

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4317990号  
(P4317990)

(45) 発行日 平成21年8月19日(2009.8.19)

(24) 登録日 平成21年6月5日(2009.6.5)

(51) Int.Cl. F 1  
**G 0 6 F 17/28 (2006.01)** G 0 6 F 17/28 Z

請求項の数 10 (全 17 頁)

<p>(21) 出願番号 特願2005-69816 (P2005-69816)                  (22) 出願日 平成17年3月11日 (2005.3.11)                  (65) 公開番号 特開2006-252323 (P2006-252323A)                  (43) 公開日 平成18年9月21日 (2006.9.21)                  審査請求日 平成20年2月19日 (2008.2.19)</p>	<p>(73) 特許権者 301022471                  独立行政法人情報通信研究機構                  東京都小金井市貫井北町4-2-1                  (74) 代理人 100130111                  弁理士 新保 斉                  (72) 発明者 内元 清貴                  東京都小金井市貫井北町4-2-1 独立                  行政法人情報通信研究機構内                  (72) 発明者 井佐原 均                  東京都小金井市貫井北町4-2-1 独立                  行政法人情報通信研究機構内                  審査官 成瀬 博之</p>
--	---

最終頁に続く

(54) 【発明の名称】 データ変換適性評価方法及びデータ変換装置

(57) 【特許請求の範囲】

【請求項1】

第1データから第2データへの機械翻訳を行うデータ変換手段と、第2データから第1データへのデータ逆変換手段とが併存するデータ変換装置を用いて、該第1データに対して、データ変換手段における変換適性を評価して変換適性値を算出するデータ変換適性評価方法であって、

該データ変換手段が、第1言語の参照テキストである第1データを機械翻訳して第2言語のテキストである変換後第2データを取得するデータ変換ステップ、

該データ逆変換手段が、翻訳結果である変換後第2データを第1言語の折り返し翻訳テキストに機械翻訳して逆変換後第1データを取得するデータ逆変換ステップ、

類似度算出手段が、該第1データと該逆変換後第1データとを取得し、逆変換後第1データのテキストにおける依存構造木に基づいて、文節内の単語はすべて隣に係り、係り文節における末尾の形態素は受け文節の先頭の形態素に係るとしたときの単語n-gramをパラメータに用いて類似度を算出する類似度算出ステップ、

出力手段が、該類似度を第1データのデータ変換手段における変換適性値として出力する変換適性値出力ステップ

を含むことを特徴とするデータ変換適性評価方法。

【請求項2】

第1データから第2データへの機械翻訳を行うデータ変換手段と、第2データから第1データへのデータ逆変換手段とが併存するデータ変換装置を用いて、該第1データに対し

10

20

て、データ変換手段における変換適性を評価して変換適性値を算出するデータ変換適性評価方法であって、

該データ変換手段が、第1言語の参照テキストである第1データを機械翻訳して第2言語のテキストである変換後第2データを取得するデータ変換ステップ、

該データ逆変換手段が、翻訳結果である変換後第2データを第1言語の折り返し翻訳テキストに機械翻訳して逆変換後第1データを取得するデータ逆変換ステップ、

類似度算出手段が、該第1言語の参照テキストに対する該折り返し翻訳テキストの類似度を測るパラメータと、該折り返し翻訳テキストに対する該参照テキストの類似度を測るパラメータとの調和平均を算出する類似度算出ステップ、

出力手段が、該類似度を第1データのデータ変換手段における変換適性値として出力する変換適性値出力ステップ

10

を含むことを特徴とするデータ変換適性評価方法。

#### 【請求項3】

第1データから第2データへの機械翻訳を行うデータ変換手段と、第2データから第1データへのデータ逆変換手段とが併存するデータ変換装置を用いて、該第1データに対して、データ変換手段における変換適性を評価して変換適性値を算出するデータ変換適性評価方法であって、

該データ変換手段が、第1言語の参照テキストである第1データを機械翻訳して第2言語のテキストである変換後第2データを取得するデータ変換ステップ、

該データ逆変換手段が、翻訳結果である変換後第2データを第1言語の折り返し翻訳テキストに機械翻訳して逆変換後第1データを取得するデータ逆変換ステップ、

20

類似度算出手段が、該第1言語の参照テキストの依存構造木を解析処理し、抽出された部分木毎に類似度を算出する類似度算出ステップ、

出力手段が、各部分木における類似度からテキスト全体の類似度が最大となる部分木集合を求めて、その類似度を第1データのデータ変換手段における変換適性値として出力する変換適性値出力ステップ

を含むことを特徴とするデータ変換適性評価方法。

#### 【請求項4】

前記データ変換適性評価方法において、

前記変換適性値出力ステップの後に、出力手段が、

30

前記テキスト全体の類似度が最大となる部分木集合の中で、類似度が最小の部分木又は所定の閾値よりも小さな類似度の部分木の少なくともいずれかを抽出し、

抽出された部分木を機械翻訳不適箇所として出力する

ことを特徴とする請求項3に記載のデータ変換適性評価方法。

#### 【請求項5】

前記データ変換適性評価方法の類似度算出ステップにおいて、前記類似度算出手段が、少なくとも前記第1言語の参照テキスト又は前記折り返し翻訳テキストのいずれかにおける単語又は単語列を、単語又は単語列を語義及び品詞により階層に分類した所定の単語クラスの分類テーブルに基づき、より上位の階層の語義又は品詞に汎化した後に、所定の類似度算出式による算出を行う

40

ことを特徴とする請求項1ないし4のいずれかに記載のデータ変換適性評価方法。

#### 【請求項6】

第1データから第2データへの機械翻訳を行うデータ変換手段と、第2データから第1データへのデータ逆変換手段とが併存するデータ変換装置であって、該第1データに対して、データ変換手段における変換適性を評価して変換適性値を算出するデータ変換装置において、

第1言語の参照テキストである第1データを機械翻訳して第2言語のテキストである変換後第2データを取得するデータ変換手段と、

該変換後第2データを、第1言語の折り返し翻訳テキストに機械翻訳して逆変換後第1データを取得するデータ逆変換手段と、

50

該第 1 データと該逆変換後第 1 データとを入力して、逆変換後第 1 データのテキストにおける依存構造木に基づいて、文節内の単語はすべて隣に係り、係り文節における末尾の形態素は受け文節の先頭の形態素に係るとしたときの単語 n-gram をパラメータに用いて類似度を算出する類似度算出手段と、

該類似度を第 1 データのデータ変換手段における変換適性値として出力する出力手段とを備えたことを特徴とするデータ変換装置。

【請求項 7】

第 1 データから第 2 データへの機械翻訳を行うデータ変換手段と、第 2 データから第 1 データへのデータ逆変換手段とが併存するデータ変換装置であって、該第 1 データに対して、データ変換手段における変換適性を評価して変換適性値を算出するデータ変換装置において、

第 1 言語の参照テキストである第 1 データを機械翻訳して第 2 言語のテキストである変換後第 2 データを取得するデータ変換手段と、

該変換後第 2 データを、第 1 言語の折り返し翻訳テキストに機械翻訳して逆変換後第 1 データを取得するデータ逆変換手段と、

該第 1 言語の参照テキストに対する該折り返し翻訳テキストの類似度を測るパラメータと、該折り返し翻訳テキストに対する該参照テキストの類似度を測るパラメータとの調和平均を算出する類似度算出手段と、

該類似度を第 1 データのデータ変換手段における変換適性値として出力する出力手段とを備えたことを特徴とするデータ変換装置。

【請求項 8】

第 1 データから第 2 データへの機械翻訳を行うデータ変換手段と、第 2 データから第 1 データへのデータ逆変換手段とが併存するデータ変換装置であって、該第 1 データに対して、データ変換手段における変換適性を評価して変換適性値を算出するデータ変換装置において、

第 1 言語の参照テキストである第 1 データを機械翻訳して第 2 言語のテキストである変換後第 2 データを取得するデータ変換手段と、

該変換後第 2 データを、第 1 言語の折り返し翻訳テキストに機械翻訳して逆変換後第 1 データを取得するデータ逆変換手段と、

前記第 1 言語の参照テキストの依存構造木を解析処理し、抽出された部分木毎に前記類似度算出を行う類似度算出手段と、

各部分木における類似度からテキスト全体の類似度が最大となる部分木集合を求めて、その類似度を第 1 データのデータ変換手段における変換適性値として出力する出力手段とを備えたことを特徴とするデータ変換装置。

【請求項 9】

前記出力手段が、

各部分木における類似度からテキスト全体の類似度が最大となる時の部分木集合の中で、類似度が最小の部分木又は所定の閾値よりも小さな類似度の部分木の少なくともいずれかを抽出し、

抽出された部分木を機械翻訳不適箇所として出力する

ことを特徴とする請求項 8 に記載のデータ変換装置。

【請求項 10】

前記類似度算出手段が、

少なくとも前記第 1 言語の参照テキスト又は前記折り返し翻訳テキストのいずれかにおける単語又は単語列を、単語又は単語列を語義及び品詞により階層に分類した所定の単語クラスの分類テーブルに基づき、より上位の階層の語義又は品詞に汎化した後に、所定の類似度算出式による算出を行う

ことを特徴とする請求項 6 ないし 9 のいずれかに記載のデータ変換装置。

【発明の詳細な説明】

【技術分野】

10

20

30

40

50

## 【0001】

本発明はコンピュータにおけるデータ処理方法に関し、より詳しくは機械翻訳等のデータ変換時に適切なデータ変換が行えるか否かを評価する変換適性値の算出方法に係る。

## 【背景技術】

## 【0002】

機械学習技術の進歩に伴って、例えば機械翻訳のようにコンピュータを用いてあるデータから異なるデータに変換処理することが広く行われている。このように一義的にデータを変換するのではなく、コンピュータが機械学習などに基づいて変換処理する場合には、その変換精度が問題となる。

機械翻訳の場合には機械翻訳後のテキストを手によって確認し、自然な言語テキストが出力されているかを判断していた。

10

## 【0003】

従来の機械翻訳精度を向上させる技術として例えば特許文献1又は特許文献2に開示される方法が知られている。

特許文献1の技術は1つの翻訳元言語で書かれた入力文から複数の翻訳システムにより翻訳先言語で書かれた翻訳結果を出力する。そして、相互に意味解析を行って翻訳結果を比較するものである。

このように意味解析などによって実質的に翻訳結果が妥当であるかを判定する方法は従来から用いられている。

しかし、意味解析を行うのも機械翻訳と同様に構成された解析モジュールであるため、適切な解析が行えない場合も多い。

20

## 【0004】

【特許文献1】特開2004-318344号公報

## 【0005】

特許文献2の技術は、あらかじめ文字列パターンと翻訳パターンを定義しておきそれぞれのパターンと条件とに合致するか否かをチェックすることで、原文により忠実な翻訳パターンを選択して高精度化を図ろうとするものである。

このように、パターンへの当てはめによって翻訳の妥当性を判断する方法も従来から用いられている。

しかし、パターン化された翻訳については高精度化が期待できるものの、全ての文に対応することは難しく、判断するための構成も複雑になってしまう問題がある。

30

## 【0006】

【特許文献2】特開2005-4402号公報

## 【0007】

機械翻訳に限らず、構文解析などの解析結果の評価や、音声合成におけるテキストデータから音声波形データへの変換結果の評価など、コンピュータを用いてデータ変換する場合にこれを評価する好適な方法が求められている。

この方法の実現に先立ち、あらかじめどのような入力データが変換処理に適しているかが判定できれば、変換処理に適した形に入力データを変形したり、変換結果の確信度に疑いが強いことを使用者に提示することが可能である。

40

## 【0008】

従って、完璧な変換技術が提供されていないデータ変換においては、上記のような変換適性を自動的に評価する方法が提供されれば、変換技術の実用性の向上や、変換精度の向上にも寄与させることができる。

## 【発明の開示】

## 【発明が解決しようとする課題】

## 【0009】

本発明は、上記従来技術の有する問題点に鑑みて創出されたものであり、第1データから第2データへのデータ変換及び、その逆方向のデータ変換が可能なデータ変換時に、変換適性を自動的に評価する技術を提供することを目的とする。

50

## 【課題を解決するための手段】

## 【0010】

本発明は、上記の課題を解決するために、次のようなデータ変換適性評価方法を提供する。

すなわち、請求項1に記載の発明は、第1データから第2データへの機械翻訳を行うデータ変換手段と、第2データから第1データへのデータ逆変換手段とが併存するデータ変換装置を用いて、該第1データに対して、データ変換手段における変換適性を評価して変換適性値を算出するデータ変換適性評価方法を提供するものである。

## 【0011】

該方法において、データ変換手段が、第1言語の参照テキストである第1データを変換して機械翻訳して第2言語のテキストである変換後第2データを取得するデータ変換ステップ、データ逆変換手段が、翻訳結果である該変換後第2データを第1言語の折り返し翻訳テキストに機械翻訳して逆変換後第1データを取得するデータ逆変換ステップを有する。

10

## 【0012】

この結果得られた第1データと逆変換後第1データとを類似度算出手段が取得して、所定の類似度算出式により類似度を算出する類似度算出ステップ、出力手段が、類似度を第1データのデータ変換手段における変換適性値として出力する変換適性値出力ステップとを含むことにより構成する。

## 【0013】

本発明は、上記類似度算出ステップにおいて、逆変換後第1データのテキストにおける依存構造木に基づく単語n-gramをパラメータに用いて類似度を算出することを特徴とする。すなわち、従来のように出力文と模範訳をフラットな単語列における単語n-gramに基づいて計算するだけでなく、依存構造木に基づいて、文節内の単語はすべて隣に係り、係り文節における末尾の形態素は受け文節の先頭の形態素に係るとして単語n-gramを数え、類似度を算出する。これにより語順が比較的自由な言語における語順の異なりに対しても有効に類似度の算出を行う。

20

## 【0014】

請求項2に記載の発明は、請求項1記載のデータ変換適性評価方法と同様に、データ変換手段が、第1言語の参照テキストである第1データを変換して機械翻訳して第2言語のテキストである変換後第2データを取得するデータ変換ステップ、データ逆変換手段が、翻訳結果である該変換後第2データを第1言語の折り返し翻訳テキストに機械翻訳して逆変換後第1データを取得するデータ逆変換ステップを有する。そして、第1データと逆変換後第1データとを類似度算出手段が取得して、所定の類似度算出式により類似度を算出する類似度算出ステップ、出力手段が、類似度を第1データのデータ変換手段における変換適性値として出力する変換適性値出力ステップとを含むことにより構成する。

30

## 【0015】

上記類似度算出ステップにおいて、第1言語の参照テキストに対する折り返し翻訳テキストの類似度を測るパラメータと共に、折り返し翻訳テキストに対する参照テキストの類似度を測るパラメータとの調和平均を算出することを特徴とする。これにより双方向の類似度を考慮する技術を提供する。

40

## 【0016】

請求項3に記載の発明は、上記と共通するデータ変換適性評価方法の類似度算出ステップにおいて、第1言語の参照テキストの依存構造木を解析処理し、抽出された部分木毎に前記類似度算出を行うと共に、変換適性値出力ステップにおいて、各部分木における類似度からテキスト全体の類似度が最大となる部分木集合を求めて該第1データの変換適性値を算出する。これにより構文を考慮しながら最適な部分木集合を得る。

## 【0017】

請求項4に記載のデータ変換適性評価方法は、変換適性値出力ステップの後に、テキスト全体の類似度が最大となる部分木集合の中で、類似度が最小の部分木又は所定の閾値よ

50

りも小さな類似度の部分木の少なくともいずれかを抽出し、抽出された部分木を機械翻訳不適箇所として出力手段から出力することを特徴とする。これにより、テキスト中の翻訳不適箇所を提示することができる。

**【 0 0 1 8 】**

請求項 5 に記載の発明は、上記各発明の類似度算出ステップにおいて、少なくとも前記第 1 言語の参照テキスト又は前記折り返し翻訳テキストのいずれかにおける単語又は単語列を、単語又は単語列を語義及び品詞により階層に分類した所定の単語クラスの分類テーブルに基づき、より上位の階層の語義又は品詞に汎化した後に、所定の類似度算出式による算出を行うことを特徴とする。これにより類義語等に対して有効な類似度の算出を行う方法を提供する。

10

**【 0 0 1 9 】**

本発明は上記のデータ変換適性評価方法を実装したデータ変換装置として提供することもできる。

すなわち、請求項 6 に記載のように、第 1 データから第 2 データへの機械翻訳を行うデータ変換手段と、第 2 データから第 1 データへのデータ逆変換手段とが併存するデータ変換装置であって、該第 1 データに対して、データ変換手段における変換適性を評価して変換適性値を算出するデータ変換装置を提供する。

**【 0 0 2 0 】**

該装置は、第 1 言語の参照テキストである第 1 データを機械翻訳して第 2 言語のテキストである変換後第 2 データを取得するデータ変換手段と、該変換後第 2 データを第 1 言語の折り返し翻訳テキストに機械翻訳して逆変換後第 1 データを取得するデータ逆変換手段と、該第 1 データと該逆変換後第 1 データとを入力して、所定の類似度算出式により類似度を算出する類似度算出手段と、該類似度を第 1 データのデータ変換手段における変換適性値として出力する出力手段とを備える。

20

**【 0 0 2 1 】**

そして、本発明のデータ変換装置では、類似度算出手段が、逆変換後第 1 データのテキストにおける依存構造木に基づく単語 n-gram をパラメータに用いて類似度を算出することを特徴とする。上記と同様に、依存構造木に基づいて、文節内の単語はすべて隣に係り、係り文節における末尾の形態素は受け文節の先頭の形態素に係るとして単語 n-gram を数え、類似度を算出する。

30

**【 0 0 2 2 】**

請求項 7 に記載のデータ変換装置では、類似度算出手段が、第 1 言語の参照テキストに対する折り返し翻訳テキストの類似度を測るパラメータと共に、該折り返し翻訳テキストに対する該参照テキストの類似度を測るパラメータとの調和平均を算出することを特徴とする。

**【 0 0 2 3 】**

請求項 8 に記載のデータ変換装置では、類似度算出手段が、言語の参照テキストの依存構造木を解析処理し、抽出された部分木毎に前記類似度算出を行うとともに、出力手段が各部分木における類似度からテキスト全体の類似度が最大となる部分木集合を求めて該第 1 データの変換適性値を算出することを特徴とする。

40

**【 0 0 2 4 】**

請求項 9 に記載のデータ変換装置では、出力手段が、各部分木における類似度からテキスト全体の類似度が最大となる時の部分木集合の中で、類似度が最小の部分木又は所定の閾値よりも小さな類似度の部分木の少なくともいずれかを抽出し、抽出された部分木を機械翻訳不適箇所として出力することを特徴とする。

**【 0 0 2 5 】**

請求項 10 に記載のデータ変換装置では、類似度算出手段が、少なくとも第 1 言語の参照テキスト又は折り返し翻訳テキストのいずれかにおける単語又は単語列を、単語又は単語列を語義及び品詞により階層に分類した所定の単語クラスの分類テーブルに基づき、より上位の階層の語義又は品詞に汎化した後に、所定の類似度算出式による算出を行う構成

50

を提供する。

【発明の効果】

【0026】

本発明は、上記構成を備えることにより、データ変換装置における変換適性を数値により自動評価を行うことが可能となる。また、機械翻訳の場合に、テキストを構成する翻訳不適の箇所を抽出することができるため、機械翻訳精度の向上にも寄与する。

【発明を実施するための最良の形態】

【0027】

以下、本発明の実施形態を、図面に示す実施例を基に説明する。なお、実施形態は下記に限定されるものではない。

図1は本発明に係る機械翻訳装置(1)の全体構成図である。本発明は公知のパーソナルコンピュータにより容易に実現することが可能であり、演算処理やテキスト処理などを司るCPU(10)によって本発明の各ステップを実行処理する。CPU(10)は周知のように図示しないメモリと協働して動作し、キーボードやマウスなどの入力手段の他、出力結果を表示する表示手段、ハードディスク等の磁気記憶手段などを備えている。

【0028】

以下はデータ変換の例として日本語から英語への翻訳処理を挙げて説明するが、翻訳処理の場合にはいかなる言語間の機械翻訳処理に適用してもよい。また、翻訳処理以外の任意のデータ変換処理に用いることもできる。

まず本装置(1)に対して日本語テキストデータ(2)を入力する。入力する方法は磁気記憶装置からの読み出しやネットワークを通じた取得、キーボードからの入力などいかなる態様でもよい。

【0029】

入力された日本語テキストデータ(2)は、まず日英機械翻訳処理部(11)において英語に機械翻訳される。機械翻訳には公知の技術を用いることができる。一般に形態素解析処理、構文解析処理、テキスト生成処理などの各処理手段と、辞書データ(形態素、品詞情報等)などを用い、これらは本装置(1)におけるCPU、メモリ、磁気記憶媒体を用いて実現される。該処理部(11)については周知技術を利用すれば良いので、詳細な構成は省略する。

【0030】

翻訳結果である翻訳後英語テキストは一時的に磁気記憶媒体に記憶された後、次の英日機械翻訳処理部(12)において再び日本語への機械翻訳処理がなされる。

該英日機械翻訳処理部(12)についても公知の機械翻訳方法を用いることができる。

【0031】

英日機械翻訳処理部(12)で生成された折り返し翻訳日本語テキストは、一旦磁気記憶媒体に記憶された後、日本語テキストデータ(2)と共に類似度算出処理部(13)に入力される。本発明ではこのように類似度算出処理部(13)において元の参照テキストと、2回の翻訳を経た折り返し翻訳テキストとの類似度を算出することに特徴がある。類似度が高いものは入力文を的確に翻訳したものと推定され、ここでは確信度(変換適性値)として値を確信度出力処理部(14)から出力処理する。すなわち、本発明では、参照文とその折り返し翻訳文の類似度が高いほど機械翻訳可能性が高いとして高い確信度が出力される。

【0032】

以下、類似度算出処理部(13)における処理を説述する。

ここで、類似度の算出には機械翻訳の自動評価手法としてよく用いられるBLEUを拡張したものをを用いる。BLEUは米国IBM社(登録商標)によって提案されている方法であり、非特許文献3などに記載されている。

BLEUは出力英文と模範訳を4-gramで評価する手法であり、一般的には模範訳を複数入力する。類似度は0から1の範囲のスコアとして得られる。

【0033】

10

20

30

40

50

【非特許文献3】Kishore Papineni, Salim Roukos, Todd Ward, and Weining. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th ACL, pp.311 - 318, 2002.

【0034】

本発明における類似度（以下、C-measure：CM値と呼ぶ）は下記の式により計算する。

【0035】

【数1】

$$CM = \frac{2 \times CM_{bleu}(B|S) \times CM_{bleu}(S|B)}{CM_{bleu}(B|S) + CM_{bleu}(S|B)} \quad 10$$

ここで、 $CM_{BLEU}(B|S)$ のSとBはそれぞれ、原文とその折り返し翻訳文を表している。Logを取ると、次の式によって表される。

【0036】

【数2】

$$\log(CM_{bleu}(B|S)) = \min\left(1 - \frac{s}{b}, 0\right) + \sum_{n=1}^N \frac{1}{N} \log p_n(B|S) \quad 20$$

この式において、sは原文の単語長、bは折り返し翻訳文の単語長、Nは考慮する単語n-gramの最大のnの値を表わす。 $p_n(B|S)$ は次の式で表わされる。

【0037】

【数3】

$$p_n(B|S) = \frac{\sum_{wn \in B} Count_{clip}(wn)}{\sum_{wn' \in B} Count(wn')} \quad 30$$

ここで、 $Count(wn')$ はBにおける単語n-gram  $wn'$ の出現頻度を表わす。 $Count_{clip}(wn)$ は、次の式で表わされる。

【0038】

【数4】

$$Count_{clip}(wn) = \min(Count(wn), Max\_Count(wn|S))$$

ここで、 $Max\_Count(wn|S)$ はSにおける単語n-gram  $wn$ の出現頻度を表わす。

40

【0039】

本発明による類似度算出処理部は、以上の計算を実行し、日本語テキストデータ(2)と英日機械翻訳処理部(12)から出力される折り返し翻訳日本語テキストとの類似度を算出する。このような類似度の算出は従来提案されていなかったものであり、類似度算出処理部(13)の計算は上記の計算式に限定されない。

【0040】

しかし、本発明は上記の計算によりBLEUとは異なる、類似度を算出するのに好適な構成を備えている。以下に説述する。詳細な実施態様を図2に示す。同一符号を伏した要素は全て上記と同様である。

【0041】

50



まず第1に、上記数式1では依存構造木に基づく単語n-gramに基づいて計算を行っている。

日本語や韓国語などいくつかの言語においては、語順が比較的自由である。例えば、「太郎と花子はテニスをした」という日本語文の文節を単位とする依存構造は「((太郎と花子は)(テニスをした))」のように表わされ、この依存構造からは、「太郎と花子はテニスをした」と「テニスを太郎と花子はした」の二種類の語順の日本語文が生成可能である。BLEUScoreはフラットな単語列における単語n-gramに基づいて計算されるため、この二文のBLEUScoreは1とはならない。

【0042】

しかし、この二文は同じ意味で、同じ訳文になると考えられるため、類似度は1となるべきである。そこで、依存構造木に基づく単語n-gramを採用している。単語の単位は形態素解析システムJUMAN(非特許文献4を参照)で定義されている形態素とし、文節内の単語はすべて隣に係り、係り文節における末尾の形態素は受け文節の先頭の形態素に係ると仮定する。

予備実験を基に、本実施例では数式2で $N=3$ と設定している。

【0043】

【非特許文献4】黒橋禎夫,長尾眞.日本語形態素解析システムJUMAN使用説明書Version3.61.京都大学大学院情報学研究科,1999.

【0044】

上記構成を用いるため、本実施例の類似度算出処理部(13)には形態素解析処理部(131)と構文解析処理部(132)を備えている。各解析技術は公知であり、形態素解析処理部(131)には磁気記憶媒体に格納された形態素辞書(134)を用い、構文解析処理部(132)で依存構造木を取得する。

【0045】

類似度算出処理部(13)の特徴の第2に、調和平均を用いていることが挙げられる。すなわちBLEUScoreは機械翻訳文における単語n-gramの精度に基づいて計算されるため、機械翻訳文と参照文を入れ替えたときに計算されるBLEUScoreは元のものとは異なる。しかし、類似度としては、入れ替えても同じ値になるのが自然である。

そこで、数式1で表わされるように、参照文に対する機械翻訳文BLEUScoreだけでなく、機械翻訳文に対する参照文のBLEUScoreも考慮するように定義している。

【0046】

第3の特徴として、類義語などの汎化が挙げられる。

BLEUScoreは表層単語に基づいて計算される。したがって、類義語は別の単語として扱われる。しかし、二つの文の違いが類義語の関係にある単語のみであった場合には、類似度は1となるのが望ましい。そこで、単語を単語クラスに置き換えて汎化している。

【0047】

ひとつの単語が複数の単語クラスに属する場合は、原文と折り返し翻訳文との間で一致する単語クラスの数が増えるように山登りの的に準最適な単語クラスの集合を探索する。

単語クラスとしては「分類語彙表」(非特許文献5を参照)の上位から5レベル目の階層を用いる。分類語彙表に収録されている単語の異なり数は101,070である。

【0048】

【非特許文献5】国立国語研究所(編)、分類語彙表(増補改訂版)、大日本図書,2004.

【0049】

さらに、接続表現や数量表現をひとつのクラスに汎化するため、品詞カテゴリが「接続助詞」あるいは「数詞」である単語は品詞に汎化し、連続する数詞はひとつの数詞に置き換えた。

そして、敬体と常体を区別しないために「接尾辞」の「ます」は無視するようにし、句読点の有無によって類似度が異なるようなことがないように句読点も無視するようにした。

10

20

30

40

50

## 【 0 0 5 0 】

これらの汎化処理のため、本実施例では単語汎化処理部 ( 1 3 3 ) を設けると共に、磁気記憶手段に上記単語クラスとして単語クラステーブル ( 1 3 5 ) を格納している。

汎化処理は単語に限らず、句・節・文といった任意の単語列に対して行っても良い。

## 【 0 0 5 1 】

次に、第 2 の実施例としてテキストを構成する各部分について、それぞれに CM 値を求めることにより、機械翻訳可能な部分とそうでない部分を特定する技術について説述する。

部分としては、与えられた文の各部分木を利用する。つまり、与えられた文におけるすべての部分木に対し CM 値を計算することを考える。ここで、与えられた文におけるすべての部分木の集合を SST とし、与えられた文そのものも SST に含まれるものとする。

10

## 【 0 0 5 2 】

図 3 には本実施例の流れを示す。与えられた文の依存構造木は JUMAN と KNP ( 非特許文献 6 を参照 ) で解析することによって得られ、部分木はその依存構造木から得られる。( S 1 )

## 【 0 0 5 3 】

【非特許文献 6】黒橋禎夫. 日本語構文解析システム KNP 使用説明書 Version 2.0b6. 京都大学大学院情報学研究科, 1998.

## 【 0 0 5 4 】

与えられた文  $s$  における任意の部分木  $st_i$  (  $S$  ) の確信度スコア  $Scr(st_i)$  を次のように定義する。( S 2 )

20

## 【 0 0 5 5 】

## 【数 5】

$$Scr(st_i) = (st_i \text{ の CM }) \times \frac{st_i \text{ の 文 節 数 }}{\text{与えられた文の文節数}}$$

そして、下記のように確信度スコアが最大となる部分木集合  $ST_{best}$  を求める。( S 3 )

## 【 0 0 5 6 】

## 【数 6】

$$ST_{best} = \operatorname{argmax}_{ST} \sum_{s_i \in ST} Scr(s_i)$$

30

ただし、ST は SST の部分集合であり、ST 中の部分木の文節は重ならないものとする。つまり、ST に含まれる部分木を単純に繋ぎ合わせると元の文  $s$  が得られるものとする。複数の部分木が同じ確信度スコアを持つ場合は、最長のものを優先する。ここで、長さは部分木中の文節数と定義する。与えられた文の CM 値がすべての部分木の中で最大となる場合は、与えられた文そのものが  $ST_{best}$  として選ばれる。最適な部分木集合は山登り法により探索する。

## 【 0 0 5 7 】

最適な部分木集合をそれぞれの CM 値とともにユーザに提示する。( S 4 )

40

以上のように実施例 2 は最終的に確信度出力処理部 ( 1 4 ) から部分木ごとの確信度を出力することにより、ユーザにいずれの部分木が機械翻訳に適性を有しているのか、提示することができる。

## 【 0 0 5 8 】

さらに実施例 2 の応用として、機械翻訳不適個所の候補を自動的に提示することもできる。該手順は図 4 に示す。

まず、最適な部分木集合を入力 (  $s_1$  ) し、部分木の CM 値がすべて閾値よりも低い場合 (  $s_2$  ) は、全部分木集合の中から CM 値が最低となる部分木を抽出 (  $s_3$  ) して機械翻訳不適個所の候補としてユーザに提示 (  $s_4$  ) する。

50

## 【 0 0 5 9 】

このような場合は、文末の文節が機械翻訳不適個所であるか、主語が欠落しているなど翻訳に必要な情報が不足している場合が多い。複数の部分木が同じCM値を持つ場合、最長のものが優先される。すべての部分木のCM値が同じ場合は、末尾の文節を優先する。

## 【 0 0 6 0 】

最適な部分木集合に、CM値が閾値を越える部分木がある場合(s5)、その部分木は機械翻訳可能であることが多く、残りの部分木のうち、CM値が閾値より低いものが機械翻訳不適個所である場合が多い。

このような場合、後者を機械翻訳不適個所の候補としてユーザに提示(s6)する。

## 【 0 0 6 1 】

さらに、全部分木集合の中に、機械翻訳不適個所の候補を含み、かつ、CM値が閾値を越える部分木がある場合、その部分木と機械翻訳不適個所の候補との差異の部分が機械翻訳不適個所である場合があるので、参考としてユーザに提示する。

## 【 0 0 6 2 】

第3の実施例として、以上による機械翻訳不適箇所が抽出された時に、例えば図1において日本語テキストを日英機械翻訳処理部(11)で機械翻訳する前に、公知の言い換え技術を用いて言い換えを行い、同義の日本語テキストを入力することで翻訳精度を向上させることもできる。

この場合、たとえば言い換え文を複数生成して、それぞれの確信度(3)を求め、最も確信度が高かった言い換え文の日英翻訳結果を最適な結果として出力することができる。

## 【 0 0 6 3 】

本方法によると、単に確信度を抽出して、ユーザに提示するだけでなく、機械翻訳装置(1)において自動的に最適な翻訳を行わせることにも寄与する。このように本発明では機械翻訳精度を向上した機械翻訳装置を提供することもできる。

## 【 0 0 6 4 】

本発明は以上の構成を備えるが、本発明の効果を示すため、次のような実験を行った。

まず第1に本発明に係るCM値と機械翻訳自動評価指標であるBLEUやNIST(非特許文献7を参照)および、人間による主観評価との関係について述べる。CM値は入力文とその折り返し翻訳文を基に計算され、BLEUscoreおよびNISTscoreは、入力文をMTシステムにより翻訳した英訳文と英語参照文を基に計算される。

## 【 0 0 6 5 】

【非特許文献7】NIST. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Technical report, NIST,2002.

## 【 0 0 6 6 】

テストセットとしては、NTT(登録商標)より配布されているMTテストセット(非特許文献8を参照)を用いた。このテストセットは、日英MTシステムの評価用に作成されたもので、3,718文の日本語文とその英訳からなる。このテストセットにおける各日本語文に対し、ひとつずつ英語参照文を選択し、自動評価に用いた。

## 【 0 0 6 7 】

【非特許文献8】池原悟,白井諭,小倉健太郎.言語表現体系の違いに着目した日英機械翻訳機能試験項目の構成.人工知能学会誌,Vol.9,No.4,71994.<http://www.kecl.ntt.co.jp/icl/mtg/resources/index.php>

## 【 0 0 6 8 】

BLEUscoreとNISTscoreの計算には、mteval(versionv11a)(非特許文献9を参照)を用いた。図5から図8に、それぞれ、CM値とBLEUscoreおよびNISTscoreとの関係、CM値と人間による主観評価(flucency,adequacy)との関係を示す。

これらのグラフは、CM値に閾値を設けて0から1の間で変化させ、各閾値に対しその閾値を越える入力文を抽出し、その英訳の自動評価値を計算することによって描いたものである。

## 【 0 0 6 9 】

10

20

30

40

50

【非特許文献9】<http://www.nist.gov/speech/tests/mt/resources/scoring.htm>

【0070】

主観評価にはテストセットの内、先頭から奇数番号の例文を950文抽出して用いた。CM値として、本発明の数式1による特徴を採用した場合と採用しなかった場合との違いが比較できるように、各図の各グラフには、CM値として、それぞれ、BLEU("bleu")、NIST("nist")、調和平均("harmonic")、依存構造木に基づく単語n-gram("tree-ngram")、汎化("generalization")およびこれらの組み合わせを採用した場合の結果を示した。

【0071】

CM値として、本発明の全特徴を採用した場合、図5から図8におけるCM値とBLEUscore、NISTscore、fluency、adequacyとの相関係数は、それぞれ、0.9911、0.8948、0.9758、0.9536と高かった。これは、CM値によって平均的に評価値の高い翻訳を選別することができることを示している。また、参照文を用意しなくても、CM値が低い文を集めることによって、機械翻訳システムにとって翻訳が難しい文の集合を自動的に収集できることも示している。

10

CM値が高く、かつ、原文と折り返し翻訳文との差異が小さいものを信頼するようにすれば、より信頼性の高い翻訳が得られると考えている

【0072】

図5と図6において、BLEUscoreとNISTscoreそれぞれに対する相関係数の平均値が最も高かったのは、「依存構造木に基づく単語n-gram+汎化」を採用した場合で、相関係数はそれぞれ、0.9848と0.9701であった。このとき、図7と図8において、人間の主観評価(fluency,adequacy)に対する相関係数はそれぞれ、0.9344、0.8999と高かった。

20

この結果は、機械翻訳可能性の自動評価においては、折り返し翻訳文における単語n-gramの調和平均よりも精度を重視する方が良いことを示している。

【0073】

次に、機械翻訳不適個所の推定が、原文の機械翻訳可能性を向上させるのに貢献するかどうかを調べる実験を行なった。テストセットの先頭の100文に対し、最適な部分木集合と機械翻訳不適個所候補を推定して被験者に提示した。

【0074】

その情報をもとに、書き換えるべきであり、かつ、書き換え可能だと判断したものについてのみ原文を書き換えさせた。CM値としては、図5から図8において相関が高かった「依存構造木に基づく単語n-gram+汎化」を採用した。被験者に提示した情報の例を図9に示す。

30

「Partialtranslation」で示されているのが最適な部分木集合であり、「Check!」で示されているのが機械翻訳不適個所の候補である。実施例では、閾値は0.5とした。

【0075】

図9の例では、「鉛筆は、」が候補として示されている。ここで、原文の「鉛筆は、2BかHBを使ってください。」を例えば「2BかHBの鉛筆を使ってください。」に書き換えることができれば、「Use the pencil of 2B or HB.」といったよりよい翻訳を得ることができる。

40

【0076】

上記100文に対する書き換えの後、機械翻訳システムで再度翻訳し、翻訳文を評価したところ、BLEUscore、NISTscoreがそれぞれ、0.1739と3.3162から、0.2161と3.6674に向上した。書き換えた文は43文であった。書き換え前後で翻訳文の質を下記に示す主観評価により調べたところ、29文について前より良くなっており、悪くなったものはなかった。

【0077】

この29文中、18文(62%)については、被験者が修正した個所とシステムが提示した機械翻訳不適個所が重なっていた。他の11文については、原文と折り返し翻訳文との差異から主語を追加するべきと判断したものが2文あり、残りは、「Partialtranslation」の折り返し翻訳が不自然であることから修正したと考えられる。書き換えの際には「Partialtra

50

nslation」の折り返し翻訳が参考になった。

【0078】

この結果は、システムの提示した情報が有効に働いていることを示している。主観評価は、同じ100文に対し、5段階で行なった。この主観評価では、各文に対し1点(とても悪い)から5点(とても良い)の点数が評価点として与えられた。3点以上が理解可能であるとした。主観評価の点数は、書き換え前後で平均値2.73点から3.52点へと向上した。書き換えた43文については、平均値1.63点(合計70点)から3.47点(合計149点)へと大幅に良くなった。

【0079】

図10は対象言語における出力の例である。翻訳とともに、CM値や他の翻訳候補が示されている。他の翻訳候補としては、各部分木の翻訳だけでなく、機械翻訳不適個所の推定により得られた部分木集合の各翻訳を単純に組み合わせて生成したものも提示している

10

例えば、図10の[Use 2B or HB.][The pencil]が後者の生成例である。これは、元の翻訳文「The pencil use 2B or HB.」よりは理解しやすくなっている。

【0080】

次に、英語側の情報をもとに誤っていると思われるところを指摘し、その指摘にしたがって適切に原文の日本語文を書き換えることができるかを調べた。まず、上述と同じ100文から原文のCM値が0.5以上のもの32文を抽出した。その中から、英語を母国語とする人が、英語としてはおかしいものの、機械翻訳不適個所の候補の英訳に問題がないものを選択し、「Subtrees」で示されている部分木の訳の情報をもとに、問題と思われる文節を指摘した。

20

【0081】

指摘できたのは、32文中7文であった。その指摘を受けて、日本語を母国語とする被験者が日本語の部分木とその折り返し翻訳の情報のみを参照しながら原文を書き換えるという作業を行なった。その結果、7文中、1文については悪くなったが、5文については評価点が良くなり、理解不能だった文は、7文中5文から1文に減った。数は少ないが、指摘した部分についてはうまく修正できることが多い。

【0082】

以上説述したように、本発明は与えられた入力文がどの程度機械翻訳可能かを表わす変換適性値(確信度)の計算方法、その確信度を用いて入力文のうち機械翻訳不適個所を推定する方法、対象言語における他の翻訳候補を生成する方法からなる。

30

それぞれ、機械翻訳結果の妥当性を数値化したり、原文のうち機械翻訳に適していない部分を推定したり、他の翻訳候補を示したりすることができ、その結果、原文を書き換えるべきかどうか、書き換えるとすればどこを書き換えるべきかを指摘することができるようになる。

さらに、翻訳結果そのものが意味不明のものでも他の候補から原文の意図を推測できるようになる。

【0083】

本実施例では日英機械翻訳処理、英日機械翻訳処理と共に本発明に係る確信度の算出処理を一体的な装置で処理しているが、各処理はそれぞれ別体の装置の集合により処理してもよい。

40

【図面の簡単な説明】

【0084】

【図1】本発明の機械翻訳装置の全体構成図である。

【図2】本実施形態に係る機械翻訳装置の全体構成図である。

【図3】第2実施例に係る処理の流れ図である。

【図4】第2実施例の応用例に係る処理の流れ図である。

【図5】CM値とBLEU scoreとの関係を示すグラフである。

【図6】CM値とNIST scoreとの関係を示すグラフである。

50

【図7】CM値とfluencyとの関係を示すグラフである。

【図8】CM値とadequacyとの関係を示すグラフである。

【図9】本発明による機械翻訳不適箇所の出力結果例である。

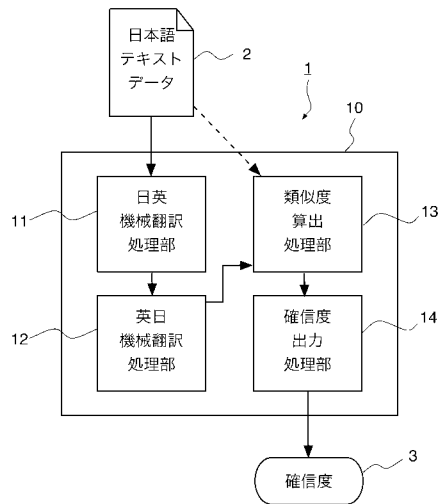
【図10】本発明による対象言語における他の翻訳候補の出力例である。

【符号の説明】

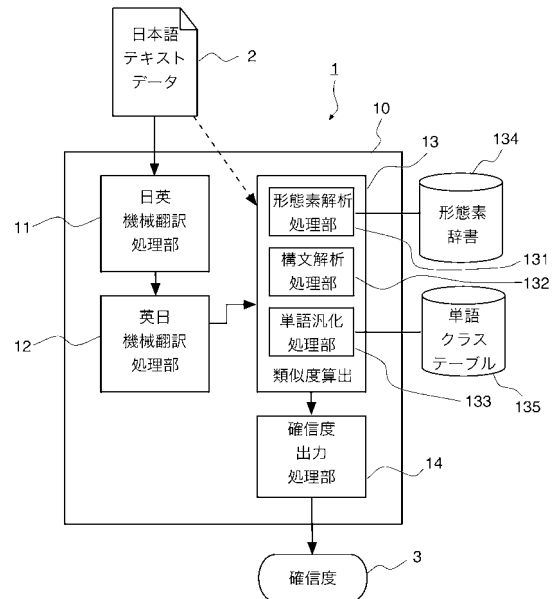
【0085】

- 1 機械翻訳装置
- 2 日本語テキストデータ
- 3 確信度
- 11 日英機械翻訳処理部
- 12 英日機械翻訳処理部
- 13 類似度算出処理部
- 14 確信度出力処理部

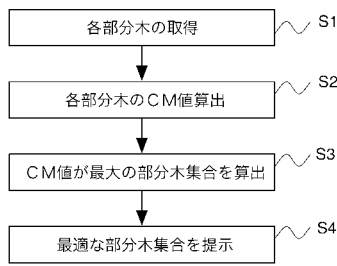
【図1】



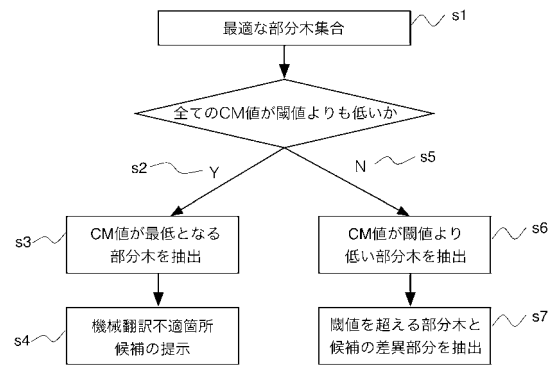
【図2】



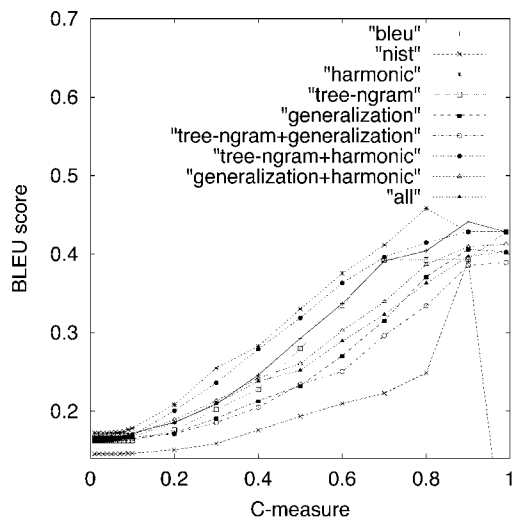
【 図 3 】



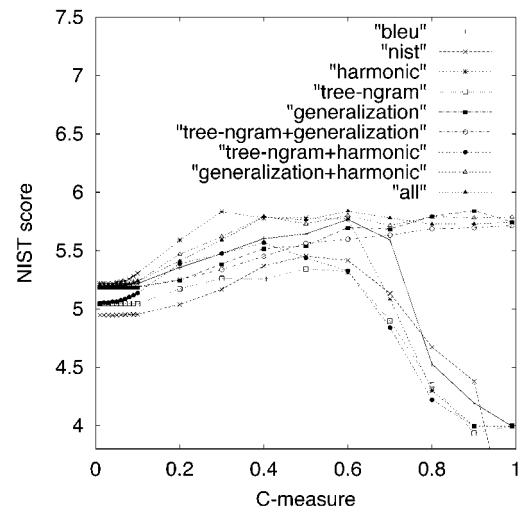
【 図 4 】



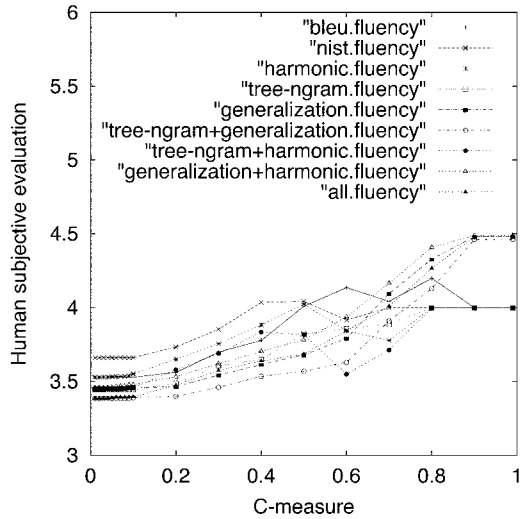
【 図 5 】



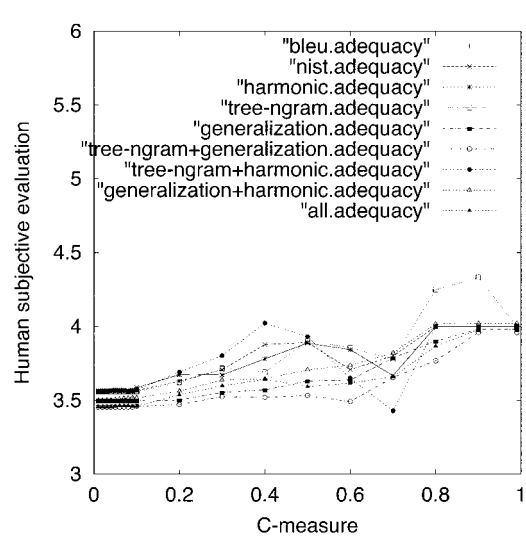
【 図 6 】



【 図 7 】



【 図 8 】



【 図 9 】

```
#ORIGINAL: 鉛筆は、2 BかHBを使ってください。
#-----
# 原文(部分木)      折り返し翻訳      確信度スコア
#-----
#--Subtrees-->
2 BかHBを使ってください。      2 BまたはHBを使ってください。      0.58
HBを使ってください。          HBを使ってください。          0.5
鉛筆は、2 Bかを使ってください。      2 Bまたは鉛筆を使います。      0.26
使ってください。              使ってください。              0.25
2 Bか使ってください。          2 Bまたは使用。              0.23
鉛筆は、2 BかHBを使ってください。      鉛筆使用2 BまたはHB。        0.22
鉛筆は、HBを使ってください。          鉛筆使用HB。                  0
鉛筆は、使ってください。            鉛筆を使ってください。        0
鉛筆は、                          鉛筆                          0
2 Bか                                それは2 Bですか?              0
HBを                                  HBの                            0
#<--Subtrees---
#-----
# 原文(部分木)      折り返し翻訳      C-measure
#-----
#--Partial translation-->
2 BかHBを使ってください。      2 BまたはHBを使ってください。      0.77
[鉛筆は、                          鉛筆                              0 ]
(鉛筆は、2 Bか使ってください。      2 Bまたは鉛筆を使います。        0.26)
#<--Partial translation---
#--Check!-->
[鉛筆は、                          鉛筆                              0 ]
#<--Check!---
EOD
```

【 図 10 】

```
#ORIGINAL: The pencil use 2B or HB.
#-----
# 原文(部分木)      翻訳      C-measure
#-----
#--Subtrees-->
1 2 3      Use 2B or HB.      0.58
2 3        Use HB.            0.5
0 1 3      2B or use a pencil. 0.26
3          Use.                0.25
1 3        2B or use.         0.23
0 1 2 3    The pencil use 2B or HB. 0.22
0 2 3      The pencil use HB.  0
0 3        Use a pencil.       0
0          The pencil       0
1          Is it 2B?          0
2          of HB            0
#<--Subtrees---
#-----
# 原文(部分木)      翻訳      確信度スコア
#-----
#--Partial translation-->
1 2 3      Use 2B or HB.      0.77
[0         The pencil          0 ]
(0 1 3     2B or use a pencil. 0.26)
[Use 2B or HB.][The pencil]
#<--Partial translation---
#--Check!-->
[0         The pencil          0 ]
#<--Check!---
EOD
```



フロントページの続き

(56)参考文献 特開平03-282776(JP,A)

(58)調査した分野(Int.Cl., DB名)

G06F 17/27 - 17/28