管理番号＃１４－０２

平成１７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１８年４月

情報通信研究機構

平成１７年度　研究開発成果報告書
（一般型）

「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。また、標準文書には、世界中の人々が様々な言語で記述し、翻訳ニーズが高い特許文書を選んだ。

我々と関連の深い文書処理システムの研究動向として、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 がある。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

上記の技術を導入することにより、Cliché では人手翻訳の 3～4 倍の効率化を達成している。この設計思想は、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。

一方、標準文書の処理という観点における関連の深い活動では、AAMT（アジア太平洋機械翻訳協会）が平成 15 年度に発足させた特許翻訳に関する研究会（AAMT/Japio 特許翻訳研究会）がある[文献 2]。この研究会では、主に、(i)ヨーロッパ特許庁(EPO)など諸外国の状況調査、(ii)特許の機械翻訳研究のリソースの構築、(iii)共通のデータセットを使った個別研究の推進、(iv)翻訳結果の自動評価手法の研究、の活動を行っている。当チーム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。

(i)においては、当社の研究員も調査団の一人として同行しヨーロッパにおける動向を視察した。ヨーロッパでは開発者とユーザによる共同開発の重要性が指摘されていたが、この考えは、協調型機械翻訳システム「訳してねっと」の開発において大いに参考になった。また、本年度は、機械翻訳の国際会議（MT-Summit IX）において AAMT/Japio が特許翻訳のワークショップを開催し、当研究員も企画メンバ、発表者として参加し、数多くの知見を得た。 (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た。このリソースは「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)は、我々の研究方針と類似する「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。また、(iv)は、近年、その重要性が指摘されており、我々も本年度は本格的に自動評価手法の導入した（4-1-4 節を参照)。

[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既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。

２　研究開発の全体計画

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2-1-1に示す。

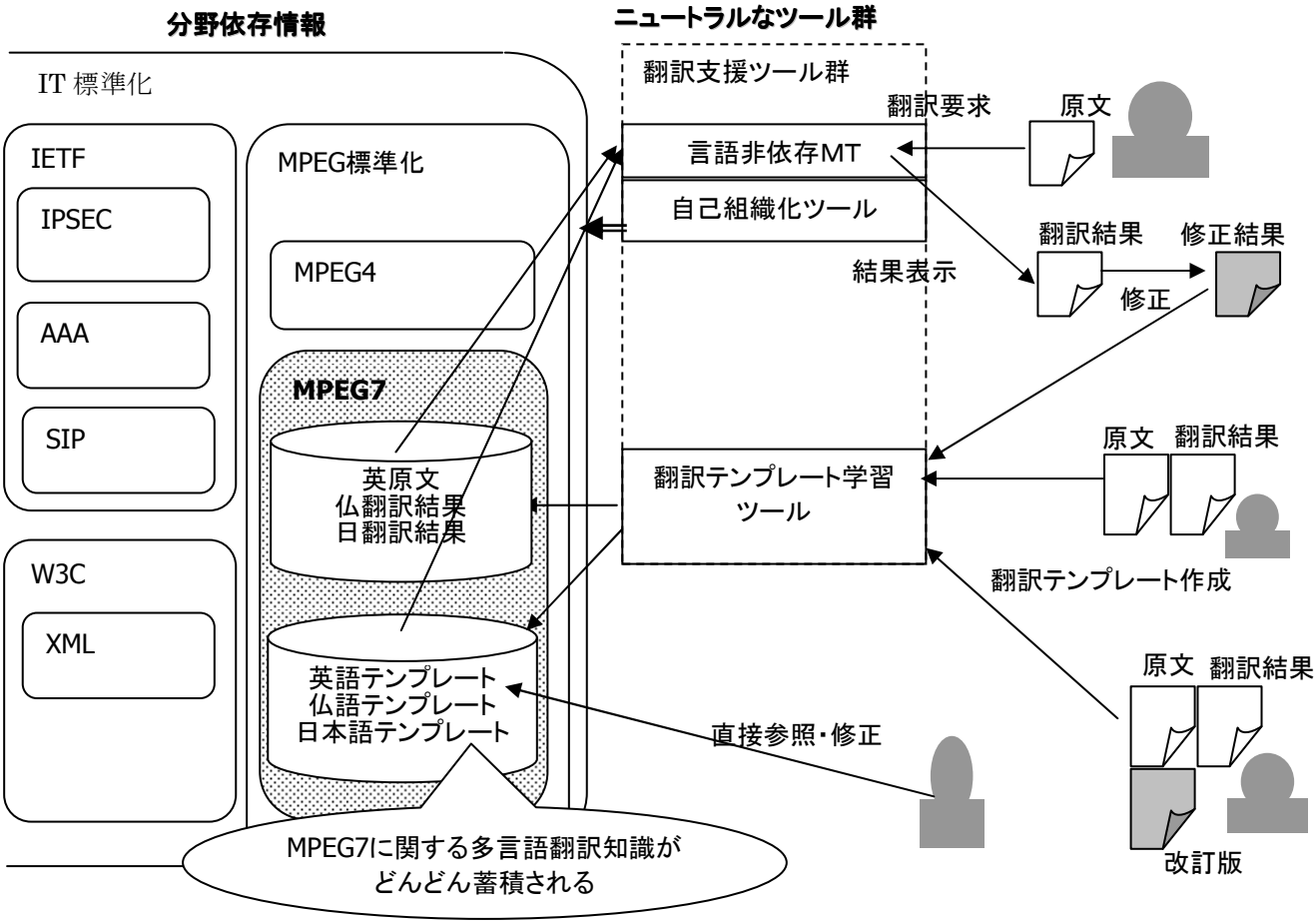


図2-1-1:　多言語標準文書処理システムの構成図及びユーザによる利用形態

２－２　研究開発目標

２－２－１　最終目標（平成 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 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

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

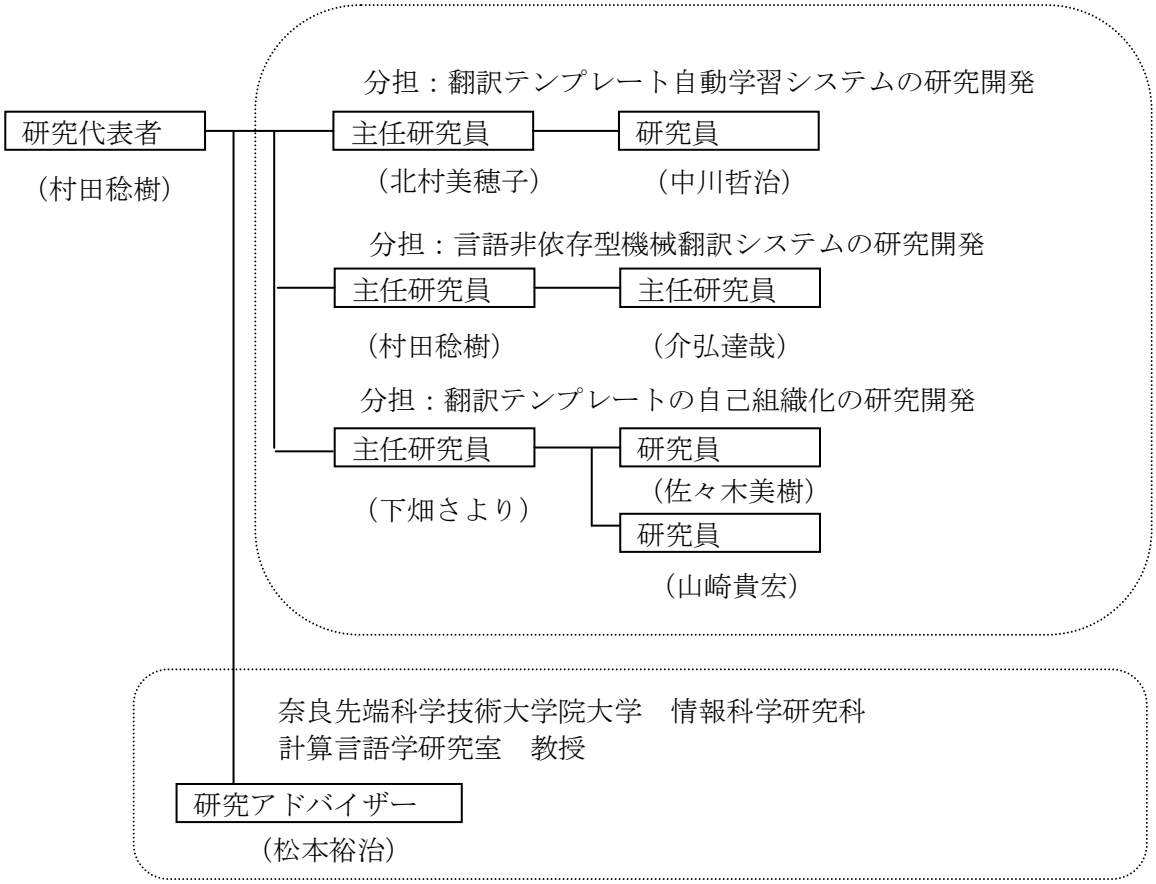
（金額は非公表）

研究開発項目	１４年度	１５年度	１６年度	１７年度	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

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）を自動的に作成する技術は、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、以下の 3 種類の手法を考案し、これらの手法を用いて既存の特許文書から翻訳テンプレートを作成した。

- 言語資源を活用した対訳表現の半自動抽出（4-1-2-(a)）
- 単語長と出現回数を考慮した対訳からの訳語検出（4-1-2-(b)）
- コンパラブルコーパスからの専門用語訳語自動抽出（4-1-3）

さらに、作成した翻訳テンプレートを「訳してねっと」上に搭載し、その翻訳テンプレートが翻訳品質に与える影響を機械翻訳の自動評価手法を用いて調査した(4-1-4)。

以下に、各項目の概要を説明する。

４－１－２　文対応済み対訳文書からの翻訳テンプレート獲得に関する研究

(a) 言語資源を活用した対訳表現の半自動抽出

手法の説明：既存の対訳文書や翻訳の用例から翻訳テンプレートを自動的に作成する技術において重要なことはその精度（作成した翻訳テンプレートの正確さ）である。抽出されたものが適切でなければ、抽出結果を人手で確認する手間等がかかり、一から人手で作成する手間とかわらないものとなる。そこで、我々は、「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」という従来手法[文献 3]を応用した精度の高い抽出方法を提案する。

この手法の特長は、**(1)**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、 **(2)**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、**(3)**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、という３点である[文献 4,5]。

実験：英日対訳文書における実験結果については、[文献 4]及び[文献 5]に結果の詳細及び考察が記載されているので、ここでは説明しない。以下、中英対訳文書を用いた実験及びその結果について報告する。

実験には、http://bowland-files.lancs.ac.uk/corplang/babel/babel.htm で公開されている中英対訳文書(以下 babel 文書と呼ぶ)を利用した。これは時事雑誌“World of English”及び “Time”から収集された 327 記事を中国語に翻訳したものである。英語は 253,633 語、中国語は 287,462 語からなり、12,176 ペアの対訳数を有する（但し対応は、1 文対 1 文とは限らず、複数文対 1 文や複数文対複数文の対応も存在する)。また、辞書は、http://www.mandarintools.com/cedict.html で公開されている中英対訳辞書を用いた。この辞書の総対訳ペア数は 55,318 ペアである。

これらの文書、辞書を用いて、図 4-1-1 に示す方法で対訳表現の自動抽出を行った。 設定された条件及びパラメータを表 4-1-1 にまとめる。

類似度計算	Log-likelihood
文節区切り	あり （但し、中国語は句読点、接続詞の前後で区切るのみ)
最低閾値(fmin)	2 回
対訳辞書利用	あり
人手確認	なし
文書分割	2000 文毎に抽出 （分割途中の最低閾値(fmerge)は 3 回とした)

表 4-1-1: 中英対訳文書による対訳表現抽出での設定条件

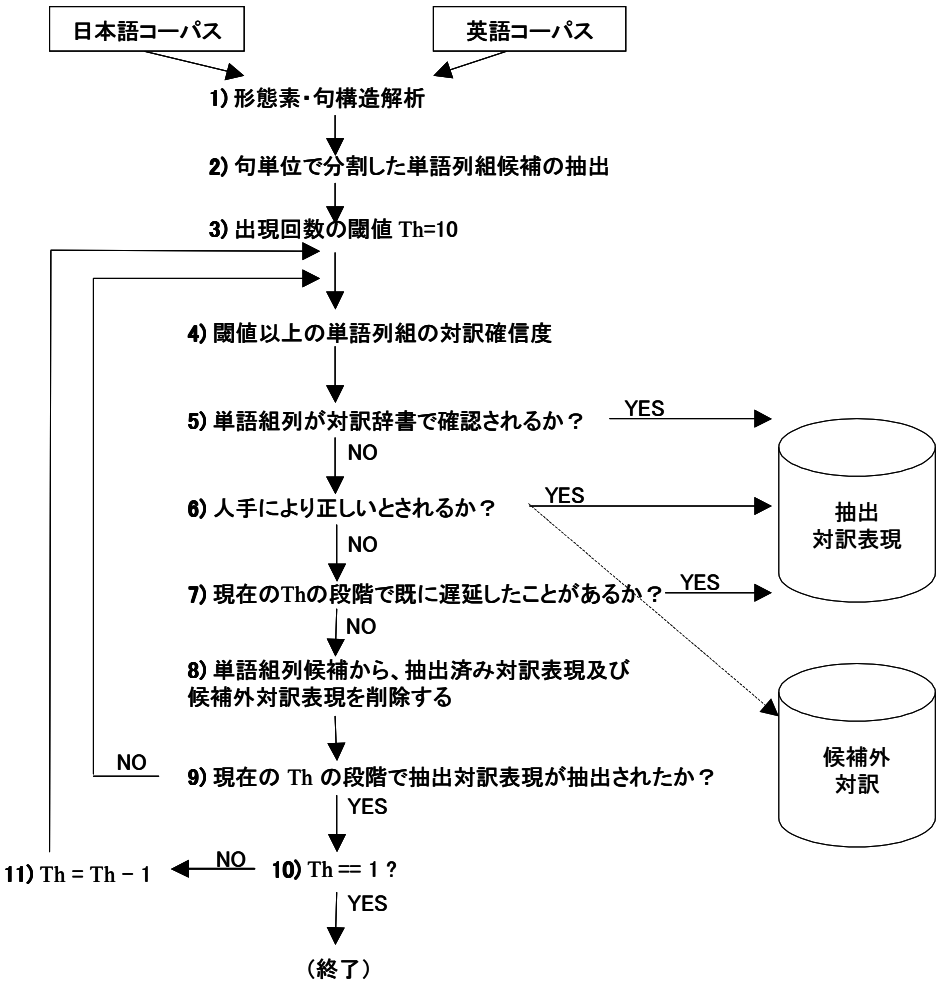


図 4-1-1：対訳表現の半自動抽出の流れ

評価は、精度とカバレッジを求めることにより行った。精度は、対訳表現抽出結果を

正解： 対訳表現をそのまま辞書として登録できる
半正解： 対訳表現のどちらか一方の一部の表現を削除すれば辞書として登録できる
不正解： 正解及び半正解以外

の三段階で評価し、抽出総数に対する正解及び半正解の割合を百分率(%)で求めた。表 4-1-2 では半正解の割合を()内に示す。

一方、カバレッジは、英語、中国語それぞれの文書において

$$\mathrm{cov}\,erage(\%) = (1 - \frac{未抽出自立語異なり語数}{文書中自立語異なり語数}) \bullet 100$$

を計算した。上記式内の「未抽出自立語異なり語数」とは各文書から対訳表現抽出後を除去した結果、残った自立語の異なり単語数である。また自立語の異なり単語数だけでなく、自立語の総出現単語数に対しても同様にカバレッジを求めた。表 4-1-2 では後者を()内に示す。

実験結果及び評価： 表 4-1-2 は、babel 文書において上記の設定条件を用いた場合における対訳表現の抽出結果である。8,000 ペアからなる英日対訳文書を用いた実験では、精度 89%、カバレッジ 85%で約 6,400 語の翻訳テンプレートを獲得することができたが[文献 5]、中日対訳文書においては、精度 75%、カバレッジ 55%となり、英日より劣る結果となった。これは、英日版では、対訳表現として有効か否かの判断に、文節区切り情報や品詞情報をきめ細かく利用しているのに対し、中日版では英日版のような言語依存の情報をほとんど利用していないためである。今後、中国語独自の言語情報を利用することに

		対訳辞書利用
抽出総数(対訳辞書登録有り語数)		5,605 ペア(2,612 ペア)
精度	全体	75.1% (78.1%)
	辞書有り語	92.4% (94.7%)
	辞書無し語	61.8% (65.3%)
カバレッジ	英語	8% (67%)
総数(異なり語数)	中国語	17% (52%)

表 4-1-2：babel 文書における対訳表現抽出結果

評価	中国語	英語	対応度
正解	奥尼尔	O_'_Neal	151.039
	星球_大战	Star_Wars	97.32
	环球影城	universal_studio	22.6718
半正解	克利夫兰_诊所	Cleveland_clinic_'s_department	25.802
	太阳_微_系統_公司	Sun_Microsystems	19.428
	在_索马里	Somalia	19.428
不正解	现在_的_意见	say_now	84.6556
	路易威登	Gucci	56.0205
	创意_总监	creative_director	19.428

表 4-1-3：babel 文書における対訳表現抽出

より、英日版と同等の精度で対訳表現を抽出できると思われる。

次に、表 4-1-3 に、babel 文書で抽出された対訳表現の例を示す。本手法では「环球影城：universal studio」のように、多くの固有名詞表現を抽出することができた。逆に不正解となった結果を分析すると、大きく３つのタイプに分けられる。「现在_的_意见：say_now」のように中国語の語句の範囲の認定が正しくないもの、「路易威登（ルイビトン）：Gucci（グッチ)」のようになんらかの関係はあるものの訳語としては正しくないもの、「创意_总监：creative_director」のように英語、中国語のそれぞれにおいて余計な語が存在するものの 3 タイプである。今後、半正解、不正解の結果をより詳細に分析することで、本手法を改良し、さらなる精度向上に努めたい。

(b)単語長と出現回数を考慮した対訳からの訳語検出

背景：上記に説明した対訳表現抽出手法は、第一言語と第二言語の対訳文書から網羅的に対訳表現を抽出する方法である。しかし、もしどちらか一方の言語の専門用語のリスト等が存在すれば、その訳語を対訳文書から統計的に検出する手法の方が有効である。

このような片言語の用語リストの訳語を発見する手法は、数多く提案されている。一般的な手法は、統計情報と既存の対訳辞書を利用する方法[文献 6,7]である。しかし、対象とする対訳文書が特許公報である場合、特許公報には専門用語が多く、既存の対訳辞書情報はあまり有用でない。特許公報を対象にする手法としては、特許公報の数情報を利用して対応関係を推定する手法[文献 8] 、統計情報と特許文書特有の構成単語の品詞情報を利用する方法[文献 9]がある。

我々が本研究で扱う対象も、[文献 8,9]と同様、特許公報である。したがって、一般的な手法で用いる既存辞書等の言語情報を利用することができない。また、我々が主に利用したい対訳関係が明瞭な特許の発明の名称（タイトル）には数情報はないので[文献 8]の手法は有効でない。さらに、[文献 9]における実験結果の精度もそれほど高くない。そこで、我々は単語長と出現回数を考慮した統計的な手法を用いた訳語検出手法を新しく考案した。

手法の概要： 図 4-1-2 に英語の用語から日本語の訳語を検出する処理の流れを示す。まず、日英対訳コーパスから訳語発見の対象とする英語用語を含む英文を検索し、その英文に対応する日本語文を形態素解析、チャンク解析をする。そしてチャンクを超えない範囲で、全ての n-gram 単語列を抽出する。ここで抽出された n-gram 単語列に対して、入力の英語用語との対応度を計算し、予め設定された閾値を超える対応度を有する日本語単語列を訳語候補とする。

対応度 *sim*("*input*"⇒"*str*") の計算方法は、単語長と出現回数を考慮した以下の式で定義する。

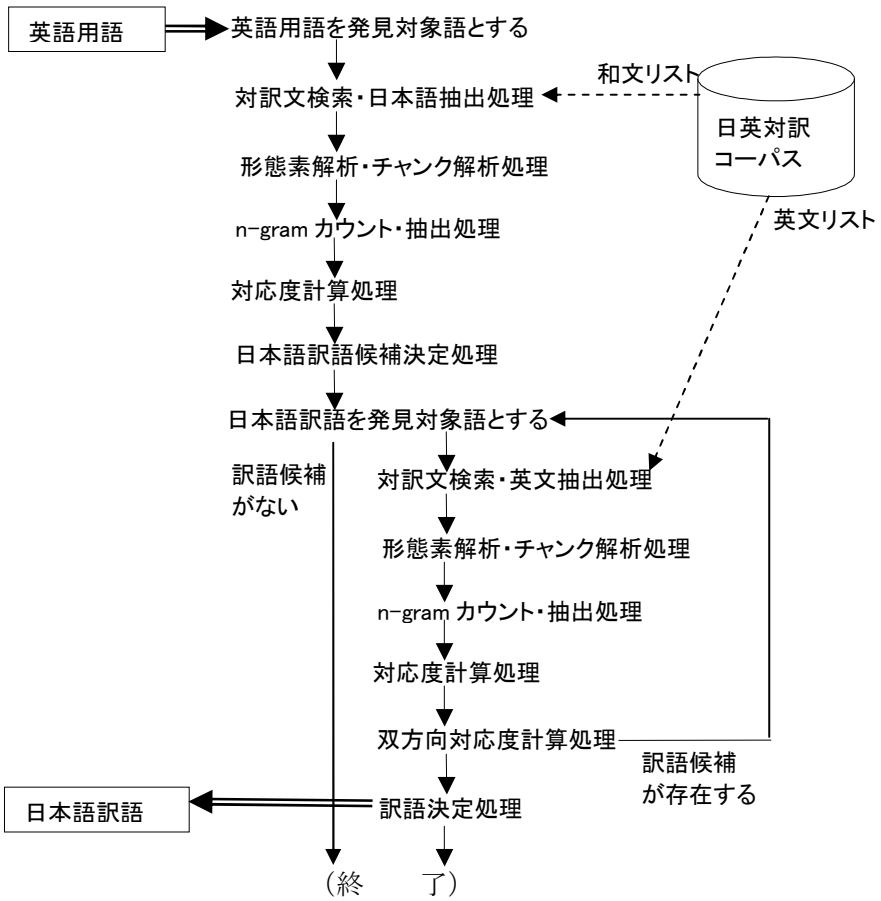


図 4-1-2：訳語検出処理の流れ

$$sim(\text{"}input\text{"} \Rightarrow \text{"}str\text{"}) = \sum_{\text{"}substr\text{"} \in \text{"}str\text{"}\textbf{の部分単語列}} \frac{conNum(\text{"}substr\text{"})}{conNum(\text{"}str\text{"})} \bullet freq(\text{"}substr\text{"})$$

$conNum(\text{"exp"})$:"exp"***の構成単語数***
$freq(\text{"exp"})$:"exp"***の出現回数***

具体例を挙げて説明する。英語の専門用語 “aspergillus oryzae”と日本語単語列「アスペルギルスオリザエ」の対応度の計算は、次のようにして求める。
「アスペルギルス　オリザエ」の部分単語列は、「アスペルギルス　オリザエ」、「アスペルギルス」”、「オリザエ」の 3 つである。したがって、その 3 つに対して $\frac{conNum(\text{"}substr\text{"})}{conNum(\text{"}str\text{"})} \bullet freq(\text{"}substr\text{"})$ を求め、それらを合計することによって $sim(\text{"}aspergillus\ oryzae\text{"} \Rightarrow \text{"}$***アスペルギルス オリザエ***$\text{"})$ を求める。この式は $freq(\text{"}substr\text{"})$ によって、部分単語列の出現回数が多いほど対応度が高くなるように設定されている。さらに、$\frac{conNum(\text{"}substr\text{"})}{conNum(\text{"}str\text{"})}$ によって、その部分単語列の構成数の大小に応じて重みを与えている。また、$\sum$ を計算することにより、構成単語数(つまり単語長)が長くなるほど、部分単語列の総数が多くなり、その総和も大きくなる。これにより、構成単語数(つまり単語長)が長いほど、対応度が高くなるという性質を対応度 $sim(\text{"}input\text{"} \Rightarrow \text{"}str\text{"})$ に与えることができる。

次に、上記の式によって計算された対応度が予め設定された閾値を超える日本語訳語候補について、逆の方向の訳語検出を行う。つまり、日本語訳語候補を入力とし、対訳文を検索し、その訳語候補を含む日本語文に対応する英文を抽出し、形態素解析・チャンク解析処理をし、チャンクを超えない範囲で全ての n-gram 単語列を抽出する。但し、正解となる日本語訳語候補の出現回数は、入力となる英語用語の出現回数とほぼ等しいことが予測されるので、対訳文を検索した際、入力となる英語用語の出現回数の 100 倍以上の出現回数を有する日本語訳語候補は予め除外しておく。

ここで抽出された n-gram 単語列の中に、入力とした英語用語が含まれており、かつ、日本訳語候補

と英語用語の対応度が、予め設定された閾値を超えるならば、 $sim("input" \Rightarrow "str")$ と $sim("str" \Rightarrow "input")$ との和を求めることにより、双方向対応度 $sim("input" \Leftrightarrow "str")$ を計算する。

$$sim("input" \Leftrightarrow "str") = sim("input" \Rightarrow "str") + sim("str" \Rightarrow "input")$$

全ての日本訳語候補に対して、上記の計算を行い、双方向対応度の高い順番から日本語訳語候補を出力する。

実験および評価：上記手法の有効性を確かめるために、本手法における訳語検出実験を行った。訳語検出に必要な専門用語リストは、既に我々が提案した隣接単語の分散に基づく用語抽出技術[文献 10]によって抽出された未知語及びイディオムを利用した。用語抽出に用いたデータは、研究用として我々が所持している米国特許 (US-patent)6 万文である。一方、訳語検出には、日本語特許とその英文抄録(PAJ)のタイトルの対訳約 350 万文と、米国特許とその日本語訳(JAU)のタイトルの対訳約 233 万文を用いた。

まず、米国特許 (US-patent)6 万件に対して、出現回数が 10 回以上の単語列を英語用語として抽出し、さらに 1 単語構成語と複数単語構成語に分けた。1 単語構成語に関しては品詞推定し、訳語検出の対象として未知語 734 語のみを抽出した。複数単語構成語は、12,591 語のうち、ランダムに選んだ 802 語を訳語検出の対象として利用した。一位正解率(%)は(一位正解数／検出数)×100、含正解率(%)は(正解を含む結果数／検出数)×100 として求めた。カバー率(%)は、(検出語数／訳語検出対象語数)×100 で求めた。

	未知語	イディオム
訳語検出対象語数	734 語	802 語
検出語数	575 語	624 語
一位正解語数	462 語	493 語
一位正解率	80.3%	79.0%
含正解数	556 語	595 語
含正解率	96.7%	95.4%
一用語当たりの訳語候補語数	4.2 語	4.6 語
カバー率	78.3%	77.8%

表 4-1-4：米国特許から抽出された用語の訳語検出結果

本実験結果を表 4-1-4 に示す。未知語、イディオム共、第一位の候補語は、約 80%の正解率であった。また、一用語当たり抽出される訳語候補語数は、4.2～4.6 語であったが、その中に正解が含まれる割合は、未知語、イディオム共、約 96%とかなり高かった。カバー率に関しては、未知語、イディオム共、78.8%前後となった。

検出できなかったイディオムの英語用語 178 語について、検出できなかった原因を調査した結果、主に 3 つの原因が存在した。第一の原因は、出現回数が 7 回以下で、かつ、検出に用いた対訳文書中の訳語が一意でなかったため、閾値を満たす訳語候補語が存在しなかったためである。大半の用語(174 語)がこれに相当した。第二の原因は、チャンク解析間違いによるもので 2 語存在した。第三は、英語用語の出現回数と、正解となる訳語候補語の出現回数が極端に異なるため、正解となる訳語候補語が予め除外されてしまった原因によるもので、これも 2 語存在した。

今後、カバー率を向上させるためには、第一の原因を解消することが有効であろう。例えば、検出できなかった英語用語は、閾値を低くし再検出する、双方向対応度の計算によって訳語候補語を検出するのではなく、英語用語から日本語候補語の対応度の計算のみで検出する等が考えられる。しかし、これらの手段は、一位正解率の低下を伴う危険性もあるので、実装には工夫が必要である。

一方、間違った結果が得られた用語は、未知語、イディオム共に、一用語当たりの訳語候補語数が 1.9 語と少なかった。間違いは、次の 3 タイプに分類される。

(1)用語の訳が訳語候補語の一部に含まれる場合（表 4-5(1)）：　未知語 12 語、イディオム 9 語
(2)用語の訳の一部のみが訳語候補になっている場合（表 4-5(2)）：未知語 1 語、イディオム 7 語
(3)対応関係がない（表 4-5(3)）：　未知語 6 語、イディオム 9 語

	英語用語	日本語訳語候補語	評価
未知語	xerography	ゼログラフィ	◎
	varifocal	可変焦点レンズ 可変焦点 バリフォーカルレンズ	○
	arylene	アリーレンスルフィドポリマー	×(1)
	backpatching	-	―
イディオム	magnetic thin film	磁性薄膜	◎
	actuator assembly	アクチュエタ アクチュエタアセンブリ 作動装置	○
	amorphous semiconductor layer	非晶質半導体	×(2)
	liquid vehicle	スルホポリエステル ヒドロアルコール エアゾールヘアースプレー	×(3)
	back focal distance	-	―

表 4-1-5：米国特許から抽出された用語の訳語検出例

評価は、◎：一位正解語、○：含正解語、×：不正解語、－：未検出語　を表す

これらの間違いの原因は、正解となるべき候補の対応度にばらつきが生じ、間違った候補が選択されてしまったと考えられる。例えば、　表 4-1-5 の英語用語“liquid vehicle”を含む対訳文の“liquid vehicle”の訳語は、「液体」、「液状の媒体」、「溶剤」、「液体賦形剤」、「液体ビヒクル」と全て異なっていた。

表 4-1-5 に、本手法で抽出された訳語候補の例を示す。この例ように、本手法を利用することにより、一般的な対訳辞書には見られない数多くの専門用語の訳語を検出することができる。

本手法は、精度が 100%でないため、検出された訳語候補は、人手で正しいか否かを確認した後、登録しなければならないという課題が残る。しかし、96%の含正解率ということは大半の場合、正解が含まれているということである。本研究で扱った特許文書では、対象とする英語用語は技術用語が大半であるので、日本語訳語候補になる語は、表 4-1-5 のように、英語のカタカナ表記(例：xerography の訳語はゼログラフィ)や英語表現と同一(例: ALU の日本語訳語候補は ALU)である場合が多い。したがって訳語が正しいかどうかを辞書で確認する手間はほとんどかからず、その分野の専門知識がない人でも容易に正解を得ることができる。

４－１－３　コンパラブルコーパスからの翻訳テンプレート獲得に関する研究

(a) はじめに

本研究は、文対応のついていない、同一分野の対訳文書(コンパラブルコーパス)を用いて、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は、文対応済み対訳文書(パラレルコーパス)と比べて文書が多数存在するので、実用化が望まれる技術である。しかし、従来手法[文献 11,12]では対応付けの精度が低い、計算量が大きいなどの問題があり、さらなる改善が望まれている。そこで、我々は既存の文対応がない対訳文書を利用した用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[文献 13]。

(b) 本手法の提案

[文献 11]では、「文書 T において用語 A と用語 B の関連度が高い場合、文書 T'において用語 A の訳語である A'と用語 B の訳語である B'も関連度が高い」という仮説に基づき、seed word リストと呼ばれる既知の訳語対と、各言語のコーパスにおける候補語との関連度を測り、関連度の高い候補語どうしを

対応付ける方法を提案している。ここで、単語の関連度とは、一定の範囲内(例えば、文、パラグラフなど)における出現文脈の類似度である。つまり、seed word リストの単語と候補語との各単言語コーパス中での共起を調べ、その共起パターンの類似度を測って、類似度の高い単語ペアを対訳語として抽出する。それに対して、我々は seed word 獲得時にも seed word 候補語どうしの共起を調べ、出現文脈が類似している単語ペアのみを seed word として採用する方法を提案した[文献 13]。つまり、対象コーパスにおいて有効な seed word のみを利用することにより、候補語の対応付けにおける精度の向上を計る。

本研究の専門用語およびその対訳候補の獲得プロセスは、1．seed word リストの作成、2．専門用語の抽出、3．seed word を用いた専門用語の訳語推定という 3 つのステップから構成される。以下では、各ステップの詳細を説明する。

seed word リストの作成：以下の手順に従って seed word リストを作成する。

1. 対訳辞書より、コーパス C_J および C_E に閾値 $freqMin$ を超えて出現するデータ対の集合 $D_{JE} = \{\langle d_J, d_E\rangle \mid freg(d_J) \geq fregMin; freq(d_E) \geq freqMin\}$ を抽出する。
2. D_{JE} の各単語 d_i に対して、d_i と同一言語の単語辞書データ $(d_1, d_2, \ldots d_n)$ との、コーパスにおける一定の範囲内(例えば、一定語数内、一文内、パラグラフ内など)での共起頻度 $Cmat(t_i)$ を求める。
3. 対訳関係にある d_{Ji} と d_{Ei} の共起マトリクスの類似度 $Sim(Cmat(d_{Ji}), Cmat(d_{Ei}))$ を求め、類似度が閾値を超えるデータの対を seed word リスト $SW_{JE} = \{\langle d_J, d_E\rangle \mid sim(d_J, d_E) \geq simMax\}$ として抽出する。類似度 $Sim(Cmat(d_{Ji}), Cmat(d_{Ei}))$ は、以下の式によって得る。

$$Sim(Cmat(w_J, w_E)) = \sqrt{\sum_{1<i<n} (Cmat(w_J) - Cmat(w_E))^2}$$

専門用語の抽出：日英各単言語コーパスより、専門用語の候補を抽出する。ここでは用語の前後に出現する単語の分散の度合いによってユニット性を、また、頻度および tf・idf によってターム性を計測する。具体的な手順は以下の通りである。

1. コーパスを形態素解析した結果から、可能な n-gram 単語列 $T_J = \{t_{J1}, t_{J2}, \ldots t_{Jn}\}$ および $T_E = \{t_{E1}, t_{E2}, \ldots t_{Em}\}$ を抽出する。
2. t_i の前後に出現する単語の種類と出現数から、t_i のユニット性を測る。
3. さらに、頻度、tf・idf によってフィルタをかけて候補語を選別する。

専門用語の訳語推定：訳語対応づけは、前節で述べた seed word 作成プロセスと同様の手段によって行なう。

1. 両言語のコーパスから抽出された T_J および T_E の各用語 t_i と、SW_{JE} の各要素 d_k との各コーパスにおける共起頻度を計数し、共起マトリクス $Cmat(t_i)$ を作成する。
2. T_J と T_E のすべての用語の組み合わせにおける共起マトリクスの類似度 $Sim(Cmat(t_{Ji}), Cmat(t_{Ei}))$ を求め、類似度の高いデータの対を対訳として抽出する。

(c) 実験及び評価

日英特許抄録(PAJ: Patent Abstracts of Japan)1 年分および米国特許抄録(USP)を用いて、専門用語辞書の抽出および訳語の抽出実験を行なった。実験に使用したデータは以下の通りである。

・PAJ コーパス
　　日英特許抄録(C12N:遺伝子分野)　11,781 件
　　日本語　38,481 文(1,694,362 単語)
　　英語　　35,343 文(1,778,663 単語)
・USP コーパス(C12N:遺伝子分野) 12,534 件
　　51,201 文(1,259,299 単語)
・正解データ
　　Japio 辞書(C12N:遺伝子分野)　4,789 件

freq(Tj)	*Tj*	*freq(Te)*	*Te*
14203	遺伝子	16929	acid
13592	配列	16918	sequence
11143	細胞	16333	gene
10013	提供	13193	cell
8899	DNA	12749	DNA
6794	酵素	11459	protein
6379	製造	11134	solution
6328	培養	8347	amino acid
5942	を コード する	7892	culture
5719	活性	6049	amino acid sequence
5405	タンパク質	6010	provide
5277	アミノ酸	5649	enzyme
5166	新規	5032	activity
5144	微生物	5016	human
4756	発現	4809	microorganism
4478	蛋白	4531	encode
4105	ポリペプチド	4422	polypeptide
3998	蛋白 質	4359	formula
3787	本 発明	4217	comprise
3679	構成	3960	nucleic acid

表 4-1-6：専門用語抽出結果の上位 20 件

Tj	*Te*
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 4-1-7：訳語対応付けの結果の例

・対訳辞書

EDICT　173,000 件

テストコーパスとして用いる PAJ は、遺伝子分野(IPC コード：C12N)の日本語特許抄録およびその翻訳で、日本語および英語で記述された「発明の名称」および「要約」の文を用いた。また、米国特許抄録についても、同様に、同じ C12N 分野のタイトルとその要約を用いた。

評価は同分野の Japio 辞書を正解データとし、Japio 辞書の見出し語のうち日本語、英語とも PAJ 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。Japio 辞書は、PAJ 作成の補助に用いられているもので、人手で作成した辞書である
専門用語の抽出では、[文献 10]に基づき、$Entropy(str) \geq 1$ かつ $Freq(str) \geq 100$ の単語列を抽出した。その結果、日本語 1143 件、英語 1098 件の用語が抽出された。抽出結果の上位 20 件を表 4-6 に示す。抽出結果と Japio 辞書の 57 件との重複は日本語 51 件(89.5%)、英語 48 件(84.2%)で、日本語、英語ともに抽出されていた見出しは 43 件(75.4%)であった。
次に、日本語、英語とも専門用語として抽出された 43 件を含む日本語 1143 件と英語 1098 件の対応付けを行った。対応付け結果の例を表 4-7 に示す。日本語の各用語に対して、英語の用語を類似度の高いものから順に対応付けたところ、対象となった 43 件における正解の平均順位は 14 位となった。また、正解の英語専門用語が第 1 候補になったものは 24 件(55.8%)、20 位以内に抽出されたものは 39 件(83.0%)であった。

次に、PAJ の日本語抄録と同分野の米国特許抄録を用いて、訳語の対応付け実験を行った。PAJ コーパスと同様の条件で単語列を抽出したところ、英語 885 件の用語が抽出された。これと PAJ で抽出された日本語 1143 件との組み合わせで見たところ、Japio 辞書の見出し語で日本語、英語ともに抽出されていた見出しは 47 件であった。PAJ、USP 両コーパスで専門用語として抽出された日本語 1143 件と英語 885 件の対応付けを行った結果、正解の英語専門用語が第 1 候補になったものは 9 件(19.1%)、20 位以内に抽出されたものは 26 件(55.3%)となった。PAJ どうしでの対応付け精度と比べて全体的に精度が低くなっているが、これは、テキストの類似性に差があるためと考えている。

４－１－４　特許翻訳における翻訳テンプレート利用効果の検証実験

(a) はじめに

我々の機械翻訳システムの特長は、辞書すなわち翻訳テンプレートの追加が容易なことである。上述した手法により獲得した翻訳テンプレートの追加は頻繁に行われる。翻訳テンプレートの追加は、翻訳品質の向上を目的とするが、思わぬ副作用により、翻訳品質の低下が起こることもある。したがって、システムのバージョンアップを行う際には、翻訳品質の低下が起こっていないかどうかを常に確認する必要がある。従来は、人手で評価していたが、人手評価には以下のような問題がある。

- 評価に時間がかかる。
- 特許がカバーする分野は広く、全ての分野の文書を網羅的に評価することができない。
- 評価者に高度なスキルが求められる。
- 評価指標の設定が難しく、点数が評価者に依存する。

上記の問題を解決するため、本研究では、機械翻訳システムの自動評価手法を導入することにした。自動評価手法とは、機械翻訳の結果と参照訳（人が作成した正しい翻訳結果）との類似度(評価値)を測定する手法のことで、その評価値を比較することにより、機械翻訳システム間の翻訳品質の優劣を決定することができる。機械翻訳システム間の相対評価には有効であることが既に報告されている[文献 14]。

(b) 自動評価の代表的な手法

以下の５つの手法を使用した。いずれも国際的なワークショップなどで使用されている代表的な手法である［文献 15］。

(1) BLEU (Bilingual Evaluation Understudy)［文献 16］
n-gram 一致数（通常 4gram）を元に算出する。0～1 の実数で値が高いほど良い。語順の正しさ、翻訳の流暢さを重視した自動評価スコアである。

(2) NIST (National Institute of Standards and Technology で開発された)［文献 17］
BLEU の変形で、n-gram 一致数（通常 5-gram）を元に算出する。個々の n-gram に情報量に基づく重みが付けられている。0～∞の実数で値が高いほど良い。機能語より内容語に重みが付き、単語訳の正しさを重視した自動評価スコアである。

(3) WER (Word Error Rate)［文献 18］
編集距離（挿入、削除、置換の操作によって２つの文が一致するまでの距離。通常、各操作を 1 として合計する)を参照訳の語数で割る。0～1 の実数で値が低いほど良い。もし十分な参照訳が与えられれば、主観評価と相関が高い。

(4) PER (Position independent word Error Rate)［文献 19］
WER の変形で、正しく翻訳されていても語順が異なる場合があるため、語順を無視してエラー率 Word Error Rate を算出する。0～1 の実数で値が低いほど良い。もし十分な参照訳が与えられれば、主観評価と相関が高い。

(5) GTM (General Text Matcher)［文献 20］
unigram の MMS(Maximum Matching size) に基づく F 値（適合率と再現率の調和平均）を元に算出する。MMS は重複なしに最大に一致した部分をできるだけ取った数の合計である。0～1 の実数で値が高いほど良い。

(c) 実験及び結果

(1) 自動評価の有効性の検証

自動評価の有効性を検証するために、自動評価と人手評価との相関関係を調べた。翻訳システムのあるバージョン(a)を基準にして、自動評価の(1)から(5)のすべての手法で(a)と有意差があるバージョン(d)と、手法によっては(a)と有意差があるバージョン(c)と、すべての手法で(a)と有意差がないバージョン(b)とを用意して、(d)(c)(b)のバージョンに人手評価を行って、結果を比較した。
有意差があるか検定するには、t-検定（t-分布を用いた検定）を行った。検定とは 2 集団間に差があるかないかを統計的に調べることである[文献 21]。

人手評価の方法： 人手評価にともなう評価のぶれを小さくするため、2 つのバージョンの翻訳結果を相対評価して、点数化した。どちらのバージョンの翻訳結果かわからないように順序を変えて提示し、2 人の評価者で同じ作業を行った。(a)と(d)、(a)と(c)、(a)と(b)の間で、(a)の方が良い場合は×を、(a)でない方が良い場合は○を、同じあるいはどちらともいえない場合は△を、付与した。×は 0 点、○は 2 点、△は 1 点にして合計して点数化した。

自動評価の方法： 自動評価は特許文書のタイトルに対して行った。アブストラクトの日本語訳は意訳された部分が多いが、タイトルは日本語訳を参照訳として利用できると判断したからである。
英日翻訳結果で自動評価を利用するために、実行時には以下の工夫をした。

- 自動評価の計算には n-gram や編集距離が必要なので、単語毎に区切られていなければならない。翻訳結果や日本語訳は語が区切れていないので、茶筌で形態素解析しておく。
- 通常の自動評価では複数の参照訳を用意する。参照訳は 1 つしかないので、評価用例文を多くすることで補う。
- BLEU, NIST では、タイトルのように 1 文書が 1 文で短いと n-gram がほとんど取れないので、

評価者	(d)	(c)	(b)	(a)
評価者Ａ	1483	1623	1682	1792
評価者Ｂ	1489	1608	1674	1792
合計	2972	3231	3356	3584

表 4-1-8：人手評価による結果

手法	(d)	(c)	(b)	(a)
BLEU	0.102	0.106	0.110	0.115
NIST	4.025	4.111	4.199	4.302
WER_1	0.778	0.772	0.762	0.749
PER_1	0.810	0.802	0.792	0.780
WER_2	0.895	0.889	0.876	0.866
PER_2	0.905	0.898	0.885	0.875
GTM_d	0.402	0.408	0.413	0.419
GTM_s	0.405	0.412	0.419	0.425

表 4-1-9：自動評価による結果

- 同じ分野の複数タイトルを１つの文書としておく。
- WER, PER では、語の区切り方や語数が影響するので、非自立語を入れない方法も試す。
- GTM では、文書単位でも文単位でも計算できるので、複数タイトルをまとめる方法も試す。

評価用例文は、特許のタイトルから分野毎に 60 文ずつ 69 分野から無作為に抽出した 4,140 文にした。BLEU, NIST は分野毎の文書単位で、WER, PER は、文単位で評価した。GTM は、分野毎の文書単位と文単位の 2 タイプを求めた(以下、表では各々GTM_d、GTM_s とした)。また WER,PER は、自立語相当のみを対象にした場合と、すべての単語を対象にした場合を求めた（以下、表では各々_1、_2 とした)。

実験結果: 表 4-1-8 に、人手評価で相対評価した結果の点数を、表 4-1-9 に、各評価指標による評価値を記す。表 4-1-8 と表 4-1-9 を見ると、人手評価でも自動評価でも(d)から(a)の順に良くなっている。このことから、自動評価の代表的な 5 つの手法はいずれも人手評価と相関関係があり、自動評価は人手評価を補うことができるといえる。

(2) 翻訳テンプレート利用による翻訳品質向上の検証

自動評価の有効性が(1)で確認できたので、次に、翻訳テンプレートの登録量によって自動評価値がどのように変化するか、翻訳テンプレートの登録が翻訳品質の向上に寄与しているかどうかを自動評価によって確認することにした。

以下のように、各コミュニティ辞書の登録する量を変えて、辞書の登録語数が異なる翻訳システムのバージョンを用意した。評価用例文と評価手法は、(1)と同じである。

1. コミュニティ辞書なし　　登録辞書総数　　0 語
2. 各コミュニティ辞書の 1/4　　308,555 語
3. 各コミュニティ辞書の 2/4　　617,152 語
4. 各コミュニティ辞書の 3/4　　925,730 語
5. コミュニティ辞書全部　　1,234,346 語

図 4-1-3 は、辞書登録総数が上記のように変化した場合の各自動評価指標における自動評価値の変化をグラフにしたものである。図 4-1-3 を見ると、どの評価指標においても、最初の値の立ち上がりがやや急であるが、その後は、ほぼ比例して評価値が上昇している。このことから、翻訳テンプレートの登録における翻訳品質の向上は、登録語数が少ない状態の方がその効果は大きいが、登録語数が多くなった場合でも、我々が作成した約 120 万語の登録の範囲では、緩やかに比例して品質が向上していくという

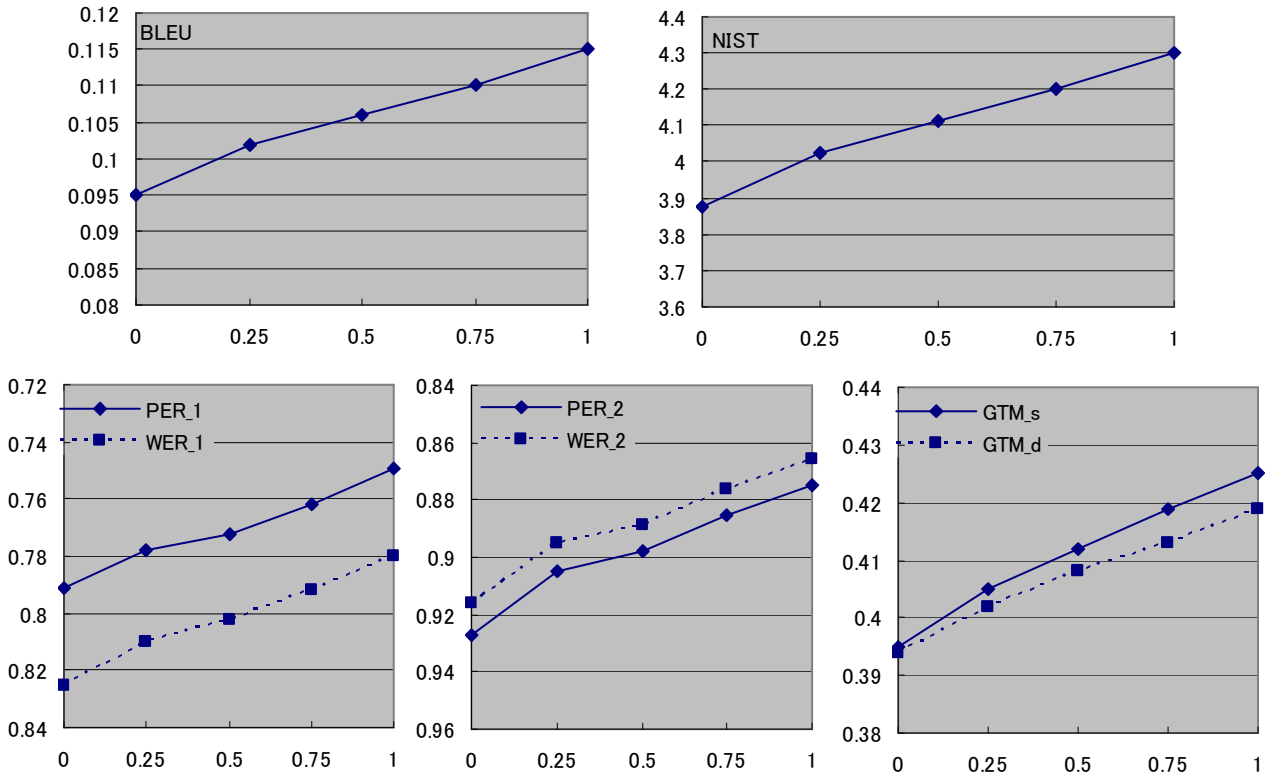


図 4-1-3：登録辞書数における各評価値の変化

日付	BLEU	NIST	WER_1	PER_1	WER_2	PER_2	GTM_d	GTM_s
2005/11 版	0.1115	4.2289	0.7574	0.7877	0.8713	0.8808	0.4146	0.4209
2006/01 版	0.1134	4.2635	0.7549	0.7857	0.8683	0.8778	0.4169	0.4228
2006/02 版	0.1146	4.3017	0.7493	0.7799	0.8660	0.8747	0.4191	0.4252
2006/03/09	0.1153	4.3169	**0.7519**	**0.7826**	0.8634	0.8731	0.4220	0.4275
2006/03(/13)版	0.1175	4.3426	0.7501	0.7809	0.8618	0.8717	0.4232	0.4285

表 4-1-10：特許翻訳システムの自動評価値の変遷

ことがわかった。

我々は、今後も引き続き翻訳テンプレートの獲得及び登録を行う予定であるが、この評価値の上昇がずっと続くのか、それとも限界があるのかは興味深い点であり、今後も継続して調査する予定である。

(3) 新バージョン開発時の翻訳品質の確認

我々のシステムは、1～2ヶ月に1度バージョンアップを行っている。その際に、翻訳テンプレートの新規の登録や文法の改良によって副作用が発生し、翻訳品質の低下が起こっていないかどうかを、上記の自動評価指標を用いて確認している。確認の際には、全ての評価指標において値が向上することを条件とし、値が下がった場合には、値が下がった原因をすぐさま突き止め、改良することでバージョンを確定している。

表 4-1-10 は、2005 年 11 月から 2006 年 3 月までに作成した特許翻訳システムの 3 つのバージョンの自動評価の結果である。2006 年 3 月版に関しては途中の版の評価値も示す。この途中の版では太字の WER_1 と PER_1 の値が低下している。我々はその原因を突き止め、改良をすることで、その 4 日後に 3 月版を完成させた。このように、自動評価を利用することによって、我々のバージョン作成時の作業効率は大きく向上した。また、辞書登録や文法改良の効果を、自動評価指標により客観的に知ることは、作業者にとって登録や改良の大きな励みとなった。さらに、他システムとの翻訳品質の比較も容易となり、今後、システムの能力を客観的に示すための有益な資料としての利用を考えている。

４－１－５　結論

最終目標の各項目を列挙し、それぞれにおける実施状況及び残された課題を最後にまとめる。

(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。

・英日対訳文書においては、考案した翻訳テンプレート獲得手法（下記(2)）を用いて、日本の公開特許とその PAJ(公開特許英文抄録)から､81 分野､約 120 万語の英日翻訳テンプレートを作成した。

・中日対訳文書においては、言語資源を活用した対訳表現抽出手法によって、フリーで公開されている対訳文書から、5,605語の翻訳テンプレートを作成した。

・専門用語の翻訳テンプレート作成するためには、第1言語の専門用語の選定、その用語に対応する第2言語の訳語選択という2つのステップがある。J社(翻訳用辞書作成会社)への聞き取り調査によると、言語のエキスパートであっても訳語選択だけで約10分/語かかるということであった。提案の翻訳テンプレート獲得技術によれば、人手を要する作業は獲得されたテンプレートのチェックのみとなり、翻訳テンプレート作成作業工数の大幅な削減が可能になる。チェックにかかる時間は約3分/語程度なので、試算の結果、獲得精度が80%と仮定しても工数を1/3に削減できることを確認した。

・自動評価の信頼性と獲得した翻訳テンプレートの有効性について検証し、次のことを確認した。
(1) 自動評価は人手評価と相関があり、人手評価を補うものとして利用可能である。
(2) 翻訳テンプレートを登録量にほぼ比例して自動評価値は上昇することを確認した。このことから、我々が獲得した約 120 万語の翻訳テンプレートは、翻訳品質の向上に寄与していることが確認できた。

さらに、翻訳テンプレートの登録や文法改良を伴う新バージョンの開発時に本自動評価指標を利用して訳質低下の有無を確認することにより、効率的なシステム開発ができるようになった。

(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。

・言語資源を効果的に利用した貪欲的統計手法による対訳表現の半自動抽出手法を考案し、英日対訳文書を対象にして実験・評価した結果、精度 89%, カバレッジ 85%で翻訳テンプレートが作成できることを確認した。また、中日対訳文書における実験においても、精度 75%、カバレッジ 55%で翻訳テンプレートが作成できることを確認した。

・隣接単語の分散に基づく用語抽出手法と、単語長と出現回数による手法を利用して、精度80%、カバレッジ76%で翻訳テンプレートが作成できることを確認した。

・上記の統計的手法によって抽出した単語列ペア及びその単語列対応度を、構造照合時に利用することによって、動画関連の国際標準文書において 1,000 文中 992 文の構造照合結果を抽出することができ、83%の精度で句単位の対応付けができることを確認した。（なお、構造照合利用型の手法については､統計的手法利用型において十分な翻訳テンプレートを作成することができたので、実際の翻訳テンプレート作成には用いていない。）

(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。

・特許文書を対象に、文対応がついていない対訳文書からの専門用語の翻訳テンプレート抽出実験を行った。PAJ(日本語特許抄録)の対訳を用いた実験と PAJ の日本語と USP(米国特許抄録)を用いた実験を行ったところ、PAJ の対訳では、20 位以内の抽出精度が 83.0%と、従来手法と比べて高い精度

で抽出できることが確認できた。一方で、PAJ の日本語と USP を用いた実験では 20 位以内の抽出精度が 55.3%と、文書によって対応付けの精度に差が出ることが明らかになった。今後、文書ペアの類似性に基づいて対応付けの確信度を提示するなど、精度の差を吸収し、実用に耐えうる出力を提供するための技術の開発が必要となる。

４－２　分野辞書の自己組織化に関する研究開発

４－２－１　序論

4-1 節で説明した技術を利用して作成された大量の翻訳テンプレートは、多種多様な分野の文書から抽出されており、様々な分野の専門用語や分野固有の表現を含んでいる。これらの翻訳テンプレートは未分類の状態で管理するのではなく、分野毎に分類して管理することにより、その後の修正や追加が簡単になる。さらに、翻訳する際、入力文に応じて、分野毎に分類された翻訳テンプレートを使い分けることにより、より正確な訳語推定が可能になる。例えば、 “……. It’s Java. ……”という文が入力された場合、それがプログラミング言語に関する文書の一部であれば、プログラミング言語の分野の翻訳テンプレートを利用して“Java”は「Java」と訳し、一方、アジアの地理に関する文書であった場合、地理分野の翻訳テンプレートを利用して「ジャワ」と訳すことによって、分野毎に訳し分けることができる。

このように、辞書の分野単位の分類、及び、翻訳時に使用する分野辞書の選択は、機械翻訳の翻訳品質を高めるためには必須の作業である。しかし、辞書の管理者やユーザが、手動でそれらを行うには作業負荷が大き過ぎる。

そこで、我々は、それらを手動で行うのではなく、自動化するための研究及び「訳してねっと」における実証実験を行った。

- 入力文書情報を利用して翻訳時に使用する分野辞書を自動的に選択するための研究(4-2-2)
  及び、「訳してねっと」における実証実験(4-2-3)
- 既存の分野体系に拠らない分野辞書の自動構築の研究(4-2-4)

以下に、各項目の研究の概要を説明する。

４－２－２　コアワードを利用した分野の自動判定の研究

(a) はじめに

我々が現在開発中の「訳してねっと」は、階層化された数多くのコミュニティ（分野）を持っている(PX を参照)。しかし、初心者のユーザにとっては、ある文書を翻訳する際どのコミュニティの辞書を用いて翻訳すべきか、また、自分が登録したい用語はどのコミュニティに登録すべきか等、その判断は難しい。そこで、本研究では、ある語又はある文書が「訳してねっと」上のどのコミュニティ（分野）にもっとも関連が深いかを自動的に判定する技術を開発する。

分野を判定する際重要なことは、その判定速度である。なぜなら翻訳要求があった時にリアルタイムに分野を判定して翻訳する必要があるからである。そこで、我々は分野に特徴的でかつ代表的な単語群（「コアワード」と呼ぶ）を予め用意しておき、入力語（又は入力文）とその単語群との類似性を測ることで、階層化された数多くの分野から最適な分野を瞬時に判定する手法を開発した。

以下、手法の概要を説明し、語及び文書の分野判定実験とその結果について述べる。

(b) コアワードと分野関連度

ある分野において、特徴的でかつ代表的な単語を「コアワード」と定義する。各分野のコアワードには、その分野にどれだけ関連が強いかを示す値が付与されている。それを分野関連度と定義する。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。

各コアワードの分野関連度は次の方法で計算する。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方法として一般的に用いられる tf*idf で求める。

tf(d,t)：分野 d に出現するコアワード t の生起頻度。一般に TF(Term Frequency)と呼ぶ。
df(t)：　コアワード t が一回以上生起する分野の数。DF(Document frequency)と呼ぶ。
idf(t)＝log(N/df(t))：　分野の数 N と、DF の逆数をかけて、対数をとる。IDF(Inverse Document frequency)と呼ぶ。

分野関連度では、語がどのくらいその分野を特定するかを idf によって反映させる。多くの分野に現れる一般的な語の場合には idf は小さくなり、逆に、特定の分野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、特定の分野で繰り返し生起する語はその分野において重要な概念であると考えるためである。しかし、ある分野に多数出現するほど大きくなる値 tf と特定の分野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、その分野が有するコアワードののべ総数が多いほど大きい値を取るという問題が発生する。したがって、分野毎に、tf*idf をコアワードのべ総数で割った以下の値を分野関連度 ref(t,d)とする。

ref(t,d)= tf(d,t)*idf(t)／all(d)
ref(t,d)：分野 d におけるコアワード d の分野関連度
all(d)：　分野 d におけるコアワードのべ総数

(c) 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。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。例えば、スポーツ分野の子分野には、最下層の分野の野球分野とサッカー分野があり、スポーツ分野は野球分野とサッカー分野の親分野で中間層の分野でもある。

分野が単層の場合、各分野におけるコアワードを作成するだけで容易に分野判定が可能であるが、上記のように階層化された分野の場合には、中間層に位置する分野のコアワード及び分野関連度はどのように求めるかという課題が残る。我々は、まず、各分野のコアワードを予め作成し、その分野が階層のどこに位置するかで分野関連度に対して特別な重みを与えることで、各分野におけるコアワードの分野関連度を求めた。以下に、コアワードを作成する方法と、作成されたコアワードに対して分野関連度を再計算する方法に分けて説明する。

コアワードの作成：階層化された全分野にコアワードを付与する必要があるが、作成するコアワードは、最下層の分野のみとし、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。親分野のコアワードをその分野からではなく、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ためである。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。

分野関連度の再計算：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方法の２通りが存在する。

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。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。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。

(d) 語の分野判定

手法の概要：上記に述べた作成方法に従って作成した各分野のコアワード群（以降、これを分野別コアワード辞書と呼ぶ）を用いて、語の分野判定を行った。手順を以下に示す（本実験及び結果の詳細は[文献 25]を参照のこと）。

まず判定の対象となる語を、コアワード検出用文書から検索し、一文内に同時に出現した自立語を全て抽出する。ここでいう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自立語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。

次に、抽出された全ての自立語に対して、上記に述べた分野別コアワード辞書を検索し、各コアワードにおける分野判定度を計算する。なお、分野判定度は、分野関連度とコアワード検索文書における出現回数をかけた値とする。

最期に、分野判定度の高い順にコアワードを順位付けし、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。

単層分野における実験及び結果：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。

階層化された分野における実験及び結果：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。

(e) 文書の分野判定

手法の概要：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○○は 10 月 22 日、新チーム名を「××」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。

上記の考え方により、対象文書における、ある分野の分野判定度は、その文書に出現する全てのコアワードの分野判定度を分野毎に合計することによって求める。上記の例で言えば、野球分野の分野判定度の計算には、「広島」「優勝」「ヤクルト」の分野判定度が関与することになる。なお、ここでの分野判定度は、語の分野判定と同様、各コアワードの分野関連度に対象文書での出現回数をかけた値とする。

実験及び結果：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。これらの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。一方、比較のために、子のコアワードは用いないでその分野の文書のみを利用して分野関連度を付与する場合も試した。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 4,239 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、それぞれ、55%、70%、85%となった。一方、その分野の文書のみを利用した場合には、それぞれ、56%、72%、87%となった。全体の精度では、文書のみを利用する方法より少し劣っていたが、中間層の分野の精度は逆に上がっていた。また、全体の精度の差もそれほど大きくなく、実際には文書が存在しない中間層の分野も多い。したがって、本手法は、階層化された分野における文書の分野自動判定に有効な手法だと言える。

４－２－３　「訳してねっと」を利用した実証実験

(a)実験の目的

自己組織化として行ってきた分野の自動判定の目的は、実際に「訳してねっと」（4-3-6 の項を参照）に搭載して翻訳品質を向上させることである。そこで、4-2-2 で述べたコアワードを利用した分野の自動判定が「訳してねっと」に搭載できることと、4-2-2 で述べたコアワードを利用して翻訳対象文書の分

野を自動判定することによって翻訳対象文書の翻訳結果の精度が向上することを検証する。

(b)「訳してねっと」への搭載

「訳してねっと」への搭載には「おすすめコミュニティ」を利用することにした。おすすめコミュニティとは、翻訳対象文書からおすすめのコミュニティを判定し、そのコミュニティおよびその上位のコミュニティの辞書を利用して翻訳する機能である。おすすめのコミュニティの判定には、おすすめコミュニティ設定ファイルを利用する。おすすめコミュニティ設定ファイルには、各コミュニティにそのコミュニティを代表する特徴的な語が複数登録されている。さらに、各語にはその特徴の度合い（以下、特徴度と呼ぶ）を示す点数が付与されている。おすすめコミュニティの判定方法は、翻訳対象となる文書において、おすすめコミュニティ設定ファイルの登録語を検索し、コミュニティ毎に特徴度の合計を計算する、というものである。合計した点数が高いコミュニティを、おすすめコミュニティとして選択する。

従来、人手により作成していたこの「訳してねっと」のおすすめコミュニティ設定ファイルを、4-2-2 で作成するコアワードと分野関連度を利用することによって自動作成する。おすすめコミュニティ設定ファイルを自動作成するにあたって、コアワードを絞り込み、分野関連度を正規化した。コアワードは、1 文字の語を除いて、分野関連度の高い方から 30 個抽出した。1 文字の語を除いたのは、現行の登録語の検索は文字列マッチで行っているため間違って検索される場合が多いからである。分野関連度は、特徴度の範囲に合わせるため、1～1000 に正規化した。

(c) 分野の自動判定による翻訳品質の向上の検証実験及び結果

4-2-2 で述べたコアワードを利用して翻訳対象文書の分野を自動判定することによって翻訳対象文書の翻訳結果の精度が向上するかどうかを、4-1-4 の自動評価手法を用いて実験した。

日英翻訳による翻訳品質向上の検証：コアワードが日本語であるので、日本語で分野の自動判定をして日英翻訳することによって、翻訳品質の向上を検証する。大まかな実験手順は以下の通りである。

１． 4-2-2 の手法を用いて特許文書から抽出したコアワードと分野関連度を利用して、おすすめコミュニティ設定ファイルを自動作成する。
２． 次の２つの方法で以下に述べる日本語の翻訳対象文書を英語に翻訳する。
　(a) １．で作成したおすすめコミュニティ設定ファイルを用いて翻訳する。
　(b) おすすめコミュニティの機能を用いないで翻訳する。
３． 自動評価手法を用いて、２．の(a)で翻訳した場合と２．の(b)で翻訳した場合の翻訳品質を比較する。

翻訳対象文書は、4-1-4 の実験で使用した評価用例文を利用した。特許のタイトルから分野毎に 60 文ずつ 69 分野から無作為に抽出した計 4,140 文である。コアワードを日本語で作成したので、評価用例文の参照訳の方を翻訳対象文書にして翻訳し、例文の英文の方を参照訳にした。4-1-4 では WER と PER を単語の区切り方の影響を考慮した 2 通りの方法で計算したが、ここでは翻訳結果と参照訳が英語で単語の区切り方を考える必要はないので、すべての単語を対象とした方のみを計算した。

結果：表 4-2-2 は２．(a)の自動判定ありで翻訳した場合と２．(b)の自動判定なしで翻訳した場合の自動評価値である。

自動判定	BLEU	NIST	WER	PER	GTM_d	GTM_s
(a)判定あり	0.0438	2.9870	1.1870	1.1439	0.2923	0.3009
(b)判定なし	0.0467	2.9029	1.1767	1.1356	0.2833	0.2892

表 4-2-2:日英翻訳における自動判定の有無による自動評価結果

自動評価の結果は、手法によるばらつきがみられた。判定なしより判定ありの方が良くなったのはNIST, GTM_d, GTM_s、悪くなったのはBLEU, WER, PERであった。

この理由を分析してみると、以下のことが言える。4-1-4によれば、NISTは単語訳の正しさを重視した自動評価スコアである。また、GTMは一致した部分の長さの和を取るので、部分訳が正しくなることで評価値が上がる。このような性質をもつNIST, GTM_d, GTM_sでは、単語訳の正しさが評価値に影響する。(a)では分野が正しく判定されたことで、翻訳結果中の未知語（辞書になく翻訳できずに原語のまま残った語）が減った。このため、評価値が良くなったと考えられる。

一方、BLEUは語順の正しさや翻訳の流暢さを重視した自動評価スコアである。WERとPERは翻訳結果を参照訳に一致するまで変換するので、翻訳結果全体が正しくなることで評価値が上がる。(a)では正しく判定されたことにより単語訳が正しくなった一方で、その副作用により文全体の正しさや流暢さは逆に減少したためだと考えられる。

英日翻訳による翻訳品質向上の検証：日英翻訳では文法的な問題があり翻訳品質の向上が語の訳レベルにとどまった。そこで、日本語であるコアワードを自動的に英語に変換して、英語で分野の自動判定をして英日翻訳することによって、翻訳品質の向上を検証する。

コアワードを自動的に英語に変換するには、コミュニティ辞書に登録済の英日方向と日英方向の翻訳テンプレートを利用する。まず、コミュニティ辞書毎にコアワードが見出しになっている翻訳テンプレートを検索して、コアワードに対応する英語の見出しを抽出する。次に、その英語の見出しをそのコミュニティのコアワードにして、元の日本語のコアワードの分野関連度と合わせて、おすすめコミュニティ設定ファイルを自動作成する。大まかな実験手順は以下の通りである。

１．4-2-2の手法を用いて特許文書から抽出したコアワードと分野関連度を利用して、上記の方法でコアワードを英語に変換し、おすすめコミュニティ設定ファイルを自動作成する。
２．次の4つの方法で以下に述べる翻訳対象文書を翻訳する。
　(a) 正解のコミュニティを示す特許文書中のIPCコードを用いて翻訳する。
　(b) １．で作成したおすすめコミュニティ設定ファイルを用いて翻訳する。
　(c) おすすめコミュニティの機能を用いないで翻訳する。特許文書に対応した文法は使用する。
　(d) おすすめコミュニティの機能を用いないで翻訳する。特許文書に対応した文法も使用しない。
３．自動評価手法を用いて、２．の(a)から(d)の方法で翻訳した場合の翻訳品質を比較する。

日本語のコアワードは特許の108分野に30個ずつの計3240個で、日本語のコアワードから変換した英語のコアワードは計1024個である。

翻訳対象文書は、4-1-4の実験で使用した評価用例文を利用した。特許のタイトルから分野毎に60文ずつ69分野から無作為に抽出した計4,140文である。

結果：表4-2-3は２．の(a)から(d)の方法で翻訳した場合の自動評価値である。

自動判定	BLEU	NIST	WER_1	PER_1	WER_2	PER_2	GTM_d	GTM_s
(a)正解	0.1175	4.3426	0.7501	0.7809	0.8618	0.8717	0.4232	0.4285
(b)判定あり	0.1048	4.1206	0.7740	0.8037	0.8936	0.9026	0.4127	0.4151
(c)文法あり	0.1038	4.0947	0.7794	0.8084	0.8985	0.9067	0.4117	0.4139
(d)文法なし	0.0804	3.3655	0.8629	0.8732	1.0432	1.0341	0.3716	0.3717

表4-2-3：英日翻訳における自動判定の有無による自動評価結果

自動評価の結果では、翻訳品質はいずれの手法でも(a)(b)(c)(d)の順に良くなっている。このことから、日本語で作成したコアワードを自動的に英語に変換して自動判定を利用することによって、英日翻訳の翻訳結果が向上することが確認できた。

４－２－４　分野辞書の自動構築の研究

(a) 研究の背景

4-2-2 で述べた手法は、翻訳にとって理想的な体系化・階層化された分野辞書が存在することを前提とした手法である。しかし、そのような分野辞書が存在しない場合、または、その分野辞書の体系が翻訳にとって有効でない場合には、既存の辞書をなんらかの規則に従って体系化また分類しないと、訳語間のコンフリクトが発生し、翻訳品質の向上は見込めない。

そこで、本研究では、4-2-2 の前提を変え、体系化・階層化された分野辞書が存在しない場合にも有効な辞書の自動分類方法を提案する。具体的には、既存の対訳文書を利用することで、大量の語から意味的・概念的に近い語の集まりを作り、この語の集まりから分野辞書を作る。さらに、分野辞書構築の際に得られた情報によって、翻訳時の辞書の適用順序の推定をすることができる。これについても検証する。

(b) 手法の説明

国際特許分類(IPC)によって分類された文書を用いて、大量の語から辞書を構成し、構成された辞書を用いて翻訳を行う際の辞書の利用順序を PLSI(Probabilistic Latent Semantic Indexing)[文献 22]を用いて決定する手法を説明する。

PLSI は情報検索分野における文書-単語行列の次元削減の手法で、潜在的な意味クラスを仮定し、仮定したクラスからの文書・単語の生成確率を推定することにより、意味的・概念的に近い語が集まるような次元削減を行う。PLSI の次元削減は文書、単語を意味クラスへまとめあげる確率的ソフトクラスタリングとなる。本提案では、仮定する意味クラス＝分野と考え、まとめあげられた語をもとに辞書を構成し、翻訳時の辞書利用順序を決定する。このとき、同一の IPC が割り当てられた文書は単語の出現傾向が同じと仮定し、それぞれの IPC に含まれる文書をひとつの文書として考える。

まず、語と IPC との共起 $P(w,ipc)$ を意味クラスからの同時発生と仮定すると、

$$P(w,ipc)=\sum_{z\in Z}P(w\mid z)P(ipc\mid z)P(z)$$

のように表される。ここで、Z は意味クラスを表す確率変数である。観測可能な語と IPC の共起から、内部状態である $P(w\mid z)$, $P(ipc\mid z)$, $P(z)$ を EM アルゴリズムによって推定する。

推定された $P(w\mid z)$ にベイズの定理を適用することで語の意味クラスに対する帰属確率 $P(z\mid w)$ を得、この帰属確率を用いて辞書の構成を行う。得られた帰属確率が一定の閾値以上ならば、分野辞書へ登録する語とし、どのクラスに帰属確率が閾値以上にならない語は一般語辞書へ登録する。

次に，IPC のクラスへの帰属確率 $P(z\mid ipc)$ から、翻訳に利用する辞書の優先順位を決定する。IPC のクラスへの帰属確率は、語のクラスへの帰属確率と同様にして求めることができる。帰属確率の高いクラスで作成された分野辞書が、翻訳を行う上で最も適していると考えられるため、翻訳時の辞書の優先順位はクラスへの帰属確率の大きさによって決定できる。これにより、IPC に対するすべての分野辞書の利用順位が決定できる。上で作成した一般語辞書の優先度は分野辞書よりも低くしておくことで、分野に特徴的な語が優先的に翻訳結果に反映されると考えられる。

(c) 実験及び結果

現在の「訳してねっと」のコミュニティ辞書へ登録されている語から、特許文献に合わせた分野辞書と一般辞書を作成し、作成した辞書を用いて米国特許文献の翻訳を行った。

まず、辞書構成と辞書の優先順位のための学習データとして、日本語特許とその英文抄録(PAJ)から 100 万件のタイトル、アブストラクトの対訳を取り出した。このデータに対し、英語、日本語の語のペアの出現頻度を得、登録語 787,223 対中、312,637 対の出現を得た。その後、語のペアと、各文書に付与されている IPC の共起から、PLSI のパラメタ推定を行い、語のクラスへの帰属確率との IPC のクラスへの帰属確率を得た。このとき、分野辞書の数による影響を調べるため、クラスの数 $|Z|$ を 100, 200 と変化させた。分野辞書への登録する際の語のクラスへの帰属確率の閾値を 0.2 とし、分野辞書を作成した。翻訳時の辞書の利用順序に関しては、各 IPC の帰属確率の高いクラスの辞書を帰属確率の高い順に従って 3 個利用し、最も低い優先度(優先順位 4 位)に一般語辞書を利用するように指定するようにした。この設定で、件の米国特許のタイトルの翻訳を行った。参考のため、同じデータに対して現在の「訳

してねっと」のコミュニティ辞書の構成を用いて翻訳を行った。このとき、辞書中の語は PLSI の推定に用いた語と同じ語(312,637 語)を用い、IPC とコミュニティ辞書の対応は人手で決定したものを用いた。それぞれの翻訳結果から無作為に 3,000 文を取り出し、自動評価手法(4-1-4 参照)によって評価したものを表 4-2-4 に示す。

	BLEU	NIST	WER_1	PER_1	WER_2	PER_2	GTM_d	GTM_s
従来手法(訳してねっと)	0.0849	3.5574	0.7875	0.8295	0.9166	0.9388	0.3772	0.3858
提案手法 (辞書数 100)	0.0880	3.6560	0.7767	0.8171	0.9032	0.9241	0.3820	0.3903
提案手法 (辞書数 200)	0.0875	3.6224	0.7821	0.8223	0.9089	0.9297	0.3805	0.3880

表 4-2-4：特許文献における自動評価結果

自動評価の結果は、いずれの評価指標においても提案手法での翻訳結果が良い値を示した。また、分野辞書の数は 100 個の場合のほうがよりよい結果であった。実験により、同じ語彙数の条件のもとで、提案手法により自動的に作成した分野辞書と、文書に対する辞書の利用順序を用いた翻訳結果は、人手によって体系づけられた分野辞書での翻訳結果と同程度かそれ以上の翻訳結果を得ることを確認した。

(d) 考察

本実験において、PAJ コーパス 100 万件からはコミュニティ辞書の 4 割程度の語を発見し、特許文献のための辞書を構成することができたが、より多くの語で構成するためには対象となる文書を増やす必要がある。しかし、文書が大きくなるに従い、PLSI の計算負荷が高くなるという問題点があり、[文献 23]のようなコーパス分割の手法を用いる必要がある。実験における、作成する辞書の数や閾値といったパラメタの最適化も検討する必要がある。

本節で提案した手法では、特許文書のようにあらかじめ分類された文書に対し、同じ分類コードに属する文書の語の分布は同じと仮定することで、翻訳時の分野辞書の優先順位を決定した。しかし、分類がなされていない文書に対しては、対応する辞書や、その優先順位を決定することができない。この問題に対しては、未知文書に対する予測分布を推定する LDA[文献 24]といった手法を用いることが考えられる。

４－２－５　結論と今後の課題

最終目標の各項目とそれぞれにおける実施状況を以下にまとめる。

(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。

・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、ユーザが階層化された数多くの分野に分類された辞書に対して、語を登録する際や文書を翻訳する際に、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。

・対訳文書を利用した辞書の自動分類及び翻訳時の利用順序の推定手法を提案した。機械翻訳の自動評価手法を利用して、本手法の導入によって翻訳品質が向上することを確認した。本手法を用いることにより、未分類の辞書の自動的に分類することができ、かつ、その辞書の利用順序をも自動的に決定することができる。

(2) 自己組織化後は、翻訳結果の精度が向上すること。

・上記(1)のコアワードを利用した分野の自動判定手法を利用して、「訳してねっと」上で翻訳品質が向上するかどうかを、4-1-4 で述べた自動判定手法によって検証した。この結果、日英翻訳における実験では、語レベルの翻訳品質が向上することを確認した。一方、英日翻訳による実験では、語レベル、文法レベル共に、翻訳品質が向上することを確認した。

４－３　言語非依存の翻訳エンジンの研究開発

４－３－１　序論

本サブテーマでは、前節 4-1 及び 4-2 で得られた成果を用いて、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。

- 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多言語化。(4-3-2)
- 多言語翻訳データベース：英日・中日・韓日翻訳システムの開発及び翻訳テンプレートの構築。(4-3-3, 4-3-4, 4-3-5)
- 協調的翻訳支援環境：支援環境の開発及び実証実験。(4-3-6)

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、中国語や韓国語への取り組みを行ない、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、英日、中日、韓日翻訳に必要なデータベースを構築し、言語非依存翻訳エンジンに実装する。

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。

以下では、それぞれのテーマにおける取り組みについて、具体的に説明する。

４－３－２　多言語に対応した形態素解析システムの研究

(a) はじめに

形態素解析は入力された文を単語に分割して品詞を付与する処理であり，自然言語を処理する様々なアプリケーションで必要とされる基本技術である。形態素解析を行う上で、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一般に多数存在するため、その曖昧性を適切に解消する必要がある。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関する情報は形態素解析器が保持していないため、正確に単語を分割して品詞タグを付与するのは非常に難しい。3 つ目の課題は、多言語化の問題である。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。

ここではこのような形態素解析に関連した課題に対して、2 つの研究を行った。一つは、未知語の問題に対処しつつ高い精度で曖昧性を解消することができ、複数の言語へ適用可能な、単語レベルと文字レベルの情報を用いた中国語・日本語・韓国語形態素解析の研究である。もう一つは、様々な言語に対して高精度で未知語の品詞を推定するための、大域的な情報を用いた未知語の品詞推定の研究である。

(b) 単語レベルと文字レベルの情報を用いた中国語・日本語・韓国語形態素解析

背景：中国語や日本語や韓国語は、英語のように明示的な単語境界を持っていないため、単語分割を正確に行うことが必要となる。これらの言語の単語分割・品詞付与を行う方法は、これまでに様々なものが研究されている。特に近年、人手で作成した規則を用いて解析を行うルールベースの手法に代わり、形態素解析を行うのに必要なパラメータを学習用のデータから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が、これらの手法にはいくつかの問題点が存在する。

手法：コスト最小法[文献 26]は、単語単位で形態素解析を行う手法であり、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うことができるが、そのままでは未知語を扱うことが困難である。また、文字タグ付け法[文献 27]は、文字単位で形態素解析を行う手法であ

り、未知語に対する解析精度は高いが既知語に対する解析精度は比較的低いことが観測されている[文献 28, 29]。そこで、既知語に対しても未知語に対しても高い精度で解析を行うために、これらの二つの手法を組み合わせた形態素解析手法を提案する。

提案手法では、入力された文に対して、単語単位の候補と文字単位の候補の両方を作成し、単語単位の候補により既知語を、文字単位の候補により未知語を処理する。図 4-3-1 に例を示す。

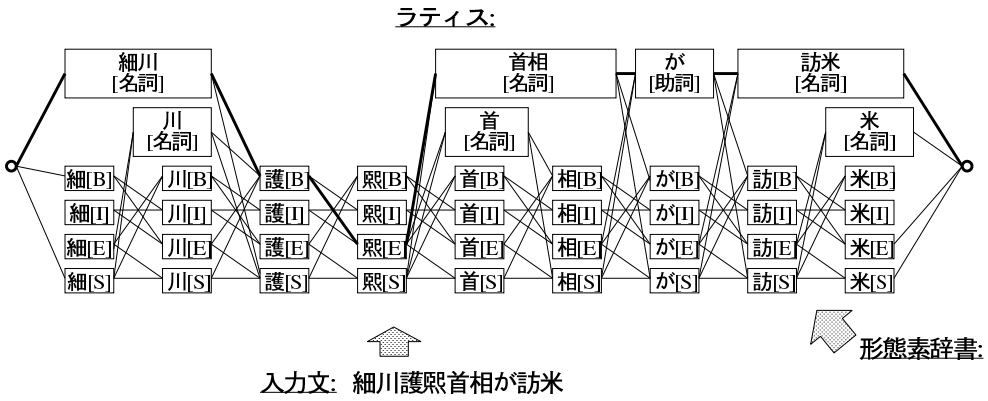


図 4-3-1 : 提案手法における解析結果の候補の生成

文字単位のノードは(1)B、(2)I、(3)E、(4)S のいずれかの文字位置タグを持っているが、これらのタグはその文字が、(1)単語の先頭に位置すること、(2)単語の中間に位置すること、(3)単語の末尾に位置すること、(4)単語を一文字で構成すること、をそれぞれ表している。提案手法では、これらの文字位置タグを、単語単位のノードが持っている品詞タグと同等に扱う。そのため、従来の単語単位で形態素解析を行う方法と同じようにして、図 4-3-1 のグラフ構造で表現された解候補の中から最適と思われる解を選択することができる。ただし、従来の形態素解析でしばしば用いられている単純な品詞 bigram モデルは、モデルの表現力が十分ではない。そのため、英語の品詞タグ付けや日本語の形態素解析では、品詞 trigram モデルや語彙情報を利用する方法が提案されている。提案手法では、以下のような品詞 unigram、品詞 bigram、品詞 trigram、単語 bigram 確率を混合したモデルを利用して、単語列と品詞列の同時出現確率を計算することにより、任意の解候補に対する尤もらしさを計算することにする。文中の単語の数を n、i 番目の単語を w_i、i 番目の単語の品詞を t_i、w_i の集合を W、t_i の集合を T とした場合に、次のような混合モデルを使用する：

$$
\begin{aligned}
P(W,T) &= \prod_{i=1}^{n} P(w_i t_i \mid w_0 t_0 \cdots w_{i-1} t_{i-1}) \\
&\equiv \prod_{i=1}^{n} P(w_i t_i \mid h) \\
&\approx \prod_{i=1}^{n} \left\{ \lambda_1 P^{\text{POS unigram}}(w_i t_i \mid h) + \lambda_2 P^{\text{POS bigram}}(w_i t_i \mid h) + \lambda_3 P^{\text{POS trigram}}(w_i t_i \mid h) + \lambda_4 P^{\text{word bigram}}(w_i t_i \mid h) \right\} \\
&= \prod_{i=1}^{n} \left\{ \lambda_1 P(w_i \mid t_i) P(t_i) + \lambda_2 P(w_i \mid t_i) P(t_i \mid t_{i-1}) + \lambda_3 P(w_i \mid t_i) P(t_i \mid t_{i-2} t_{i-1}) + \lambda_4 P(w_i t_i \mid w_{i-1} t_{i-1}) \right\}
\end{aligned}
$$

ただし、λ_i は補間係数であり、leave-one-out 法により求めることができる[文献 30]。またその他の確率については、最尤推定により求めることができる。また、h はマルコフ過程の履歴である。

提案手法では、中国語と日本語を処理することができる。しかし、韓国語では縮約等の複雑な形態素の変化が頻繁に起きるため、提案手法を直接適用することはできない。そこで韓国語に対しては、入力文に対してまず綴り復元処理を行い形態素の原形を復元した後、中国語や日本語と同様に単語分割と品詞タグ付けを行うことにより、形態素解析を行う。

実験：中国語のコーパス(AS、CTB、HK、PK)、日本語のコーパス(EDR、KUC、RWC)、および韓国語のコーパス(MKT、KTB)を使用して、提案手法の有効性を調べた。使用したコーパスの統計情報を表1に示す。評価指標としては以下のように計算されるF値を使用し、既存の形態素解析システムとの比較

を行った(韓国語については比較可能なシステムが存在しなかったため比較を行っていない)。

$$\text{F値}=2RP / (R+P)$$
$$R = \text{(解析結果中の正解単語数)} / \text{(テストデータ中の単語数)}$$
$$P = \text{(解析結果中の正解単語数)} / \text{(解析結果中の単語数)}$$

実験結果を表4-3-1に示す。提案手法は、複数のコーパスにおいて既存手法よりも高いF値を得た。

コーパス	訓練データの単語数	テストデータの単語数(既知語/未知語)	辞書中の単語数	品詞の数
AS	5,806,611	11,985(11,727/258)	146,212	(64)
CTB	250,841	39,922 (32,706/7,216)	19,730	(64)
HK	239,852	34,955 (32,463/2,492)	23,747	(64)
PK	1,121,017	17,194 (16,005/1,189)	55,226	(64)
EDR	2,452,891	2,652,156 (2,600,051/52,105)	82,410	15
KUC	198,514	31,302 (29,926/1,376)	1,870,461	42
RWC	840,879	93,155 (93,085/70)	315,602	69
MKT	69,323	7,177 (6,623/554)	6,989	32
KTB	86,034	7,118 (6,733/385)	3,397	33

表4-3-1：実験に使用したコーパスの統計情報

コーパス(言語)	F 値	
	既存手法	提案手法
AS （中国語）	0.961	**0.972**
CTB （中国語）	**0.881**	0.874
HK （中国語）	0.940	**0.950**
PK （中国語）	0.951	**0.954**
EDR （日本語）	0.946	**0.950**
KUC （日本語）	**0.987**	0.985
RWC （日本語）	0.991	**0.993**
MKT （韓国語）	N/A	0.950
KTB （韓国語）	N/A	0.967

表 4-3-2：形態素解析の実験結果

(c) 大域的な情報を用いた未知語の品詞推定

背景: 未知語の品詞をできるだけ正確に推定することは、高精度な品詞タグ付けを行う上で必要であり、また単語辞書を自動的に作成するような場合にも重要である。未知語の品詞推定に関して、これまでに様々な研究が行われている[文献31,32]。これらの既存手法の多くでは、未知語の品詞は局所的な情報、つまり未知語の前後の単語や未知語自身の情報(語尾や文字種等)のみを用いて推定されている。しかしながら、局所的な情報のみでは品詞推定が困難な場合が存在する。例えば、名詞のように使われている未知語があった場合、それが普通名詞であるか固有名詞であるかを、局所的な情報のみを用いて判断するのは困難な場合がある。しかし、もしそのような曖昧な語と同じ語形を持つ未知語が、文書中の別の箇所で、大きな手がかりとなる局所的な情報(人名に付く敬称など) と共に出現していれば、そのような情報は曖昧な語の品詞を推定する上で役に立つ。

別の例として、サ変名詞に関する問題が挙げられる。サ変名詞は普通名詞のように使うことができるが、単語の末尾に「する」を付けることにより動詞として使うこともできる。名詞のように使われている未知語が、サ変名詞か普通名詞かのどちらであるかを判定することはしばしば困難である。この問題は、「可能性に基づく品詞の問題」として指摘されている[文献33]。可能性に基づく品詞とは、その品詞が単語の個々の事例の性質をあらわすのではなく、その単語が持つことが可能な全ての性質を表すような品詞である。例えばサ変名詞は、名詞として使われることも可能であり、「する」が末尾に付いて動詞として使われることも可能であるが、個々の事例はこれらの全ての性質を一度に有しているわけではない。このように、可能性に基づく品詞を持つ単語においては、個々の事例は全ての可能な性質の中

の一部の性質を持っているだけなので、一つの事例の局所的な情報のみからその品詞を推定するのは難しい場合がある。しかしながら、サ変名詞なのか普通名詞なのか曖昧である未知語が存在した場合、もし同じ語形を持つ未知語が文書中の別の箇所で「～する」という形で出現していれば、その曖昧な未知語の品詞はサ変名詞である可能性が高いと判断することができる。

以上のような課題に対処するために、局所的な情報だけではなく大域的な情報も利用した未知語の品詞推定手法を提案する。

手法: 提案手法では、同じ語形を持つ文書中の全ての未知語を同時に考慮してモデル化を行う。そして、そのような未知語の品詞は相互に影響しあい、なおかつ各未知語の品詞は局所的な文脈の影響も受けると考える。このような未知語の品詞の性質をモデル化するために、ボルツマン分布を使用する。

以下の説明では、文書中に同一の語形を持つ未知語がK回出現するとする。未知語がとりうる品詞はN種類あるとし、各品詞は1からNの整数で表現されることとする。k番目に出現した未知語の品詞をt_kで表し、k番目に出現した未知語の局所的文脈(未知語の前後の単語や品詞等)をw_kで表すことにする。また$\mathbf{w}$と$\mathbf{t}$を、それぞれw_kとt_kの集合とする。$\lambda_{i,j}$を品詞iとjの間における相互作用の強さを表す重みとする。そして、$\mathbf{w}$が与えられた場合に未知語の品詞が$\mathbf{t}$であるエネルギーを次のように定義する:

$$E(\mathbf{t} \mid \mathbf{w}) = -\left\{ \frac{1}{2} \sum_{k=1}^{K} \sum_{k'=1;k' \neq k}^{K} \lambda_{t_k, t_{k'}} + \sum_{k=1}^{K} \log p_0(t_k \mid w_k) \right\},$$

ここで、$p_0(t|w)$は局所的文脈wのみを用いて計算される品詞tの初期分布(局所的モデル) であり、最大エントロピーモデル等の任意の統計的モデルを用いて計算されるものとする。上記の式の右辺は、2つの要素から構成されている。一つは大域的な品詞間の相互作用を表す項であり、もう一つは局所的な文脈による影響を表す項である。このようなエネルギーを持つボルツマン分布を考えることにより、次のような$\mathbf{t}$の確率分布が得られる。

$$P(\mathbf{t} \mid \mathbf{w}) = \frac{1}{Z(\mathbf{w})} p_0(\mathbf{t} \mid \mathbf{w}) \exp\left\{ \frac{1}{2} \sum_{k=1}^{K} \sum_{k'=1;k' \neq k}^{K} \lambda_{t_k, t_{k'}} \right\},$$

$$Z(\mathbf{w}) = \sum_{\mathbf{t}} p_0(\mathbf{t} \mid \mathbf{w}) \exp\left\{ \frac{1}{2} \sum_{k=1}^{K} \sum_{k'=1;k' \neq k}^{K} \lambda_{t_k, t_{k'}} \right\},$$

$$p_0(\mathbf{t} \mid \mathbf{w}) \equiv \prod_{k=1}^{K} p_0(t_k \mid w_k)$$

提案手法では、この確率分布に基づいて未知語の品詞推定を行う。以下では、この確率モデルに基づいて未知語の品詞を推定する方法と、モデルのパラメータを訓練データから獲得する方法について説明する。

まず、テストデータ$\mathbf{w}$と初期分布$p_0(t|w)$とモデルのパラメータ$\lambda_{i,j}$が与えられた場合に、テストデータ中の未知語の品詞$\mathbf{t}$を求めることを考える。ここでは、最大事後周辺確率推定を用いて$\mathbf{t}$を求める。つまり、未知語の品詞推定結果の解$\mathbf{t}'=\{t'_1, \cdots, t'_K\}$を、次のようにして求めることにする。

$$t'_k = \arg\max_{t} P_k(t \mid \mathbf{w}),$$

ここで、$P_k(t|\mathbf{w})$は局所的な文脈の集合$\mathbf{w}$が与えられた場合におけるk番目の未知語の品詞の周辺確率であり、以下のように確率分布$P(\mathbf{t}|\mathbf{w})$に関する期待値として計算することができる：

$$\begin{aligned} P_k(t \mid \mathbf{w}) &= \sum_{t_1, \cdots, t_{k-1}, t_{k+1}, \cdots, t_K} P(\mathbf{t} \mid \mathbf{w}), \\ &= \sum_{\mathbf{t}} \delta(t_k, t) P(\mathbf{t} \mid \mathbf{w}). \end{aligned}$$

このような期待値は、確率分布から生成された有限個のサンプルを用いて近似的に計算することができる。つまり、確率分布 $P(\mathbf{t}|\mathbf{w})$から生成された M 個のサンプル$\{t^{(1)}, \cdots, t^{(M)}\}$を用いることにより、品詞の周辺分布は次のように近似することができる：

$$P_k(t \mid \mathbf{w}) \approx \frac{1}{M}\sum_{m=1}^{M}\delta(t_k^{(m)}, t)$$

ここでサンプルを生成する手法としては、マルコフ連鎖モンテカルロ法[文献34]の一種であるギブスサンプリングを使用した。

次に、L 個の事例からなる訓練データ $\{\langle \mathbf{w}^1, \mathbf{t}^1\rangle, \cdots, \langle \mathbf{w}^L, \mathbf{t}^L\rangle\}$ と初期分布 $p_0(t|w)$が与えられた場合に、モデルのパラメータ$\Lambda=\{\lambda_{ij}\}$を推定することを考える。ここでは、最大事後確率推定によりΛ を求める。つまり、次のような目的関数 $\mathcal{L}_\Lambda$を定義し、この値を最大化するΛを求める。

$$\mathcal{L}_\Lambda = \log\prod_{l=1}^{L} P(\mathbf{t}^l \mid \mathbf{w}^l) + \log P(\Lambda)$$

ここで、パラメータの事前分布 $P(\Lambda)$には Gaussian prior[文献 35]を使用し、準ニュートン法を用いて最適解を求めることにする。また、解析時と同様にあらゆる可能な品詞に対する数え上げが必要になるため、ギブスサンプリングを用いて近似的に計算を行う。

提案手法を用いて未知語の品詞推定を行う場合、ラベル無しデータを利用した半教師あり学習が容易に実現できると思われる。すなわち、テストデータに単純にラベル無しデータを結合し、テスト時にその結合されたデータをまとめて解析することにより、容易にラベル無しデータを利用することができると思われる。このようにテストデータの量を増やせば、有用な局所的な情報を持った事例の数も増加すると思われるが、そのような事例の品詞は容易に予測することができ、他の事例の品詞を予測する際に大域的な情報として利用できる可能性がある。

実験：中国語のコーパス(CTB、PFR)、日本語のコーパス(EDR、KUC、RWC)、および英語のコーパス(GEN、SUS、WSJ)を使用して、提案手法の有効性を調べた。これらのコーパスの統計情報を表4-1-3に示す。各コーパスは、training、test、unlabeledの3つに分割した。trainingデータを用いてモデルのパラメータを学習させ、testデータ中の未知語の品詞をどれだけ正しく推定できたかを評価した。ただし、trainingデータ中に含まれない単語を未知語と定義した。また、unlabeledデータは、半教師あり学習の実験に使用した。

実験結果を表4-3-4に示す。この表の中で、**局所**、**局所＋大域**、**局所＋大域＋ラベル無しデータ**は、それぞれ局所的な情報のみを使用した場合、提案手法により局所的な情報と大域的な情報を利用した場合、提案手法により局所的な情報と大域的な情報を利用してさらにラベル無しデータも利用した場合の未知語の品詞推定結果を表す。提案手法により、局所的な情報だけではなく大域的な情報も利用することで、局所的な情報しか使用しない場合に比べて未知語の品詞推定精度を向上させることができた。また、ラベル無しデータも利用した場合、いくつかのコーパスでは精度が向上したが、CTB、KUC、WSJコーパスでは逆に精度の低下が見られた。

言語	略称	コーパス	品詞の候補数	テストデータ中の未知語数
中国語	**CTB**	Penn Chinese Treebank	28	749
	PFR	PFRコーパス	39	27,774
日本語	**EDR**	EDRコーパス	15	24,178
	KUC	京大コーパス	36	2,477
	RWC	RWCPコーパス	55	11.177
英語	**GEN**	GENIAコーパス	36	7,775
	SUS	SUSANNEコーパス	90	5,760
	WSJ	Penn Treebank WSJ	33	4,253

表4-3-3：実験に使用したデータ

コーパス(言語)	精度		
	局所	局所＋大域	局所＋大域＋ラベル無しデータ
CTB(中国語)	0.7423	**0.7717**	0.7704
PFR(中国語)	0.6499	**0.6690**	**0.6785**
EDR(日本語)	0.9639	**0.9643**	**0.9651**
KUC(日本語)	0.7501	**0.7634**	0.7562
RWC(日本語)	0.7699	**0.7785**	**0.7787**
GEN(英語)	0.8836	**0.8837**	**0.8863**
SUS(英語)	0.7934	**0.7957**	**0.7979**
WSJ(英語)	0.8345	**0.8368**	0.8352

表4-3-4：未知語の品詞推定の実験結果

４－３－３　英日翻訳、標準文書翻訳の研究

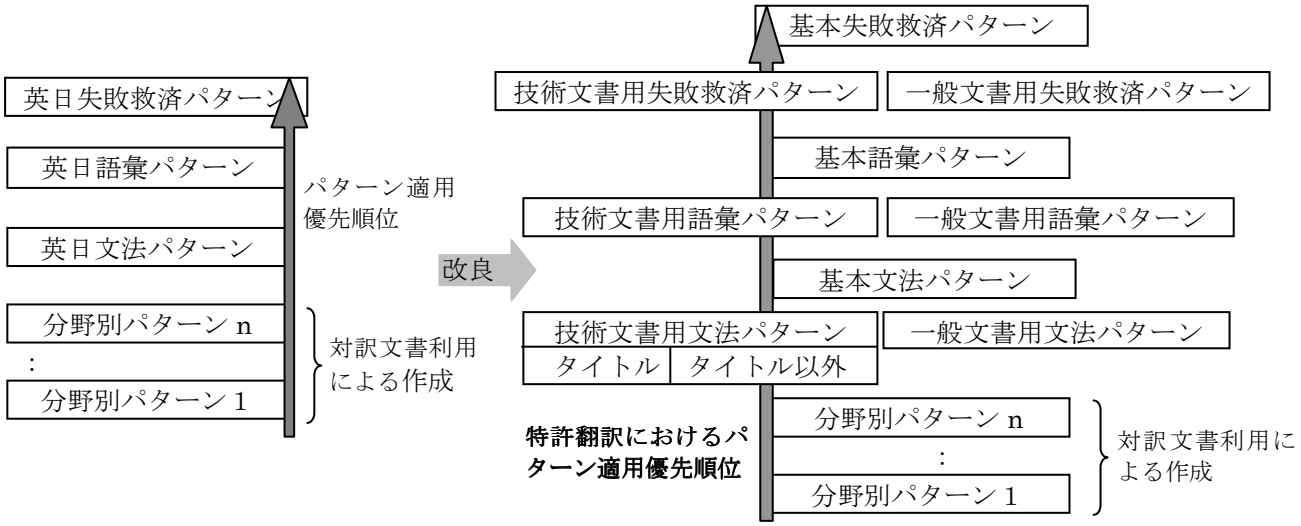


図 4-3-2：英日文法・語彙パターンの構成概略図

本研究テーマ開始前の英日翻訳の文法・語彙パターンは、図 4-3-2 の⇒の左のように単純な構成をしていた。このような構成の場合、各分野における専門用語や特許特有の表現は、分野別パターン 1～n に登録しなければならない。しかし、例えば、和語動詞よりサ変動詞の訳語が望ましい、外来語はカタカナ語表記が望ましい、というような技術文書全般にみられる翻訳の特徴に関する翻訳パターンを分野別パターンに登録することは非常に難しい。また、Web ページ等は口語的な表現や非文が多くあるため、Web ページ翻訳用の特別な語彙パターンを用意する必要があるが、このような口語向けのパターンは技術文書の翻訳に副作用を及ぼし、逆に訳質が低下するという問題が発生する。

上記の課題を解決するために、図 4-3-2 の⇒右のように、システムが有する文法・語彙パターンを、技術文書翻訳用と一般翻訳用に分類し、それぞれを使い分けることによって、より適切な翻訳結果が得られるように翻訳パターンの構成を改良した。

具体例を挙げて説明すると、 “controller”は、一般文書用語彙パターンには「会計監査」、「コントローラ」の２つの訳語を有する。しかし、技術文書用語彙パターンには、「コントローラ」の訳語しか持たない。 “have”の訳語は、一般文書用語彙パターンでは、「持つ」であるが、技術文書用語彙パターンは「有する」である。

さらに、技術文書用文法パターンにおいては、タイトルかそうでないかで、使用するパターンを変える。一般に、タイトルの場合、名詞句で終わる可能性が高く、タイトルでない場合は主部と動詞句をもつ文である可能性が高い。したがって、タイトルの場合、名詞句の優先順位を上げ、そうでない場合は

文の優先順位を上げる。このような、辞書の切り替えは、モードの切り替えによって容易に行なうことができ、今後の辞書の拡張（さらなる分類項目の追加）も容易になっている。

最後に、テーマ研究開始前と現在の英日文法・語彙パターンの登録数を表 4-3-5 にまとめる。

2002 年 10 月		2006 年 3 月		
失敗救済パターン	340	基本用	350	460
		一般用	40	
		技術用	70	
語彙パターン (*注:データ数)	95,000	基本用	98,000	131,000
		一般用	20,000	
		技術用	13,000	
文法パターン	2,000	基本用	2,600	3,110
		一般用	150	
		技術用	360	

表 4-3-5：英日翻訳パターン数

４－３－４　中日翻訳システムの研究

(a) はじめに

中日翻訳システムは 2003 年度から新しく研究開発を開始した。まったく何もない状態から開発を始めたが、言語に依存しない部分は既存の翻訳エンジンおよび形態素解析などを流用できたため、短期間で業界トップレベルの翻訳品質を得ることができた。

(b) システム構成

翻訳の流れを図 4-3-3 に示す。図中の形態素解析は 4-3-2 で述べたような手法で作成した。構文解析生成および形態素生成部分は既存のモジュールをほとんど変更することなしに使うことができた。

形態素解析の学習に使用するタグつき中文コーパスに関しては、中国科学院から品詞タグのついた約 100 万語のコーパスを購入した。さらにうまく形態素解析ができない文に関しては、再学習用コーパスを作成し、学習させた。

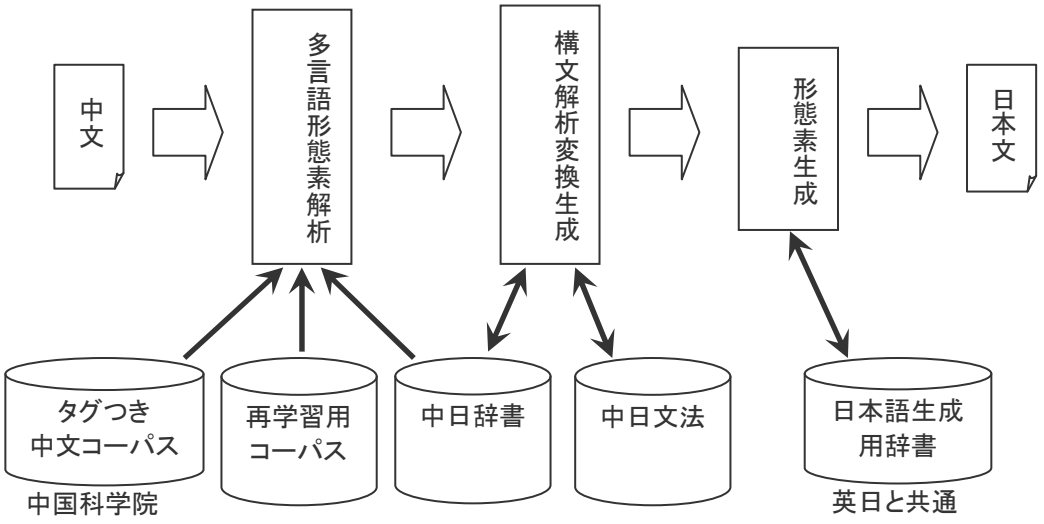


図 4-3-3：中日翻訳の流れ

(c) 文法・辞書

翻訳パターンおよび辞書パターンの例を以下に示す。

- 文法の例　[zh:S [1:NP] [2:VP]]
  [ja:S [1:NP] は [2:VP]];
- 辞書の例　[zh:Adj 漂亮]
  [ja:Ks 美しい];
- 特殊構文用パターンの例
  [zh:VP 不但 [1:VP] , 而且 [2:VP]]
  [ja:VP [1:VP] だけでなく [2:VP]];
  [zh:S 你们好]
  [ja:S みなさん、こんにちは];

このように、文法、辞書ともにパターンで記述されている。中日翻訳用の文法数と辞書数を表 4-3-6 に示す。文法パターンは東京外語大の評価例文を参考にして、約 40 の文法項目に対応した。辞書は中国科学院のコーパスの中に出現した語および品詞を抽出し、日本語訳を付けていく形で約 12 万語作成した。また、人名地名辞書およびコンピュータ辞書 10 万語を業者より購入し、計 22 万語の辞書を持つシステムになった。

文法数		600
辞書数	基本辞書	120000
	人名地名	80000
	コンピュータ	20000

表 4-3-6：中日翻訳の文法数と辞書数

品詞は表 4-3-7 のように 19 種類に分類した。ユーザ辞書には、名詞、動詞、形容詞、副詞、方位詞、数詞、時間詞、量詞のみを登録できるようにした。

品詞	例	システム辞書	ユーザ辞書
名詞	书,笔	○	○
動詞	看，走，能	○	○
形容詞	大，白	○	○
副詞	都，不	○	○
方位詞	里，上，下	○	○
数詞	百，几	○	○
時間詞	今天，现在	○	○
介詞	在，从	○	
量詞	册，个	○	○
語気詞	呢，吗	○	
助詞	的，了	○	
代詞	这边，哪些，我	○	
連詞	还是，无如	○	
感嘆詞	呸，噢	○	
接頭語	非，无	○	
接尾語	们，业	○	
擬声語	阿嚏，哈哈	○	
記号	。，、，(，"，！	○	
文	你们好，对不起	○	

表 4-3-7：中日翻訳における品詞体系

(d) 評価

派遣社員を使って簡易評価を行った。Web ページのニュース記事、コンピュータマニュアル、特許タイトルおよびアブストラクトの 3 分野、6 種類（それぞれ 100 文）を評価例文とした。システム名は伏せておき、それぞれの文に、１～５の点数を付けてもらった。

評価結果を表 4-3-8 に、評価例文の一部を表 4-3-9 に示す。数字は点数の平均値を表している。この

結果、既存の翻訳システムと大差ない翻訳品質になったことがわかった。

例文		沖電気		A 社		B 社	
ニュース記事	政治	2.91	2.98	3.17	3.16	2.58	2.58
	技術	3.05		3.15		2.57	
コンピュータマニュアル	hp	3.29	3.21	3.37	3.16	2.52	2.61
	三星	3.13		2.94		2.70	
特許タイトル・アブストラクト	G06 計算機	3.10	2.94	2.57	2.46	1.59	1.75
	D04 繊維	2.78		2.35		1.91	

表 4-3-8：中日翻訳システムの比較評価結果

原文	沖電気	A 社
中国､欧盟和美国都是国际社会的重要成员,在国际事务中负有重大的责任,扮演着重要的角色。	中国,EU とアメリカはみな国際社会の重要な構成員で、国際事務の中で重大な責任を負って、重要な役を扮しています。	中国、EU のむつまじく楽しい首都は国際社会の重要な成員で、国際的事務の中で重大な責任を負って、重要な役を演じている。
中国的文化是世界上最丰富多彩的文化之一。	中国の文化は世界の上で最も豊富な多彩な文化のひとつです。	中国の文化は世界で最も多彩な文化の 1 つだ。
做政治家比做物理学家难多了	政治家になることは物理学者になることよりずっと難しいです	政治家をするのは物理学者をするより多いにくくなった
辞书检索部检索辞书部,汇集单词各构成文字的同音异义字排列顺序号。	辞書検索部分は辞書部分を検索して、単語おのおのが文字を構成する同音異義文字排列のシーケンス番号を集約します。	辞書の検索部は辞書部を検索して、集まる単語はそれぞれ文字の同音の異なる義の字の整列順序号を構成する。
将从上述键盘手段得到的拼音符号和声调与从上述数据库文件手段得到的拼音符号和声调进行比较的比较手段	上述のキーボード手段から入手するピンイン記号と声調と上述のデータベース文書手段から入手するピンイン記号と声調で比較するのを行なって手段を比較します	上述のキーボードの手段の得るピンインの記号と声調から上述のデータベースファイルの手段の得るピンインの記号と声調からと比較の比較の手段を行う
新收购电脑业务的协同效益及供应链效率较预期提前显现，集团强调整合进程顺利,初见成效。	新しいコンピュータ業務を買い集める協同効果と利益およびサプライチェーン効率は比較的繰り上げを予期して現れる、グループ強調は過程を相当することは順調で、初めて効果を見ます。	新しくコンピュータの業務のを買い付けて効果と利益に協力しておよび鎖に効率を供給する比較的に早めにはっきりと現れると予想して、グループは過程を統合するのが順調であることを強調して、初めて効果が現れる。

表 4-3-9：中日翻訳の翻訳例

４－３－５　韓日翻訳システムの研究

(a) はじめに

韓日翻訳システムは､最終年度１年での開発となったため､言語非依存型翻訳エンジンの特徴を生かし、短期間で動作させることを目指した。4-3-2 で述べたように、韓国語形態素解析は、日本語や中国語と共通の単語分割と品詞タグ付け処理に、縮約等の韓国語特有の形態素変化に対応するための綴り復元処理を追加することにより実現された。また、日本語形態素生成は、英日翻訳、中日翻訳と共通の日本語生成モジュールを利用した。以下では、辞書パターンおよび文法パターン(以下、これらを併せて翻訳パターンと呼ぶ。）の開発を中心に、韓日翻訳システムの研究開発の詳細を説明する。

(b)韓日翻訳システムの開発工程

評価例文の作成：韓日翻訳システムの研究に際して、翻訳パターン開発の基礎となる評価例文を作成し

た。評価例文は、単文を中心とする 393 文(平均文長 4.4 eojeol[2])で、韓国語の基本的な文法を網羅することを主眼に選定を行った。

辞書パターンの作成：韓国語形態素解析では利用したコーパスの品詞体系[3]にあわせて、入力文を 54 品詞に分類、タグ付けしているが、翻訳処理部では形態素解析の誤りを吸収し、文法を簡素化するために、名詞、代名詞、動詞(形容詞を含む)、副詞、限定詞、助詞、終助詞、記号の 8 品詞のみを用いることとした。そして、評価例文に出現する単語を中心に、名詞 796 語、用言(動詞、形容詞)864 語など、全品詞で約 2500 語の辞書パターンを作成した。

文法パターンの作成：韓国語と日本語は語順や文法構造が類似しており、(例１)のように、単語を韓国語から日本語に置き換えるだけで翻訳できる文も多い[文献 36]。このことから、我々はまず、入力された韓国語単語列を、語順を入れ替えずに日本語に置き換える文法パターンを作成した。

(例１)
내가 어제 그 책방에서 산 책은 아주 재미있었다.
私が 昨日 あの 本屋で 買った 本は とても 面白かった。

実験及び評価：上記の文法・辞書セットを用いて評価例文 393 文を翻訳したところ、全体のほぼ 90%にあたる 348 文が正しく解析されていることを確認できた。解析に失敗した主な理由は以下の通りである。

- 形態素解析の失敗、形態素解析と辞書パターンの不整合によるもの
- 品詞の多義性解消に失敗したもの
- 翻訳パターンの不足によるもの

また、今回の評価例文は単文が中心であること、使用した翻訳パターンが少ないことから、顕在化しなかったが、今後、実用レベルの翻訳を行う場合に想定される問題もある。以下に、主な課題を挙げる。

- 助詞の訳し分け
  - 韓国語と日本語では助詞の使い方が類似しているが、対応がつかない場合もある。例えば、韓国語では、「A の B」というときの「の」に当たる助詞は使わないことが多いので、(例２)の文を単純に翻訳すると「私たち父は大学教授です。」となる。あるいは、格助詞「が」にあたる助詞の「이」の使い方が日本語と異なるため、(例３)の文は「ソウルが どこですか？」となり、日本語として不自然な表現になる。このような現象に対応するには、前後の並びを見るなどして、適切な翻訳を行うようにしなければならない。

(例２)
우리 아버지는 대학 교수입니다.
私たち(の) 父は 大学 教授です。

(例３)
서울이 어디에요?
ソウルが どこですか？

- 同音異義語の訳し分け
  - 韓国語では 19 個の子音字母と 21 個の母音字母を組み合わせて作るハングル文字のみで表現される。もともと漢字由来の単語や外来語もハングル文字に置き換えているため、漢字・ひらが

[2] eojeol は、韓国語における分かち書きの単位で 1eojeol およそ 1～4 語の単語で構成される。
[3] http://kibs.kaist.ac.kr/beginner/kbase2-e.htm

な・カタカナを用いる日本語と比べて同音異義語が多い。特に、漢字由来の単語は、(例４)に示すように、同音異義語になりやすい。今回の実験では、限られた語彙の使用であったため問題にはならなかったが、今後翻訳パターンの増加させていく際には、品詞推定処理および語彙的曖昧性解消処理の精度を向上させる枠組が必要になると考えている。

（例４）

사고

史庫/四苦/事故/社告/思考

・ 品詞体系の見直し
  - ➢ 今回の研究では、品詞数を８つに絞ってシステムを構築した。しかしながら、今後語彙を増やし、複雑な文法制御を行っていくためには、細分化された品詞体系と辞書情報が必要になる。先行して開発された英語や中国語の品詞体系を参考に、韓国語の特徴にあわせた品詞体系を検討していく。

４－３－６　「訳してねっと」協調的翻訳支援環境の研究

(a) はじめに

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 4-3-4)上に実装した。実装した主な機能は以下の８項目である。以下に各機能の概要を示す。

1. コミュニティ管理・作成機能
2. 他のコミュニティの辞書参照機能
3. コミュニティ専用の辞書の登録・検索・更新機能
4. 文書や URL を共有知識として管理する文書管理機能
5. 各種文書の翻訳及び翻訳結果の修正機能
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. ユーザ間の情報伝達手段としてのメッセージ機能
8. コミュニティに関する情報の多言語表示機能

図 4-3-4：訳してねっとのトップページ

(b) コミュニティ管理・作成機能

「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 4-3-5 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。

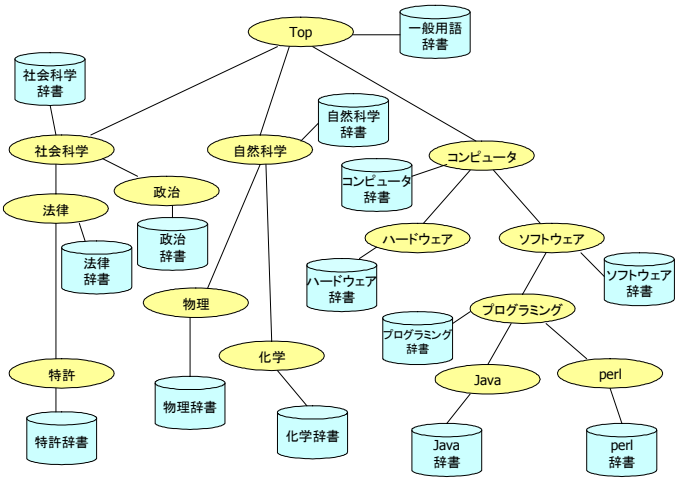


図 4-3-5：「訳してねっと」のコミュニティ構成

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。

(c) 他のコミュニティの辞書参照機能

より高精度の翻訳を行うために、翻訳に使用する辞書を自由に選択することもでき、例えば図のように、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。

図 4-3-6：使用時書の選択画面

(d) コミュニティ専用の辞書の登録・検索・更新機能（辞書機能）

辞書登録画面を図に示す。図 4-3-7 のように、英語見出しと日本語見出しを入力するだけで、簡単に登録できる。意味などの詳細情報も登録したい場合は、オプションボタンを押すことによって、登録できる。

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。

図 4-3-7：辞書登録の画面

(e)文書や URL を共有知識として管理する文書管理機能

図 4-3-8 のように、よく翻訳するサイトや文書をコミュニティごとに登録することができ、一度登録すると翻訳ボタンを押すだけで翻訳することができるようにした。

図 4-3-8：Jakarta コミュニティのメインページ

(f)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）

翻訳は英語から日本語、日本語から英語、中国語から日本語の３通りが可能で、テキスト翻訳、Web 翻訳、ファイル翻訳が選択できる。テキスト翻訳はユーザが翻訳したいテキストを入力して翻訳する。

Web 翻訳はユーザが翻訳したいページの URL を入力して翻訳する。ファイル翻訳はユーザのコンピュータ上の Microsoft Word, Excel, PowerPoint や XML, HTML などのファイルを翻訳することができる。また図 4-3-9 のように翻訳結果を修正することができるようにした。これによって、修正した部分を自動的に辞書に取り込む機能も作成した。

(g)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

本システムを使用する場合、ユーザタイプ・ユーザロールによって、アクセス権を設定・変更できるようにした。ユーザタイプは本システムを使用する場合に与えられる役割で、ユーザロールはコミュニティ参加した場合に与えられる役割である。表 4-3-9 にユーザタイプ、表 4-3-10 にユーザロールの概要を

示す。又、図 4-3-10 にユーザタイプ、ユーザロールの役割を表す概念図を示す。

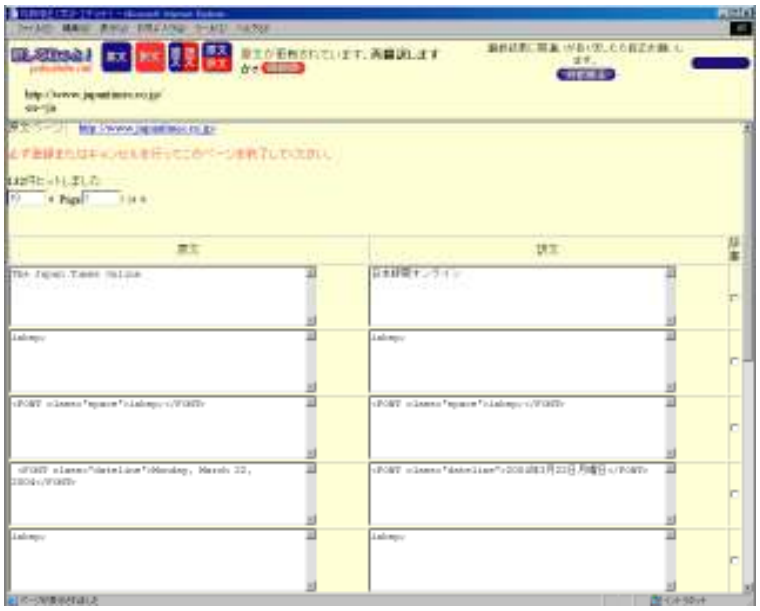


図 4-3-9：ポストエディット画面

項番	ユーザタイプ	内容
1	管理者	本システムを管理するユーザ。
2	オペレータ	コミュニティの追加/修正・ユーザの登録・ユーザのコミュニティへの登録/削除等を行うユーザ。
3	エキスパート	単語の承認を行うユーザ。コミュニティに参加しなくても承認は可能とする。
4	一般	本システムを利用するユーザ。
5	ゲスト	本システムを利用するユーザ。コミュニティに参加することはできない。

表 4-3-9：ユーザタイプ

項番	ユーザロール	内容
1	リーダ	コミュニティを管理するユーザ。コミュニティ内の辞書の登録・承認や、掲示板の削除などができる。
2	コミッタ	単語の承認を行うユーザ。参加しているコミュニティのみで承認可能とする。
3	メンバ	コミュニティに参加しているユーザ。コミュニティ内の辞書登録や掲示板への投稿ができる

表 4-3-10：ユーザロール

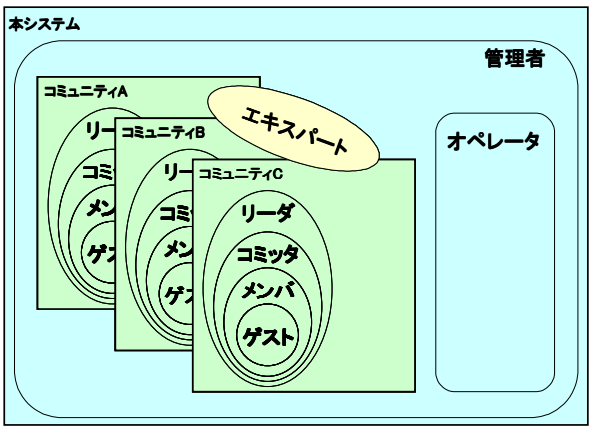


図 4-3-10：ユーザタイプ・ユーザロールの役割のイメージ

(h) ユーザ間の情報伝達手段としてのメッセージ機能

掲示板の機能や Q&A の機能をコミュニティごとに設けて、メンバ同士でコミュニケーションが取れるよ

うにした。また、コミュニティリーダーがコミュニティメンバに同時にメールを送信できる機能も搭載した。

(i)コミュニティに関する情報の多言語表示機能

表示する言語はユーザごとに設定画面から変更できるようにした。図 4-3-11 は英語を選択した場合の例である。また、掲示板などに投稿された文も自動的に翻訳を行ない、指定した言語で表示できるようにした。

図 4-3-11：英語表示の例

(j)「訳してねっと」の公開

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た。2006 年 1 月 31 日より中日翻訳版の公開も開始した。

ユーザによる辞書登録は、現在までに数千語におよび、翻訳品質の向上に貢献している。

フィードバックの画面を図に示す。ユーザは「はい」か「いいえ」をクリックして、送信ボタンを押すと、開発者にその情報が送られるようになっている。1 日に数１０通のフィードバックが寄せられる。コメント欄に丁寧に問題点を書いてくれているものも多く、翻訳品質の改良に非常に役に立っている。

図 4-3-12：ユーザフィードバックの画面

４－３－７　結論と今後の課題

最終目標の各項目とそれぞれにおける実施状況を以下にまとめる。

(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。

・多言語に対応した形態素解析システムを実現するために、単語レベルと文字レベルの情報を用いた中国語・日本語・韓国語形態素解析の研究を行った。この提案手法により，単語単位の候補と文字単位の候補を同時に考慮して解析することで、明示的な単語境界を持たない中国語や日本語、韓国語に対して高い精度で単語分割と品詞付与を行えることを確認した。さらに、大域的な情報を用いた未知語の品詞推定の研究を行った。この提案手法により，同じ語形を持つ未知語の文書中での全ての出現を同時に考慮した確率モデルを利用することで、未知語の品詞推定精度を向上させられることを確認した。

・協調型翻訳支援環境である「訳してねっと」は、一般公開を行い、ユーザの辞書登録とフィードバックにより、翻訳品質の向上に非常に役立っている。今後は、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と同時に、ビジネス化へ向けた機能も追加して行く予定である。

(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。

・中日翻訳システムの研究では、ゼロからシステムを開発し、短時間で市販パッケージソフト程度の翻訳品質を有するシステムを構築することができた。

・韓日翻訳システムの研究では、言語非依存の機械翻訳エンジンの特長を生かし、短期間でのシステム稼動を実現した。また、単文を中心とする評価例文に対して翻訳実験を行い、韓日翻訳が正しく動作することを確認した。

・英日翻訳は標準文書（特に特許文書）に特化したシステムを開発し、高い翻訳精度を実現した。

４－４　総括

平成 14 年度の下期から約 3 年半、開始当初は「多言語標準文書処理システム」の設計方針及び各項目の研究方針を確定するのに時間を有したが、方針の確定後は、図 4-4-1 に示す「多言語標準文書処理システム」の全体像にしたがって、研究開発を順調に進めることができた。

最も注目すべき成果は、「多言語標準文書処理システム」の中核である協調的翻訳支援環境として 2004 年 9 月には、訳してねっと英日版を、2006 年 1 月には訳してねっと中日版を一般公開したことである（http://www.yakushite.net）。現在、ユーザ数は約 2,000 人、1 日の平均ページビュー数は、約 11,000 である **(図 4-4-1:全体(1)(3))**。

また、我々は、標準文書として、翻訳要求が高まっている特許分野を選定し、本研究で開発した翻訳テンプレート自動獲得手法を用いて、特許文書から、81 分野、約 120 万語の英日翻訳テンプレートを作成した。この翻訳テンプレートは英日特許翻訳において実際に利用している。機械翻訳の自動評価手法により、自動獲得した翻訳テンプレートは翻訳品質の向上に寄与していることも検証した**(図 4-4-1:ア)**。

自己組織化研究においては、分野辞書の自動構築方法、及び、入力文書の分野の自動判定方法を提案し、「訳してねっと」における実証実験により、本手法が翻訳品質向上に有益であることを確かめた**(図 4-4-1:イ)**。

多言語翻訳システムの構築には、言語に依存しない翻訳エンジン、及び、各言語の翻訳テンプレートが必要となる。我々は、多言語に対応した形態素解析システムを開発し、日本語、英語、中国語、韓国語の動作を確認した**(図 4-4-1:イ)**。また英日翻訳テンプレート開発の経験を活かして、ゼロから中日、韓日の翻訳テンプレートを構築した。中日翻訳においては、市販パッケージソフト程度の翻訳品質を有するシステムを構築することができた**(図 4-4-1:全体(2))**。

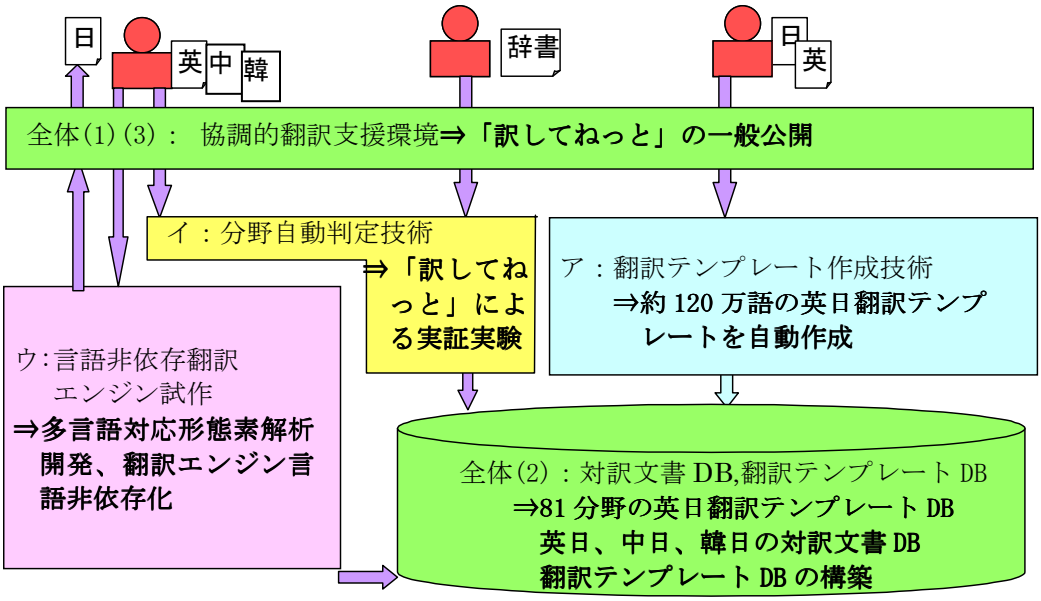


図 4-4-1：多言語標準文書処理システムの最終目標及びその成果

本研究テーマは、数多くの研究成果を出しただけでなく、この研究成果を応用したビジネスも進行中である。例えば、本研究成果はリコーテクノシステムズ株式会社殿の日米統合概念検索サービス「RIPWAY[4]」(http://www.ripway.net)に利用されている。これは、研究成果である特許文書における翻訳品質の高さ（特に専門用語の適訳性）が高く評価されたためである。その他、翻訳テンプレート自動獲得モジュールの提供など、具体的な案件もある。

今後、本研究の成果が、より世の中の役に立つよう、積極的なビジネス展開を図っていきたい。

参考文献

[文献1] 潮田他(2003). “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

[文献 2] “平成 16 年度 AAMT/Japio 特許翻訳研究会 報告書”, 財団法人 日本特許情報機構, 平成 17 年 3 月

[文献 3] 北村美穂子, 松本裕治(1997). “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, 38(4), pp.727-736

[文献 4] 北村美穂子(2003). “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

[文献 5] 北村美穂子, 松本裕治(2006). “言語資源を活用した実用的な対訳表現抽出”自然言語処理, 13(1), pp.3-25

[文献 6] 熊野明, 平川秀樹(1994). “対訳文書からの機械翻訳専門用語辞書作成” 情報処理学会論文誌, 35(11), pp.2283-2290

[文献 7] 出羽達也(2001). “対訳文書から自動抽出した用語対訳による機械翻訳の訳語精度向上” NL-144-1 pp.1-7

[文献 8] 福井雅敏, 樋口重人, 藤井敦, 石川徹也(2001). “日米対応特許コーパスを用いた対訳抽出手法” NL-145-4, pp.23-28

[4] RIPWAY はリコーテクノシステムズ株式会社の登録商標である。

[文献 9] 高橋博之, 川崎立八, 牧田光晴, 樋口重人, 藤井敦, 石川徹也(2003). “日米特許公報を用いた対訳辞書および翻訳メモリの構築” NL-155-7, pp.39-46

[文献 10] Shimohata, S., Sugio, T. and Nagata, J.(1997). “Retrieving Collocations by Co-occurrences and Word Order Constraints” In Proceedings of 35th Annual Meeting of the Association for Computational Linguistics, pp.476-481

[文献 11] Fung, P. and McKeown, K.(1997). “Finding Terminology Translations from Non-parallel Corpora” In Proceedings of 5th Annual Workshop on Very Large Corpora, pp.192-202.

[文献 12] Rapp, R.(1999). “Automatic Identification of Word Translations from Unrelated English and German Corpora” In Proceedings of 37th Annual Meeting of the Association for Computational Linguistics, pp.519-526.

[文献 13] Shimohata, S.(2005). “Finding Translation Candidates from Patent Corpus” In Proceedings of Workshop on Patent Translation (MT Summit X), pp.50-54

[文献 14] 今村賢治, 隅田英一郎, 松本裕治(2004). “機械翻訳自動評価指標の比較” 第 10 回言語処理学会 年次大会(NLP2004), pp.452-455

[文献 15] Matthias Eck and Chiori Hori(2005). “Overview of the IWSLT2005 Evaluation Campaign” http://www.is.cs.cmu.edu/iwslt2005/proceedings.html

[文献 16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu(2002). “BLEU: a Method for Automatic Evaluation of Machine Translation” In Proceedings of ACL-2002, pp.311-318

[文献 17] George Doddington(2002). “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics” In Proceedings of the Human Language Technology Conference(HLT-2002), pp.128-132

[文献 18] Sonja Nieβen, Franz Josef Och, Gregor Leusch and Hermann Ney(2000). “An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research” In Proceedings of LREC 2000, pp.39-46

[文献 19] Franz Josef Och(2003). “Minimum Error Rate Training in Statistical Machine Translation” In Proceedings of ACL-2003, pp.160-167

[文献 20] Joseph P. Turian, Luke Shen, and I. Dan Melamed(2003). “Evaluation of Machine Translation and its Evaluation” MT Summit IX, pp.386-393

[文献 21] 金山 博, 荻野紫穂(2002). “翻訳精度評価手法 BLEU の日英翻訳への適用” 情報処理学会研究報告, NL-154, pp.131-P136

[文献 22] Thomas Hofmann(1999). “Probabilistic Latent Semantic Indexing” In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, pp.50-57

[文献 23] 高野祐介, 村松哲, 森辰則(2003). “空間分割型 PLSI を用いた言語横断情報検索”言語処理学会第 9 回年次大会発表論文集, pp.389-392

[文献 24] David M. Blei, Andrew Y. Ng, and Michael I. Jordan(2001). “Latent Dirichlet Allocation” Neural Information Processing Systems 14, pp.993-1022

[文献 25] 佐々木美樹, 北村美穂子, 下畑さより, 中川哲治(2003). “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

[文献 26] 黒橋禎夫, 長尾眞(1998). “日本語形態素解析システム JUMAN version 3.61”京都大学大学院情報学研究科

[文献 27] 山本幹雄, 増山正和(1997). “品詞・区切り情報を含む拡張文字の連鎖確率を用いた日本語形態素解析” 言語処理学会第 3 回年次大会発表論文集, pp.421-424

[文献 28] 吉田辰巳, 大竹清敬, 山本和英(2003). “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

[文献 29] Richard Sproat and Thomas Emerson(2003). “The First International Chinese Word Segmentation Bakeoff” In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp.133-143

[文献 30] Thorsten Brants(2000). “TnT --- A Statistical Part-of-Speech Tagger” In Proceedings of the 6th Applied Natural Language Processing Conference and the 1st Meeting of the North American Chapter of the Association of Computational Linguistics, pp.224-231

[文献 31] Andrei Mikheev(1997). “Automatic Rule Induction for Unknown-Word Guessing” Computational Linguistics, Vol.23, No.3, pp.405-423

[文献 32] Chao-jan Chen, Ming-hong Bai and Keh-Jiann Chen(1997). “Category Guessing for Chinese Unknown Words” In Proceedings of NLPRS'97, pp.35-40

[文献 33] Masayuki Asahara(2003). “Corpus-based Japanese morphological analysis,” Nara Institute of Science and Technology, Doctor's Thesis

[文献 34] 伊庭幸人, 種村正美, 大森裕浩, 和合肇, 佐藤整尚, 高橋明彦(2005). 統計科学のフロンティア 12 計算統計 II マルコフ連鎖モンテカルロ法とその周辺, 岩波書店

[文献 35] Stanley Chen and Ronald Rosenfeld, “A Gaussian Prior for Smoothing Maximum Entropy Models” Technical Report CMUCS-99-108

[文献 36] (2002). “多言語機械翻訳システムの評価研究 共同研究報告書 第１分冊” 東京外国語大学, 国際情報化協力センター, pp.387-393

５ 参考資料・参考文献

５－１ 研究発表・講演等一覧

1. 「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」、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、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

2. “Collaborative Translation Environment `Yakushite.Net'”（デモ展示), ACL(Association for Computational Linguistics) 2003

3. 「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」、松永聡彦、北村美穂子、村田稔樹、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

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

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

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

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

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

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

10.「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

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

12.「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

13.”Finding Translation Candidates from Patent Corpus”, Sayori Shimohata, On-line Proceedings of MT Summit X Workshop on Patent Translation, 2005

14.”A Pattern-Based Machine Translation System - Yakushite Net MT Engine”, Miki Sasaki and Toshiki Murata, On-line Proceedings of International Workshop on Spoken Language Translation, 2005

15.「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、情報処理学会論文誌、Vol.46、No.11、pp.2714- 2727、November 2005

16.“NTCIR-5 CLIR Experiments at Oki”, Tetsuji Nakagawa, In Proceedings of the 5th NTCIR Workshop, pp.104--109, December 2005

17.「言語資源を活用した実用的な対訳表現抽出」、北村美穂子、松本裕治、自然言語処理、13(1)、pp.3-25、2006

18.「大域的な情報を用いた未知語の品詞推定」、中川 哲治, 松本 裕治、情報処理学会研究報告、2006-NL-172、March 2006