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[†]
[†]沖電気工業株式会社研究開発本部ユビキタスシステムラボラトリ
[*]神戸大学大学院

1．はじめに

機械翻訳において高品質な翻訳結果を得るためには、辞書の充実が必須である。特に特許の文章は一般の文章と比べて文が長い、専門用語が多い、といった特徴がある。このような文章において、専門用語辞書を構築することは、適切な訳語の生成にとどまらず、構文解析の曖昧性解消や翻訳時間の短縮にも大きな効果があると考えられる。

そこで我々は、日本語特許抄録および対応する公開特許英文抄録(PAJ: Patent Abstracts of Japan)より自動的・半自動的に抽出した大量の専門用語辞書を用いて、特許翻訳における辞書登録の有効性を検証した。

2．特許文の特徴と翻訳における問題

図 1 に、日本語特許抄録および対応する PAJ の例を示す。この例からも分かるように、特許の文章には以下のような特徴がある[1]。

(1) 文が長い

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。

(2) 専門用語が多い

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。

3．専門用語辞書の構築

3.1 特許の分類

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、図 1 の特許の IPC コードは “C12N 11/00”で、分類の詳細は表 1 のようになっている。

例) Ｃ (1)　１２ (2)　Ｎ (3)　１１ (4)　／　００ (5)

表 1 IPC コードの例

(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

CARRIER FOR BIOREACTOR, METHOD FOR PRODUCING THE SAME AND METHOD FOR USING THE SAME CARRIER
PROBLEM TO BE SOLVED: To obtain a carrier for a bioreactor having water swellability, capable of controlling the degree of volume swelling or the degree of volume swelling and specific gravity, uniformly flowable in a reactional vessel and further having a high physical strength and to provide a method for using the carrier.
SOLUTION: This carrier for the bioreactor comprises (A) a water-swellable thermoplastic resin, (B) a resin compatible with the component (A) and, as necessary, (C) an inorganic filler and has the degree of volume swelling controlled within the range of 120-3,000% or further specific gravity during swelling with water controlled within the range of 1.02-2.12. The method for denitrifying treatment of organic wastewater comprises using the carrier for the bioreactor as a carrier for immobilizing microorganisms in the method for nitrifying and denitrifying the nitrogen in the organic wastewater with the microorganisms.

【発明の名称】バイオリアクター用担体、その製造方法及び該担体の使用方法
【国際特許分類第7版】
C12N 11/00
B29C 47/30　ZAB …
【要約】
【課題】 水膨潤性を有し、かつ体積膨潤度又は体積膨潤度と比重が制御され、反応槽内における均一流動が可能であって、物理的強度の高いバイオリアクター用担体及びその使用方法を提供すること。
【解決手段】 (A)水膨潤性熱可塑性樹脂と、(B)該(A)成分に対する相溶性樹脂と、場合により(C)無機フィラーを含み、かつ体積膨潤度が１２０～３０００%に、あるいはさらに、水膨潤時の比重が１．０２～２．１２の範囲に制御されたバイオリアクター用担体、並びに、有機性排水中の窒素を微生物により硝化･脱窒素する方法において、微生物固定化用担体として、上記バイオリアクター用担体を用いる有機性排水の脱窒素処理方法である。

図 1 日本語特許抄録と対応するＰＡＪの例

3.2 辞書獲得

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)より用語の抽出および辞書化を行った。辞書獲得の詳細についてはここでは触れないが、日本語特許抄録と対応する PAJ を対訳コーパスとして、統計的手法に基づく対訳表現抽出手法[4][5]により、自動的・半自動的に獲得した。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。

3.3 翻訳環境

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™」[1]上に展開した。図 2 に訳してねっとの概念図を示す。

「訳してねっと™」は、インターネット上の翻訳環境で、ツリー型のディレクトリ構造を持ち、各ディレクトリがさまざまな分野(コミュニティ)

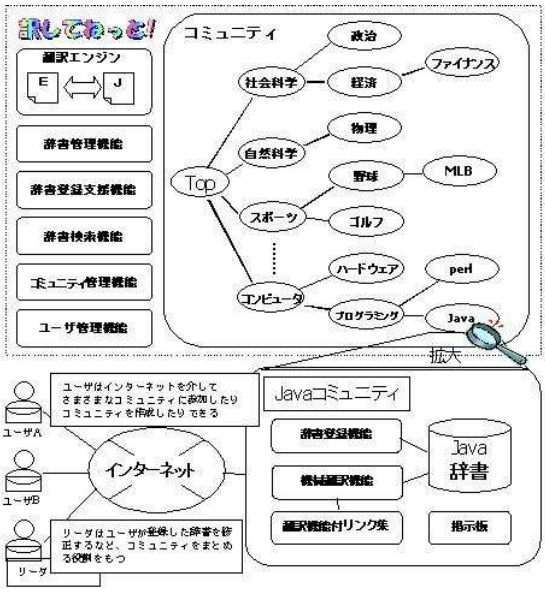

図 2 訳してねっとの概念図

[1] http://www.yakushite.net

(例 1)
E: … the organic compound is **three-dimensionally crosslinked** ….
J(旧): … 有機的な合成物は … **dimensionally** に **crosslinked** の 3 です。
J(新): … 有機化合物は …**三次元的に相互リンクされます**。

(例 2)
E: A **spacer arm** having 300-5,000,000 number-average molecular weight is bonded to a carrier prepacked …
J(旧): **スペーサー**は … 分子のウェイトは予め包装されるキャリアに接合されるという 300-5,000,000 数平均を持っていることを**武装させます**。
J(新): 300-5,000,000 数平均分子量付き**スペーサーアーム**は … 予め包装されるキャリアに接合されます。

(例 3)
E: The treating water is discharged from a **treating water outlet tube** 9.
J(旧): 水を扱うことは**排水口を扱うこと地下鉄** 9 から放出されます。
J(新): 処理水は**処理水排出管** 9 から放出されます

図 3 翻訳結果の例

に対応する。各コミュニティでは、コミュニティ辞書の作成およびコミュニティ辞書を使った翻訳を行うことができる[2][3]。

コミュニティ辞書を使った翻訳は、各コミュニティの辞書だけでなく、そのコミュニティの上位概念の辞書も使って翻訳する。例えば、Java コミュニティで翻訳を行うと、Java コミュニティの辞書はもちろん、木構造の上位に位置する「プログラミング」コミュニティや「コンピュータ」コミュニティの辞書も参照して翻訳を行うことになる。

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。

4. 評価

4.1 評価方法

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

4.2 評価結果

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 2 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。

表 2 評価結果

	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

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 3 に翻訳結果の例を示す。

(例 1)では、”crosslink” が未登録語だったことと、”three-dimentionally”がひとかたまりの副詞と認識できなかったことから、構文解析に失敗していたが、これらの用語が辞書登録されたことにより、正しく解析できるようになった。

(例 2)では、”arm” が動詞と認識され、誤って解析されていたものが、”spacer arm” を辞書登録することにより、正しい構文が認識されるようになった。

(例 3)では、”treating + water outlet + tube”と解析されていた名詞句が、”treating water”と “outlet tube” を辞書登録したことにより、正しい句構造に認識されるようになった。

訳が悪化したものとしては、(例 3)の場合とは反対に、名詞句の一部を登録することにより、全体が名詞句と認識できなくなったもの、限定的な場面でしか使われない訳語を登録した結果、文脈によって不適切な訳になったものがほとんどであった。このような問題は、さらに辞書登録を増やすこと、専門用語辞書と既存の文法・辞書との整合性を高めることで解決できると考えている。

3. まとめ

本稿では、特許翻訳における訳質向上を目的に、日英特許抄録からの専門用語の獲得とその結果得られた辞書を使った翻訳の結果について述べた。その結果、専門用語辞書を多数登録することにより、翻訳品質が向上することが分かった。今後は、研究を進めている対訳コーパスからの対訳表現の自動的・半自動的抽出手法[4][5]の実用化とも絡めて、大規模な専門用語辞書の構築と翻訳品質の更なる向上を目指していく。

今回は辞書の問題についてのみ言及したが、特許文には用語だけでなく、表現形式や文パターンの出現頻度にも偏りがあり(例えば、命令文や疑問文はほとんど出現しないなど)、文法規則についてもカスタマイズが必要である。今後は、これらの面にも検討、調査を行い、総合的な特許翻訳環境の実現に取り組みたいと考えている。

また、今回獲得した辞書は機械翻訳だけではなく、特許における多言語検索などにも利用できると考えており、今後の展開を模索中である。

本研究は、情報通信研究機構平成 14 年度基盤技術研究促進制度に係る研究開発課題「多言語標準文書処理システムの研究開発」の一環として行われている。

謝辞：データを提供していただいた財団法人日本特許情報機構殿、および、有益な議論をしてくださったアジア太平洋翻訳協会 AAMT/Japio 研究会のメンバーに深く感謝いたします。

参考文献

[1] 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

[2] Shimohata, S., Kitamura, M., Sukehiro T., and Murata, T.: “Collaborative Translation Environment on the Web”. In proceedings of *the MT Summit VIII*, pp331-334, 2001.

[3] Sukehiro, T., Kitamura, M. and Murata, T.: “Collaborative Translation Environment `Yakushite.Net' “, In proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, pp769-770, 2001

[4] Kitamura, M. and Matsumoto, Y.: “Practical Translation Pattern Acquisition from Combined Language Resources”, In proceedings of IJCNLP-04: The First International Joint Conference on Natural Language Processing, pp.652-659, 2004.

[5] 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 発表論文集，pp.528-pp.531，1998