスが効率化されることを実証する。

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 ・改版文書利用型翻訳テンプレート作成 ・構造照合による対訳対応付け ・統計的手法による対訳汎化技術 ・専門用語の対訳対応付け イ．翻訳テンプレート自己組織化の研究開発 ・翻訳テンプレート自動格納技術 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 ・翻訳環境の開発			→	→		
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った。

(1) 既存の対訳文書、及び、その改版文書を入力手段とする翻訳テンプレート作成システムの仕様を確立した。

(2) 構造照合技術を利用して、対訳の対応付けが精度 80%を実現した。

(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。

(4) 文対応がついていない対訳文書について、専門用語を対応付ける手法を確立し、専門用語対応付けの精度上位 5 位以内 80%を実現した。

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った。

(1) 新しく作成した翻訳テンプレートを最適な分類・階層の DB に格納できる手法を確立し、上位 5 位で 88%の精度が得られることがわかった。

ウ．言語非依存の翻訳エンジンの研究開発を行った。

(1) 言語非依存の翻訳エンジンの設計を完了した。

(2) 国際標準等､3 分野の対訳文書DB及び翻訳テンプレートDBの構築に着手した。

(3) 国際標準等の翻訳プロセスを効率化するための協調的翻訳支援環境の基本開発を完了し、簡単な翻訳プロセスのデモを行った。

5　研究開発実施状況（平成 14 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

文書の電子管理が日常化した現在､保存された文書から言語知識を自動獲得する技術や､保存された文書を利用して言語処理を行うための技術が注目されている。
　その中でも翻訳知識（例えば、各言語の文法規則や他の言語への変換規則など）を扱う研究は

(a) 体系化、規則化が難しい
(b) 世界中の数多くの言語対において翻訳要求がある

という理由から、翻訳知識の自動獲得や翻訳用例を用いた機械翻訳システムの研究開発が強く望まれている。

このような要求に応えるために、本サブテーマでは、性質が異なる 3 種類の翻訳済文書（以下太字で示す）を用いた以下の研究開発を行っている。本年度は、各方式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。なお、本年度は、英日・日英の対訳文書を題材としたが、各試作システムの基本アルゴリズムは言語非依存であり、他言語への適用を考慮したものとなっている。

ア. 改版文書利用型翻訳 : **翻訳済文書が存在する標準文書やマニュアル等の文書**において、原本が改版された場合に改版前の翻訳済文書の知識を利用して機械翻訳を行う方式

イ. 文対応付き対訳文書からの翻訳テンプレートの抽出 : **文対応がついた対訳文書**から統計的手法や構造照合手法を用いて翻訳テンプレート（パターンベース機械翻訳の辞書）を自動作成する手法

ウ. 文対応なし対訳文書からの専門用語対訳辞書の抽出 : **文対応がついていない対訳文書**（完全な対訳でなくても構わない）から統計的手法を用いて、二言語間での専門用語を自動的に対応付ける手法

5-1-2　改版文書利用型翻訳の研究

(1) 研究の内容

本年度は、改版文書利用型翻訳の方式について検討し、仕様策定を行った。また、改版文書利用型翻訳方式のコア技術となる“原文書と改版文書の文対応付けシステム”の試作、実験を行った。

・改版文書利用型翻訳の基本方式

翻訳を行おうとする改訂後の原文書に対し、その改定前の翻訳済文書データベースを準備する。まず、改訂前の原文書と改訂後の原文書の差分情報の検出処理を行う。次に、改訂後の原文書の各文に対して、検出された差分情報から

(a) 改訂前の翻訳結果を直接利用する
(b) 改訂前の翻訳結果を部分的に利用し、変更された部分のみ機械翻訳で翻訳する
(c) 一文を機械翻訳システムで翻訳する

のいずれかの翻訳方式を選択し、その翻訳方式を用いて翻訳結果を出力する。最後に、そこで作成された新しい改版翻訳済文書から、翻訳済文書データベースを再生成する。

上記の処理は、過去の翻訳済文書を既存の翻訳知識として利用して翻訳する一方で、そこで作成された翻訳済文書を新しい翻訳知識として格納するという方式に見るように、翻訳知識の再利用を主眼においた方式となっている。

この改版文書利用型翻訳の要素技術の中でも、改版前の原文書と改版後の原文書間の文レベルの対応付け技術は、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。したがって、今年度はこの文対応付け方式のシステム試作を行っ

た。

・原文書と改版文書の文対応付け方式

改版前と改版後の文書中の文を対応付ける最も簡単な方法として、改版前の原文書を一文毎に区切り、各文をデータベースに格納しておき、改版後の原文書の各文をデータベースから検索するという方法が挙げられる。しかし、その方法では、同一文を対応付けるのは容易であるが、類似文の対応付けは、検索範囲を限定しないとその判断が難しい。したがって、単なる検索だけではなく、その文が原文中のどの章に位置するか、また、その前後の文脈にどれだけ一致しているか等の情報を鑑みながら、文の一致度を判断することが望ましい。

また、文書の章や段落単位に文の対応付けを行う手法もあるが[文献 1]、この場合も章全体が別の章に移動したり、章全体が削除されたりした場合、差分を検知することができないという課題がある。

上記の課題の解決方法として、章、段落や文の前後関係を考慮した以下の文の対応付け方式を考案した。

まず、改訂前と改訂後の文書を、章、節、段落の各レベルでブロック単位に区切り、各ブロック単位で改訂前文書と改訂後文書の対応付けを行う。各ブロック単位の対応付けは、文の一致数、タイトル名の一致度、対応するブロックの連続度等から求めるブロック類似度をもとに行う。類似度の高い対応付けを優先的に決定し、決定された対応付けの結果を利用して、類似度の低い対応付けを行う。対応の付かなかったブロックは対応付けを保留し、大きなブロックから小さいブロックへと対応付けの範囲を徐々に狭めていく。最も小さなブロック単位（段落単位）の対応付けの終了後、その段落内での文単位の対応付けを行う。

文単位の対応付けは、形態素数の一致度及び対応する文の連続度をもとにした文類似度をもとに行う。対応する文が存在しない場合は、新しく追加された文と認定する。また、ブロックの対応付けで保留されていた未対応のブロックに関しては、ここで、文単位に分割し、文レベルの対応付け処理で対応する文を探す。

・実験とその結果

文対応付け方式の動作検証をするために、システムを試作し、ＩＳＯ／ＩＴＵの国際標準化に関する草案文書及びその改定後の文書で実験を行った。その結果、改版後に良く見られる文書構成の変更によって章単位で文が移動した場合や、新しく章が追加された場合でも、適切に差分を検知することができることを確認した。

(2) 研究の効果

① 改版文書利用型翻訳方式の仕様を策定した。本仕様に基づいた改版文書利用型翻訳システムを構築することで、翻訳作業の効率化を図ることができる。

② 改版文書利用型翻訳方式のコア技術となる差分検出システムを試作し、実験の結果、文脈や章の構成を考慮した高精度の差分検知ができることを確かめた。本差分検知方式を搭載することで、改版文書利用型翻訳の精度の向上が期待できる。

5-1-3　構造照合技術と統計的手法の融合による翻訳知識獲得の研究

(1) 研究の内容

構造照合技術と統計的手法を融合した構造照合ツールを試作・実験し、その有効性を確かめた。また、統計的手法から得られる情報を用いて、構造照合結果から翻訳テンプレートを作成する手法の基本設計を行った。

構造照合技術[文献 2]とは、対訳文を各言語で構文解析し、構文解析結果に対し、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技術である。一方、統計的手法[文献 3]は、大量の文対応済の対訳文書での同時出現頻度から、言語間の単語列の対応付けを行うという手法である。

両者の手法の融合を考えた場合、1 つの課題がある。それは、構造照合技術は対訳文が 1 文でも良いのに対して、統計的手法は 1 万文以上の文を必要とする点である。実用化を考えた場合、統計的手法に必要な文数はより少ないことが望ましい。この問題を解決するために「文数の少ない文書における対訳表現の自動獲得」の研究開発を行い、その実験を行った。

次に、上記の統計的手法と構造照合技術を融合した「構造照合ツールの試作」を行い、その実験を行った。

さらに、構造照合ツールの結果を翻訳テンプレートとして利用するための「構造照合結果からの翻訳テンプレート作成手法の基本設計」を行った。それぞれの詳細について以下に示す。

・文数の少ない文書における対訳表現の自動獲得

文数の少ない対訳文書でも精度良く対訳表現を抽出するために、従来手法に対して

(a) 係り受け解析結果による句単位の情報を利用する
(b) 既存の対訳辞書を利用する
(c) 人手によって抽出過程途中の対訳表現をチェックする

という 3 つの改良を行い、その評価実験を行った。

対訳文 1,000 文をもつ対訳文書を利用した実験では、対訳文書中に 1 回しか出現しなかった対訳表現を抽出した場合でも、84%の精度と 78%のカバレッジで抽出できることが確認され、新手法の有効性が確認された（従来手法では精度は 12%であった。カバレッジは求めていないが新手法と同程度と予測される)。

・統計的係り受け解析結果を利用した構造照合ツールの試作

従来の構造照合手法は、複文や重文のような複数の動詞を持つ対訳文を対象にすることはできない。これは、本手法の実用化を考えた場合、大きな課題となる。この要因として、構文解析時での失敗と構造照合の計算コストの高さ、の問題が挙げられる。前者の問題は、構文解析に用いられる解析装置を翻訳で用いるような精密な解析装置でなく、統計手法による係り受け解析装置を用いることで、構文解析時の失敗をゼロにした。一方、後者の問題は、構造照合のアルゴリズムをトップダウンの探索手法から、統計的手法により抽出した対訳表現を手がかりとしたボトムアップの探索手法に変更することによって、低コスト化を図った。

これら 2 つの改良を試みた結果、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された。

・構造照合結果からの翻訳テンプレート作成手法の基本設計
構造照合結果を機械翻訳で用いる翻訳テンプレートとして利用するためには、構造照合結果を複数の部分木に分解し、さらに、語彙の一部を変数化した部分木をテンプレートの形式に変換する必要がある。本処理の自動化ツールの基本設計を行った。
設計手法は、まず、木構造の状態と品詞情報を利用して、構造照合結果を部分木単位に分解する。次に、各部分木に対して、その部分木を構成する各語彙の意味、品詞、及び、統計的手法で得られた語彙対の確信度により、変数化（汎化）する語彙を決定する。最後に、その部分木の主辞の品詞から、翻訳テンプレートの品詞を決定した後、翻訳テンプレートに変換する。机上シミュレーションによって本手法の有効性を確かめた。

(2) 研究の効果

① 従来技術に３点の改良を施すことにより、対訳文数が少ない対訳文書でも統計的手法を利用して対訳表現を精度良く抽出できることを確かめた。従来技術では、１万文以上の対訳文書が必要で、その実用化は収集文数が課題になっていたが、１千文程度の文書でも抽出可能になったことにより、統計的手法を利用した対訳表現抽出の実用化の見通しがついた。

② 統計的手法と構造照合技術を融合した試作システムを開発した。本試作システムは、従来技術では不可能であった長文に対しても構造照合することができ、高い精度で構造照合することができる。

③ 構造照合結果から翻訳テンプレートを自動作成するためのツールの基本仕様書を作成した。これにより、試作システムの開発を早期に着手することができる。

5-1-4　文対応のついていない対訳文書からの専門用語の自動対応付けに関する研究

(1) 研究の内容

既存の翻訳済文書から対訳表現を抽出する研究には、5-1-3 で説明したような既に文の対応がついた文書を対象とする研究の他に、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研究がある。後者の手法は、抽出対象となる文書が多数存在するので、前者の手法に比べて実用的な手法であるが、従来手法では、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［文献 4］（以下、Fung の手法と呼ぶ）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。)

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、次の手法を考案した。

・提案手法
我々は、Fung の手法に対して

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

の２つの改良を行った。
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。

・実験とその結果
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では、以下のような結果を確認することができた。

(a) 上位５位以内に正解の対訳表現が含まれていた割合は 82%であった。
(b) 従来手法と比べて、1.2 倍ほど多くの対訳表現を抽出することができた。
(ただし、単語リストに追加したのは、獲得された対訳表現のうち正解の表現のみ(1 位とは限らない)である。)

(2) 研究の効果

① 内容や分野が同じという制約しかない対訳文書から、既存の研究より高い精度で、対訳表現を自動抽出するという技術を考案し、試作システムによる実験結果から本手法の有効性を確かめることができた。提案手法は、本方式の実用化のための大きなステップとなる。

5-1-5　結論と今後の課題

本サブテーマは、上述した通り、当初予定した目標を達成することができた。各項目別に結論と今後の課題を以下に示す。

・改版文書利用型翻訳

改版文書利用型翻訳の仕様を検討し、本翻訳の処理で中心となる改定後、改定前の文書の差分検出方式を試作し、実験を行った結果、本方式の有効性を確かめた。この検証結果により改版文書利用型翻訳の仕様を構築した。

今後は、本仕様に基づき、改版文書利用型翻訳の試作システムを完成させ、翻訳方式の有効性の評価を行う予定である。また、構造照合技術との融合を鑑みて、検出した差分情報を元にした翻訳テンプレートの自動作成手法についても検討を開始する予定である。

・構造照合技術と統計的手法の融合による翻訳知識獲得の研究

対訳文数が少ない対訳文書からの対訳表現の統計的自動抽出手法を確立し、新しい構造照合試作システムを開発し、本システムの有効性を検証した。また、本構造照合結果を利用した翻訳テンプレート自動作成システムの基本設計を終了した。

今後は、本システムを試作し、対訳文書から翻訳テンプレートを自動的に作成する一連のシステムの統合を計画している。

・文対応がない対訳文書からの専門用語の自動対応付けに関する研究

文対応がない対訳文書からの専門用語の自動対応付け手法を考案し、限定された条件での小規模な実験ではあったが、精度の高い結果が得られた。

今後は、文書の量を増やし、実用レベルで本手法が有効であることを実証することを予定している。また、単語リストの重み付けから対訳表現の再帰的な抽出のプロセスについても、完全に自動化し、処理時間の短縮を図る予定である。

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。

本年度は、このような目的を達成するために必要な要素技術を検討し、研究方針を確立した。以下に、本年度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。

本年度は、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。

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

ある語の分野を自動的に判定する従来研究には、各分野の単語リストを分野判定済文書からあらかじめ作成しておき、その単語リストの出現回数を利用して分野を判定する方法がある[文献 5]。

本研究では、単語リストをあらかじめ作成するのではなく、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、分類したい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分類したい語を自動的に分類する。以下に、コアワードの自動作成とコアワードを利用した自動分類とに分けて、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。

・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。今回は、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 の値を用いて分野関連度を計算した。

・コアワードを利用した分野の自動判定

ある語の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、分野判定したい語と共起す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

みの文書を利用しても構わない)。
　分野判定したい語を、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を、コアワードに付与されている分野関連度の値が大きい順に順位付けをする。最後に、最も大きい値を持つコアワードが属する分野を、その語の分野として判定する。
　なお、ここで用いる分野関連度は、コアワードの自動作成で得られた値に対し、次の2種類の補正

(a) その分野が保有している全てのコアワードの総出現数で tf*idf の値を割る
(b) コアワードの出現回数を分野関連度に掛ける

を施した。(a)は、tf*idf の値が、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を解消するためである。また、分野判定対象語と共起するコアワードの出現回数が多ければ多いほど、分野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。このため、(b)のコアワードの出現回数を分野関連度に掛けることでこの問題を解消した。

・実験とその結果
我々が現在開発中の「訳してねっと」[資料 6][文献 7]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。分野判定したい語を検索する文書は、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。

(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」を利用した分野判定方式は、未分類の文書 (「コアワード検索用文書」と呼ぶ) を準備するだけで精度の向上が図れる等、数多くの利点がある。今後の自己組織化の基礎技術としての利用が期待できる。

5-2-3　結論と今後の課題

本サブテーマは、上述した通り、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、階層化された分野における分野自動判定方式の設計、及び、訳語の情報を利用した分野自動判定方式の設計である。また、現在の実験では、分野判定語として単語を利用しているが、句や言い回しなどの表現も利用し、実用化に向けての検討も行う予定である。

5-3　言語非依存の翻訳エンジンの研究開発

5-3-1　序論

言語非依存の翻訳エンジンの研究は、本研究テーマの実用化に必要不可欠な技術である。文書からの言語知識の獲得に関する研究によって得られた多言語翻訳に必要な翻訳テンプレートを、分野辞書の自己組織化に関する研究によって適切な分野ごとに蓄積し、それらを用いて実際に翻訳を行うのが本研究開発である。

翻訳を行うためには、まず翻訳エンジンが必要である。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することにより、指定された言語間の翻訳を行う。したがって、翻訳エンジン自体、言語に非依存である必要がある。言語に非依存であるという意味は、翻訳エンジンのプログラム部分を変えることなく、翻訳テンプレートを切り替えるだけでさまざまな言語対の翻訳が行えるということである。

次に、翻訳エンジンが用いる多言語の文書をどのような形で多言語翻訳データベースに格納すればよいかを定める必要がある。翻訳や知識獲得にとって効率的かつ必要な格納方法およびデータはどんなものかを研究する。

翻訳エンジンおよび多言語翻訳データベースを用いて、翻訳プロセス全体を効率良く行うのが、協調的翻訳支援環境である。国際標準等の技術文書の翻訳には、標準を議論して策定する国際標準化委員、および、各国内の当該技術分野に精通する技術者、各国語用に翻訳する翻訳者、翻訳プロセスを管理・運用するコーディネータなど複数のグループが協調して作業を行う必要がある。これを実現するのが協調的翻訳支援環境である。この環境は、ネットワーク上から簡単に利用することができ、必要なデータを効率良く蓄積し、参加者間のコミュニケーションを取りながら、翻訳作業を支援する必要がある。

上記に述べた要求に応えるために、本サブテーマでは、以下の３つのテーマにさらに区分し、研究開発を行った。

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。特に言語非依存形態素解析システムについての方式検討

イ．多言語翻訳データベース：データベースのおおまかな設計、および、データベース構築の着手

ウ．協調的翻訳支援環境：支援環境の全体設計および基本開発、簡単な翻訳プロセスのデモ

5-3-2　言語非依存形態素解析システムの研究

(1) 研究の内容

多言語形態素解析器を実現するために必要な確率モデルについて検討し、複数の n-gram モデルを結合する手法に基づいたシステムを試作・評価し、この手法の有効性を確かめた。

形態素解析において、形態素の区切りの同定と品詞決定の際に生じる曖昧性の解消は大きな課題である。この曖昧性の解消に関して従来から多くの方法が試みられており、特に近年は最大エントロピー法を用いた手法[文献8]などのコーパスに基づいた手法が盛んに研究され、様々な言語の形態素解析に適用されている。

コーパスに基づいた手法の一つである品詞 n-gram を用いた方法は、単純なモデルでありながら計算量が非常に少なく、英語や日本語の実用的な形態素解析器に広く用いられている。しかしながら、この品詞 n-gram モデルを用いて多言語形態素解析器を実現する場合、いくつかの課題がある。

(a) 品詞 n-gram モデルでは語彙の情報を利用できないため、精密な解析を行うのが難しい。助詞などの機能語は特徴的な振る舞いをすることが多いが、それに対処するには品詞の情報だけではなく形態素に関するより具体的な情報である語彙自体を考慮する必要がある。
(b) 多量の品詞を扱う言語体系においては、品詞 n-gram モデルではパラメータの数が大きくデータの過疎性が顕著になるため、信頼性のあるモデルを推定することができない。この問題に対処するには、形態素に関してより抽象的な情報が利用できるとよい。
(c) ラベル無しデータを適切にモデルに取り込むのが難しい。コーパスに基づく形態素解析器を学習させる場合、人手でラベルを付与した多量の学習用データが必要とされるが、このようなラベル付きデータの作成は非常に人手を要する。ラベルが付与されていないラベル無しデータは比較的入手が容易であるため、多言語形態素解析を実現する際には、ラベル付きの学習データが少量しか入手できない場合でも、ラベル無しデータを利用して精度を向上させられることが望ましい。

以上のような課題に対処するために、[文献 9]では誤り駆動方式に基づいて品詞の語彙化やグループ化を行う方法を提案している。しかしながら、この方法では一部の品詞に対してのみ処理を行っているため、形態素に関する十分な情報を利用できていないと思われる。

・提案手法

上記の(a)と(c)の課題に対して、特別な調整を行わなくても複数の言語に対して高い性能を持つような多言語形態素解析器を実現するために、従来の品詞 n-gram を拡張し、複数の n-gram モデルを利用して形態素解析を行うことを考える。

本手法では、品詞 n-gram 確率の他に、語彙自体の情報を持った単語 n-gram 確率、およびラベル無しデータから教師無し学習により学習させたクラスを用いたクラス n-gram 確率を各モデルで条件付けて結合して単語列の出現確率を計算する。各 n-gram モデルの確率は、leave-one-out 法に基づき学習データから自動的に計算する。このように、形態素に関する具体的な情報と抽象的な情報を統合して利用することで、精密で

頑健な形態素解析が可能になることが期待できる。今回実験に使用したデータの品詞体系が小規模だったこともあり、(b)の課題に対する検討と実験は行えなかったが、品詞の階層を考慮した n-gram モデルを同じ枠組みで結合することにより、(b)の課題に対しても対処できると考えられる。

・実験とその結果

上記の手法に基づいた英語用品詞タグ付けシステムを試作し、新聞記事データ(Penn Treebank WSJ コーパス, (ラベル付き)学習データは 211,727 トークン、テストデータは 47,377 トークン、ラベル無しデータは 700,617 トークン)を使用して評価を行った。

実験の結果、従来の品詞 n-gram のみを使用した方法に比べてエラー率を 4%ほど低下させることができた。また、Baum-Welch アルゴリズムを用いてラベル無しデータから学習させたクラスの n-gram を利用することで、解析精度が向上することも確認できた。

(2) 研究の効果

① 複数の n-gram モデルを結合させて用いることにより、言語に特化したチューニングを行うことなく高い精度で形態素解析時の曖昧性の解消を行うことができた。

② この品詞付与の曖昧性解消は形態素解析の最も基本となる要素技術であり、多言語形態素解析器の開発を一歩進めることができた。

5-3-3 多言語翻訳データベースの研究

(1) 研究の内容

本研究では、まず、複数言語で書かれた文書を管理するためのフォーマットとして現在普及しつつある、TMX[資料 10]及び XLEFF[資料 11]について調査した。TMX は翻訳支援ツール等で利用されているフォーマットであり、一方、XLEFF は、システムのローカライズの分野で注目されているフォーマットである。両者は、共に XML に準拠しており、その仕様が公開されている。データの相互利用における効果やその普及を考慮した場合、我々が設計する多言語翻訳データベースにおいても、これらのフォーマットに相互変換できることが必要とされる。

次に、翻訳エンジンにおける利用という側面から、多言語翻訳データベースに必要とされる情報及び機能を洗い出し、その仕様を検討した。必要とされる情報の主なものとしては、翻訳単位（一文の切れ目）の情報、その文を構成している単語やイディオムの情報(形態素の情報や、単語やイディオムに関する翻訳テンプレートの情報も含む)、5-1-2 の改版文書利用型翻訳で利用される旧版文書としての情報（タイトルや段落の情報等）が挙げられる。

さらに、今回我々が対象とする標準文書の性質、その翻訳の性質、標準文書から抽出される翻訳テンプレートの性質について調査した。調査には、ISO 標準文書、動画関連の国際標準文書、FCC(米国通信委員会)の規則文書の３種類の英語文書を利用した。調査方法は、まず、各標準/規則文書に対して、翻訳単位(一文)に区切り、人手で翻訳してみる。その際に、今の技術レベルでの機械翻訳で翻訳可能か、機械翻訳に今後どのような改良が必要とされるか、を検討する。さらに、各文から、翻訳テンプレートを作成し、その有効性を検証することも行った。

また、多言語化という観点から、中国語、韓国語、アラビア語、フランス語、ロシア語等の品詞・文法体系、文書の特徴についての情報収集も平行して行った。

これらの調査、検討の結果から、最終的に、XML に準拠した多言語文書データベースの基本仕様を策定した。現在、上記の調査の際に作成された３種類の標準/規則文書の英日対訳文書や翻訳テンプレートを元にして、データベースの作成に着手している。

(2) 研究の成果

① 多言語翻訳データベースの基本仕様を策定し、標準文書におけるデータベースの作成に着手した。本データベースの作成を今後も進めることにより、標準文書における機械翻訳用のデータ収集を加速することができる。

② 標準文書の特徴、機械翻訳で翻訳する場合の課題の検証結果を利用し、本課題の解決を図ることにより、より標準文書に適した機械翻訳システムの開発が可能となった。

③ 多言語に関する情報収集、調査を行ったことにより、今後拡張すべき英日以外の言語の基礎資料が整った。

5-3-4　協調的翻訳支援環境の研究

(1) 研究の内容

協調的翻訳支援環境の概略設計を行った。これは我々がすでに研究を進めているコミュニティ型機械翻訳サイト「訳してねっと」[資料 6][文献 7][文献 12]をベースにした。「訳してねっと」は、分野ごとにコミュニティを設置し、メンバー間で協調して辞書を登録することで翻訳品質を向上させる機能を有する。

我々は、「訳してねっと」を利用することにより、２つの調査を行った。第一に、実際の翻訳実験である。「訳してねっと」の中から２分野を選択し、それぞれの分野において、実際の翻訳者が利用することにより、詳細な評価を行った。第二に、協調的翻訳支援環境がどのような場面で必要とされており、どういった機能が必要なのかを知るために、翻訳者・翻訳初心者・システム管理者の３者に対し、アンケート及びグループインタビューを行った。これら２つの調査および我々の考察を元に、協調的翻訳支援環境に必要な機能を洗い出した。

また、システムの実装面では、協調的翻訳支援環境を実現するための３つの機能（翻訳機能、翻訳知識蓄積・管理機能、および、コミュニケーション機能）の統合について検討した。例えば、コミュニケーションを行うために掲示板を考える。掲示板ではさまざまな言語が利用されるため、すべてのメッセージをそれぞれのユーザの母語に翻訳する必要がある。ただし、翻訳する場合には、そのユーザが所属している分野の辞書を用いて適切な専門用語を訳出する必要がある。その専門用語は翻訳知識として蓄積・管理しているものである。このように、協調的翻訳支援環境で用いられる各コンポーネントは、その裏側でさまざまな機能に深く結びついている状態でなければならない。本研究により、データベース周辺および翻訳に関わる細かな処理を上位アプリケーションから隠蔽し、統一的なインタフェースを提供するとともに、翻訳アプリケーション作成を容易にする協調的翻訳支援フレームワークを考案した。

上記の検証及び研究結果により、基本的な協調的翻訳支援環境を協調的翻訳支援フレ

ームワーク上に実現した。本支援環境は、以下の８つの機能を持つ。

(a) ユーザ間の情報伝達手段としてメッセージ機能
(b) コミュニティ専用の辞書の登録・検索・更新機能（辞書機能）
(c) 他のコミュニティの辞書参照機能
(d) 文書やURLを共有知識として管理する文書管理機能
(e) 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)
(f) ユーザを限定したりアクセス権を設定するユーザ管理機能
(g) コミュニティに関する情報の多言語表示機能
(h) サブコミュニティ作成機能

これらの基本開発を完了し､国際標準に関する文書において簡単な翻訳プロセスのデモを行った。

(2) 研究の効果

① 協調的翻訳支援環境が有する各機能の必要用件を確認した。また、支援環境の実現には、フレームワークを中間層に設ける方法が必要不可欠であることを確認し、その基本開発を完了した。本フレームワークを利用した翻訳支援環境を利用することにより、複数の人間が協調して翻訳する際の作業の効率化を図ることができる。

② 今後、本フレームワークを基本とした開発を推進することにより、利用ユーザの適性に合わせて様々な形態で支援を行うことができる翻訳支援システムの完成が期待できる。

5-3-5　結論と今後の課題

本サブテーマは、上述した通り、当初予定した目標を達成することができた。各項目別に今後の課題を以下に示す。

・言語非依存形態素解析システムの研究
多言語形態素解析器を実現するために必要な確率モデルについて検討し､複数のn-gramモデルを結合する手法に基づいたシステムを試作・評価し、この手法の有効性を確かめた。本年度は英語のデータを利用して提案手法の実験を行ったが、今後、日本語や中国語等の他の言語や、大規模な品詞体系を持つデータでの検証を進めたい。
　また、形態素解析に関する課題の一つとして、未知語(形態素解析システムの辞書中に存在しない単語)処理の問題がある。固有名詞や新語などの未知語は頻繁に出現するが、未知語処理はしばしば言語に依存した情報を必要とする。そのため、今後多言語に対応した未知語処理システムを作成し､今回検証したモデルと結合して一つの形態素解析システムにまとめることが今後の課題である。

・多言語翻訳データベースの研究
本研究テーマでは、多言語標準翻訳データベースの基本仕様を策定し、標準文書におけるデータベースの作成に着手した。
　多言語翻訳データベースの今後の課題は、多言語への展開、及び、他システムとの連

携による実際の運用である。多言語への展開は、現在中国語への適用について調査を開始している。一方、他システムとの連携に関しては、次の２つ連携がある。

(a) 翻訳テンプレート学習結果を本データベースに格納し、また本データベースを利用して学習するという翻訳テンプレート学習システムとの連携
(b) 本データベースの情報を辞書として利用して機械翻訳システムが翻訳するという機械翻訳システムとの連携

各研究テーマとの関連を深めて、研究を遂行していきたい。

・協調的翻訳支援環境の研究

協調的翻訳支援環境の基本設計及び開発を行い、国際標準に関する文書において簡単な翻訳プロセスのデモを行った。

今後の課題としては、協調型翻訳支援環境を実際に運用することにより、その問題点や改良点を明らかにし、協調型翻訳支援フレームワークおよび協調型翻訳支援環境の完成度を高めることがある。さらに、他の研究項目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、多言語翻訳データベースモジュールを取り込み、多言語標準文書処理システムとしての試作システムを開発する予定である。

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

(a) 翻訳テンプレート学習に関する研究開発
(b) 分野辞書の自己組織化に関する研究開発
(c) 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。個別の研究課題と研究テーマの全体像との関係を図３に示す。

このように、研究開発の初年度となる本年度は、研究テーマ全体の方針及び研究体制を確立し、各サブテーマにおける基本課題の解決に対して重点的に取り組み、期待通りの成果が得られた。まとめると以下のようになる。

(a) 翻訳テンプレート学習に関しては、改版前の翻訳文書、文対応付き翻訳文書、文対応関係のない翻訳文書という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システム及び機械翻訳システムの実現の見通しをつけた。

(b) 分野辞書の自己組織化に関しては、登録辞書の分野判定のためのキーとなる語である「コアワード」を媒介して分野を自動判定する手法を考案した。この技術を基礎として、自動分類、自動階層化の研究を促進するが期待できる。

(c) 言語非依存翻訳エンジンの研究では、言語の非依存の最も課題となる形態素解析システムの設計を完了し、多言語翻訳に必要な翻訳エンジンの全体設計を終えた。また、翻訳エンジンが利用する多言語翻訳データベースに関しても基本設計を終え、現在データベース作成を進めている。また、フレームワークという概念を利用した

協調型支援環境の開発を完了し、多言語標準文書処理システムの基本的な枠組みの開発を完了した。

本年度、獲得した成果物及び知見を基礎にして、今後は、各サブテーマの連携も視野に入れながら、本テーマの研究及びシステム開発を進める予定である。

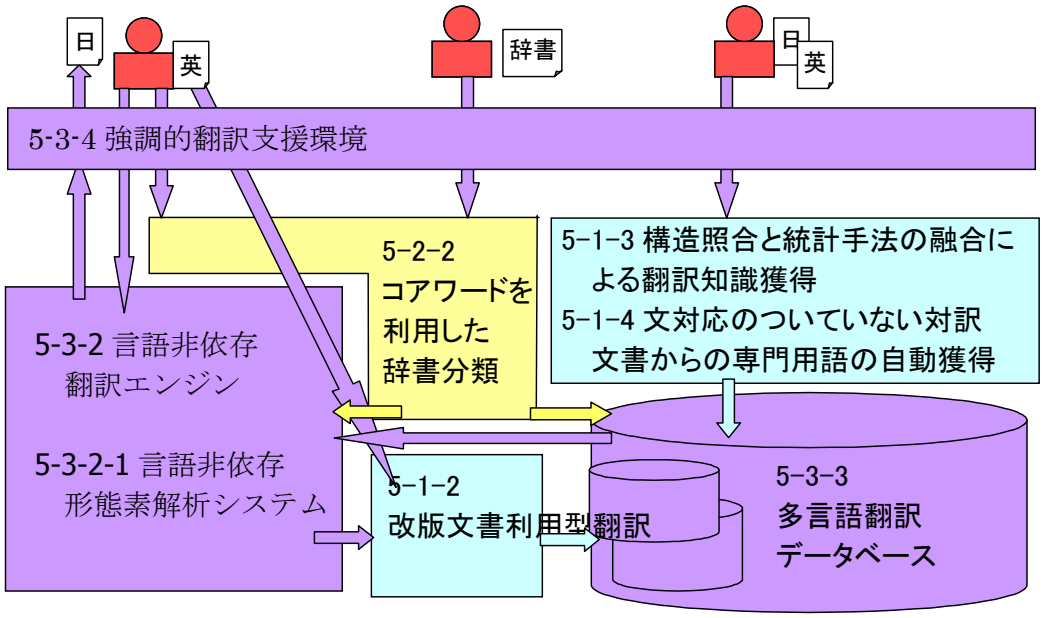


図 3 各研究テーマの全体像

参考資料、参考文献

［文献 1］介弘、下畑、松下：改版文書の翻訳を支援する差分翻訳システム、情報処理学会第 51 回全国大会、1995

［文献 2］北村、松本：対訳コーパスを利用した翻訳知識の自動獲得、情報処理学会論文誌、vol.37、No.6、pp.1036－1040、1996

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

［文献 4］Fung. P., 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 5］神山、伊藤：自律的語彙拡充を行う機械翻訳システム、情報処理学会第 65 回全国大会、2003

［資料 6］http://www.yakushite.net/

［文献 7］Shimohata, S., et al.： Collaborative Translation Environment on the Web, Proceedings of the MT Summit VIII, 2001

［文献 8］Ratnaparkhi, A.: A Maximum Entropy Model for Part-of-Speech Tagging, Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.133－142 , 1996

［文献 9］浅原、松本裕治： 形態素解析のための拡張統計モデル、情報処理学会論文誌、Vol.43、No.3、pp. 685－695、2002

［資料 10］http://www.lisa.org/tmx/

［資料 11］http://www.oasis-open.org/committees/xliff/documents/xliff-specification.htm/

［資料 12］Shimohata, S., et al.: Machine Translation System PENSEE: System Design and Implementation, Proceedings of the MT Summit VII, pp.380－384, 1999

（添付資料）

１　研究発表、講演、文献等一覧

北村美穂子、村田稔樹、介弘達哉、下畑さより、佐々木美樹、松永聡彦、中川哲治：
コミュニティ型機械翻訳サイト「訳してねっと」 の基盤技術とその展開、
情報処理学会第 65 回全国大会講演論文集（特別トラック「言語バリアフリー技術」）、
No.5、pp.319-322、2003