中国語特許文書から文パターンを抽出する一方法

A extraction method of sentential patterns from Chinese patent documents

山梨英和大学人間文化学部人間文化学科教授 **江原　暉将**

PROFILE: 1967 年早稲田大学理工学部電気通信学科卒。同年 NHK 入局。2003 年、諏訪東京理科大学教授。2009 年より現職。アジア太平洋機械翻訳協会 (AAMT) ／ Japio 特許翻訳研究会委員。

1 はじめに

機械翻訳において正確な翻訳を行うためには、適切な訳語を出力することと入力言語の構文構造を出力言語の構文構造に適切に変換することが必要である。前者は語彙的曖昧性解消、後者は構造的曖昧性解消と呼ばれ、機械翻訳の大きな課題とされている [1]。特許文書では、専門用語が多く使われ、かつ文長が長く複雑な構文構造を持つという特徴がある。これらの特徴は、語彙的曖昧性解消と構造的曖昧性解消を共に困難にさせ、機械翻訳の精度を低くする要因となっている。

専門用語に対して適切な訳語を出力するために、専門用語対訳辞書の充実が有効で、AAMT/Japio 特許翻訳研究会 [1] の報告書でも何回も取り上げられている [たとえば 2,3,4,5]。筆者らが行っている規則方式機械翻訳の結果に統計的自動後編集を加えるハイブリッド方式の機械翻訳システムでも、統計的自動後編集の主な役割は専門用語の訳語の改善にある [6]。

一方、複雑な構文構造を正確に捉え、出力言語の構文構造に適切に変換する方法の一つとして、文パターンの利用がある。特許文は構文が複雑であるが、類似した構文が繰り返して出現するという特徴がある。このような場合、頻出する文パターンを用意しておき、翻訳に利用することで構造的曖昧性解消を適切に行うことができる。実際、[7,8,9,10] などで本手法が提案されており有効性が示されている。

文パターンの利用でまず問題になるのが、頻出する文パターンをどのようにして収集するのかということである。人手で収集する方法はコストがかかるという難点がある。一方、構文解析器を利用して自動または半自動で収集する方法が考えられるが、そもそも構文解析の精度が低いから文パターンを利用するのであり、この方法では本末転倒になってしまう。

日本語の文パターンを抽出する方法として、文末表現に着目する方法がある [11]。日本語は典型的な主要素後置型言語 [2] であるため、文末に着目することで文パターンが抽出できる。例えば「本発明は魚釣用スピニングリールに関し、釣糸の巻取り操作時に於ける糸撚れの蓄積を防止し、併せて釣糸のダメージの軽減を図った魚釣用スピニングリールを提供することを目的とする。」という文の文末に着目することにより「・・・を提供することを目的とする。」という文パターンが得られる。この方法は日本語では有効であるが、主要素後置型言語ではない英語や中国語に適用することはできない。

本文では、中国語の頻出する文パターンを抽出する一方法として、2 段階の方法を提案する。まず、日本語と中国語の文対応コーパスを用意し、その日本語部分の文パターンを文末表現から求める。次に、対応する中国語部分の文に共通する表現に着目して中国語の文パターンを求める。次節では、その具体的方法を示す。

1　http://aamtjapio.com/

2　文は単語や文節などの要素が修飾しあって構成されているが、修飾する要素を従要素、修飾される要素を主要素と呼ぶ。日本語は主要素が従要素より文末側に位置するので「主要素後置型」と呼ばれる。

2 中国語文パターンの抽出

以下のステップで中国語頻出文パターンを抽出する。

a） 日中文対応コーパスを作成する。

b） 上記コーパスの日本語部分から文末表現に着目して日本語文パターンを抽出する。

c） 共通の日本語文パターンを持つ複数の日中文対において、その中国語部分に共通する表現を利用することで中国語文パターンを抽出する。

これらを順に説明する。

2.1 日中文対応コーパスの作成

日中のパテントファミリーを用いて、日中文対応コーパスを作成する。本文で用いる元データはパテントファミリーの要約部分であり、138,065 件の日中文書対応からなる[3]。まず、この文書対応コーパスから文対応コーパスを作成しなければならない。その方法は、各文書の中国語部分を市販の機械翻訳システムを用いて、日本語に翻訳し、この機械翻訳結果と対応する日本語部分とを比較して対応スコアの高い文対を求めるものである。ここで、文 A と文 B の対応スコア $s(A, B)$ を以下のようにして計算する。文 A に含まれるキーワードの集合を K_A とし、文 B に含まれるキーワードの集合を K_B とする。また、一般的に集合 S の要素数を $\#(S)$ とすると、

$$s(A, B) = \frac{2 \times \#(K_A \cap K_B)}{\#(K_A)+\#(K_B)}$$

で対応スコアが計算される。$0 \leq s(A, B) \leq 1$ が成立する。ここで、キーワードとしては、漢字またはカタカナを含む単語とする。対応スコアを求める機械翻訳文の元となった中国語文と対応する日本語文を対にすることで文対応コーパスが得られる。この文対応付けの具体例を付録 1 に示す。

文対応コーパスにおいて対応スコアが 0.5 以上の対応のみを求めて最終的な文対応コーパスとした。得られた文対は 43,404 である。

3 本コーパスは、日本語部分が 453,616 文で構成され、中国語部分が 394,744 文で構成されている。

2.2 日本語文パターンの抽出

得られた文対応コーパスから日本語部分を抽出し、文末表現を用いて日本語文パターンを求めた。その結果、付録 2 に示すようなデータが得られた。元データが要約部分であるため「提供する。」や「特徴とする。」およびそれらの類似パターンが上位にきている。付録 2 に示すパターンだけで 43,404 件のうち 8,851 件（約 20%）をカバーしている。

2.3 中国語文パターンの抽出

次に、日本語の各文パターン毎に中国語部分を調査して共通する表現を調べる。ここでは、紙幅の関係で、付録 2 で最も頻度の多い日本語文パターンとして

・提供することを目的とする。

・提供することを目的としている。

の 2 種のパターンを対象に中国語文パターンを調査した結果について述べる。これらのパターンを含むデータは 440 件であり、全データの約 100 分の 1 である。日本語側でこれらのパターンを含むデータについて、中国語側で「提供」をキー表現として、その前の部分と後の部分を調べた。その結果、前部分に共通性が見られ、105 通りの異なりパターンが得られた。前部分のパターンのうち度数 3 以上のパターンを付録 3 に示す。度数 3 以上のパターンは異なり度数が 14 であり、延べ度数は 283 である。したがって、14 通りのパターンだけで、全件数 440 のうち 64% をカバーしていることが分かる。

中国語のパターンでは「**本发明的目的在于提供**」で始まる文パターンが最も多く、67 件ある。

しかしながら、「**本发明以提供**」で始まるパターンも 5 件有り、この場合には、「**本发明以提供・・・为目的。**」のように、文頭と文末にパターンが分離してしまう。例えば、「**本发明以提供可提高处理量及异物与良质纤维的分离效果的筛分装置为目的。**」（本発明は、特に、処理量及び異物と良質繊維の分別効果を向上させることができるスクリーン装置を提供することを目的としている。）の例がある。

また、「**本发明涉及磁性石榴石单晶和使用该单晶的法拉第转子，提供抑制了晶体缺陷发生的磁性石榴石单晶和提高消光比的法拉第转子。**」（本発明は、磁性ガーネット

単結晶およびそれを用いたファラデー回転子に関し、結晶欠陥の発生を抑えた磁性ガーネット単結晶及び、消光比を向上させたファラデー回転子を提供することを目的とする。）のように「**本发明**」で始まっても「**提供**」との間に「**涉及磁性石榴石单晶和使用该单晶的法拉第转子,**」が挿入され、分離されるパターンもある。

3 まとめと今後の課題

機械翻訳の２大課題である語彙的曖昧性解消と構造的曖昧性解消のうち、後者に関して、その精度を向上できる可能性がある文パターンについて述べた。主要素後置型言語である日本語では、文末表現に着目することで、文パターンが得られる。このことを利用して、日中文対応コーパスから中国語の文パターンを求める方法を提案した。

特に「・・・提供することを目的とする。」という日本語文パターンに着目して、対応する中国語文パターンを調べた。得られた文パターンを用いることで機械翻訳の精度を向上できる可能性がある。例えば「**本发明涉及电阻器及其制造方法，其目的在于提供在向安装基板安装时，在安装面积中焊接面积所占比例减少的电阻器及其制造方法。**」（本発明は、実装基板に実装した際の実装面積を低減できる抵抗器およびその製造方法を提供することを目的とする。）という中国語の文を、ある機械翻訳サイトで翻訳したところ「当発明は抵抗器と製造の方法に関連して、その目的は基の板をインストールするにインストールする時提供することにあって、インストールの面積の中で溶接の面積は割合の減らす抵抗器と製造の方法を占めました。」となり、中国語において「**提供**」より後の部分全体が「**提供**」の目的語であるということが認識できていない。これに対して、「**本发明涉及**【名詞句 $_1$】(,)（**其**）**目的在于提供**【名詞句 $_2$】。⇒本発明は、【名詞句 $_1$】に関し、【名詞句 $_2$】を提供することを目的とする。」という中国語の文パターンとその日本語訳の文パターンを用いることで正確な翻訳ができる可能性がある[4]。

4 【名詞句】は名詞句を現す変数を表し、() は、括弧内の表現が任意選択であることを示す。

本文では、特許文献の中で「要約」部分のみを対象にして考察した。また、日本語側の文パターンも「提供することを目的とする。」のみに着目して考察した。今後の課題としては、以下のようなことが考えられる。

- 特許文献の各段落や分野によって頻出する文パターンが異なる可能性がある。これらを網羅的に調査して文パターン辞書を構築する。
- 文パターン辞書を実際の機械翻訳システムに組み込んで利用する。
- 日本語でも文末表現以外の文パターンが考えられる。例えば「基体粒子上に、金属硫化物、金属フッ化物、金属炭酸塩又は金属燐酸塩からなる薄膜の少なくとも１層を含む多層膜を成膜してなることを特徴とする多層膜被覆粉体。」では、「ことを特徴とする」というパターンが文末ではなく、文の中間に存在する。このような場合でも文パターンを効率的に抽出する方法を考案する。

4 謝辞

本実験で用いた日中対応コーパスは AAMT/Japio 特許翻訳研究会での研究目的で Japio から提供されたものである。使用を許諾していただいた Japio および AAMT/Japio 特許翻訳研究会に感謝する。

本文中の中国語関連部分に関して、Japio の王向莉研究員に校閲していただいた。同研究員に感謝する。

参考文献

[1] 江原暉将、田中穂積：機械翻訳における自然言語処理、情報処理、自然言語処理技術の応用特集号、Vol.34, No.10, pp.1266-1273, Oct., 1993.
[2] 梁冰ほか：対訳特許文を用いた同義対訳専門用語収集における推移的方式の評価、平成 23 年度 AAMT/Japio 特許翻訳研究会報告書、pp.2-7, March, 2012.
[3] 範暁蓉、二宮崇：語学学習サイトウェブページからの対訳語抽出、平成 23 年度 AAMT/Japio 特許翻訳研究会報告書、pp.8-14, March, 2012.
[4] 豊田樹生ほか：日英パテントファミリーにおける対訳文対非抽出部分を利用した専門用語訳語推定、平成 24 年度 AAMT/Japio 特許翻訳研究会報告書、pp.2-9, March, 2013.
[5] 梶博行ほか：複数の 2 言語辞書とコンパラブルコーパスからの多言語辞書の生成、平成 24 年度 AAMT/Japio 特許翻訳研究会報告書、pp.10-17, March, 2013.
[6] 江原暉将：規則方式と統計的後編集を組み合わせた特許文の日英機械翻訳（その 2）、平成 21 年度 AAMT/Japio 特許翻訳研究会報告書、pp.56-60, March, 2010.
[7] 船守茉美：中国公開特許公報の日本語への機械翻訳、特技懇 262 号、pp.3-10, Aug., 2011.
[8] Minah Kim：Current Status of Korea’s Machine Translation for Patent Domain Users, 第 1 回特許情報シンポジウム資料 , Dec., 2010.
[9] Wang Dan：Making Effective Use of Machine Translation for Patent Documents: Practice of CPIC, 第 2 回特許情報シンポジウム資料 , Nov., 2012.
[10] Jin’ichi Murakami et. al：Pattern-Based Statistical Machine Translation for NTCIR-10 PatentMT, Proceedings of the 10th NTCIR Conference, pp.350-355, June, 2013.
[11] 特許庁：日本語特許出願書類の中国語への機械翻訳に関する調査報告書、Feb., 2011.

付録 1 文対応例

下記の文書は日本語部分が 3 文から成り、中国語部分も 3 文から成る。本文記載の手法によって文対応を取ると日本語部分の 1 文目は中国語部分の 1 文目に、日本語部分の 2 文目は中国語部分の 2 文目に、日本語部分の 3 文目は中国語部分の 3 文目に対応する。

日本語部分：ハイブリッド・カーにおける多相交流回転機による駆動装置（内燃機関）へのアシストおよび電池への充電を電池の状態に応じて行えるようにする。車両に搭載された電池の放電時および充電時の電流、電圧を測定し、プログラム制御回路がこの測定された電流、電圧の情報に基づき、電池と車両の駆動装置に連結された多相交流回転機との間の交流直流もしくは直流交流変換を行うインバータを介して駆動装置に補助動力を与えるときの放電電流を制御する。電池の充電状態に応じて充放電電流が制限されるので、電池の充電量の多少にかかわらず一様に充放電が行われることが避けられ、充電効率を高めるとともに、電池の使用寿命を増大することができる。

中国語部分：**本发明能够根据电池的状态来进行由混合式汽车多相交流电机产生的、给驱动装置（内燃机）的加速以及给电池的充电。对车辆上装载的电池其放电时及充电时的电流、电压进行测定，程序控制电路根据此测得的电流、电压信息，通过用于实行电池与连接于车辆驱动装置的多相交流电机之间的交流变直流或直流变交流变换的变换器，对在向驱动装置提供辅助动力时的放电电流进行控制。由于是根据电池的充电状态来限制充放电电流，避免了不论电池充电量的多少而同样实行的充放电，因此在提高充电效率的同时，还能够使电池的寿命延长。**

中国語部分の機械翻訳結果：当発明は電池の状態によって混じる適当な自動車の多相の交流電気機械の発生の、から駆動装置（ディーゼル・エンジン）の加速およびに電池の充電に行うことができる。車の上で積載する電池のその放電する時と充電する時の電流、電圧に対して測定を行って、プログラムの制御回路はこのによって電流、電圧の情報を測って、電池を実行して車の駆動装置の多相の交流電気機械間のにつながって次第に変化する直流に交際するあるいはとずっと次第に変化して変換の変換機を交流することに用いるを通じて（通って）、

駆動装置に動力に協力することに提供する時の放電する電流に対してコントロールを行う。なことが電池の充電する状態によって放電する電流を担当することを制限しにくるため、電池が電気量のいくらを担当して同様に実行したのが担当して放電するのであろうとを免れて、そのため充電する効率を高める同時に、また電池の寿命に延長することができる。

付録 2 頻出する日本語文パターン

（第 1 フィールドは文末表現の文節数、第 2 フィールドは頻度、第 3 フィールドは文末表現を示す）

5|27| 製造方法を提供することを目的とする。
5|21| その製造方法を提供することにある。
4|406| 提供することを目的とする。
4|77| 提供することを課題とする。
4|41| 有することを特徴とする。
4|38| 含むことを特徴とする。
4|34| 提供することを目的としている。
4|31| 製造方法を提供することにある。
4|29| ことを目的とするものである。
4|24| なることを特徴とする。
4|22| 備えることを特徴とする。
4|22| 含有することを特徴とする。
4|20| 形成されていることを特徴とする。
3|628| ことを特徴とする。
3|522| ことを目的とする。
3|210| 提供することにある。
3|162| その製造方法を提供する。
3|109| 行うことができる。
3|104| ことを課題とする。
3|101| 得ることができる。
3|80| ことが可能となる。
3|75| 製造方法を提供すること。
3|72| 防止することができる。
3|68| することができる。
3|64| 図ることができる。
3|62| ことを特徴としている。
3|48| 適用することができる。
3|42| 電子機器を提供すること。
3|41| ことを目的としている。
3|39| 低減することができる。
3|39| 向上させることができる。
3|38| 提供することができる。
3|38| 高めることができる。
3|35| 提供を目的とする。
3|31| 目的とするものである。
3|30| 半導体装置の製造方法を提供する。
3|29| 抑えることができる。
3|29| 実現することができる。
3|28| 抑制することができる。
3|27| 形成することができる。
3|25| 確保することができる。
3|24| ことを可能とする。
3|22| 製造することができる。
3|21| 方法を提供すること。
2|1667| ことができる。
2|970| 提供すること。
2|639| 特徴とする。
2|577| 目的とする。
2|414| 製造方法を提供する。
2|271| ことにある。
2|124| ことができるようにする。
2|123| 提供することである。
2|116| 課題とする。
2|106| 可能となる。
2|105| 方法を提供する。
2|104| ことが好ましい。

付録 3 中国語部分において「提供」より前の部分とその頻度

（NULL は前部分が空、つまり「提供」が文頭にあることを示す。）

67 本发明的目的在于
66 本发明的目的是
66 本发明
20 NULL
15 本发明目的在于
13 本发明的目的在于，
10 本发明的目的是，
5 本发明目的是

5 本发明以
4 可
3 目的是
3 目的在于
3 本发明的目的旨在
3 以