(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2006-252323
(P2006-252323A)
(43)公開日　平成18年9月21日(2006.9.21)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/28	(2006.01)	G06F 17/28	Z	5B091

審査請求 未請求　請求項の数 14　ＯＬ　（全 16 頁）

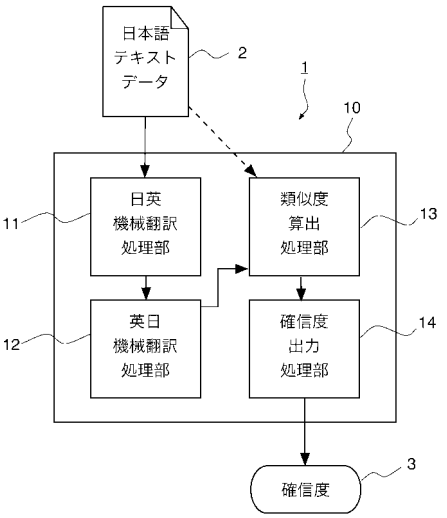
(21)出願番号　特願2005-69816(P2005-69816)
(22)出願日　平成17年3月11日(2005.3.11)

(71)出願人　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

Ｆターム(参考)　5B091 AA04　CD11　DA07　DA08　EA14

(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)【要約】
【課題】　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。
【解決手段】　データ変換手段１１により第１データ２を変換して変換後第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１２により該変換後第２データを逆変換して逆変換後第１データを取得するデータ逆変換ステップ、第１データ２と逆変換後第１データとを類似度算出手段１３に入力して、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ステップ、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３として出力手段１４から出力する変換適性値出力ステップを含む。
【選択図】　図１



【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、
該データ変換手段により第１データを変換して変換後第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により該変換後第２データを逆変換して逆変換後第１データを取得するデータ逆変換ステップ、
該第１データと該逆変換後第１データとを類似度算出手段に入力して、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ステップ、
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手段から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】
前記データ変換が機械翻訳であって第１言語の参照テキストから第２言語のテキストへの機械翻訳を行う前記データ変換ステップの後に、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データを取得する前記データ逆変換ステップが行われる
ことを特徴とする請求項１に記載のデータ変換適性評価方法。
【請求項３】
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、
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
ことを特徴とする請求項２に記載の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の類似度算出ステップにおいて、
前記第１言語の参照テキストに対する前記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を用いる
ことを特徴とする請求項２又は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、
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２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】
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い、
前記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
ことを特徴とする請求項２ないし５のいずれかに記載のデータ変換適性評価方法。
【請求項７】
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、

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手段から出力する
ことを特徴とする請求項６に記載のデータ変換適性評価方法。
【請求項８】
第１データから第２データへのデータ変換手段と、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データを変換して変換後第２データを取得するデータ変換手段と、
該変換後第２データを逆変換して逆変換後第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
を備えたことを特徴とするデータ変換装置。
【請求項９】
前記データ変換が機械翻訳であって、データ変換手段が第１言語の参照テキストから第２言語のテキストへの機械翻訳を行うと共に、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
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、
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
ことを特徴とする請求項９に記載のデータ変換装置。
【請求項１１】
前記類似度算出手段が、
前記第１言語の参照テキストに対する前記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を用いる
ことを特徴とする請求項９又は１０に記載のデータ変換装置。
【請求項１２】
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９ないし１１のいずれかに記載のデータ変換装置。
【請求項１３】
前記類似度算出手段が、
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い、
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
ことを特徴とする請求項９ないし１２のいずれかに記載のデータ変換装置。
【請求項１４】
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１３に記載のデータ変換装置。

【発明の詳細な説明】
【技術分野】
【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人手によって確認し、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。
【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。
【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。
【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。
【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。
【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、データ変換手段により第１データを変換して変換後第２データを取得するデータ変換ステップ、データ逆変換手段により該変換後第２データを逆変換して逆変換後第１データを取得するデータ逆変換ステップを有する。
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に入力して、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ステップ、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手段から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。
【００１２】
請求項２に記載の発明は、上記のデータ変換が機械翻訳であって第１言語の参照テキストから第２言語のテキストへの機械翻訳を行う前記データ変換ステップの後に、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データを取得する前記データ逆変換ステップが行われることを特徴とする。これにより本発明を機械翻訳に適用することができる。
【００１３】
請求項３に記載の発明は、上記の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。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。
【００１４】
請求項４に記載の発明は、上記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を用い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。
【００１５】
請求項５に記載の発明は、上記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。
【００１６】
請求項６に記載の発明は、上記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７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。

【００１８】

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８に記載のように、第１データから第２データへのデータ変換手段と、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。

【００１９】

該装置は、第１データを変換して変換後第２データを取得するデータ変換手段と、該変換後第２データを逆変換して逆変換後第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たことを特徴とする。

【００２０】

また、請求項９の発明は、上記のデータ変換が機械翻訳であって、データ変換手段が第１言語の参照テキストから第２言語のテキストへの機械翻訳を行うと共に、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データを取得する構成である。

【００２１】

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、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。

【００２２】

請求項１１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を用いることを特徴とする。

【００２３】

請求項１２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を提供する。

【００２４】

請求項１３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。

【００２５】

請求項１４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。

【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】
【非特許文献３】Kishore Papineni,Salim Roukos,Todd Ward,and Weiing. Zhu. BLEU:a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 4

0th 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】

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

ここで、$\mathrm{CM_{BLEU}(B|S)}$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】

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

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】

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

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。$\mathrm{Count_{clip}(wn)}$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。

【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】

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最も多くなるように山登り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。

【００５０】

これらの汎化処理のため、本実施例では単語汎化処理部（１３３）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（１３５）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。
【００５１】
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。
【００５２】
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。（Ｓ１）
【００５３】
【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Version2.0b6.京都大学大学院情報学研究科,1998.
【００５４】
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。（Ｓ２）
【００５５】
【数５】

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

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STbestを求める。（Ｓ３）
【００５６】
【数６】

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

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STbest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。
【００５７】
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。（Ｓ４）
以上のように実施例２は最終的に確信度出力処理部（１４）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。
【００５８】
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（ｓ１）し、部分木のＣＭ値がすべて閾値よりも低い場合（ｓ２）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（ｓ３）して機械翻訳不適個所の候補としてユーザに提示（ｓ４）する。
【００５９】
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。

【００６０】

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。

【００６１】

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。

【００６２】

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。

【００６３】

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。

【００６４】

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。

【００６５】

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

【００６６】

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。

【００６７】

【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

【００６８】

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(fluency,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。

【００６９】

【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。

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、BLEUsocre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。

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。

「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。

【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「Partialtranslation」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。

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。

【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。

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。

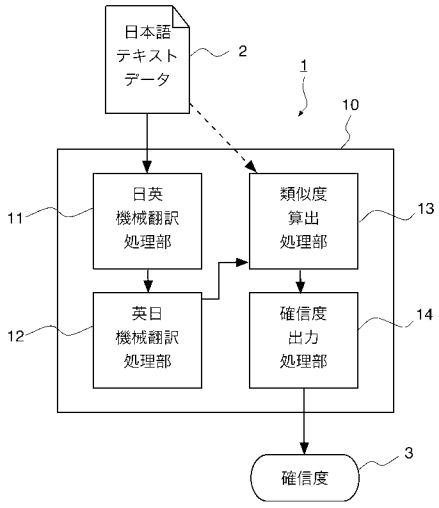
【図面の簡単な説明】

【００８４】

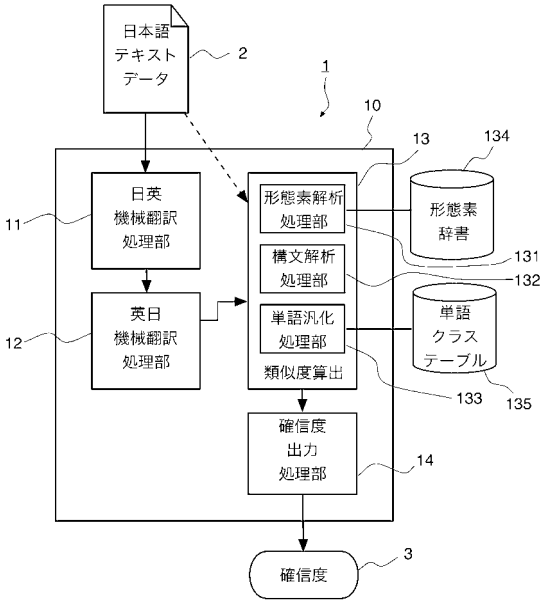
【図１】本発明の機械翻訳装置の全体構成図である。
【図２】本実施形態に係る機械翻訳装置の全体構成図である。
【図３】第２実施例に係る処理の流れ図である。
【図４】第２実施例の応用例に係る処理の流れ図である。
【図５】ＣＭ値とBLEU scoreとの関係を示すグラフである。
【図６】ＣＭ値とNIST scoreとの関係を示すグラフである。
【図７】ＣＭ値とfluencyとの関係を示すグラフである。
【図８】ＣＭ値とadequancyとの関係を示すグラフである。

【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0。
【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

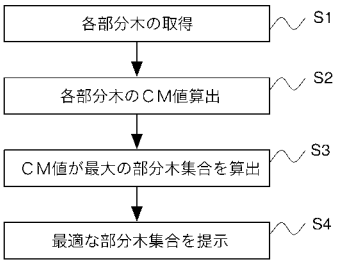
【図１】



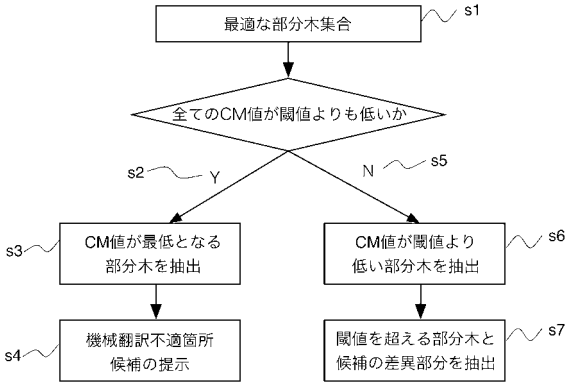
【図２】



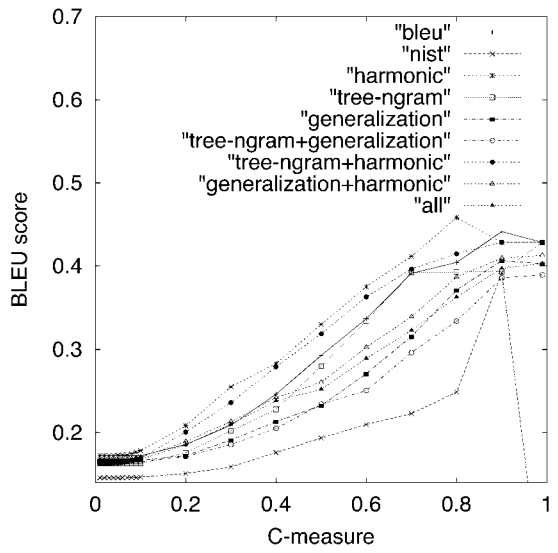
【図３】


各部分木の取得
S1
各部分木のＣＭ値算出
S2
ＣＭ値が最大の部分木集合を算出
S3
最適な部分木集合を提示
S4

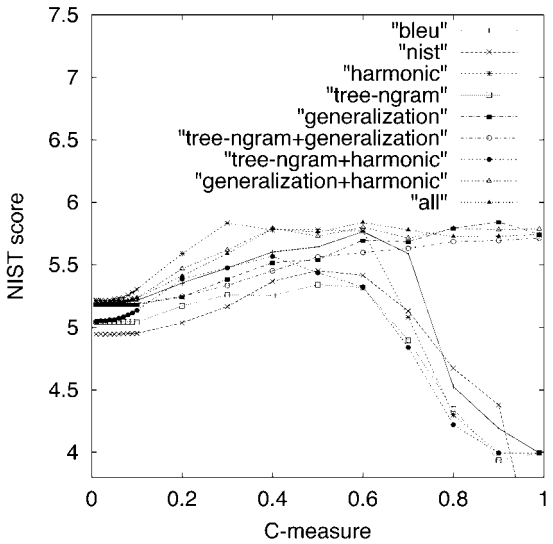
【図４】


最適な部分木集合
s1
全てのCM値が閾値よりも低いか
s2
Y
N
s5
s3
CM値が最低となる
部分木を抽出
CM値が閾値より
低い部分木を抽出
s6
s4
機械翻訳不適箇所
候補の提示
閾値を超える部分木と
候補の差異部分を抽出
s7

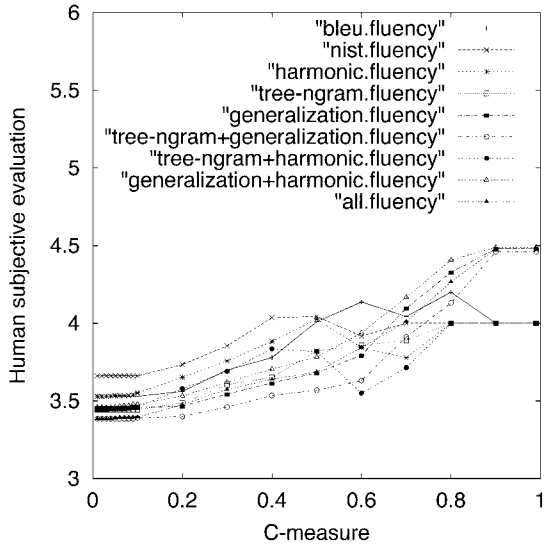
【図５】


BLEU score
C-measure
"bleu"
"nist"
"harmonic"
"tree-ngram"
"generalization"
"tree-ngram+generalization"
"tree-ngram+harmonic"
"generalization+harmonic"
"all"

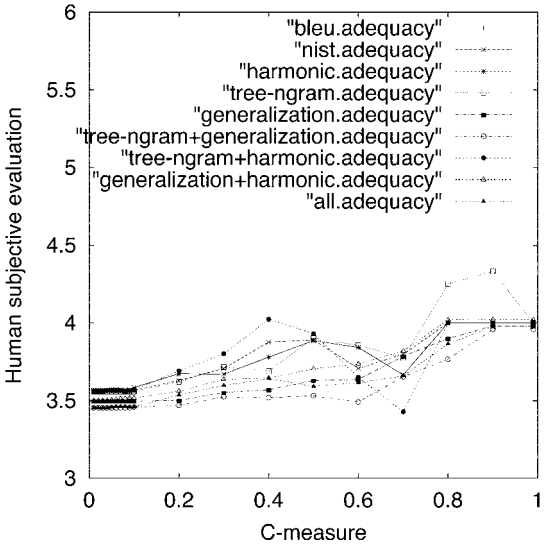
【図６】


NIST score
C-measure
"bleu"
"nist"
"harmonic"
"tree-ngram"
"generalization"
"tree-ngram+generalization"
"tree-ngram+harmonic"
"generalization+harmonic"
"all"

【図７】



【図８】



【図９】

```
#ORIGINAL: 鉛筆は、２Ｂ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-->
２ＢかＨＢを使ってください。      ２Ｂ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、２Ｂか使ってください。    ２Ｂ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
２Ｂか使ってください。            ２Ｂまたは使用。                  0.23
鉛筆は、２ＢかＨＢを使ってください。 鉛筆使用２Ｂ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
２Ｂか                            それは２Ｂ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-->
２ＢかＨＢを使ってください。      ２ＢまたはＨＢを使ってください。  0.77
[鉛筆は、                         鉛筆                              0  ]
(鉛筆は、２Ｂか使ってください。   ２Ｂ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
```