管理番号＃１４－０２

平成１６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

委託先：沖電気工業㈱

平成１７年５月

情報通信研究機構

平成１６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。

今年度は、上記の標準文書として「特許文書」を選択した。その理由は、特許文書は多種多様な分野を含んでいるため、その訳し分けや特殊な表現において翻訳が難しいとされており、かつ、翻訳需要も多いためである。わが国においても、平成 15 年度から AAMT（アジア太平洋機械翻訳協会）が、特許翻訳に関する研究会（AAMT/Japio 特許翻訳研究会）を発足させ、特許文書における自然言語処理技術の重要性を唱えている。（当社からも研究員を派遣し、技術調査及び、技術交流を行っている。）この研究会の本年度の活動は、以下の４つである。

i)　ヨーロッパ特許庁(EPO)など諸外国での状況の調査
ii)　特許の機械翻訳研究のためのリソースの構築
iii)　共通のデータセットを使った個別研究の推進
iv)　翻訳結果の評価手法

以下に、これらを簡単に紹介する。

i)に関しては、当社の研究員も調査団の一人として同行し、諸外国の動向を視察し、多くの知見を得た。特に、ユーザと開発者が一体となって機械翻訳システムを開発することの重要性は、協調型機械翻訳システム「訳してねっと」の理念と通じ、その手法において大いに参考になった。

ii)に関しては、当研究会が構築したリソースの提供を受け、「多言語標準文書処理システム研究開発」の翻訳パターン獲得の実験に利用している。

iii)における個別研究で、最も注目されている研究は、翻訳辞書の構築に関する研究である。特許翻訳においては、専門用語に関する辞書をいかに効率良く構築するかが、翻訳品質を左右する重要な要因となる。我々の研究方針と同じく、本研究会でも、いくつかの翻訳辞書の自動獲得に関する研究結果が報告されている。

iv)に関しては、今後、評価のためのアラインメントデータや様々なアノテーション付きのデータが提供される予定である。これらのデータも我々の研究に利用可能であると思われる。
来年度も引き続き、研究員を派遣し、積極的な技術交流を行う予定である。

参考文献：平成 16 年度 AAMT/Japio 特許翻訳研究会 報告書　平成 17 年 3 月
財団法人 日本特許情報機構

２　研究開発体の全体計画

２－１　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。

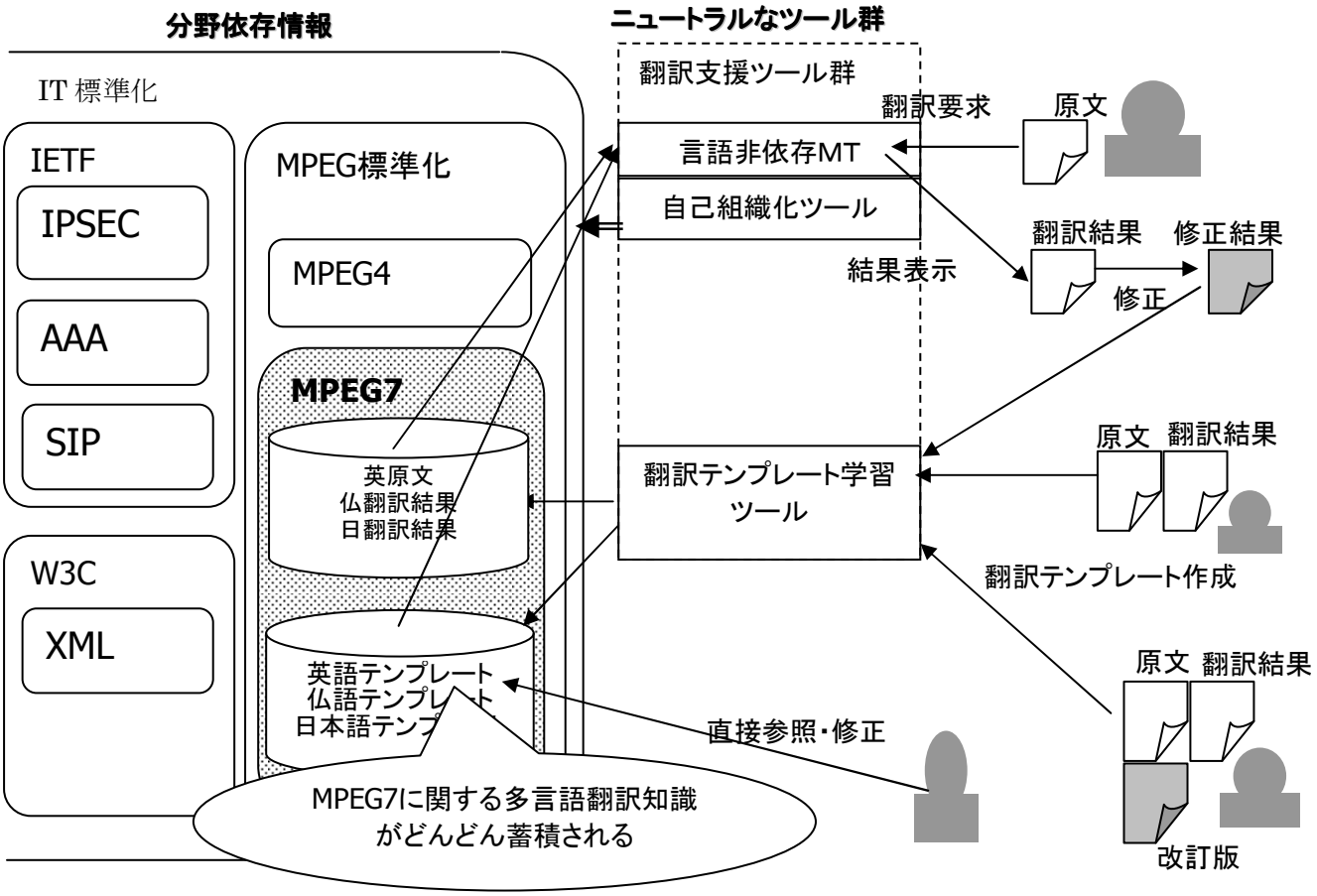


図3-1　多言語標準文書処理システムの構成図及びユーザによる利用形態

２－２　研究開発目標

２－２－１　最終目標（平成 17 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。

２－２－２　中間目標（平成 15 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

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。

２－３　研究開発の年度別計画

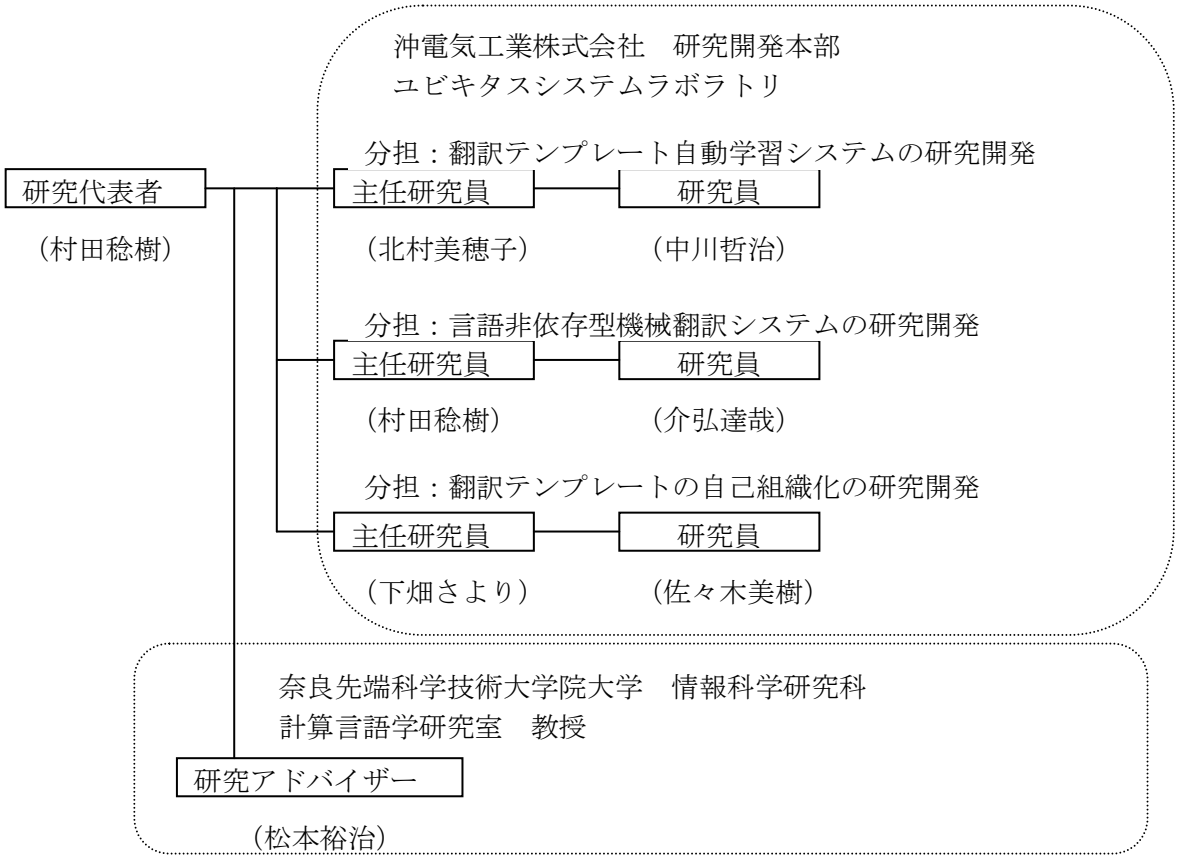
（金額は非公表）

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。

３　研究開発体制

３－１　研究開発実施体制



４　研究開発実施状況

４－１　翻訳テンプレート学習に関する研究開発

４－１－１　序論

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。昨年度までの研究で我々は、対訳辞書を自動的に作成する方法として、パラレルコーパスを用いる方法とコンパラブルコーパスを用いる方法の 2 つを提案し、実験レベルである程度の精度を達成できることを示した。

今年度は、第一に、これらの方式を特許文書という実課題に適用し、さらに実用性の検証を行うとともに、大規模な翻訳テンプレートの構築を行った。また、翻訳結果の自動評価手法を応用したテンプレート自動獲得技術の提案を行った。第二に、これらのテンプレートを効果的に構築し、利用するための調査及び開発を行った。以下に、これらの各研究について述べる。

４－１－２　特許翻訳のための翻訳テンプレート構築とその評価

(1) コンパラブルコーパスからの翻訳テンプレート自動獲得技術の評価実験

・提案手法の概要

図 4-1-1 は、コンパラブルコーパスからの対訳抽出の簡単な例を示している。各候補語と seed word をつなぐ線はコーパス中での共起頻度(文や節など、あらかじめ定めた範囲に同時に出現する回数)を表しており、実線は共起頻度が高いことを、点線は共起頻度が低いことを示している。例えば、「機械」は「技術」や「科学」と共起することが多く、「経済」や「市場」とは共起することが少ない。同様に、”machine” は「技術」に対応する “technology” や「科学」に対応する “science” と共起することが多いが、「経済」に対応する “economy” や「市場」に対応する “market” とは共起することが少ない。このような場合、「機械」と “machine” は共起パターンが類似しているので翻訳対として抽出し、翻訳テンプレート化して辞書登録する。

提案手法の具体的な手順は以下の 3 つのステップからなる。

step1 seed word リストの作成
step2 単言語コーパスからの候補語の抽出
step3 候補語の対応付け

seed word リストは、単語の類似度を測る際の重要な基準となる。Fung & McKeown[1]は、対象コーパスに一定頻度以上出現していて、かつ、英語と日本語が 1 対 1 で対応する単語ペアのみ(あるいは、英日方向に 1 つの訳語候補しか持たない単語ペアのみ)を seed word として抽出した。しかしながら、この方法では、対象コーパス中での語の使われ方を考慮していないことや、利用できる seed word の数が少なくなることから、seed word リストの質に問題があり、対応づけの精度に悪影響がある。

本提案では、両言語のコーパスに一定以上の出現頻度のある翻訳対を抽出した後、翻訳対の共起パターンの類似度を両言語で比較し、類似度の高い翻訳対のみを seed word とする。これにより、両言語でコーパス内での振舞いが類似する seed word が抽出されるようになり、候補語対応付けの精度向上が実現した。

それぞれのコーパスからの候補語の抽出には様々な方法が考えられるが、今回は各コーパスから可能な n-gram 文字列をすべて抽出し、頻度、tf・idf などによってフィルタをかけて候補語を選別する方法[文献 2][文献 3]を用いた。候補語どうしを対応付けて翻訳対を獲得する際には、各言語の候補語と seed word のコーパスにおける共起パターンを取得し、共起パターンの類似度を比較して、類似度の高い候補語対を翻訳対として抽出した。類似度の評価には、最も一般的なユークリッド距離を用いた。

・実験および評価

実験は、Japio(財団法人日本特許情報機構)殿より提供の平成 15 年度公開の日英特許抄録および専門用語辞書(以下、Japio 辞書と呼ぶ)を用いて行った。実験に使用したデータの詳細は以下の通りである。

コーパス
　日英特許抄録(C12N:遺伝子分野)　11781 件
　日本語　38481 文
　英語　　35343 文
正解データ
　Japio 辞書(C12N:遺伝子分野)　4789 件
対訳辞書
　EDICT

評価は同分野の Japio 辞書を正解データとし、Japio 辞書のエントリのうち見出し語、訳語ともコーパスに 100 回以上出現する 57 件が本手法により抽出されているか、正しく対応付けられているかを調べた。

候補語抽出プロセスにおいて、見出し語、訳語とも候補語として抽出されていたエントリは 43 件(75.4%)であった。また、訳語対応付けプロセスにおいて、日本語の候補語に対して、英語の候補語を類似度の高いものから順に対応付けたところ、正解の平均順位は 14 位となった。また、正解の訳語が第 1 候補になったものは 24 件、10 位以内に抽出されたものは 33 件であった。以上のことから、全体としての再現率と抽出精度はそれぞれ、57.9%、76.7%となる。

表 4-1-1 に各コーパスから抽出された候補語の例を示す。なお、本表には例が少なかったが、本手法では ”amino acid sequence” のように複数の単語からなる候補語も、単語と同様に多数抽出されている。また、表 4-1-2 に各日本語の候補語に対して、類似度の高い英語の候補語を抽出した例を示す。太字が Japio 辞書における訳語を示している。日本語候補語の対訳候補として、正解訳語や関連語が抽出されていることが分かる。

対応づけ誤りの主な原因は、以下の 2 点である。

1) 同義語内での優先順位が違う。
2) 分野専門用語ではないものが候補語に含まれている。

対応付け誤りの多くが、1) の原因によるものである。例えば、「変異」の訳語候補に正解が含まれていないが、Japio 辞書における訳語は “variation” であった。本手法の結果では、”variation”に対応する形容詞 ”variant” が第 1 候補訳語として抽出されているが、Japio 辞書にはなかった。ただ、実際に翻訳を行う際には、「変異」の訳として “variant” を用いることも考えられ、必ずしも間違いとは言い切れ

ない。このような派生語や同義語の扱いをどのように吸収していくかが今後の課題である。 2)の問題についても、今回は用語の専門性を測る手段として tf・idf を用いたが、「新規」「本 発明」や “comprise” のように、分野の専門用語ではなく特許用語として登録すべきものも含まれてしまった。候補語抽出プロセスにおいて、このような語はあらかじめ除去しておかなければならない。さらに強力なフィルタリングを行う必要があると考えられる。

(2) 特許文書からの大規模翻訳テンプレート構築

・特許文書の特徴

特許の文章には以下のような特徴がある[文献 4]。

- 文が長い
- 専門用語が多い

一般に特許文は、文章が非常に長くなる傾向がある。例えば、遺伝子分野(IPC:C12N)の 2004 年出願の全データ 11781 件から抄録部分の形態素数を集計したところ、日本語では一文が平均 57 形態素(105 文字)、英語では平均 44 形態素であった。読みやすい文の長さの目安が 50 文字程度といわれていることからも、特許の文が長くて理解しにくいものであることが分かる。文の長さとも関連して、特許文では並列構造が多く、係り受け関係が複雑であるという特徴もある。

また、特許では分野が細分化されており、それぞれの分野に多くの新語や専門用語が存在する。これらの語は、分野によって訳語が決まっていたり、複合語で 1 つの概念を表す訳語に変換する必要があったりするので、一般の単語辞書のみで翻訳すると、文法的には間違っていなくても意味の通らない文になってしまう。また、専門性の高い単語の中には辞書に未登録のものも多く、構文解析の失敗を招く原因となっている。

以上のような特徴から、特許文の翻訳においては、解析に失敗したり、訳語選択に誤りが生じたりする可能性が高い。そこで我々は、特許文の特徴を踏まえ、機械翻訳での品質構造を目的として、大規模な専門用語辞書を構築することにした。これは、専門用語を正しく認識することにより、用語を適切に翻訳するだけでなく、構文解析の曖昧性解消や翻訳速度の向上にも寄与するという考えに基づいている。以下では翻訳テンプレートの自動獲得技術に基づき構築した専門用語辞書の概要とその評価結果について述べる。

・特許文書からの翻訳テンプレート獲得

特許には国際特許分類(IPC: International Patent Classification) と呼ばれる記号が付与されている。IPC は発明に関する全技術分野を段階的に細分化したもので、技術分野をＡ－Ｈの８つの「セクション」に分け、各セクションをクラス、サブクラス、メイングループ、サブグループに階層的に展開したものである。例えば、以下に示す IPC コードの分類の詳細は表 4-1-3 のようになっている。

例）Ｃ　１２　Ｎ　１１　／　００

(1)セクション =C	化学; 冶金
(2)クラス =12	生化学；…；酵素学；突然変異または遺伝子工学
(3)サブクラス =N	微生物または酵素；その組成物…; 突然変異または遺伝子工学；培地
(4)メイングループ =11	担体結合または固定化酵素；担体結合または固定化微生物
(5)サブグループ =00	なし

表 4-1-3 IPC コード分類の詳細の例

我々は、IPC のサブクラスまでの情報を用いて分野を設定し、日本語特許抄録および対応する PAJ10 年分(1994 年～2003 年)より翻訳パターンの抽出および辞書化を行った。翻訳テンプレートの獲得には以下の 2 種類の方法を用いた。

・ タイトル対訳からの翻訳テンプレート自動抽出
タイトルは 1 対 1 の対応付けが取れているので、パラレルコーパスからの翻訳テンプレート自動抽出技術により翻訳テンプレートの候補を抽出し、抽出結果を人間がチェックした。

・ アブストラクト対訳からの翻訳テンプレート自動抽出
アブストラクトは内容としてはほぼ一致するが、文として 1 対 1 の対応が取れていないので、コンパラブルコーパスからの翻訳テンプレート自動抽出技術により翻訳テンプレートの候補を抽出し、抽出結果を人間がチェックした。

獲得した専門用語辞書はのべ約 110 万件、異なり約 77 万件である。

・翻訳実験

獲得した専門用語は、Web サイト型の協調型機械翻訳サイト「訳してねっと™ 」 上に展開した。今回作成した専門用語辞書は、訳してねっと™ 上の対応するコミュニティの辞書に追加する形で登録した。また、翻訳を行う際には、特許抄録の IPC コードから対応するコミュニティを選択し、該当コミュニティ、および、その上位コミュニティの辞書を使って翻訳を行うようにした。

実際に特許抄録(PAJ)を使って、専門用語辞書登録を行ったことで翻訳結果がどのように変化したかを調べた。翻訳は英日方向で、実験の環境は以下の通りである。

・対象文書
　特許抄録(C12N:遺伝子分野) 1000 文
・対応する専門用語辞書(生物コミュニティ)
　7567 件
・上位辞書
　20301 件

実験の結果、変化があった文は 1000 文中 871 文で、多くは専門用語登録による訳語

の変化であった。表 4-1-4 は専門用語辞書を利用した場合とそうでない場合との解析成功率および翻訳時間について比較したものである。解析成功率は若干上昇したが、翻訳時間はほとんど変わらなかった。

解析成功率が上がった原因としては、複数語からなる専門用語を登録したことにより曖昧性が減った(解析候補数が減った)ことと、未登録語を登録したことにより解析不能文が減ったことが大きな原因であると考えられる。また、予想に反して翻訳時間が短縮しなかった原因としては、現状では、専門用語辞書の適用数が一文平均 3 語程度であることから、飛躍的な短縮に結びつかなかったものと考えられる。翻訳品質、翻訳速度の向上のためには、更なる辞書の登録が必要である。

	辞書あり	辞書なし
解析成功率	83.3%	81.7%
翻訳時間	8905 秒	8833 秒

表 4-1-4 翻訳実験の結果

翻訳品質の面では、適切な訳が生成されるようになっただけでなく、以下のような効果があった。

・未知語で構文が崩れていたものが正しく解析できるようになった。
・長い専門用語登録で係り受け構造が正しく認識できるようになった。

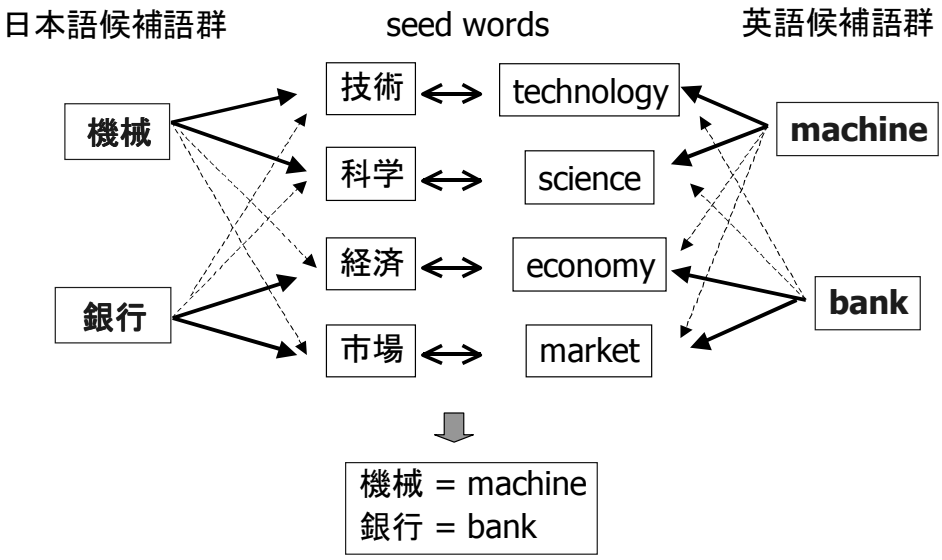


図 4-1-1 コンパラブルコーパスからの対訳抽出の概念図

日本語候補語	出現頻度	英語候補語	出現頻度
遺伝子	14203	acid	16929
配列	13592	amino acid	8347
細胞	11143	gene	16333
DNA	8899	sequence	16918
酵素	6794	protein	11459
製造	6379	DNA	12749
培養	6328	cell	13193
を コード する	5942	solution	11134
活性	5719	amino acid sequence	6049
タンパク質	5405	microorganism	4809
アミノ酸	5277	culture	7892
新規	5166	encode	4531
微生物	5144	polypeptide	4422
発現	4756	formula	4359
ポリペプチド	4105	comprise	4217
蛋白 質	3998	nucleic acid	3960
本 発明	3787	vector	3921
ヒト	3555	enzyme	5649
核酸	3535	bacterium	3952

表 4-1- 1 抽出された候補語の例

日本語候補語	英語候補語
培養	**culture**
	cultured
	objective
	culture medium
	medium
タンパク質	**protein**
	express
	expression
	DNA
	sequence
アミノ酸	protein
	amino acid
	formula
	DNA
	DNA encode
発現	gene
	express
	expression
	protein
	transducing
領域	**region**
	domain
	construct
	link
	gene
変異	variant
	mutation
	sequence of formula
	mutant
	gene
分子	bond
	acid
	solution
	terminal
	molecule

表 4-1- 2 訳語対応付けの結果の例

(3) 翻訳結果自動評価方法を用いたパラレルコーパスからの翻訳テンプレート自動獲得技術の改良

近年、翻訳結果の自動評価技術についての研究が盛んに行われている[文献 5][文献 6]。我々は、これらの自動評価技術を、単に評価のために用いるのではなく、翻訳テンプレート自動獲得技術に応用することを試みた。具体的には、翻訳結果の評価結果を用いることにより、翻訳テンプレート自動獲得技術における各種のパラメータの自動調整を行うというものである。

・提案手法の概要

我々が昨年度開発したパラレルコーパスからの翻訳テンプレート自動獲得手法は、

対訳文書における原言語の表現と目的言語の表現の同時出現頻度に基づく統計情報を利用して、対訳文書が有する原言語と目的言語の方言の対（翻訳パターン）を自動的に抽出し、翻訳パターン辞書を作成する技術である。この手法は、１つの課題を有する。それは、正解可能性の高い翻訳パターンから順に抽出する手法（以下、貪欲的手法と呼ぶ）であり、正解が保証されなくなった時点で処理を終了しなければならないが、どこで処理を終了すべきかの判断が難しいことである。この技術を実際利用する場合には、予備実験で抽出された翻訳パターンを人手で確認し、終了条件を設定しなければならなかった。

翻訳結果の自動評価手法を用いて、この終了条件の設定を行う。この手法は、対訳表現の出現回数の閾値を徐々に下げていくことで、貪欲的手法を実現している。したがって、この閾値をさらに下げるか、または、処理を終了するかを、自動評価の結果が良くなったか、悪くなったかで決定する。

手法を以下にまとめる。

(1)抽出対象とする対訳文書を翻訳し、BLEU 等の翻訳結果自動評価手法[文献 5]を用いて、評価点を算出する。

(2)ある閾値で抽出された翻訳パターンを実際に機械翻訳システムに登録し、その翻訳結果を、(1)と同様の手法で評価し、評価点を算出する。

(3)(2)で求めた評価点が、前回の評価点より高くなった、または等しい場合、(2)の翻訳パターンの登録が有効だとみなし、閾値を下げて、抽出処理を続ける。評価点が下がった場合、その翻訳パターンの登録は悪影響を及ぼすとみなし、処理を中止する。かつ、評価点を下げた翻訳パターンの登録も取り消す。

上記の処理により、使用者に終了条件設定の負担をかけることなく、入力文書に最適な終了条件を自動設定することができる。

４－１－３　効果的な翻訳テンプレート作成・利用方法の検討

(1) 翻訳テンプレートの人手による作成と自動獲得ツールによる作成の比較

4-1-2 の項で述べたように、特許翻訳では、大量の専門用語に関する翻訳テンプレートの登録が翻訳品質を支えている。それらの大量の翻訳テンプレートは、我々が保有する翻訳テンプレート自動作成技術を用いて作成しているが、抽出されたテンプレートの全てを登録すると、間違った登録によって誤訳を招いたり[1]、構文解析に悪影響を及ぼす登録[2]もあったりすることから、自動抽出した翻訳テンプレートは、以下のような手順で登録している。

(1) 自動獲得した翻訳テンプレートを辞書に登録する。
(2) 誤り可能性が高い翻訳テンプレートを、機械的に抽出し、人手で確認する。
(3) 実際に間違っていたテンプレートを辞書から削除する。

[1]抽出結果が正しくないだけでなく、元文書のスペルミスなども多かった。

[2]例えば、現在分詞形や過去分詞形の動詞を名詞として登録すると、動詞としての働きの優先度が下がり、常に名詞として翻訳されてしまう。例えば、”using/利用”が登録された場合、”system using electron”は“電子を利用するシステム”ではなく”システム利用電子“となる。

なお、(2)で要する時間は、300～500 語/１日(6.5h)である。

我々は、自動抽出の効率性を確かめるために、(2)において機械的に抽出した結果を確認するのに要する時間（以下、「自動抽出＋確認」と呼ぶ）と、人間が対訳文書を参照し、一から翻訳テンプレートを作成する時間（以下、「人手抽出」と呼ぶ）がどれくらい違うのかを調査した。

翻訳テンプレート自動抽出で用いた文書と同じ日英特許抄録(PAJ)の対訳文書を翻訳者に渡し、抽出作業を行ってもらった。その結果、300～600 語/１日(6.5h)となり、我々の予想に反して、確認作業にかかる時間とほぼ同じ結果になった。

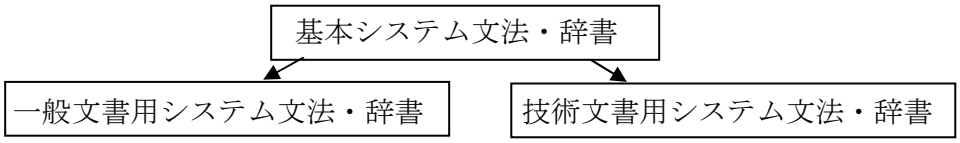
これは、数字上では、「自動抽出＋確認作業」と「人手抽出」は人間の作業負荷が変わらない結果となった。しかし、両者の実際の作業内容は、大きく異なっていた。「自動抽出＋確認作業」は、誤り可能性が高い翻訳パターンなので、翻訳者は、そのパターンが誤っていることを前提に、様々な確認作業を行う。確認作業は規則的な作業でないので多くの時間を要する。一方、「人手抽出」は、その対訳文書が正しいことを前提にして、機械的に対応をみつけ、対応箇所をピックアップするので、その作業は規則的なものとなり、コツさえつかめれば高速に処理することができる。その際、翻訳者は、その対訳文書が間違っている可能性があるとか、それを翻訳パターンとして登録することによって悪影響を及ぼすか、などを考慮しなかった。

これらのことから言えることは、我々が行っている「自動抽出＋確認作業」は、全件を人手で抽出した場合でも、しなければならない作業であるということである。また、「自動抽出＋確認作業」で翻訳者が行っている確認作業は、検索システムや専門用語辞書などを横断的に用いて行う、極めて知的な作業であり、機械が模倣することはできない。人間は知的作業を行うことに集中できるという点からも、翻訳パターンの自動獲得の重要性を再認識することができた。

(2) 効果的な翻訳テンプレート利用方法の検討

自動抽出された翻訳テンプレートは、基本的には、各分野のコミュニティ辞書に登録されるが、抽出された翻訳テンプレートの中には、分野単位で登録するより、技術分野において一般的に有効である翻訳テンプレートも存在した。これらの翻訳テンプレートを活用するために、コミュニティ辞書だけでなく、システム辞書も分類、階層化した。

現在の英日翻訳システムの辞書構成は、以下のようになっている。



例えば、 “controller”は、一般文書用システム辞書には「会計監査」、「コントローラ」の２つの訳語があるが、 技術文書用システム辞書には、「コントローラ」しかない。このように、自動抽出した翻訳テンプレートだけでなく、従来からシステムが有する翻訳テンプレート（辞書）においても、一般文書用と技術文書用のように、用途別に辞書を分類することで、より正確な翻訳結果を出力することができる。

４－１－４　結論と今後の課題

上記の研究成果により、以下のことが明らかになった。

- コンパラブルコーパスからの翻訳テンプレート自動獲得技術においては、対象を特許文書に絞った評価を行った。遺伝子分野の特許抄録 1 年分を用いた実験では、人間が作成した辞書と比較して、用語の抽出精度 75%、10 位以内の訳語対応付け精度 76.7%を達成することができ、機械翻訳用専門用語辞書構築の支援ツールとして十分有効であることを示した。

- 各自動獲得方式の実用化の目処が立ったことから、実際に、特許コーパスを用いて大量の翻訳パターン獲得を行った。獲得した翻訳パターンを用いて翻訳を行い、翻訳結果が向上したことを確認した。

- 自動評価手法を応用したパラレルコーパスからの翻訳テンプレート自動獲得技術を提案した。

- 翻訳パターン作成において、自動獲得ツールを利用しそれを人手で確認する時間と、対訳文書を利用して一から人手で抽出するのにかかる時間は、ほぼ同じであった。しかしながら、前者の確認作業は、翻訳者のノウハウを生かした知的作業を行うのに対し、後者の抽出作業は、知的作業は行われず、単純作業に終始してしまう結果となった。人間が知的作業に集中できるという点からも翻訳パターンの自動獲得の重要性を再認識することができた。

- 英日翻訳文法・辞書において、翻訳テンプレートのさらなる分類・階層化を行った。その結果、特許文の翻訳品質が大幅に向上することを確かめた。

今後は、用語抽出及び訳語対応付けの精度のさらなる向上を図るとともに、人間工学的な面から辞書構築プロセス全体を見直し、効率的に高品質な辞書を作成する方法について検討を行っていく予定である。

４－２　多種多様な分野辞書の自己組織化に関する研究開発

４－２－１　序論

ユーザが多種多様な分野辞書を利用して翻訳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、状況に応じて辞書を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。そのため、自己組織化という技術を研究開発している。本テーマでいう自己組織化とは、ユーザの作業を軽減するために、ユーザに代わってシステムが自動的に辞書の構築及び辞書の選択を自動化する技術である。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。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、さらに下層に分類すべきサブ語群を発見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、階層の適切な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),(b)については既に取り組み、その結果、今後の研究の柱となる「コアワード」を利用して語の分野を自動判定する手法を考案した。本年度は、上記の(c)について取り組み、その結果、コアワードを利用して、階層化された分野において文書の分野を自動判定する手法を考案した。

４－２－２　コアワードを利用した文書の分野自動判定の研究

(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」では、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ができる。この場合、ユーザが選択した分野が不適切であると、適切な分野を選択した場合に比べて十分な翻訳品質が発揮できないこともある。しかし、文書を登録あるいは翻訳する際に、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荷が高い。このため、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。

本年度は、これまで研究してきた、分野に特徴的でかつ代表的な単語「コアワード」を利用して、階層化された分野において文書の分野を自動判定する手法を試みた。この手法は、コアワードを分野に特徴的な文書から自動的に作成して、前もって全ての分野に付与しておき、判定時には、分野判定したい文書に存在するコアワードを利用して、文書の分野を自動的に判定する、というものである。

まず、コアワードと分野関連度の定義を説明し、次に、階層化された分野に対してコアワードの分野関連度を付与する手法について説明する。その後、階層化された分野における文書の分野自動判定の手法について説明し、最後に、実験とその結果について述べる。

・コアワードと分野関連度の定義

分野に特徴的でかつ代表的な単語を「コアワード」と定義する。コアワードに付与する値を分野関連度と定義する。分野関連度は分野に関連する度合いを示す値で、分野関連度の値が大きいほど、分野に関連する度合いが強いとする。

各分野のコアワードは、分野毎に既に分類されている文書を利用して作成する。分野毎に既に分類されている文書を形態素解析し、形態素解析を行った結果の品詞が、名詞、動詞、形容詞、形容動詞、未知語になった単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。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)。

語がどのくらい文書を特定するかを 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。

・階層化された分野におけるコアワードと分野関連度

階層化された分野とは、分野が下に行くほど詳細になるように階層が木構造になっている分野のことである。図 4-2-2(a)の上部は階層化された分野の例である。直接上にあるのが親で、直接下にあるのが子である。ある分野の直接上にある分野がその分野の親分野であり、ある分野の直接下にある分野がその分野の子分野である。子分野がないのが最下層の分野で､親分野も子分野もあるのが中間層の分野である｡図 4-2-2(a)では、野球分野とサッカー分野がスポーツ分野の子分野で最下層の分野であり、スポーツ分野が野球分野とサッカー分野の親分野で中間層の分野である。

本手法では、階層化された全ての分野に前もってコアワードの分野関連度を付与しておき、各コアワードの分野関連度を利用して分野を自動判定することによって、分野判定時の処理を軽減する、という方針を採っている。図 4-2-2(a)の下部はコアワードの例である。以下に、コアワードを付与する方法と分野関連度を付与する方法に分けて説明する。

階層化された分野に対してコアワードを付与するには、基本的には、最下層の分野のコアワードのみを文書から作成する。親分野のコアワードは、直下の子分野のコアワードすべてとする｡理由は､親分野は子分野すべてを含むと考えてよいからである。親分野のコアワードを分類された文書からではなく子分野から作成する理由は、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困難で労力を要することだからである。例えば、サッカーの文書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関係ない文書であればサッカー分野にあるのが適切であるからである。

階層化された分野に対してコアワードの分野関連度を付与する方法は、最下層にある文書のみを利用する場合と中間層にある文書を利用する場合に分けて説明する。

最下層にある文書のみを利用する場合には、親分野のコアワードの分野関連度は、子分野のコアワードに付与された分野関連度の偏り具合を考慮して、コアワード毎に子分野の分野関連度から計算する。考え方を以下に述べる。あるコアワードの分野関連度が、いずれかの子分野で突出している場合には、そのコアワードの親分野での分

野関連度を、「突出している子分野」、「親分野」、「突出していない子分野」の順に値が大きくなるように設定する。子分野のコアワードに付与された分野関連度に偏りがない場合には、そのコアワードの親分野での分野関連度を、すべての子分野よりも値が大きくなるように設定する。さらに、親の分野関連度は、直下の子の分野関連度と整合を取るだけではなく、他の分野の分野関連度とも整合が取れるように設定する。

中間層にある文書を利用する場合には、以下のような問題がある。もし、中間層に分類された文書を利用してコアワードを作成して親のコアワードとすると、子にのみ含まれるコアワードが親に反映されない。しかし、中間層に分類された文書を親のコアワード作成時には利用しないで、子のコアワードのみから親のコアワードを作成すると、子に含まれないコアワードが親に反映されない。そこで、以下の様に考える。下層に子があるにもかかわらず中間層の分野に分類される文書というのは、子に対して、複数の子に該当する全般的な文書であるか、いずれの子にも該当しないその他というべき文書であるか、のどちらかであると考える。例えば、スポーツ分野の下に野球分野、サッカー分野がある場合、親であるスポーツ分野にある文書は「野球とサッカーの両方の内容を含むスポーツ」と「野球もサッカーも含まないその他の内容のスポーツ」からなっている、と考える。前者の分野を「全般」、後者の分野を「その他」と呼ぶ。「その他」分野は下層にあるべきなので、親にある文書は子のコアワードを作成する際に「その他」分野の文書として子に加えて、子のコアワードを作成し分野関連度を計算する。次に、親は、子のすべてを含むべきであるため、「その他」と子すべてを利用して、コアワードを作成し分野関連度を計算する。その後、「その他」は親から派生した本来存在しない分野であるから、「その他」の分野関連度が作成した親に反映されるように、更に親の分野関連度を設定する。

・文書の分野自動判定

コアワードを利用した文書の分野自動判定の考え方について以下に述べる。例えば、「来季からのプロ野球参入を目指す楽天は 10 月 22 日、新チーム名を「東北楽天ゴールデンイーグルス」に決めたと発表した。」という文書では、チーム名は新語であるが、「野球」という語によって、野球分野であると判定することができる。しかし、例えば、「打たれ強いボクサーのような広島の執念が、優勝マジック点灯に王手をかけているヤクルトに再び「待った」をかけた。」という文書には、「ボクサー」のように他の分野の方でより特徴的である語や、「マジック」のように複数の分野で特徴的な語などがあり、野球分野に判定できるような決定的に特徴的な語はない。「広島」や「ヤクルト」もチーム名の略称であって複数の意味がある。このような場合には、「広島」「優勝」「ヤクルト」と合わせて考えて、野球分野であると判断するのが妥当である。そこで、以下のように考える。

ある文書が分野に関連する度合いを示す値を、文書の分野判定度とする。文書の分野判定度が高いほど、文書がその分野に関連する度合いが高いとする。コアワードの分野関連度に出現回数をかけた値をコアワードの分野判定度とする。文書の分野判定度は、判定したい文書に存在するすべてのコアワードの分野判定度を分野毎に合計した値とする。図 4-2-2(b)は文書の分野判定の例である。

・実験とその結果

我々が現在開発中の「訳してねっと」が所有する特許の分野を用いて、本手法の有効性を検証した。特許の分野は、特許文献を、文書中にあるIPCコードの1番目の分野に人手で分けて、階層化された分野にしたものである。特許の分野では最も深い層は3層目である。例えば、1層目に自然科学、2層目に自然科学の物理、3層目に自然科学の物理の核物理がある。3層目が最下層であるが1層目や2層目までの分野もある。特許分野の分

野数は89分野（1層目12分野、2層目44分野、3層目33分野）であり、文書があるのは76分野（1層目4分野、2層目39分野、3層目33分野）である。特許文書はタイトルとアブストラクトの部分を利用した。前述のように、分類された文書から最下層のコアワードを作成して分野関連度を計算し、子から親へ、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の分野関連度を付与した。分類された文書がある中間層のコアワードの分野関連度は、文書のみを利用して付与した場合と、文書と子を利用して親を作成して付与した場合とで、実験した。分野を判定する文書は、特許文書からランダムに抽出した4239文書で、文書中にあるIPCコードの1番目の分野を正解とした。

その結果、上位 1 位、上位 2 位以内、上位 5 位以内に正解が含まれた精度は、文書のみを利用した場合には、それぞれ、56%、72%、87%、親を作成した場合には、それぞれ、55%、70%、85%であった。親を作成した場合には、最下層の分野の精度が下がったが、中間層の分野の精度が上がった。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。

(2) 研究の効果

コアワードを利用した文書の分野自動判定の手法を確立した。階層構造の各分野に前もって自動的にコアワードの分野関連度を付与しておき、各コアワードの分野関連度を利用して分野を自動判定することによって、文書の分野判定時の処理を軽減することができる。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。これにより、文書を翻訳する際に、多種多様な分野に分類された辞書に対して、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分野を選定することができる。

４－２－３　コアワードを利用した辞書の不整合検知の研究

(1) 研究の内容

辞書の不整合検知とは、分野毎の辞書からその分野には不適切な辞書データを不整合として検知することである。不整合検知を行うことは、適切な分野に辞書データがあるようにする助けになるので、翻訳品質の向上につながる。

辞書の不整合検知を、コアワードを利用した分野自動判定の研究から、検討を行った。

・不整合検知

コアワードの分野関連度の定義や、コアワードを利用して辞書データの分野を自動判定する手順は、これまでの研究の通りである。辞書の不整合を検知する考え方は、以下である。ある分野に登録済の辞書データを分野判定し、上位にその分野が含まれないならばその分野に関連が弱いということなので不整合とみなす、という考え方である。

・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。毎日新聞(1995 年)の記事からコアワードの分野関連度を付与し、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抽出し、「訳してねっと」で登録されている分野を正解とした。その結果、自然科学分野に登録済であった「アデノシンデアミナーゼ」が不整合として正しく検知された（生物化学分野が正しい)。不

整合として検知されたが不整合とは言い切れないものには、データに複数の意味があるもの（「スーパーボール」：スポーツ分野に登録済だが玩具の意味の記事が多く不適合となった）や他の分野との関連の方が強かったもの（「ホームシアター」：趣味分野に登録済だが経済・ビジネスの記事が多く不整合となった）などがあった。

(2) 研究の効果

コアワードを利用した辞書の不整合検知の手法を検討し有効性を確認した。不整合を、分野を判定して上位にその分野が含まれないならばその分野に関連が弱い、と考えた。他の分野との関連の方が強いために不整合として検知されたものもあったので、上位下位という相対的な基準ではなく分野関連度を用いた絶対的な基準が設定できないかということも検討したい。

４－２－４　分野自動判定手法を利用した翻訳結果の評価について

(1) 研究の内容

翻訳結果を評価する目的は、システムが適切な分野を自動判定して翻訳することによって翻訳結果が向上するかを確認することである。そこで、文書を、自動判定された分野に関連する辞書を用いて翻訳し、辞書を用いずに翻訳した場合と比較して、評価を行った。

・実験とその結果

我々が現在開発中の「訳してねっと」が所有する特許の分野を用いて、翻訳結果の評価を行った。辞書を用いて翻訳したことによって、辞書を用いずに翻訳した翻訳結果から変化した文数の割合を、辞書ヒット率とする。翻訳に用いる辞書は、翻訳する文書の分野を自動判定して１位になった分野に関連する辞書である。関連する辞書は、その分野から特許分野の最上層までの親の辞書すべてとする。自動判定された分野が正解ではない場合の影響を確認するために、正解が上位１位、上位 2～5 位以内、それ以外に含まれた場合に分けて評価した。文書と正解の分野と自動判定された分野は、文書の分野自動判定の研究での実験結果からランダムに抽出して利用した。

その結果、正解が上位１位、上位 2～5 位以内、それ以外に含まれた場合の辞書ヒット率は、1 層目の分野が正解である文書の場合は、それぞれ、85%、85%、82%、2 層目や 3 層目の分野が正解である文書の場合は、それぞれ、75%、73%、77%であった。これらからは以下のことがいえる。1 層目の分野が正解である文書の場合は、正解が上位に含まれない分野の辞書を用いると辞書ヒット率が低下するが、2 層目や 3 層目の分野が正解である文書の場合は、正解が上位に含まれることと辞書ヒット率との関係は薄い。これは、2 層目や 3 層目の分野では、上層で、関連する辞書が一致することで、適切な辞書をいくらか使用することができるからだと推定される。

(2) 研究の効果

文書の翻訳結果を、自動判定された分野に関連する辞書を用いて翻訳した場合と、用いずに翻訳した場合とで、辞書ヒット率を比較して、評価を行った。1 層目の分野では、正解が上位に含まれる分野の辞書を用いると辞書ヒット率が高いことがわかった。他の視点からの評価も検討したい。

４－２－５　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。以下に項目毎の結論を示す。

1. コアワードを利用した文書の分野自動判定の手法を確立した。これにより、文書を翻訳する際に、多種多様な分野に分類された辞書に対して、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分野を選定することができる。

2. コアワードを利用した辞書の不整合検知の手法を検討し有効性を確認した。これにより、適切な分野に辞書データが登録されているようになれば、翻訳品質を向上させることができる。

3. 文書の翻訳結果を、自動判定された分野に関連する辞書を用いて翻訳した場合と、用いずに翻訳した場合とで、辞書ヒット率を比較して、評価を行った。これにより、分野によっては、正解が上位に含まれる分野の辞書を用いると辞書ヒット率が高いことがわかった。

4. 今後の課題は、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。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。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。特に、自動的に抽出して獲得した翻訳テンプレートをより適切に自動階層化することが望まれている。

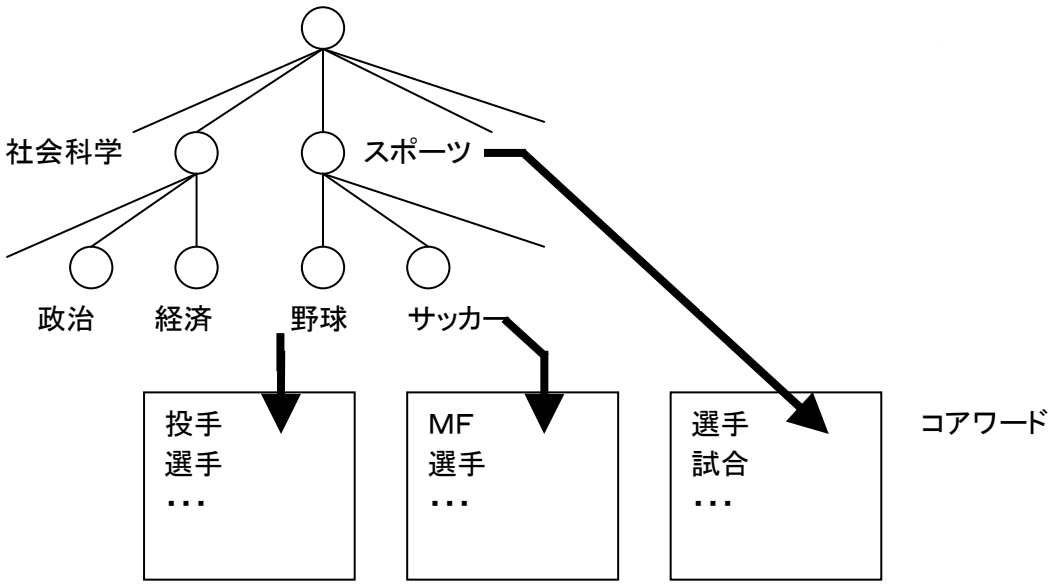


図 4-2-2(a)　階層化された分野とコアワードの例

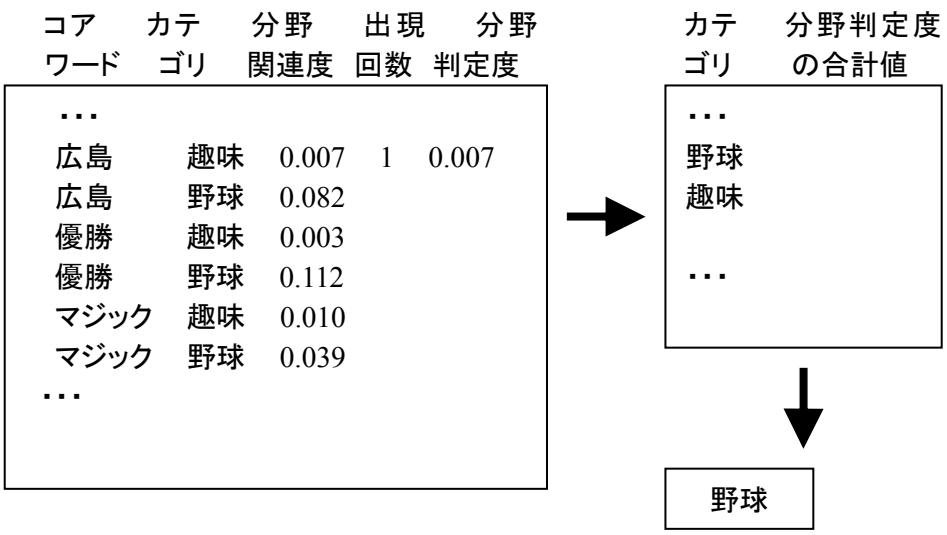


図 4-2-2(b)　文書の分野判定の例

４－３　言語非依存の翻訳エンジンの研究開発

４－３－１　序論

近年、経済などのグローバル化によって、英語以外にも様々な言語の翻訳の需要が増大している。翻訳システムの開発には莫大な工数が必要であるが、最小限の工数で色々な言語に対応できる翻訳システムを開発できるプラットフォームが求められている。ここでは、我々が開発した言語非依存の翻訳エンジン上で中日翻訳システムを開発し、短期間で開発できるかどうかを検証する。また、エンジンの問題点を洗い出し、ブラッシュアップを図る。

４－３－２　中日翻訳システムの研究開発

(1) 中国語に対応した多言語形態素解析器

昨年度行った中国語形態素解析の研究[文献 7]を踏まえ、翻訳エンジン上に中国語形態素解析器を組み込んだ。この際、中国語に特化した形態素解析器として実装するのではなく、他の言語への拡張が容易な多言語形態素解析器として実装し、中国語以外に日本語と英語にも対応させた。

多言語形態素解析器の基本構造を図 4-3-2 に示す。入力された文は、まず(1)インラインタグ処理部で処理され、文中の修飾要素が除去・保管される。次に、(2)前向き確率計算部で処理が行われるが、ここではまず(a)文正規化部が呼ばれて文中の文字の正規化が行われる。次に(b)解候補生成部が呼ばれ、単語辞書等を使用して解の候補を生成する。そして、(c)出現確率計算部が呼ばれ、解候補中の各要素に対する確率の計算が行われる。その後、(3)N-best 解探索部で処理が行われ、解候補の中から確率に基づいて複数の尤もらしい解(N-best 解)が選択される。最後に、(4)出力グラフ変換部によって、得られた N-best 解がその後の構文解析器内部で利用可能なデータ構造(グラフ)へ変換される。

日本語依存モジュール、英語依存モジュールともに、中国語依存モジュールと同じ構成となっているが、各言語に応じた正規化処理や解候補の生成を行う。

将来は、韓日翻訳に必要となる韓国語の形態素解析を行うため、この多言語形態素解析器を韓国語に対応させる予定である。韓国語の場合、出現確率の計算等は中国語や日本語と同様に行うことが可能である。しかしながら、韓国語は分かち書きを行う膠着言語であり、また縮約等の複雑な形態素の変化が起こるため、解候補の生成は中国語や日本語や英語とは異なった処理を行う必要がある。現在は韓国語依存モジュールの研究開発を行うとともに、韓国語形態素解析の評価に必要なデータ等の準備を行っている。

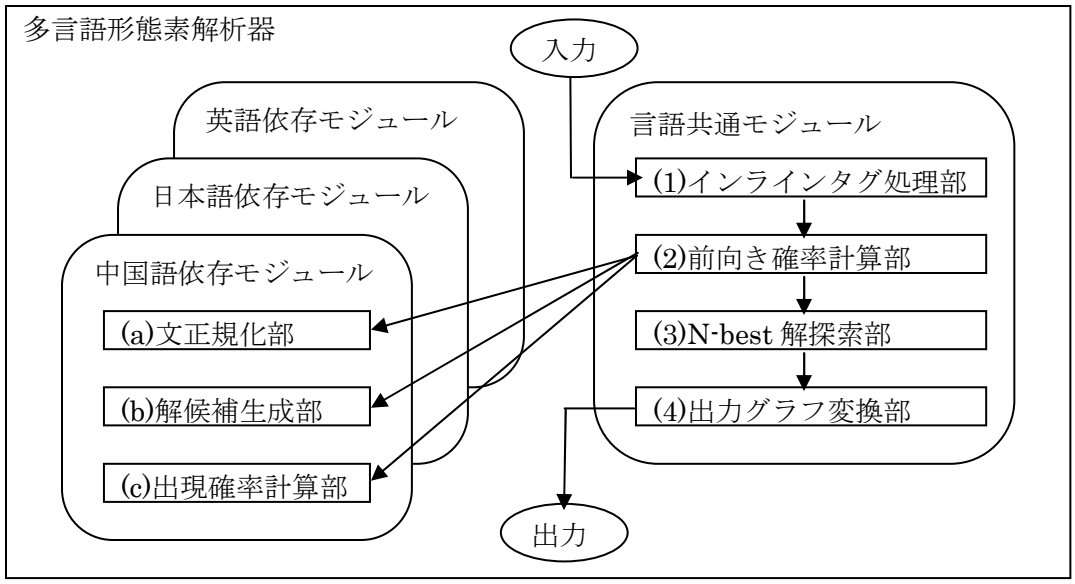


図 4-3-2　多言語形態素解析器の基本構造

(2)　中国語品詞タグ付きコーパス

中国科学院に依頼し、上述の形態素解析器で使用するパラメータの学習に必要な、中国語品詞タグ付きコーパスを準備した。このコーパスの作成は、以下の手順で行った。

1. 生コーパスの作成

インターネット上から、約 1,500,000 語の文書を収集した。文書の主な内容はここ数年分の中国の新聞で、約 1 ヶ月かけて 1 人の作業者により行われた。
その後、不必要な図表等の情報を除去し、集められた文書の整形を行い、生コーパスを作成した。

2. 単語分割と品詞の付与

中国科学院で開発された中国語形態素解析システムを使用して生コーパスを解析し、単語を分割して品詞情報を付与した。この形態素解析システムは、10,000,000 語からなるコーパスを用いてパラメータの学習を行ったもので、500,000 語からなる辞書を持ち、50 種類の品詞タグを使用する。

3. コーパスの修正

形態素解析システムにより自動的に単語分割と品詞付与を行った後、8 人の作業者に

よるコーパスの修正作業を 2 回行った。この作業には約 480 人・時間を費やした。これらの作業者は全員、このコーパスで用いる単語と品詞の基準に精通しておりコーパス修正の経験を持っている、自然言語処理の専門家である。

以上の作業により、1,089,000 語からなる中国語品詞タグ付きコーパスを作成した。このコーパスを用いて、上述の多言語形態素解析器で中国語の解析を行う際に必要となるパラメータの学習を行った。

(3) 辞書パターン

タグ付きコーパス中に出現する語を中心に約 9 万語の中国語見出しを抽出し、これらに日本語訳および意味素性などを付与する作業を行った。また、この辞書をシステムに組み込み、Web 上のさまざまな文書を翻訳し、うまく翻訳できなかった語約 1 万語を辞書に追加登録した。これにより、約 10 万語の辞書パターンが作成できた。

(4) 翻訳パターン

東京外国語大学の先生により作成された評価例文[文献 8]約 1000 文が翻訳できるように順次翻訳パターンを作成していった。約 500 パターンの文法を作成し、評価例文約 900 文が正しく翻訳できるようになった。作成した翻訳システムの評価結果を表 4-3-2 に示す。評価した文書は Web 上に載っていた、プレスリリースと時事ニュースの記事の２つである。これら文書を基本用語、専門用語、基本文法、専門文法の４項目に対して 20 点満点で評価した。一人の翻訳者による主観評価であるため、正確な評価結果とは言えないが、基本的な文の翻訳に関しては、市販の翻訳ソフトと大差がないレベルにあることが確かめられた。

表 4-3-2 訳質評価

	基本用語	専門用語	基本文法	専門文法
訳してねっと	14	11	12	9
K 社システム	13	8	10	8

４－３－３　協調的翻訳支援環境の研究開発

協調利用型機械翻訳システム「訳してねっと」を改良し９月末に一般公開を行った。今までに約 2,000 人のインターネット上のユーザがシステムにユーザ登録し、彼らによって約 3,000 語の辞書データが登録され、翻訳品質の向上に寄与している。また、ユーザからのフィードバックを収集・分析し、システム改良を行った。主な改良点は以下のとおりである。

(1) 専門用語抽出機能 – 文書中から、専門用語を自動的に抽出することができ、ユーザの辞書登録を支援する機能。

(2) おすすめ翻訳機能 – 入力された文書の分野を自動的に判定し、適切なコミュニティの辞書を使って翻訳する機能。

(3) 分類作成機能 – 翻訳の訳し分けに必要な意味分類などの分類をユーザが作成できる機能。コミュニティごとに作成することができ、わかりやすい語にしておくことで、一般ユーザも語を登録しやすくなる。

(4) 辞書承認機能 －管理者やコミュニティリーダが承認した辞書だけを翻訳に使うようにすることで、いたずらで登録されたものや、間違って登録された辞書データによって翻訳品質の悪化を防ぐことができる。

さらに、中日版「訳してねっと」の環境も構築し、他の言語にも簡単に拡張できることがわかった。中日版「訳してねっと」は 2005 年度中に一般公開する予定である。

４－３－４　結論と今後の課題

中国科学院にタグ付きコーパスに関し協力を依頼し、システム開発を行った結果、基礎的な構文に関しては他社製品と同等の翻訳品質を持った中日翻訳システムを構築することができた。開発した中日翻訳システムが実利用できるレベルに達したことが明確になっただけでなく、我々のシステムは、利用する翻訳知識を替えるだけで、多数の言語を扱うことが可能な多言語翻訳システムであることが立証された。

今後は、韓国語などの他の言語に対する翻訳システムを開発し、言語非依存の翻訳エンジンを改良していくとともに、「訳してねっと」を利用して、一般ユーザの力も借りることによって、短期間で様々な言語対の翻訳システムが構築できるような研究開発を行う予定である。

４－４　総括

今年度は、JAU(米国特許和文抄録)等、特許に関する対訳テキストを購入し、「特許文書」を対象とした大規模な翻訳テンプレート学習実験、及び、自己組織化における文書分野判定実験を行った。「特許文書」のような多くの分野を有し、かつ、専門的な用語を数多く含む文書において、我々の手法の有効性を確かめられたことの意義は大きい。他の種類の文書（例えば、医学文書）への適用可能性が確かめられ、今後、さらなる応用分野についても検討したいと考えている。

また、自己組織化に研究に関しては、今年度の成果によって、「登録語の分野自動判定」、「翻訳対象文書の分野自動判定」、「登録辞書の不整合検知」など、階層的な分野辞書を効果的に利用するためのツール群が整った。さらに、今年度は、自己組織化技術と翻訳技術を融合し、自動判定された翻訳対象文書の分野情報を利用することにより、翻訳品質が向上することが確かめられた。我々は234分野という大規模な階層構造を有するコミュニティ辞書を持っている。それらを効果的に扱う自己組織化研究は非常に重要である。来年度は、翻訳技術だけではなく、翻訳知識獲得技術との融合を図る研究を進める予定である。

最後に、多言語翻訳においては、既存の英日・日英システム構築の知見を生かし、かつ、中国科学院にタグ付きコーパスに関し協力を依頼し、システム開発を行った結果、予定より早く、基礎的な構文に関しては業界トップクラスの翻訳品質をもつ中日翻訳シ

ステムを構築することができた。開発した中日翻訳システムが実利用できるレベルに達したことが明確になっただけでなく、我々のシステムは、利用する翻訳知識を替えるだけで、多数の言語を扱うことが可能な多言語翻訳システムであることが立証された。

今後は、これらの知見を生かし、韓日翻訳システムを開発する予定である。また、中日翻訳システムの公開、中日対訳文書を利用した翻訳知識獲得研究等、多言語に展開した研究開発を行いたいと考える。

また、「訳してねっと」に関しては、β版として限定公開していたシステムを、9/29に協調利用型翻訳システム「訳してねっと正式版」として一般公開した。現在、安定稼働中であり、運用上、大きな問題がないことが確かめられた。ユーザからのフィードバックに基づき改良し続けている。

以上のように、本年度は、各研究テーマは、現実的なデータを利用した実証実験および評価を行い、その有効性を確かめた。その一方で、現実的な課題に対処するための問題を明らかにした。最終年度である来年度は、全ての研究テーマを連携させ、より実用的な多言語標準文書処理システムの実現（図 4-4）を目指して研究開発を進めていく予定である。

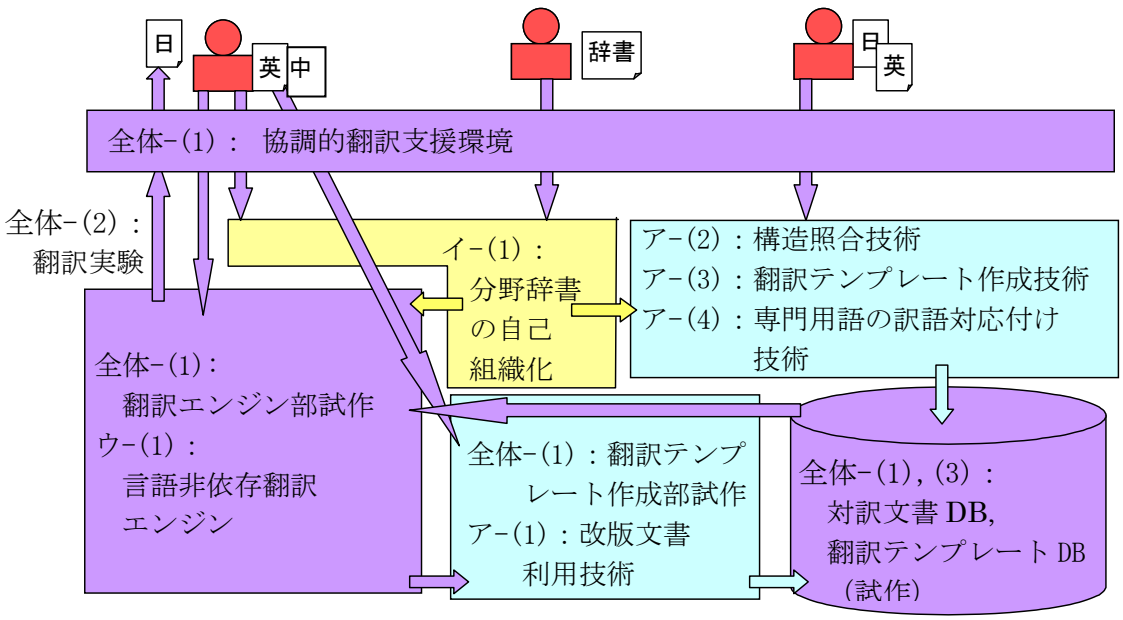


図 4-4: 各目標の関連および多言語標準文書処理システム全体像

参考文献

[文献 1] Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In The 5th Annual Workshop on Very Large Corpora, pages 192--202

[文献 2] Sayori Shimohata, et al.: "Retrieving Collocations by Co-occurrences and Word Order Constraints", Proceedings of 35th Annual Meeting of the Association for Computational Linguistics, pp. 476--481, 1997.

[文献 3] 下畑, 山本: "IDF を利用した n-gram 文字列の分類", 言語処理学会第 4 回年次大会, 1998.

[文献 4] 藤井, 岩山, 神門:"NTCIR-4 における類似特許検索テストコレクションの構築", 情報処理学会研究報告, 2004-NL-159, pp.45-52, Jan. 2004

[文献 5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In Proc. of the 40th ACL. pp.311-318, 2002

[文献 6] 金山, 荻野:"翻訳精度評価手法 BLEU の日英翻訳の適用", 情報処理学会研究報告, 2004-NL-154, pp.131-136, 2003

[文献 7] Tetsuji Nakagawa: Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information, In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, pp. 466-472, August 2004.

[文献 8]多言語機械翻訳システムの評価研究, 東京外国語大学、国際情報化協力センター, 2002

５　参考資料・参考文献

５－１　研究発表・講演等一覧

本年度は、以下の 3 件の研究発表を行った。

1.「単語レベルと文字レベルの情報を用いた中国語・日本語単語分割」中川哲治、松本裕治、情報処理学会研究報告、2004-NL-162、pp.197-204、2004

2. “Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information”, Tetsuji Nakagawa, In Proceedings of COLING-2004, pp.466-472, 2004

3. 「特許翻訳における専門用語辞書構築」下畑さより、山崎貴宏、坂本仁、北村美穂子、村田稔樹、言語処理学会第 11 回年次大会論文集、pp.356-359、2005