2009－NL－190 (3)
2009／3／25

産業日本語の構想と特許文の言い換え実験[1]

熊野 明*1 安原 宏*2 渡邊 豊英*3
*1(株)東芝 研究開発センター
*2セマンティック・コンピューティング研究開発機構
*3(財)日本特許情報機構

「産業日本語」とは、計算機による自然言語処理への入力を前提とした日本語の仕様で、人間・計算機の双方にとって理解しやすいものを目指している。産業日本語で文書を作成するためには、通常の日本語を産業日本語に言い換えるオーサリングシステムが必要である。非明晰な日本語に対して産業日本語への言い換えを支援するオーサリングシステム向けソフトを開発し、特許明細書の実文を使って実験した。今回、産業日本語オーサリング用実験システムとして、(1)日英機械翻訳システムの日本語解析モジュール、(2)日本語言い換えモジュール、(3)英日機械翻訳システムの日本語生成モジュールを組み合わせて実現した。解析・生成モジュールと言い換えモジュールの間は、CDLと呼ばれる概念記述言語の表現形式で記述したデータで受け渡しを行った。特許文書中の非明晰な文を5種類の言い換え規則を使って行った実験で、産業日本語への言い換えが可能であることを確認した。

"Technical Japanese" Platform; Experiment of Paraphrasing Patent Text

Akira Kumano*1, Hiroshi Yasuhara*2, Toyohide Watanabe*3
*1Corporate Research and Development Center, Toshiba Corp.
*2Institute of Semantic Computing
*3Japan Patent Information Organization

"Technical Japanese" (TJ) is designed for the natural language processing by computers, and it should be understandable for human and computers. In order to paraphrase documents in TJ, we need an authoring system which paraphrases usual Japanese sentences into TJ sentences. The software for the authoring system to paraphrase unclear Japanese sentences into TJ has been developed, and we have experimented using patent documents. The authoring system is realized with (1) Japanese analysis module in J-to-E MT system, (2) Japanese paraphrasing module, and (3) Japanese generation module in E-to-J MT system. The data described in CDL, concept description language, is used to transfer between analysis/generation modules and the Japanese paraphrasing module. The experiment using this software by five kinds of rules has shown that it is possible to paraphrase patent texts into TJ.

1 本研究は、(財)ＪＫＡの機械工業振興事業補助金の交付を受けて行う(財)機械システム振興協会の委託事業により実施したものである。

1. 背景

日本の産業活動は、産業情報の創成・交流・活用のサイクルの上に成り立っている。そして、産業情報は、多様な分野で制作・交換・利用される各種の産業技術文書(産業技術ドキュメント)として具現化される。

産業日本語(Technical Japanese)[1]は、日本の産業技術文書を表現する基幹メディアとなるべきものである。

日本の産業活動の活性化は、日本の産業情報サイクルの活性化に直接裏付けられ、産業情報サイクルは、産業技術文書の制作・交換・利用のサイクルの上に成り立つ。そして、現在、日本の産業情報は、その旧弊を打ち破るべく、以下に掲げる３つの大きな変革が求められている。

① グローバル化に向けて
　グローバルに流通でき、グローバルな情報力を持つ産業情報への変革

② 分野間交流、分野融合に向けて
　産業間、業界間、分野間で流通でき、分野融合を促進できる高い情報価値を持った産業情報への変革

③ 人間の活動と ICT 機能との緊密な連携処理に向けて
　人間による情報活動と ICT による情報処理の間で緊密な連携が達成でき、高度な知的生産性を実現できる産業情報への変革

2. 産業日本語の構想

これら産業情報に求められる旧弊を打破するためには、産業技術文書を表現する日本語に対して旧弊の日本語から新たな産業日本語へと脱皮することが必要である。すなわち、産業日本語として、旧来からの閉じた日本語から以下のような要件を満たす開かれた日本語へと脱皮することが求められることになる。

① 世界に開かれた日本語
　日本国内だけではなく、世界で通用する日本語という視点からに日本語である。あるいは、諸外国語に正確に容易に翻訳できる日本語の用法という視点からの日本語である。

② 分野間に開かれた日本語
　特定の分野に閉じたものではなく、分野共通に用いられ、読み手に対し正確に情報を伝達し得る日本語という視点からの日本語である。あるいは、分野間の日本語用法の違いを説明できる共通の日本語という設定に支えられた日本語という視点からの日本語である。

③ 人間からコンピュータへと開かれた日本語
　人間だけではなく、コンピュータによる情報処理にも用いられる日本語という視点からの日本語である。コンピュータによるマルチメディア処理や知識処理に連携できる日本語という視点からの日本語である。

なお、日本語を大きく技術指向の日本語と文化指向の日本語に分けてみる視点も必要である。技術指向の日本語は、客観的な観点から明晰に論理的に情報を表現するための日本語である。技術指向の日本語の内で、利用期間に永続性があり利用範囲に拡がりのある産業情報を表現するための日本語が産業日本語である。産業日本語は、様々な使用目的ごとに様々な仕様が作られる。その様々な仕様に対する共通の枠組みを規定するのが共通基盤仕様である。産業日本語をモデル化する形式で共通基盤仕様をまとめる。そして、共通基盤仕様から派生する具体的な産業日本語として 4 つの事例を挙げる。特許版産業日本語(特許版 TJ)、日英機械翻訳産業日本語(翻訳 TJ)、文書検索産業日本語(検索 TJ)、図式産業日本語(図式 TJ)の４つである。なお、産業日本語のモデル化を記述するのに CDL[2]、ないしは、CDL の記述形式を用いる。

3. 産業日本語オーサリングシステム

産業日本語オーサリングシステムとは、非明晰な日本語を入力として、計算機との対話を含む処理によって、明晰な日本語、日英機械翻訳産業日本語に変換出力するものである。既存の日英機械翻訳システムの日本語解析部、CDL 言い換えエンジン、英日機械翻訳システムの日本語生成部を利用し、図１に示す構成の産業日本語オーサリングシステム用実験ソフトを開発し、特許の実文を使って動作実験を行った。

基本的な流れは以下の通りである。

(1) 非明晰テキスト(プレーンテキスト)を日英機械翻訳用日本語解析エンジンで解析し、その解析結果である機械翻訳内部表現を出力する。
(2) (1)の出力を内部表現・CDL 変換部で CDL に変換して、CDL.jpn.sf テキスト①として出力する。
(3) CDL.jpn.sf の言い換えエンジンを用いて CDL.jpn.sf テキスト①に対して言い換えを行い、CDL.jpn.sf テキスト②として出力する。
(4) (3)の出力を CDL・内部表現変換部で日本語生成エンジン用機械翻訳内部表現に変換して出力する。
(5) (4)の出力をもとに英日機械翻訳用日本語生成エンジンで日本語テキスト(プレーンテキスト)を生成し、明晰テキストとして出力する。

図２には、日本語文を解析した結果を CDL.jpn.sf テキストで出力した例を示す。

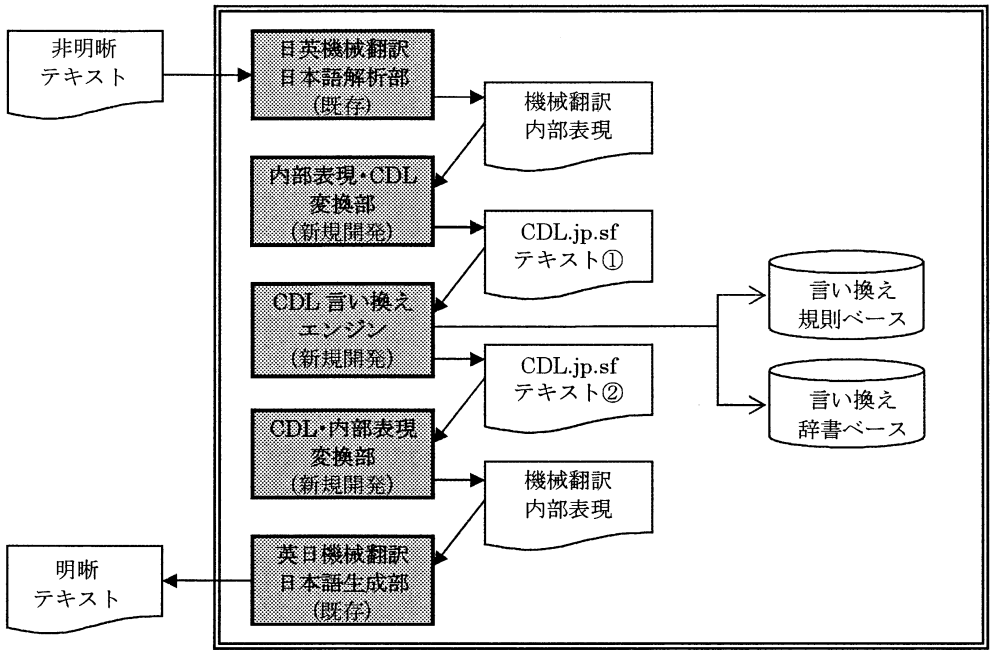

図１．産業日本語オーサリングシステム

(a) 入力

以下の説明では、第１の言語を日本語とする。

(b) 解析結果 (CDL.jpn.sf)

```
{#0 以下    pos=<名詞>      fw=<の>;}
{#1 説明    pos=<サ変名詞>  fw=<では、>;}
{#2 第      pos=<接頭辞>    fw=<>;}
{#3 １      pos=<数字>      fw=<の>;}
{#4 言語    pos=<名詞>      fw=<を>;}
{#5 日本語  pos=<名詞>      fw=<と>;}
{#6 する    pos=<動詞> inf=<動さ> fw=<。>;}
[#0      ノ格      #1]
[#1      デ格      #6]
[#2      連体      #3]
[#3      連体      #4]
[#4      ヲ格      #6]
[#5      ト格      #6]
```

図 2. CDL.jpn.sf の例

4. 言い換え処理

言い換え処理システムは、図 1 に示すように日本語解析処理の出力である CDL.jpn.sf を入力し、それを言い換えエンジンが新たな CDL.jpn.sf に変換して日本語を改善する。その結果を日本語生成処理に渡す。言い換え変換に際して言い換え規則を用いる。

4.1. 言い換え規則の表現

言い換え規則は、規則ベースと辞書ベースに分かれる。規則ベースは、構文的な情報を利用した言い換え変換の集合であり、辞書ベースは語彙的な情報による言い換え変換の集合である。

規則の記述は図 3.に示すように、＜自然言語による記述＞、＜CDL.jpn.sf 構造を想定した処理記述＞、＜実行関数＞の 3 種類の記述を併記すること(前 2 者は省略可能)とした。まず、＜自然言語による記述＞は、形式的な言語表現に通じていない言語専門家でも解読できるようにしておく。これは規則のメンテナンスで重要な情報源となる。＜CDL.jpn.sf 構造を想定した処理記述＞は、自然言語記述を形式化した記述で、言い換え処理を関数実装するときに CDL グラフを処理するときの方法を設計情報として記述しておく。これを元に＜実行関数＞を記述する。

```
{#pproot 言い換え規則;
  {#pproot1  規則ベース;
      {#r<pattern>-<number> <規則名>;
      {<自然言語による記述>;}
      {<CDL.jpn.sf 構造を想定した処理記述>;}
      {<実行関数>;}} ...
  }
  {#pproot2  辞書ベース;
      {#d<number>  <見出し語>;
      {<言い換えタイプ>;}
      {<タイプ毎の言い換えデータ>;}} ...
  }
}
```

図 3. 言い換え規則のスキーマ表現

言い換え規則は、表 1 に示すように大きく 5 つのクラスに分類した。その中でクラス 1、2、3 は

単純な規則ではなく、種々の条件下で変化するもので規則ベースとした。上述したように規則は形式的な記述に限定することはしていない。クラス4、5は変換パターンが単純であるので4.2に示すように変換方法を直接辞書ベースに記述することとした。

表1. 言い換え分類と処理方式

言い換えクラス	処理方法	CDL上での処理
1．連用節分割 (節間分割)	文節数が一定以上のときに連用節を分離する。 基本的処理は、用言で切り取り、終止形にする。用言付属の接続助詞等によって、分割後の文に接続詞を付与する。	連用修飾のアークを削除し、分割した節を文形式のCDLにする。新たな文IDには、元の文IDにサブ番号を付加する。(主節と従属節の順序や結束表現等は今後の課題である。)
2．連体節分割 (節間分割)	文節数が一定以上のときに連体節を切り出す。 被修飾された体言は切り出された節の格成分に入れる。元の位置の体言には「その」等の指示詞を追加する。連体修飾のパターンは非制限的用法/制限的用法の分類があるがここでは非制限的なものとする。	注目する体言をコピーし、連体節に埋め込む。その際の関係概念名は係り受け関係で記載されているときはそれを使用し、記載されていないときはデフォルトとして「ハ格」とする。注目体言には、「その」または「この」等を結合させる。(連体節の分割は、主節と従属節の順序や指示詞の指定、結束表現等、課題が多い。)
3．共起語句変換 (節内変換)	「コミュニケーションを取る→コミュニケーションする」、「AをBとする→BはAである」、「Nしか、Vしない→Nだけ Vする」など、文内の単語や句の構文的な共起関係に基づいて言い換える。	規則のキーワードとなる関係概念名等をトリガーにしてグラフ変換及び被修飾ノードの処理を行う。
4．単独語/連語↔句節 (語変換)	難解語、複合語や臨時一語を、句や節にして言い換える。 一般的には、名詞は名詞句に置き換える。サ変名詞は「する」が付くか否かで節あるいは句に置き換える。難解語は「(説明文)」のように補足説明として挿入する。	対象の句の CDL.jpn.sf 表現が、辞書ベースに記述されている。 原文中の該当する語ノードを置き換える。
5．形態素レベル (語変換)	冗長語の削除等、構文的な変換が必要でないもので、辞書に登録されている通りに置換したり、削除する。	対象の語の CDL.jpn.sf の置換(削除)パターンが、辞書ベースに記述されている。 原文中の該当する語ノードを置換(削除)する。

4.2. 言い換え規則の辞書ベースの例

① 難解語

フィールド数：2

フィールド名：hw(見出し語), def

例：

hw	def
合決	(貼り合わせる板の厚さをそれぞれ半分ずつ欠きとること)

② 複合語

フィールド数：3

フィールド名：hw, pnum(分割数), cdl

例：

hw	pnum	cdl
高圧力	2	「高い圧力」の CDL.jpn.sf 表現： {#0　高 pos=<形容詞>　fw=<い> ;} {#1　圧力 pos=<名詞>　fw=<> ;} [#0　連体　#1]

③ 臨時一語

フィールド数 ：4

フィールド名 ：hw, wnum(連語数), word_list(連接する語), cdl

例：

hw	wnum	word_list	cdl
冷媒	3	「減圧」「時」	「冷媒を減圧する時」の CDL.jpn.sf 表現: {#9000 冷媒 pos=<名詞> fw=<を> ;} {#9001 減圧 pos=<サ変名詞> fw=<した>;} {#9002 時 pos=<名詞> fw=<> ;} [#9000 ヲ格 #9001] [#9001 連体 #9002]

5. 実験

実験では、実際の特許明細書 2 件の約 100 文を解析し、言い換え処理が必要と判断した文に対して言い換え処理を行った。その言い換え処理は、表 1 に示した以下の 5 種類である。

(1) 連用節分割（節間分割）
(2) 連体節分割（節間分割）
(3) 共起語句変換（節内変換）
(4) 単独語/連語⇔句節（語変換）
(5) 形態素レベル（語変換）

以下に、各言い換え処理の代表的な例を挙げ、(a)入力文、(b)出力文、(c)結果のデータで示す。

5.1. 連用節分割（節間分割）

(a) 入力

語アライメント方式の翻訳モデルの生成では、ソース文に含まれる単語の集合の各々について個別に翻訳語を生成してターゲット単語の集合を生成し、さらにそれらターゲット単語の、翻訳文内での位置を決定する事により翻訳を行なう、という戦略を採っている。

(b) 出力

語アライメント方式の翻訳モデルの生成では、ソース文に含まれる単語の集合の各々について個別に翻訳語を生成しターゲット単語の集合を生成する。
語アライメント方式の翻訳モデルの生成では、それらターゲット単語の翻訳文内の位置を決定する事により翻訳するいう戦略を採る。

(c) 結果

長文を連用節で分割することにより、構造の明確な日本語文を出力することができた。

一連の内容であることを示すために、分割した第 2 文にも、「語アライメント方式の翻訳モデルの生成では」を補っている。

生成文法が不十分なために、一部不自然な表現を出力している。

5.2. 連体節分割（節間分割）

(a) 入力

統計的機械翻訳では、第 1 の言語の文と第 2 の言語の文との多数の対訳文を含む対訳コーパスを用いた学習により予め翻訳モデルを作成しておき、この翻訳モデルを用いて翻訳を行なう。

(b) 出力

学習が第 1 言語の文と第 2 言語の文の多数の対訳文を含む対訳コーパスを利用した。
統計的機械翻訳では、予め学習することで翻訳モデルを作成しこの翻訳モデル用いて翻訳する。

(c) 結果

長い名詞句の連体修飾部分を独立文として分割することにより、構造の明確な日本語文を出力することができた。

5.3. 共起語句変換（節内変換）

(a) 入力

情報手段の進歩により、海外の人々と外国語でコミュニケーションを取る機会が増えている。

(b) 出力

情報手段が進歩することで人々と海外の外国語でコミュニケーションする機会が増えている。

(c) 結果

「コミュニケーションを取る」という句を「コミュニケーションする」という簡潔な表現に言い換え、機械処理の容易な日本語文を出力することができた。

5.4. 共起語句変換（節内変換）

(a) 入力

以下の説明では、第 1 の言語を日本語とする。

(b) 出力

以下の説明では、第 1 言語が日本語である。

(c) 結果

「第 1 の言語を日本語とする」という書き言葉的な句を「第 1 の言語が日本語である」という明晰な表現に言い換え、機械処理の容易な日本語文を出力することができた。

5.5. 単独語/連語⇔句節（語変換）

(a) 入力

暗渠長手方向に相隣る暗渠用ブロック 10 同士を互いに合決で接続する。

(b) 出力

互いに合決（貼り合わせる板の厚さをそれぞれ半分ずつ欠きとること）で相隣る暗渠用ブロック 10 同士を暗渠長手方向に接続する。

(c) 結果

難解語「合決」に対して語義を示す分を補い、人間にとって理解容易な日本語文を出力することができた。

5.6. 単独語/連語⇔句節（語変換）

(a) 入力

冷媒減圧時の冷媒流動音が室内に伝播する現象が著しい。

(b) 出力

冷媒を減圧した時の冷媒流動音が室内に伝播する現象が著しい。

(c) 結果

臨時一時語「冷媒減圧時」に対して意味を明確にした言い換えを行い、機械にとって理解容易な日本語文を出力することができた。

5.7. 形態素レベル（語変換）

(a) 入力

比較的簡単な構成でかつ簡単な操作にて蓋の開閉操作を行い得る様に構成する事を目的とするものである。

(b) 出力

比較的簡単な構成でかつ簡単な操作で蓋の開閉操作を行う得る様構成する事を目的とする。

(c) 結果

意味的には不要な「するものである」という表現を簡潔な表現に言い換え、機械にとって理解容易な日本語文を出力することができた。

6. まとめ

実験では、特許の実文に対する 5 種類の言い換え規則によって、概ね正しい日本語表層文を出力することができた。産業日本語オーサリングシステムの基本機能が確認できた。

また、1 文ずつのオーサリングとは別に、テキストファイルをバッチ的に言い換え処理するための実験を、ソフト系の特許文 50 文と機械系の特許文 50 文を対象にして長文分割を実施した。いずれの場合も、言い換え後の出力は言い換える前の文に対して、明晰な表現であるといえる。

今回の実験では、1 文に対して自動的に 1 種類の言い換えしか行わなかったが、複数の言い換えが可能な場合もある。

(1) 言い換えの可否

言い換えを行うか否かをユーザが対話的に指定する機能も有効である

(2) 複数言い換えの選択

複数種類の言い換えが可能な文に対して考慮すべきである

(3) 言い換えの繰り返し

一度言い換えた結果に対して、さらに他の言い換えが可能な場合、自動的/半自動的に 2 つ目の言い換えを行う

産業日本語オーサリングシステムは、システムとユーザとの対話によって明晰な日本語を作り上げるものである。したがって、産業日本語オーサリングシステムの実現には、上述したような場合を考慮して、有効なユーザインタフェースを設計する必要がある。

7. 今後の課題

言い換え出力には、一部の単語の語尾活用が正しくない事例があった。これは、日本語解析モジュールで解析されたデータに含まれる属性(進行、受身、など)の表現が、日本語生成モジュールで利用するデータに期待する属性の表現と一致しないものがあったからである。これはデータ変換時に属性名・属性値を適切に変換し、解析モジュールと生成モジュールで属性を共有することにより解決するものである。

また、言い換えた部分以外で、語順が変わってしまうものもあった。これは、CDL.jpn.sf が、表層の語順を反映したデータであるのに対して、生成モジュールはその語順を利用しないで文法の記述に従って語順を決めるからである。これは、生成モジュールの処理方針にかかわる問題である。原文の語順や格助詞を最大限再現することが求められるなら、解析モジュール・生成モジュールと、言い換えモジュールのインタフェースを変更する必要が生じる。

今回は解析モジュールで係り受け解析まで行う中で、一部意味解析に及ぶ処理も含んでいる。これは、解析モジュールが日英機械翻訳用に設計・構成されたものであるからである。正しい英訳を出力するために、早期の段階で一部意味的な処理を行っている。原文の語順や格助詞を最大限再現するなら、少し浅い解析処理に抑え、意味的な変換処理を省いた解析結果を CDL.jpn.sf に出力するべきである。この場合、生成モジュールも表層表現を最大限利用する機能を追加する必要がある。

いずれも、今後の産業日本語オーサリングシステムの本格開発、実用化、運用を目指して、検討していく。

参考文献

[1] 産業日本語特集Japio誌ネット座談会：特許版産業日本語の取り組みと期待－特許情報を高価値化する産業日本語－, Japio 2008 YEARBOOK, 財団法人日本特許情報機構,2008, <http://www.japio.or.jp/00yearbook/yearbook2008.html>

[2] 石塚 満、自然言語テキストの共通的概念記述、人工知能学会誌 Vol.21 No.6 (2006.11)