産業日本語オーサリングに向けた特許文の言い換え

株式会社東芝　研究開発センター
知識メディアラボラトリー主任研究員
熊野　明

PROFILE
1982 年東京工業大学工学部卒業。同年東京芝浦電気（株）（現、（株）東芝）入社。現在、研究開発センター知識メディアラボラトリー主任研究員。情報処理学会、人工知能学会、言語処理学会会員。2007 年度から特許版産業日本語委員会委員。
akira.kumano@toshiba.co.jp　TEL 044-549-2239

1　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、計算機との対話を含む処理を利用して、非明晰な日本語を明晰な日本語、日英機械翻訳産業日本語に変換して出力するツールである。

産業日本語オーサリングシステムの基本機能を確認するため、日英機械翻訳システムの日本語解析エンジン、日本語の計算機用表現である CDL[1] の言い換えエンジン、英日機械翻訳システムの日本語生成エンジンを利用し、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た。さらに、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。

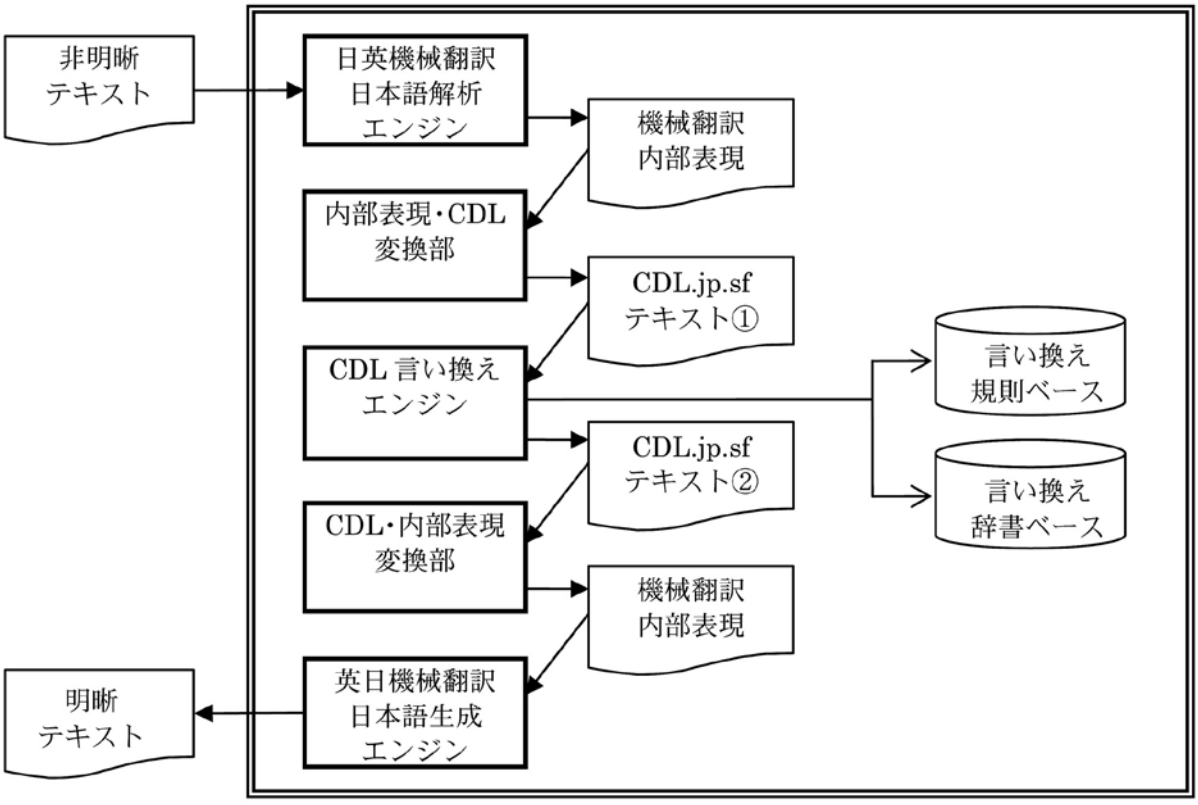

図 1　産業日本語オーサリングシステム

(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 から図 5 には、産業日本語オーサリングシステムによって言い換え処理を行うデータ例を、その処理の順に示す。

図 2 は、入力の日本語テキスト例

「**以下の説明では、第 1 の言語を日本語とする。**」

に対する、解析結果の機械翻訳内部表現の例である。

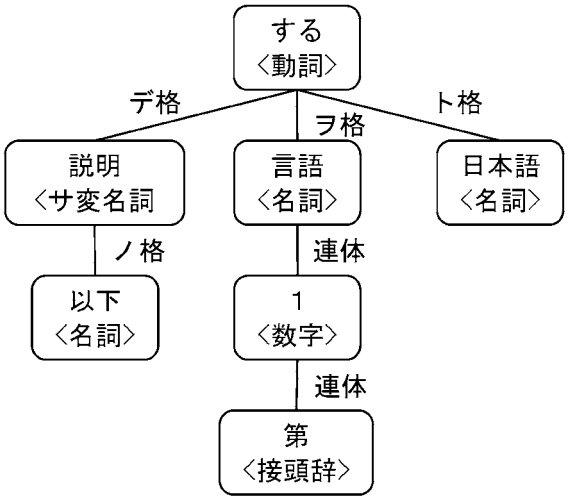

図 2　解析結果の内部表現の例

図 3 は、図 2 の解析結果を CDL.jpn.sf テキストで出力した例である。

CDL.jpn.sf は、日本語文の構造をノード・アーク構造で表現したものである。前半の { } で囲った部分は各ノードの情報で、ノード番号、見出し語、品詞 (pos)、活用種類 (inf)、機能語 (fw) からなる。後半の [ ] で囲った部分はノード間の関係を表すアークの情報で、ノード番号 1、関係、ノード番号 2 からなる。

```
{#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 3　解析結果の CDL.jpn.sf 出力例

図 4 は、図 3 の CDL.jpn.sf テキストを CDL 言い換えエンジンで言い換えた結果を CDL.jpn.sf テキストで出力した例である。

```
{#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=<は>;}
{#5 日本語 pos=<名詞>   fw=<>;}
{#6 である pos=<助動詞> fw=<。>;}
[#0     ノ格     #1]
[#1     デ格     #6]
[#2     連体     #3]
[#3     連体     #4]
[#4     ガ格     #6]
[#5     連格     #6]
```

図 4　言い換え結果の CDL.jpn.sf 出力例

図 5 は、図 4 の言い換え結果を変換出力した機械翻訳内部表現の例である。

図 5 の内部表現から英日機械翻訳用日本語生成エンジンで生成処理を行い、

「**以下の説明では、第 1 言語が日本語である。**」

という言い換え結果を出力する。

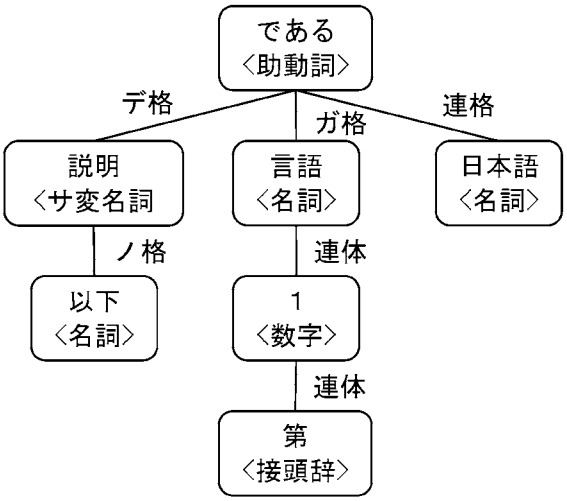

図５　言い換え結果の内部表現の例

2　言い換え処理

言い換え処理エンジンは、CDL.jpn.sf を入力し、それを新たな CDL.jpn.sf に変換して出力するが、この変換は、産業日本語用言い換え規則を用いて行う。

2.1　言い換え規則の表現

言い換え規則には、規則ベースに記述したものと、辞書ベースに記述したものがある。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。

言い換え規則の表現形式を、図６に示す。

```
{#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６　言い換え規則の表現

規則ベースの言い換え規則表現は、＜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。

＜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にも規則の意味が理解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である。＜ 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。

辞書ベースの言い換え規則表現は、＜見出し語＞、＜言い換えタイプ＞、＜タイプ毎の言い換えデータ＞の３種類の記述からなる。

＜見出し語＞は、言い換えを必要とする表現に含まれている代表語である。＜言い換えタイプ＞は、言い換えの種類を表すもので、難解語、複合語、臨時一語、などがある。＜タイプ毎の言い換えデータ＞には、実際に必要な言い換え表現などを記述する。

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あり、規則ベースとした。クラス４、５は変換パターンが単純であるので、変換方法を直接辞書ベースに記述することとした。

2.2　辞書ベースの言い換え規則記述表現

辞書ベースの言い換え規則の例を示す。

① 難解語

例：

見出し語	言い換え表現
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

例：

見出し語	言い換え後の構造
高圧力	「高い圧力」の CDL.jpn.sf 表現： {#0　高 < 形容詞 > +< い >;} {#1　圧力 < 名詞 >;} [#0　連体　#1]

③臨時一語

例：

見出し語	連接語	言い換え後の構造
冷媒	「減圧」「時」	「冷媒を減圧した時」の CDL.jpn.sf 表現： {#901 冷媒 < 名詞 >+< を >;} {#902 減圧 < サ変 >+< した >;} {#903 時 < 名詞 >;} [#901 ヲ格　#902] [#902 連体　#903]

3 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。

3.1 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。

表 1　言い換え分類と処理方式

	クラス	変換の種類	処理方法
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。 文途中の用言で分割し、終止形にする。用言付属の接続助詞等によっては、分割後の文に接続詞を付与する。
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。 連体修飾を受け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と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。
3	共起語句変換	節内分割	「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。
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」が付く場合は節に、付かない場合は句に置き換える。 難解語は具体的な説明文を挿入句のように補足説明として挿入する。
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、一部を削除する。

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。

日本語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。

3.2　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。

3.3　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、計算機処理の容易な日本語文を出力することができた。

3.4　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（あいじゃくり）」に対して語義を示す文を補い、専門家以外の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。

3.5　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、計算機にとって理解容易な日本語文を出力することができた。

3.6　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、計算機にとって理解容易な日本語文を出力することができた。

4 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、テキストファイルの文全体を連続的に言い換え処理する実験を、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判断した。

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。実用的なオーサリングシステムを実現するためには、次のようなことを考慮する必要がある。

(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 / 半自動的に次の言い換えを行う

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。

5 今後の課題

言い換え出力には、言い換えた部分以外で、語順が変わってしまうものがあった。これは、今回利用した日本語解析エンジンと日本語生成エンジンの性質によるものである。

日本語解析エンジンでの係り受け解析には、一部意味解析に及ぶ処理も含んでいる。これは、日本語解析エンジンが日英機械翻訳用に設計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、早期の段階で意味的な処理を行っている場合があるため、表層の語順情報を復元できない場合がある。

また、言い換えエンジンが出力する 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、英日機械翻訳用日本語生成エンジンはその語順を利用せず、生成用文法のみに従って語順を決定する。

入力文の語順や格助詞を再現することが求められるなら、日本語解析エンジン・日本語生成エンジンと、CDL 言い換えエンジンのインタフェースを変更する必要が生じるだろう。

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、これらの課題を検討していく。

参考文献

[1] 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)

本研究は、平成 20 年度に (財) 機械システム振興協会が (財) ＪＫＡの競輪補助金の交付を受けて (財) 日本特許情報機構に委託した財源をもとに実施したものである。