(19)日本国特許庁（ＪＰ）　(12)特　許　公　報（Ｂ２）　(11)特許番号
特許第3388393号
（Ｐ３３８８３９３）
(45)発行日　平成15年３月17日(2003.3.17)　(24)登録日　平成15年１月17日(2003.1.17)

(51)Int.Cl.[7]　識別記号　ＦＩ
G06F 17/27　G06F 17/27　M

請求項の数２　（全７頁）

(21)出願番号　特願平11－238579
(22)出願日　平成11年８月25日(1999.8.25)
(65)公開番号　特開2001－67357(Ｐ2001－67357Ａ)
(43)公開日　平成13年３月16日(2001.3.16)
審査請求日　平成11年８月25日(1999.8.25)

特許法第30条第１項適用申請有り　村田真樹・馬青・内元清貴・井佐原均「用例ベースによるモダリティの日英翻訳」情報処理学会研究報告99－ＮＬ－130－16，Ｖｏｌ.99，Ｎｏ.22，ｐ.121－ｐ.128(1999.03.05)

(73)特許権者　301022471
独立行政法人通信総合研究所
東京都小金井市貫井北町４－２－１
(72)発明者　村田　真樹
兵庫県神戸市西区岩岡町岩岡558－２
郵政省通信総合研究所関西支所内
(72)発明者　内元　清貴
兵庫県神戸市西区岩岡町岩岡558－２
郵政省通信総合研究所関西支所内
(72)発明者　馬　青
兵庫県神戸市西区岩岡町岩岡558－２
郵政省通信総合研究所関西支所内
(74)代理人　100082669
弁理士　福田　賢三　（外２名）

審査官　和田　財太

最終頁に続く

(54)【発明の名称】データベースを利用したテンス、アスペクトあるいはモダリティに関する翻訳装置

(57)【特許請求の範囲】
【請求項１】　第一の言語から第二の言語へのデータベースを利用した翻訳装置で、第一の言語に属する複数の用例と第二の言語に属する複数の用例からなり且つ個々の第一の言語に属する用例は、第二の言語に属する用例との間に少なくとも一つ以上の対応付があり、テンス、アスペクトあるいはモダリティ情報が付加されたことを特徴とする第一のデータベースを備え、
第一の言語に属する第一の用例と、第一のデータベースの第一の言語に属する第二の用例との間の文末からみて連続する共通の文字列の数を用いて第一のデータベースの第一の言語に属する第二の用例との間の類似性を評価した値を導く手段を備え、
前記の手段は、（１）該類似性を評価した値が高いほど類似性が高いとしてその類似性の高い順で、第一の言語

に属する第一の用例に対する第一の言語に属する第二の用例群を第一のデータベースから予め決められた数だけ選択するという第一の方法で選択し、（２）第一の言語に属する用例の、第二の言語に属する用例への対応から、第二の言語に属する第一の用例群を第一のデータベースから選択し、（３）この選択された第二の言語に属する第一の用例群を代表するテンス、アスペクトあるいはモダリティについて、そのテンス、アスペクトあるいはモダリティを個々の用例のテンス、アスペクトあるいはモダリティの多数決で決定するという第二の方法で決定し、（４）この決定されたテンス、アスペクトあるいはモダリティを、第一の言語に属する第二の用例の翻訳のテンス、アスペクトあるいはモダリティとして用いる、という構成を備えることを特徴とする、データベースを利用したテンス、アスペクトあるいはモダリティに

関する翻訳装置。

【請求項２】　第一の言語から第二の言語へのデータベースを利用した翻訳装置で、第一の言語に属する複数の用例と第二の言語に属する複数の用例からなり且つ個々の第一の言語に属する用例は、第二の言語に属する用例との間に少なくとも一つ以上の対応付があり、テンス、アスペクトあるいはモダリティ情報が付加されたことを特徴とする第一のデータベースを備え、

第一の言語に属する第一の用例と、形態素解析を行なって形態素を認識し、各形態素についてシソーラスの分類番号を付して、シソーラスの分類番号による構成に変換された該第一の言語に属する第一の用例と、同様に変換された第一のデータベース内の第一の言語に属する第二の用例との間の、文末からみて連続する共通の文字列の数を用いることを特徴とする第一の言語に属する第二の用例との間の類似性を評価した値を導く手段を備え、

前記の手段は、（１）該類似性を評価した値が高いほど類似性が高いとしてその類似性の高い順で、第一の言語に属する第一の用例に対する第一の言語に属する第二の用例群を第一のデータベースから予め決められた数だけ選択するという第一の方法で選択し、（２）第一の言語に属する用例の、第二の言語に属する用例への対応から、第二の言語に属する第一の用例群を第一のデータベースから選択し、（３）この選択された第二の言語に属する第一の用例群を代表するテンス、アスペクトあるいはモダリティについて、そのテンス、アスペクトあるいはモダリティを個々の用例のテンス、アスペクトあるいはモダリティの多数決で決定するという第二の方法で決定し、（４）この決定されたテンス、アスペクトあるいはモダリティを、第一の言語に属する第二の用例の翻訳のテンス、アスペクトあるいはモダリティとして用いる構成を備えることを特徴とする、データベースを利用したテンス、アスペクトあるいはモダリティに関する翻訳装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、データベースを利用した翻訳装置に関するものである。特に、時制などが文末に表現される言語、例えば日本語、から他の言語、例えば英語、に翻訳する際に問題となる文末表現のテンス（時制）、アスペクト（相）あるいはモダリティ（様相）を翻訳する時に用いる。

【０００２】

【従来の技術】従来の技術の例を、データベースを用いた翻訳装置における翻訳方法の従来例を第一の従来技術に、用例を基にしたデータベースを用いた翻訳装置での翻訳方法の従来例を第二の従来技術に、用例間の類似性を見る方法として文末から数えた一致文字列の数を用いた従来例を第三の従来技術として以下に示す。

【０００３】まず、データベースを用いた翻訳装置の翻訳方法の第一の従来技術を図２のフローチャートに示す。図２のフローチャートでは、次の四段階の手続を示している。

【０００４】従来の文末表現の日英翻訳は、人手で作成した規則によってなされてきた。このため、まず、次の作業を行う必要があった。

【０００５】１）解析以前に予め、人手による規則集の作成をする。例えば、連用形＋動詞「いる」ならば、アスペクトが「進行相」となる。このような規則を、他の組み合わせに対しても作成し、規則集を作成する。また、テンス（時制）、アスペクト（相）あるいはモダリティ（様相）についても、その規則集を作成する。この様な人手により規則集を作成する場合は、規則の不備が残ってしまい、常にメンテナンスを続けて洗練化する必要がある。

【０００６】次に解析作業として、

２）解析における手続１入力文の翻訳のための入力文の形態素解析や構文解析を行う。例えば「希望をいだいている。」が入力文の場合は、下記のような結果を得る。

希望　＜名詞＞を　＜助詞＞いだいて　＜動詞＞＜連用形＞いる　＜動詞＞

【０００７】ここで、形態素解析部や構文解析部を変更すると、上記の規則集にも影響があり、適切な翻訳を維持するためには、上記の規則集にも変更すべき点が発生してしまう。

【０００８】３）解析における手続２形態素解析や構文解析の結果と、規則を照合して、テンス（時制）、アスペクト（相）あるいはモダリティ（様相）を確定する。上記の場合、文末表現が、＜連用形＞＋動詞「いる」の形になっているので、予め作成した規則により、「進行相」と確定される。

【０００９】続いて、次の様に文全体を構成するため、合成作業を行う。

４）解析における手続３テンス、アスペクトあるいはモダリティの翻訳以外の部分は、従来の既によく知られた翻訳方法のどの方法を用いてもよく、それらのどれかを用いて翻訳し、テンス、アスペクトあるいはモダリティの翻訳は、上記の方法により翻訳し、これらを合成することにより、文全体の翻訳を完成する。

【００１０】ここで示した人手で作成した規則によって翻訳する方法では、規則のメンテナンスに多大な人的資源を投入する必要性があるという欠点がある。

【００１１】次に用例を基にしたデータベースを用いた翻訳装置での翻訳方法の先行例である第二の従来技術を示す。

【００１２】本発明の方法でも、用例を集めたデータベースを利用しており、用例ベース手法に分類される。この用例を基にした手法を日英翻訳に利用した従来例としては、報告書（ＥｉｉｃｈｉｒｏＳｕｍｉｔａ、Ｈｉｔｏｓｈｉ Ｉｉｄａ、ａｎｄ Ｈｉｄｅｏ Ｋｏｈｙａ

ma、Ｔｒａｎｓｌａｔｉｎｇ ｅｘａｍｐｌｅｓ：Ａ ｎｅｗ ａｐｐｒｏａｃｈ ｔｏｍａｃｈｉｎｅ ｔｒａｎｓｌａｔｉｏｎ、Ｔｈｅ ｔｈｉｒｄ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅ ｏｎＴｈｅｏｒｅｔｉｃａｌ ａｎｄ ＭｅｔｈｏｄｏｌｏｇｉｃａｌＩｓｓｕｅｓｉｎ Ｍａｃｈｉｎｅ Ｔｒａｎｓｌａｔｉｏｎ ｏｆ Ｎａｔｕｒａｌ Ｌａｎｇｕａｇｅ、Ｎｏ.3、（ＴＭＩ、１９９０）、ｐｐ.２０３－２１２）がある。ただし、この技術は課題が「ＡのＢ」であって、「ＡのＢ」の日英翻訳では名詞Ａと名詞Ｂの意味情報を複雑に組み合わせて利用していた。この研究と本発明とは以下の点で異なっている。

【００１３】１）請求項１の第一のデータベースに相当する用いるデータベースの構成、
２）用例間の類似性を評価する方法、
３）ひとつの用例の中で注目する位置。

【００１４】最後に、第三の従来技術として、用例間の類似性を見る方法として文末から数えた一致文字列の数を用いた従来例を示す。

【００１５】本発明でも、文末から数えた一致文字列の数を、類似性を評価した値（類似度）として用いているが、この方法自体には従来例がある。文末の省略表現の補完を行なう研究（村田真樹、長尾真、日本語文章における表層表現と用例を用いた動詞の省略の補完、言語処理学会誌、Ｖｏｌ．５、Ｎｏ．１、（１９９８））があり、文末一致文字列の文字数を類似度とする用例ベース利用を利用して行っている。

【００１６】この従来例と本発明とは、以下の点において異なっている。
１）従来例では、対象とする問題が省略表現の補完であり、本発明の様に、異なる言語間の翻訳ではない。
２）請求項１の第一のデータベースに相当する用いるデータベースの構成が言語及び項目において相異している。

【００１７】以上の従来技術は、本発明の方法とは部分的に一致しているに過ぎず、これらの先行技術を単に組み合わせても、本発明の翻訳方法を容易に思いつくものではない事は明らかである。

【００１８】
【発明が解決しようとする課題】従来のデータベースを利用した翻訳装置における翻訳方法では、従来の文末表現の翻訳は、人手で作成した規則によってなされてきた。しかし、人手で作成した規則によって翻訳する方法では、翻訳精度を向上させるために行う規則のメンテナンスに多大な人的資源を投入する必要性があるという欠点があった。

【００１９】本発明は上記に鑑み提案されたもので、人手による規則集の作成をする必要が無くデータベースを利用した翻訳の知識が無くても翻訳精度の向上を図ることができるデータベースを利用した翻訳装置を提供することを目的とする。

【００２０】
【課題を解決するための手段】上記目的を達成するために本発明で用いる手段を、フローチャートで説明すると図１の様になる。また、以下に本発明のよる方法を簡潔に記述する。

【００２１】１）解析以前に予め、第一の言語の用例と、それに対応する第二の言語の用例を集めたデータベースを作成する。また、この際に第二の言語の用例のテンス、アスペクトあるいはモダリティの分類を付与する。この付与は、人手で行っても良いし、既によく知られている形態素および構文解析システムを補助として用いることも出来る。

【００２２】２）解析における手続１入力文の翻訳のための検索で、文末からの一致文字列が最も長い用例を、上記のデータベースから検索する。検索方法は、よく知られた２分検索の方法を使うことが出来る。

【００２３】３）解析における手続２翻訳の確定で、手続1から取り出した用例の英訳側の動詞部分のテンス、アスペクトあるいはモダリティの分類を、入力文のテンス、アスペクトあるいはモダリティと確定する。

【００２４】４）解析における手続３翻訳文の構成方法で、テンス、アスペクトあるいはモダリティの翻訳以外の部分は、従来の翻訳方法のどの方法を用いてもよく、それらのどれかを用いて翻訳し、テンス、アスペクトあるいはモダリティの翻訳は、上記の方法により翻訳し、これらを合成することにより、文全体の翻訳を完成する。

【００２５】従って、上記目的を達成するために、請求項１に記載の発明は、第一の言語から第二の言語へのデータベースを利用したテンス、アスペクトあるいはモダリティに関する翻訳装置で、第一の言語に属する複数の用例と第二の言語に属する複数の用例からなり且つ個々の第一の言語に属する用例は、第二の言語に属する用例との間に少なくとも一つ以上の対応付があり、テンス、アスペクトあるいはモダリティ情報が付加されたことを特徴とする第一のデータベースを備え、第一の言語に属する第一の用例と、第一のデータベースの第一の言語に属する第二の用例との間の文末からみて連続する共通の文字列の数を用いて第一のデータベースの第一の言語に属する第二の用例との間の類似性を評価した値を導く手段を備え、前記の手段は、（１）該類似性を評価した値が高いほど類似性が高いとしてその類似性の高い順で、第一の言語に属する第一の用例に対する第一の言語に属する第二の用例群を第一のデータベースから予め決められた数だけ選択するという第一の方法で選択し、（２）第一の言語に属する用例の、第二の言語に属する用例への対応から、第二の言語に属する第一の用例群を第一のデータベースから選択し、（３）この選択された第二の言語に属する第一の用例群を代表するテンス、アスペク

トあるいはモダリティについて、そのテンス、アスペクトあるいはモダリティを個々の用例のテンス、アスペクトあるいはモダリティの多数決で決定するという第二の方法で決定し、（４）この決定されたテンス、アスペクトあるいはモダリティを、第一の言語に属する第二の用例の翻訳のテンス、アスペクトあるいはモダリティとして用いる、という構成を備えることを特徴としており、用例を基にした方法でテンス、アスペクトあるいはモダリティを適切に翻訳するものを提案している。

【００２６】また、請求項２に記載の発明は、データベースを利用したテンス、アスペクトあるいはモダリティに関する翻訳装置であり、その類似性を評価する構成の特徴は、形態素解析を行なって形態素を認識し、各形態素についてシソーラスの分類番号を付して、シソーラスの分類番号による構成に変換された該第一の言語に属する第一の用例と、同様に変換された第一のデータベース

内の第一の言語に属する第二の用例との間に、文末からみて連続する共通の文字列の数を用いることを特徴とする、第一の言語に属する第二の用例との間の類似性を評価した値を導く手段を備えている。

【００２７】

【発明の実施の形態】以下にこの発明の実施の形態を詳細に説明する。先ず第１の実施形態を、表１を用いて説明する。

【００２８】今「彼は私の知り合いだ」の時制を翻訳することを考える。このとき日英の翻訳対を大量に集めたデータベースに対して「彼は私の知り合いだ」と文末からの文字列一致が多い用例を上位から１０個集めたものが表１のものだったとする。

【００２９】

【表１】

		日　本　語　文	分類	英　語　文
入力文		彼は私の知り合いだ	現在	I am acquainted with him.
番号	類似度	用　例　デ　ー　タ		
1	6	彼とは長年の知り合いだ	現在完了	I have known him for a long time.
2	6	ふたりは長い間の知り合いだ	現在	The two are acquaintances of long standing.
3	1	彼とは１０年余の顔見知りだ	現在完了	I have known him for over ten years.
4	1	彼らは多年の知己だ	現在	They are friends of many years' standing.
5	1	彼はこのクラブの恩人だ	現在	He is a benefactor of this club.
6	1	彼は私の命の恩人だ	現在	I owe him my life.
7	1	彼はかたい人だ	現在	He is reliable.
8	1	彼はだれにも人当たりのいい人だ	現在	He is affable to everybody.
9	1	なんと男振りのいい人だ	現在	He is a gentle-mannered person.
9	1	なんと男振りのいい人だ	現在	What a handsomelooking man he is!

【００３０】表の類似度は文末からの一致文字列の数を示している。また、ここでｋ近傍法を用いる。ｋ近傍法とは１個の最も類似した用例を用いるかわりに、類似度の上位から順に取り出したｋ個の用例の多数決により求める方法である。

【００３１】類似度が等しい用例がある場合はｋの値に関わらず類似度が等しい用例はすべて用いて多数決を行なう必要がある。さらに、ここでは処理の簡単のため、用例は多くても１０個しか調べないこととする。

【００３２】また、上記の表１のうち、分類の欄は英語文の該当する動詞句より求まるものであるがこの部分はよく知られた処理プログラムを用いて自動で行なっても良いし、データベースを作成する際に人手であらかじめ分類を記入しておいてもよい。

【００３３】まず、ｋ＝１の場合を考える。このとき最も類似度の大きい１個の用例を用いて解析するわけだが、ここでは１番と２番が同じ類似度のため、１と２番の用例を用いて解析を行なう。これで多数決を行なうと分類は「現在完了」が１、「現在」が１と意見がわかれ、意見がわかれたときには先にあがった分類を解とすると決めておくと、解は先に上がった「現在完了」となり、不正解となる。

【００３４】次に、ｋ＝３の場合を考える。このとき最も類似度の大きい３個を選ぶわけだが、３番の用例以降はすべて類似度が等しいので、１０個すべての用例を用いることになる。これで多数決を行なうと分類は「現在完了」が２、「現在」が８と意見は分かれるが、数の大きい「現在」となり、これは正解の「現在」と一致し正解となる。

【００３５】次に、ｋ＝５、７、９の場合も同様に１０個の用例すべてが用いられ解は「現在」となり、これも正解となる。

【００３６】この問題ではシステムは、ｋ＝１のとき、誤った解を出力し、ｋ＝３、５、７、９のときに正しい解を出力するということになる。ｋの値については装置を実際に作成する時に適切なものを選択するとよい。この方法によるｋの値は通常、多数決の都合上、奇数が望ましく、さらに３あるいは５で充分な場合が多い。データベースの用例が増えるに従って、より小さいｋの値を用いる事ができる。

【００３７】次に第２の実施形態を、表２を用いて説明する。今「彼は私の知り合いだ」の時制を翻訳することを考える。このとき日英の翻訳対を大量に集めたデータベースに対して「彼は私の知り合いだ」と文末からの文字列一致が多い用例を上位から１０個集めたものが次の表２のものだったとする。

【００３８】

【表２】

		日　本　語　文	分類	英　語　文
入力文		彼は私の知り合いだ	現在	I am acquainted with him.
番号	類似度	用　例　デ　ー　タ		
１	25	彼とは長年の知り合いだ	現在完了	I have known him for a long time.
２	24	ふたりは長い間の知り合いだ	現在	The two are acquaintances of long standing.
３	11	彼とは１０年余の顔見知りだ	現在完了	I have known him for over ten years.
４	11	彼らは多年の知己だ	現在	They are friends of many years' standing.
５	10	彼はこのクラブの恩人だ	現在	He is a benefactor of this club.
６	10	彼は私の命の恩人だ	現在	I owe him my life.
７	10	彼はかたい人だ	現在	He is reliable.
８	10	彼はだれにも人当たりのいい人だ	現在	He is affable to everybody.
９	10	なんと男振りのいい人だ	現在	He is a gentle—mannered person.
９	10	なんと男振りのいい人だ	現在	What a handsomelooking man he is!

【００３９】表２の類似度は、入力文の形態素解析を行なって形態素を認識し、各形態素についてシソーラスの分類番号を付して、シソーラスの分類番号による構成に変換された入力文を用意し、また日英の翻訳対を大量に集めたデータベースに対しても同様な変換を行ったものを用意し、これらの変換された後の文について、文末からみて連続する共通の文字列の数を示している。

【００４０】解析は、第１の実施形態と同じくｋ近傍法を用いることにする。

【００４１】まず、ｋ＝１の場合を考える。このとき最も類似度の大きい１番の用例だけを用いて解析を行なう。１番の用例は分類が「現在完了」なので正解の分類「現在」と異なり、不正解となる。

【００４２】次に、ｋ＝３の場合を考える。このとき最

も類似度の大きい３個を選ぶわけだが、３番の用例と４番の用例の類似度が等しいので、４番の用例までの四つの用例を用いることになる。これで多数決を行なうと分類は「現在完了」が２、「現在」が２と意見がわかれ、解は先に上がった「現在完了」となり、これもまた不正解となる。

【００４３】次に、ｋ＝５の場合を考える。このとき最も類似度の大きい５個を選ぶわけだが、５番の用例以降はすべて類似度が等しいので、１０個すべての用例を用いることになる。これで多数決を行なうと分類は「現在完了」が２、「現在」が８と意見はわかれるが、数の大きい「現在」となり、これは正解の「現在」と一致し正解となる。

【００４４】次に、ｋ＝７、９の場合も同様に１０個の用例すべてが用いられ解は「現在」となり、これも正解となる。

【００４５】この問題ではシステムは、ｋ＝１、３のとき、誤った解を出力し、ｋ＝５、７、９のときに正しい解を出力するということになる。ｋの値については装置を実際に作成する時に適切なものを選択するとよい。この方法によるｋの値は通常、多数決の都合上、奇数が望ましく、さらに７あるいは９で充分な場合が多い。この場合も、データベースの用例が増えるに従って、より小さいｋの値を用いる事ができる。

【００４６】上記の実施形態に示されるように、本発明

の方法では、用例を集めたデータベースを整備して行くことによって翻訳精度の向上を図ることができ、従って、人手による規則集の作成をする必要が無くメンテナンスが容易であり、また、データベースを利用した翻訳の知識が無くても翻訳精度の向上を図ることができる。

【００４７】

【発明の効果】この発明は上記した構成からなるので、以下に説明するような効果を奏することができる。

【００４８】請求項１に記載の発明では、用例を基にした翻訳が可能となり、人手による規則集の作成をする必要が無くメンテナンスが容易であり、また、文末からみて、連続する共通の文字列の数であることとすることにより、簡単に類似性を評価することが出来、データベースを利用した翻訳の知識が無くても翻訳精度の向上を図ることができるようになった。

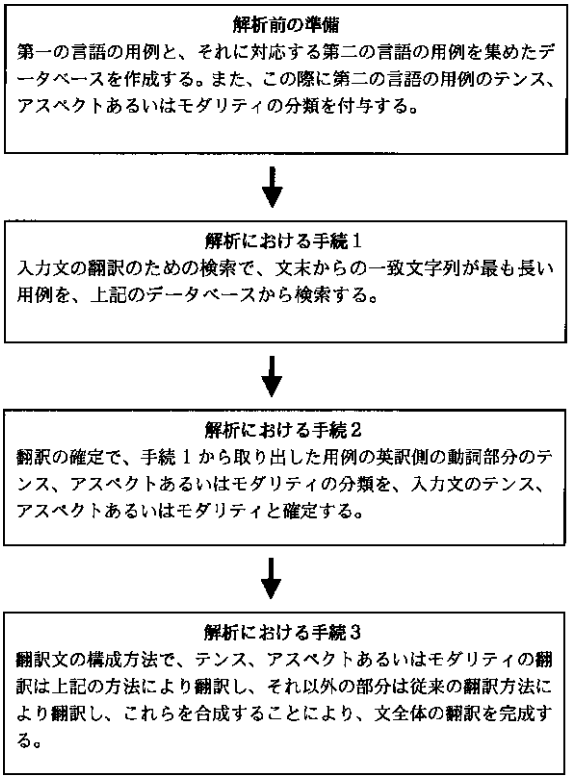
【００４９】さらに、請求項２に記載の発明では、意味上の類似性を用いて類似性を評価することが出来るようになり、意味上からも適切な翻訳ができるようになった。

【図面の簡単な説明】

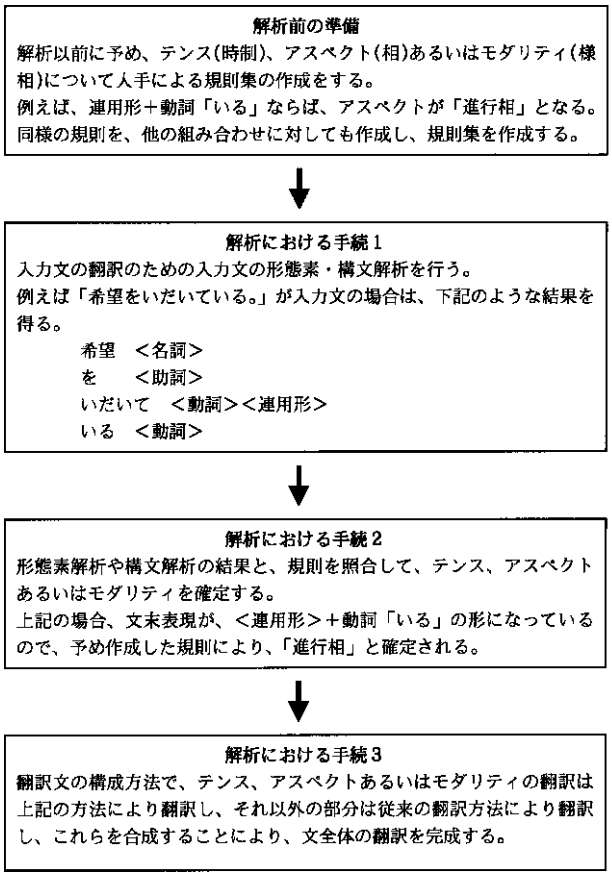
【図１】本発明における課題を解決するための手段を示すためのフローチャートである。

【図２】従来の技術における手段を示すためのフローチャートである。

【図１】



【図２】



フロントページの続き

(72)発明者　井佐原　均
兵庫県神戸市西区岩岡町岩岡558－２
郵政省通信総合研究所関西支所内

(56)参考文献　特開　平６－309352（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06F　17/21 - 17/28
ＪＩＣＳＴファイル（ＪＯＩＳ）