(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2003－16067

（Ｐ２００３－１６０６７Ａ）

(43)公開日　平成15年１月17日(2003.1.17)

(51)Int.Cl.[7]	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ (参考)
G06F 17/28		G06F 17/28	Z 5B091
			C

審査請求　有　請求項の数２　ＯＬ　（全14頁）

(21)出願番号　特願2001－201010(Ｐ2001－201010)

(22)出願日　平成13年７月２日(2001.7.2)

特許法第30条第１項適用申請有り　2001年３月９日　社団法人電子情報通信学会発行の「電子情報通信学会技術研究報告　信学技報Ｖｏｌ.100，Ｎｏ.698」に発表

(71)出願人　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ー２－１

(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内

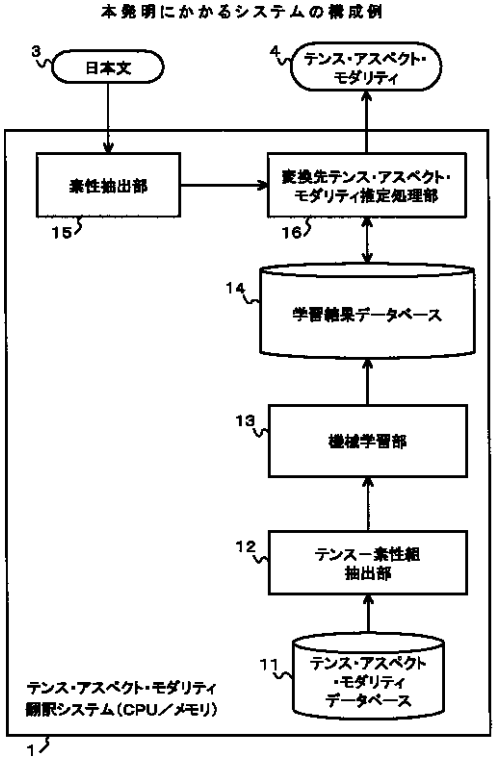
(72)発明者　馬　青
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内

(74)代理人　100119161
弁理士　重久　啓子　（外２名）

最終頁に続く

(54)【発明の名称】テンス・アスペクト・モダリティ翻訳処理方法，テンス・アスペクト・モダリティ翻訳システム

(57)【要約】
【課題】　機械翻訳の際に変換先のテンス・アスペクト・モダリティを精度よく翻訳する。
【解決手段】　テンス－素性組抽出部12は，テンス・アスペクト・モダリティ・データベース11から，事例ごとにテンス・アスペクト・モダリティ(以下「テンス等」という) と素性の集合の組とを抽出する。機械学習部13は，抽出した組からどのような素性のときにどのようなテンス等となりやすいかを学習し，その結果をデータベース14に保存する。素性抽出部15は，訳出したい日本文3 が入力されると素性の集合を抽出する。変換先テンス・アスペクト・モダリティ推定処理部16は，学習結果データベース14を参照し，その素性の集合から日本文3 の素性の場合にどのようなテンス等になりやすいかを推定し，推定したテンス等4 を出力する。



1

【特許請求の範囲】

【請求項１】　コンピュータにより一の言語から他の言語へ翻訳処理をする際に，変換元言語から変換先言語のテンス・アスペクト・モダリティを翻訳する方法であって，予め備えられた変換元言語の事例と当該事例のテンス・アスペクト・モダリティとを記憶するデータベースから，テンス・アスペクト・モダリティと当該事例または当該事例に関連するデータから抽出した，複数の形式の素性の集合とからなるテンス－素性組を事例ごとに抽出する過程と，前記テンス－素性組のうちのすべてまたはいくつかの素性を用いて，機械学習法により変換先言語のテンス・アスペクト・モダリティを判定するための学習データを作成し保存する過程と，変換元言語の入力文から，当該入力文の素性の集合を抽出する過程と，前記入力文の素性の集合のうちのすべてまたはいくつかの素性をもとに，前記学習データを参照して前記入力文のテンス・アスペクト・モダリティを推定する過程とを備えることを特徴とするテンス・アスペクト・モダリティ翻訳処理方法。

【請求項２】　請求項１に記載のテンス・アスペクト・モダリティ翻訳処理方法において，前記機械学習法として，決定リスト法，最大エントロピー法，またはサポートベクトルマシン法のいずれか一の手法を用いることを特徴とするテンス・アスペクト・モダリティ翻訳処理方法。

【請求項３】　コンピュータにより一の言語から他の言語へ翻訳処理をする際に，変換元言語から変換先言語のテンス・アスペクト・モダリティを翻訳するシステムであって，予め備えられた変換元言語の事例と当該事例のテンス・アスペクト・モダリティとを記憶するデータベースから，テンス・アスペクト・モダリティと当該事例または当該事例に関連するデータから抽出した，複数の形式の素性の集合とからなるテンス－素性組を事例ごとに抽出するテンス－素性組抽出手段と，前記テンス－素性組のうちのすべてまたはいくつかの素性を用いて，機械学習法により変換先言語のテンス・アスペクト・モダリティを判定するための学習データを作成し保存する機械学習手段と，変換元言語の入力文から，当該入力文の素性の集合を抽出する素性抽出手段と，前記入力文の素性の集合のうちのすべてまたはいくつかの素性をもとに，前記学習データを参照して前記入力文のテンス・アスペクト・モダリティを推定する変換先テンス・アスペクト・モダリティ推定処理手段とを備えることを特徴とするテンス・アスペクト・モダリティ翻訳システム。

【請求項４】　請求項３に記載のテンス・アスペクト・モダリティ翻訳システムにおいて，前記機械学習部では，決定リスト法，最大エントロピー法，またはサポートベクトルマシン法のいずれか一の手法を用いて学習することを特徴とするテンス・アスペクト・モダリティ翻訳システム。

2

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は，コンピュータによる翻訳システムの分野において，特に，機械学習法を用いてテンス（時制），アスペクト（相），またはモダリティ（様相）を翻訳する翻訳処理方法および翻訳システムに関するものである。

【０００２】

【従来の技術】近年，WWW（World Wide Web）などのインターネットの発展とともに機械翻訳の必要性が高まり続けている。この機械翻訳において，テンス・アスペクト・モダリティは翻訳が難しい問題として知られている。

【０００３】テンス・アスペクト・モダリティとは，動詞部分の時制（テンス），進行形や完了形などの相（アスペクト），または英文における助動詞相当句である様相（モダリティ）についての情報である。

【０００４】従来，テンス・アスペクト・モダリティの表現は，人手により作成されたルールによって扱われていた。しかしながら，近年では，用例ベース（ｋ近傍法）の方法などのコーパスデータにもとづくアプローチでも処理されるようになってきた。用例ベースの方法では，集めた実例ごとに，どの場合にどの時制などを使うかを記したデータを対応づけた用例データベースを用意しておき，入力された文と良く似た実例に基づいてテンス・アスペクト・モダリティを翻訳するといったことが行なわれていた。［参考文献１］村田真樹　馬青　内元清貴　井佐原均，用例ベースによるテンス・アスペクト・モダリティの日英翻訳，人工知能学会誌，Vol.16，No.1，2001参考文献１に記載されている研究では，日本文から英文への機械翻訳のテンス・アスペクト・モダリティの判定の際に，日本文のテンス・アスペクト・モダリティは文末に表されることに着目して，入力された日本文の文末の所定の長さの文字列と，予め用意したコーパスデータとの類似度をｋ近傍法により判断してテンス・アスペクト・モダリティを決定する手法を用いている。ｋ近傍法とは，最もよく似た一つの事例の代わりに，最もよく似たｋ個の事例を用い，このｋ個の事例での多数決によって分類先を求める手法である。

【０００５】

【発明が解決しようとする課題】しかし，人手でルールを記述し，このルールをもとにテンス・アスペクト・モダリティを分類する方法では，人的資源の問題や，人手による作業の精度などの問題がある。

【０００６】また，入力文とよく似た実例を使う手法では，入力文と実例の類似度を定義する必要があり，例えば文末の文字列のように類似度を定義することができるような平易な情報しか扱うことができなかった。そのため，参考文献１に記載された研究の手法において，文末の文字列の情報のみによってテンス・アスペクト・モダ

リティの分類を判定することで，判定結果の精度が低くなる場合が生じる。

【０００７】例えば，実例データ「もう行きました。」のテンス・アスペクト・モダリティが「過去完了」である場合に，「昨日行きました。」という文が入力されたとする。この入力文の正しいテンス・アスペクト・モダリティは「過去」であるにもかかわらず，文字列「ました＄（＄＝文末）」の表示の類似度から，実例データと同様に「過去完了」と判定されてしまう場合がある。

【０００８】したがって，参考文献１の研究の手法のように文末の一致する文字列だけでなく，例えば，この場合の「昨日」のように，文末の文字列とは異なる形式の情報を合わせて用いることが有効であると考えられる。

【０００９】しかし，テンス・アスペクト・モダリティを解析するための情報（素性）として，形態素情報（形態素素性），意味解析情報（単語素性など），構文解析情報（構文解析素性）などの異なる形式の素性を組み合わせて用いることが有効であるとしても，参考文献１の研究で用いたｋ近傍法のような類似度を定義する必要がある判定手法では，複数の形式の素性を用いることができないという問題があった。

【００１０】本発明は，上記問題点の解決を図り，変換元言語の事例やその事例に関連するデータから抽出した異なる形式の素性を取り扱うことができる機械学習手法を用いて，どのような素性の場合にどのようなテンス・アスペクト・モダリティになるかを学習し，その学習結果を用いて入力文の変換先のテンス・アスペクト・モダリティを精度よく翻訳できる手段を提供することを目的とする。

【００１１】

【課題を解決するための手段】上記課題を解決するため，本発明に係る方法は，予め備えられた変換元言語の事例と当該事例のテンス・アスペクト・モダリティとを記憶するデータベースから，テンス・アスペクト・モダリティと当該事例または当該事例に関連するデータから抽出した，複数の形式の素性の集合とからなるテンス－素性組を事例ごとに抽出する過程と，前記テンス－素性組のうちのすべてまたはいくつかの素性を用いて，機械学習法により変換先言語のテンス・アスペクト・モダリティを判定するための学習データを作成し保存する過程と，変換元言語の入力文から，当該入力文の素性の集合を抽出する過程と，前記入力文の素性の集合のうちのすべてまたはいくつかの素性をもとに，前記学習データを参照して前記入力文のテンス・アスペクト・モダリティを推定する過程とを備えることを特徴としている。

【００１２】本発明では，従来の手法のように，文末の文字列のように単一の種類の素性だけを用いてテンス・アスペクト・モダリティの解析を行うのでなく，文字列の他，一文全体の形態素素性，意味的素性，構文的素性，前文のテンス・アスペクト・モダリティ，または対訳データの該当する構成部分データなど，二以上の異なる形式の素性を任意に用いて解析処理を行う点が，従来の手法と異なる。

【００１３】また，本発明では，多くの形式の素性を自由に用いることができる，類似度を設定する必要のない種々の機械学習手法を用いて解析処理を行う点が，類似度の定義を必要とする従来のｋ近傍法のような手法による判定と異なる。

【００１４】以上の本発明に係る処理方法またはシステムは，その処理過程や手段，構成，要素をコンピュータに実行させるプログラムによっても実現することができる。このプログラムは，コンピュータが読み取り可能な，可搬媒体メモリ，半導体メモリ，ハードディスクなどの適当な記録媒体に格納することができ，これらの記録媒体に記録して提供され，または通信インタフェースを介して種々の通信網を利用した送受信により提供される。

【００１５】

【発明の実施の形態】以下に，本発明の実施の形態を図を用いて説明する。本実施の形態では，日本文から英文への翻訳に際しテンス・アスペクト・モダリティを翻訳する場合を例とする。

【００１６】図１に，本発明のシステム構成例を示す。図１中，１は本発明に係るテンス・アスペクト・モダリティ翻訳システムを表す。テンス・アスペクト・モダリティ翻訳システム１はＣＰＵ，メモリなどで構成され，テンス・アスペクト・モダリティ・データベース１１，テンス－素性組抽出部１２，機械学習部１３，学習結果データベース１４，素性抽出部１５，変換先テンス・アスペクト・モダリティ推定処理部１６を持つ。

【００１７】テンス－素性組抽出部１２は，予め用意しておいたテンス・アスペクト・モダリティ用コーパスであるテンス・アスペクト・モダリティ・データベース１１から，事例ごとに，テンス・アスペクト・モダリティと事例の素性の集合との組を抽出する手段である。

【００１８】機械学習部１３は，テンス－素性組抽出部１２で抽出されたテンス・アスペクト・モダリティと素性の集合との組から，どのような素性のときに，どのようなテンス・アスペクト・モダリティになりやすいかを機械学習法により学習し，その学習結果を学習結果データベース１４に保存する手段である。

【００１９】素性抽出部１５は，入力された日本文３から素性の集合を抽出し，それらを変換先テンス・アスペクト・モダリティ推定処理部１６へ渡す手段である。

【００２０】変換先テンス・アスペクト・モダリティ推定処理部１６は，学習結果データベース１４を参照して，渡された素性の集合の場合に，変換先の言語においてどのようなテンス・アスペクト・モダリティになりやすいかを推定し，日本文３の変換先のテンス・アスペクト・モダリティ４を出力する手段である。

【００２１】本発明の処理の流れの概略を説明する。図２は，図１に示すシステムの処理フローチャートである。

【００２２】図２に示す処理を開始する前に，テンス・アスペクト・モダリティ・データベース１１を予め用意しておく。テンス・アスペクト・モダリティ・データベース１１は，機械翻訳用の日英の対訳コーパスであり，日本語と英語の対訳データにテンス・アスペクト・モダリティの情報が付与されている。

【００２３】日英の対訳データに付与するテンス・アスペクト・モダリティの分類として，例えば以下のものを用いる。以下の分類は，対訳の英語文の動詞がどのような形になっているかによって定められる。

（１）各助動詞相当語句（be able to，be going to，can，have to，had better，may，must，need，ought，shall，used to，will の12種類）がつくかどうかと，{現在形，過去形} と {進行形，進行形でない} と {完了，完了でない} のすべての組み合わせ（助動詞相当語句が複数つく場合も許している。）：2^{15}種類

（２）命令形（１種類）

（３）名詞句（１種類）

（４）分詞構文（１種類）

（５）動詞省略（１種類）

（６）間投詞，挨拶文など（１種類）

（７）日本語と英語で動詞の対応がとれない場合（１種類）

（８）作業不可（１種類）

ただし，上記の分類のうち，「（３）名詞句」から「（８）作業不可」までの６つの分類はテンス・アスペクト・モダリティの分類としては扱う必要がないか，もしくはテンス・アスペクト・モダリティの翻訳を行なう必要がないと思われるので，本形態では省略している。

【００２４】また，これらの分類は，「英語の主節の動詞部分」と「日本語の主節の動詞に対応する英語の動詞部分」の二か所にふられる。しかし，日英翻訳において日本語のテンス・アスペクト・モダリティに対応するのは「英語の主節の動詞部分」であろうと考えられるので，本発明に係るテンス・アスペクト・モダリティ翻訳システム１では，日本文３を与えて変換先の「英語の主節の動詞部分」のテンス・アスペクト・モダリティの分類を推定し，推定結果であるテンス・アスペクト・モダリティ４を出力することとしている。

【００２５】ステップＳ１：まず，テンス－素性組抽出部１２により，用意されたテンス・アスペクト・モダリティ・データベース１１から，各事例ごとに，テンス・アスペクト・モダリティと事例の素性の集合との組を抽出する。

【００２６】テンス－素性組抽出部１２では，素性の集合として，文字列素性，形態素素性，単語素性，構文的素性，一前文のテンス・アスペクト・モダリティの情報，英文対訳データの動詞部分など，種々の形式の素性のうち，所定の素性を抽出することができる。

【００２７】図３に抽出する素性の集合とテンス・アスペクト・モダリティの組を示す。図３に示すように，テンス－素性組抽出部１２により，テンス・アスペクト・モダリティ・データベース１１の事例「もう登録しました。」から，テンス・アスペクト・モダリティと，文字列素性「もう登録しました＄」，「う登録しました＄」，・・・，「た＄」，単語素性「もう」，「登録」「し」「まし」「た」などの素性の集合との組を抽出する。なお，ここでは，抽出した文字列素性には入力された文全体の形態素列と区別できるように末尾に＄をつけている。また，文末表現の正規化のため，句点などは消している。

【００２８】ステップＳ２：続いて，機械学習部１３により，抽出されたテンス・アスペクト・モダリティと素性の集合との組から，どのような素性のときにどのようなテンス・アスペクト・モダリティになりやすいかを機械学習し，その学習結果を学習結果データベース１４に保存する。

【００２９】機械学習では，例えば，所定の長さの文末の文字列素性，事例の全文の形態素素性，単語素性のうち，いくつかの素性を用いて処理を行ってもよい。

【００３０】機械学習の手法は，種々の形式の素性の集合を扱うことができるような機械学習法であればよく，例えば，以下に示すような決定リスト法，最大エントロピー法，サポートベクトルマシン法などを用いる。

【００３１】（１）決定リスト法

決定リスト法は，あらかじめ設定しておいた素性f_j（∈Ｆ，$1 \leqq j \leqq k$）のうち，いずれか一つの素性のみを文脈として各分類の確率値を求めて，その確率値が最も大きい分類を求める分類とする方法である。

【００３２】ある文脈ｂで分類ａを出力する確率は以下の式によって与えられる。

【００３３】

$$p(a \mid b) = p(a \mid f_{max}) \qquad (1)$$

ただし，f_{max} は以下の式によって与えられる。

【００３４】

【数１】

$$f_{max} = argmax_{f_j \in F} \; max_{a_i \in A} \; \tilde{p}(a_i|f_j) \qquad (2)$$

【００３５】また，

$\tilde{p}(a_i|f_j)$

【００３６】

【数２】

【００３７】は，素性f_jを文脈に持つ場合の分類a_i

の出現の割合である。

【００３８】具体的には，各素性ごとに，どのようなテンス・アスペクト・モダリティの分類になるのかの確率を求めておき，入力文のすべての素性のうち最大確率の素性の分類を用いてテンス・アスペクト・モダリティの分類を推定する。

【００３９】決定リスト法にもとづく分類は簡便ではあるが，ある一つの素性のみを文脈としてテンス・アスペクト・モダリティの分類の推定を行なうので，機械学習の手法としては少々貧弱なものとなっている。

【００４０】（２）最大エントロピー法

最大エントロピー法は，あらかじめ設定しておいた素性f_j（$1 \leqq j \leqq k$）の集合をＦとするとき，以下の式（３）を満足しながらエントロピーを意味する式（４）を最大にするときの確率分布ｐ（ａ，ｂ）を求め，その確率分布にしたがって求まる各分類の確率のうち，もっとも大きい確率値を持つ分類を求める分類とする方法である。

【００４１】

【数３】

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b) = \sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b) \qquad (3)$$

$$for\ \forall f_j\ (1 \leq j \leq k)$$

$$H(p) = -\sum_{a\in A,b\in B} p(a,b)\ log\,(p(a,b)) \qquad (4)$$

【００４２】ただし，Ａ，Ｂは分類と文脈の集合を意味し，g_j（ａ，ｂ）は文脈ｂに素性f_jがあって，なおかつ分類がａの場合１となり，それ以外で０となる関数を意味する。また，

【００４３】

【数４】

$$\tilde{p}(a,b)$$

【００４４】は，既知データでの（ａ，ｂ）の出現の割合を意味する。

【００４５】式（３）は確率ｐと出力と素性の組の出現を意味する関数ｇをかけることで出力と素性の組の頻度の期待値を求めることになっており，右辺の既知データにおける期待値と，左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として，エントロピー最大化(確率分布の平滑化) を行なって，出力と文脈の確率分布を求めるものとなっている。最大エントロピー法の詳細については，以下の参考文献２および参考文献３を参照されたい。

［参考文献２］Eric Sven Ristad, Maximum Entropy Modeling for Natural Language,(ACL/EACL Tutorial Program,Madrid,,1997)

［参考文献３］Eric Sven Ristad, Maximum Entropy Modeling Toolkit, Release 1.6beta,(http://www.mnemonic.com/software/memt,1998)

本手法の利用は，もともと訳出対象となる入力データの素性と同一な素性が既知のコーパス中に多数存在することは稀であり，コーパスでの素性の出現確率をそのまま使用することはできないが，等価な状態は等価な確率を持つことを前提に確率分布を平滑化すれば，コーパスにおける素性出現の期待値を用いることができると考えられることによる。

【００４６】（３）サポートベクトルマシン法

サポートベクトルマシン法は，空間を超平面で分割することにより，２つの分類からなるデータを分類する手法である。図４にサポートベクトルマシン法のマージン最大化の概念を示す。図４において，白丸は正例，黒丸は負例を意味し，実線は空間を分割する超平面を意味し，破線はマージン領域の境界を表す面を意味する。図４（Ａ）は，正例と負例の間隔が狭い場合（スモールマージン）の概念図，図４（Ｂ）は，正例と負例の間隔が広い場合（ラージマージン）の概念図である。

【００４７】このとき，2 つの分類が正例と負例からなるものとすると，学習データにおける正例と負例の間隔（マージン）が大きいものほどオープンデータで誤った分類をする可能性が低いと考えられ，図４（Ｂ）に示すように，このマージンを最大にする超平面を求めそれを用いて分類を行なう。

【００４８】基本的には上記のとおりであるが，通常，学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や，超平面の線形の部分を非線型にする拡張（カーネル関数の導入）がなされたものが用いられる。この拡張された方法は，以下の識別関数を用いて分類することと等価であり，その識別関数の出力値が正か負かによって二つの分類を判別することができる。

【００４９】

【数５】

$$f(\mathbf{x}) = sgn\left(\sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x}) + b\right) \quad (5)$$

$$b = -\frac{max_{i,y_i=-1}b_i + min_{i,y_i=1}b_i}{2}$$

$$b_i = \sum_{j=1}^{l}\alpha_j y_j K(\mathbf{x}_j,\mathbf{x}_i)$$

【００５０】ただし，ｘは識別したい事例の文脈（素性の集合）を，x_i とy_j （ｉ＝１，…，ｌ，$y_j \in \{1, -1\}$）は学習データの文脈と分類先を意味し，関数ｓｇｎは，

ｓｇｎ（ｘ）＝１（ｘ≧０）　　（６）

－１（otherwise）

であり，また，各α_i は式（８）と式（９）の制約のもと式（７）を最大にする場合のものである。

【００５１】

【数６】

$$L(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x_i},\mathbf{x_j}) \quad (7)$$

$$0 \le \alpha_i \le C \ (i = 1,...,l) \quad (8)$$

$$\sum_{i=1}^{l}\alpha_i y_i = 0 \quad (9)$$

【００５２】また，関数Ｋはカーネル関数と呼ばれ，様々なものが用いられるが，本形態では以下の多項式のもの

Ｋ（ｘ，ｙ）＝（ｘ・ｙ＋１）^{d}　　（１０）

を用いる。

Ｃ，ｄは実験的に設定される定数である。後述する具体例ではＣはすべての処理を通して１に固定した。また，ｄは，１と２の二種類を試している。ここで，$\alpha_i > 0$ となるx_i は，サポートベクトルと呼ばれ，通常，式（５）の和をとっている部分はこの事例のみを用いて計算される。つまり，実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない。

【００５３】

【００５４】なお，拡張されたサポートベクトルマシン法の詳細については，以下の参考文献４および参考文献５を参照されたい。

［参考文献４］Nello Cristianini and John Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods,(Cambridge University Press,2000)

［参考文献５］Taku Kudoh, Tinysvm:Support Vector machines,(http://cl.aist-nara.ac.jp/taku-ku//software/Tiny SVM/index.html,2000)

サポートベクトルマシン法は，分類の数が２個のデータを扱うもので，通常これにペアワイズ手法を組み合わせて用いることで，分類の数が３個以上のデータを扱うことになる。

【００５５】ペアワイズ手法とは，Ｎ個の分類を持つデータの場合，異なる二つの分類先のあらゆるペア（Ｎ（Ｎ－１）／２個）を作り，各ペアごとにどちらがよいかを2値分類器（ここではサポートベクトルマシン法によるもの）で求め，最終的にＮ（Ｎ－１）／２個の２値分類器の分類先の多数決により，分類先を求める方法である。

【００５６】本形態における２値分類器としてのサポートベクトルマシンは，サポートベクトルマシン法とペアワイズ手法を組み合わせることによって実現するものであり，以下の参考文献６により工藤氏が作成したＴｉｎｙＳＶＭを利用している。

［参考文献６］工藤拓　松本裕治，Support vector machineを用いたchunk 同定，自然言語処理研究会，2000-NL-140,(2000)

具体的には，テンス・アスペクト・モダリティの各分類ごとに最大マージンの超平面を求めておく。そして，入力された未知の素性が，例えばテンスが現在（正例）であるか過去（負例）であるかなどについて，それぞれについて正例の領域と負例の領域のどちらに属するかを判定していき，その判定結果の多数決により，最終的にテンス・アスペクト・モダリティを推定する。

【００５７】ステップＳ３：テンス・アスペクト・モダリティを訳出したい日本文３が入力される。

【００５８】ステップＳ４：素性抽出部１５では，テン

11

ス－素性組抽出部１２での処理とほぼ同様に，入力された日本文３から素性の集合を取り出し，それらを変換先テンス・アスペクト・モダリティ推定処理部１６へ渡す。図５に，抽出した素性の集合の例を示す。入力された日本文「もう行きました。」から，文字列素性「もう行きました＄」，「う行きました＄」，・・・，「た＄」および単語素性「もう」，「行き」，「まし」，「た」などが抽出される。

【００５９】ステップＳ５：変換先テンス・アスペクト・モダリティ推定処理部１６では，学習結果データベース１４をもとに，渡された素性の集合の場合にどのようなテンス・アスペクト・モダリティになりやすいかを特定し，特定したテンス・アスペクト・モダリティ４を出力する。例えば，「過去，完了」，「過去，完了，進行形」，「現在，shall 付」，「過去；be able to付」などのデータを出力する。

【００６０】入力された日本文「もう行きました。」の素性の集合について，学習結果データベース１４に記憶された学習結果を使用すると「過去完了」でよいとわかるので，テンス・アスペクト・モダリティ４として「過去完了」を出力する。この場合に，従来のように単純に文末の文字列だけを用いて推定したときは，「ました＄」の表現の一致からテンスが「現在完了」と推定されてしまう。しかし，本発明では，文字列以外に全文の単語素性のうちのいくつかの素性を用いて学習した結果を参照することにより「過去完了」と正しく推定できる。

【００６１】また，入力された日本文３が「昨日行きました」である場合には，同様に，従来の手法では「ました＄」の表現の一致からテンスが「現在完了」と推定されてしまう。しかし，学習結果データベース１４に「昨日登録しました。」にテンスが「過去」であるという学習結果が記憶されていれば，「昨日」という単語素性をもとに正しく「過去」と推定することができる。

【００６２】以上では，主に素性の集合として形態素素性と単語素性を用いた場合を例に説明したが，テンス－素性組抽出部１２では，形態素素性や単語素性だけでなく，一前文（事例）のテンス・アスペクト・モダリティを素性として抽出してもよい。これは，テンス（時制）は継続しやすいという性質を利用するものである。すなわち，一前文に付与されたテンスが「現在」であれば，対象事例も「現在」で記述したほうがよいと学習するものである。特に論文の実験の記述で時制を統一する必要がある場合には有用である。

【００６３】また，テンス－素性組抽出部１２では，事例の英文対訳データの該当する語句（動詞部分）などを素性として抽出してもよい。これは，訳出される英語文の構造がかわると用いるべきモダリティも変化する場合があることにもとづく。

【００６４】以下の事例で説明する。

例文１）和文：彼は質実な生活を｛送っている｝

12

英訳：He {lives } a sober and simple life.

例文２）和文：彼は惰性的に怠惰な生活を｛送っている｝

英訳：He {is leading} a lazy life out of habit.

例えば，例文１の「送っている」のモダリティは「現在形」であるが，例文２の「送っている」のモダリティは「進行形」である。これらはほとんど意味の同じ文であり同じモダリティを持っていると考えてもよいものだが，訳出に用いる動詞を「live」と「lead」とかえただけでこのような違いが出てくる。

【００６５】なお，この場合に，入力される日本文３に仮訳された英文が付与されることになり，素性抽出部１５では，日本文３に付与された英文の動詞部分が素性として抽出される。

【００６６】高品質な処理を行ないたい場合，このように，素性の取り出しの際に，日本語側だけでなく英文対訳データのうち機械翻訳システムの構造解析部が想定している英語側の構造（あるいは動詞）を取り出すということが，テンス・アスペクト・モダリティの翻訳の向上に有効である。

【００６７】以下，本発明を一実施例によりさらに詳細に説明する。

【００６８】〔テンス・アスペクト・モダリティ・データベース〕図６に，実施例で用いるテンス・アスペクト・モダリティ・データベース１１である対訳コーパスの一部を示す。この対訳コーパスは，例えば，以下の参考文献７にもとづいて作成する。

［参考文献７］村田真樹　内山将夫　内元清貴　馬青　井佐原均，機械学習を用いた機械翻訳用モダリティコーパスの修正，言語処理学会第７回年次大会，(2001)

図６中，英語側の文には以下の二か所のタグが付与されている。

・英語の主節の動詞部分を＜ｖ＞，＜／ｖ＞のタグで囲む。

・日本語の主節の動詞に対応する英語の動詞部分を＜ｖｊ＞，＜／ｖｊ＞のタグで囲む。

【００６９】また，日本語側の文の先頭に“ｃ”や“ｄ”といった記号がふられているが，これらはこの対訳データのテンス・アスペクト・モダリティを意味する。例えば，“ｃ”はｃａｎを，“ｄ”は過去形を意味する。

【００７０】図６に示すコーパスの一つめのデータには“，”があるが，これは＜ｖｊ＞を用いるときに使われるもので，“，”の左に＜ｖ＞で囲まれた動詞に対するテンス・アスペクト・モダリティが，右に＜ｖｊ＞で囲まれた動詞に対するテンス・アスペクト・モダリティが記述される。なお，このコーパスでは現在形の出現が多いのでその場合はタグをふらなかった。このため，“，”の左右が空欄となってこの部分には“，”だけが付与されている。

13

【００７１】また，「日本語に対応する英語の動詞部分」と「英語の主節の動詞部分」が一致する場合は「英語の主節の動詞部分」のタグのみ付与した。また，「日本語に対応する英語の動詞部分」の方はそれほど綿密にタグ付与は行なっておらず，「日本語に対応する英語の動詞部分」と「英語の主節の動詞部分」が一致しない場合にもタグ付与をしなかった場合もある。

【００７２】日英の対訳データに付与するテンス・アスペクト・モダリティの分類として，前述した分類のうち，以下の（１）および（２）を用いた。

（１）各助動詞相当語句（be able to，be going to，can，have to，had better，may，must，need，ought，shall，used to，will の12種類）がつくかどうかと，{現在形，過去形}と{進行形，進行形でない}と{完了，完了でない}のすべての組み合わせ（助動詞相当語句が複数つく場合も許している。）：2^{15}種類

（２）命令形（１種類）

本発明に係るシステムで扱うテンス・アスペクト・モダリティの分類は英語の表層表現に基づいて定めたものであり，日本語文だけを与えてこの分類を推定できれば，モダリティ表現の日英翻訳ができあがる。このため，本例では，原則として，テンス・アスペクト・モダリティの分類を示すタグと日本語文のみを用いている。また，前述したように，これらの分類は，「英語の主節の動詞部分」と「日本語の主節の動詞に対応する英語の動詞部分」の二か所にふられるが，日本文３を与えて「英語の主節の動詞部分」のテンス・アスペクト・モダリティの分類を推定することを問題設定としている。

【００７３】本例では，テンス・アスペクト・モダリティ・データベース１１として，以下の二種類の対訳コーパスを用いた。

・K社和英辞典の例文（事例総数は39,660個，分類の総数は46個）

・白書データ（事例総数は5,805 個，分類の総数は30個）

これらのコーパスは，人手により確認しながらタグづけを行ない，さらに参考文献７および以下の参考文献８に示すコーパス修正の方法を利用して作成しており，非常に高精度なものとなっている。

［参考文献８］村田真樹　内山将夫　内元清貴　馬青　井佐原均，決定リスト，用例ベース手法を用いたコーパス誤り検出・誤り訂正，自然言語処理研究会，2000-NL-136(2000)

【００７４】〔抽出する素性〕本例では，日本文３の入力を与えられたときにテンス・アスペクト・モダリティ４として分類を出力する。このため，素性は入力される日本文３から取り出すことになる。ここでは，素性集合として以下の三種類のものに対して処理を行った。

（１）素性集合Ｆ１

日本語文末の１～１０gramの文字列と入力された文全体の形態素列を素性の集合とする。

14

例：「ない＄」「しなかった＄」「今日」「は」「走る」

この場合に，素性の数は，K社データで２３０，１３４個，白書データで２５，９５８個となる。

（２）素性集合Ｆ２

日本語文末の１～１０gramの文字列を素性の集合とする。

例：「ない＄」「しなかった＄」

この場合に，素性の数は，K社データで１９９，１９９個，白書データで１６，６１０個となる。

（３）素性集合Ｆ３

入力された文全体の形態素列を素性の集合とする。

例：「今日」「は」「走る」

この場合に，素性の数は，K社データで３０，９３５個，白書データで９，３４８個となる。

【００７５】入力された文を形態素列に分解するのには，ＪUMANを用いた。ＪUMANの詳細な説明については，以下の参考文献９に記載されている。

［参考文献９］黒橋禎夫　長尾真，日本語形態素解析システムＪUMAN使用説明書，version 3.6（京都大学大学院工学研究科，1998）

素性集合Ｆ１は，素性集合Ｆ２と素性集合Ｆ３との組合わせである。素性集合Ｆ２は，上記の参考文献１の研究を参考にして作成したものであり，日本語文においてテンス・アスペクト・モダリティを示す表現は文末の動詞にあらわれることが多いことから，日本語文の文末の文字列を素性としている。素性集合Ｆ３は，「明日」「昨日」などの副詞もテンス・アスペクト・モダリティを示す表現であり，用いるべきだと考えて作成したもので，入力された文全体の形態素列とするものである。

〔機械学習によるテンス・アスペクト・モダリティの分類〕本例では，機械学習の手法として，決定リスト法，最大エントロピー法，サポートベクトルマシン法を用いた。さらに，本発明に係るシステムで用いる機械学習法と従来の手法との処理結果の比較のためにｋ近傍法を用いた処理も行った。

【００７６】ｋ近傍法は，素性集合だけでなく事例同士の類似度を定義する必要がある。しかし，本例では素性集合Ｆ１と素性集合Ｆ３は入力された文全体の形態素列をも素性の集合とするので，類似度の定義が困難である。そのため，ｋ近傍法では素性集合Ｆ２だけを用いることにする。素性集合Ｆ２での類似度の定義としては，事例間で一致した文字列の最長がｘ－ｇｒａｍのとき，類似度をｘとすることにした。

【００７７】なお，他の機械学習の手法としては，他にＣ４．５などの決定木学習を利用する方法があるが，本例では，種々の問題で決定木学習手法が他の手法に比べて劣っていること，また，本例で扱う問題は属性の種類の数が多くＣ４．５が実行できるまで属性の数を減らす

15

と精度が落ちるであろうことの二つの理由により用いていない。

【００７８】〔第１の例〕まず，K社和英辞典の例文のデータを用いた処理を行なった。その処理結果の精度を図７に示す。本例では，クローズとオープンの二種類の処理を行なった。オープンの実験は１０分割のクロスバリデーションで行なった。図７の括弧内の数字はクローズでの精度を意味する。

【００７９】この処理結果から以下のことがわかる。

・決定リスト法は素性集合Ｆ２を用いるときに，ｋ近傍法と同程度の精度を得ている。

・最大エントロピー法は，ｋ近傍法または決定リスト法に比べて高い精度を得ている。

・サポートベクトルマシン法は，常に他の手法に比べて高い精度をあげている。

・素性集合の比較としては，最大エントロピー法および決定リスト法では，素性集合Ｆ２が最も精度が高く，素性集合Ｆ１のように形態素の情報を追加すると逆に精度が下がる結果となっている。これは，素性が増えても不要な素性も増えるために精度が低下したものと思われる。

・サポートベクトルマシン法での素性集合の比較では，素性集合Ｆ１で最も高い精度をあげている。これは，サポートベクトルマシン法では形態素の情報の追加が効果があったことを意味する。他の手法では形態素の情報の追加では逆に精度が下がったので，サポートベクトルマシン法では不要な素性を除去し有用な素性を選択する素性選択の能力も他の手法に比べて高いと推測される。

【００８０】この結果に対し，手法の理論的な側面からは以下のような説明をつけることができる。

・決定リスト手法は，ある一つの素性のみから解を求める方法のため，不要な素性が多い場合その不要な素性のみを文脈として解を求めてしまいがちになり，不要な素性が多い場合精度が低下する。

・最大エントロピー法は，常にほとんどすべての素性を用いるので，不要な素性が多い場合には精度が低下する。

・これらに対し，サポートベクトルマシン法では，サポートベクトルとなる事例のみを用いそれ以外の事例を用いないといった事例を捨てる操作があるため，多くの不要な素性をこの事例とともに捨てることになり，不要な素性が多くてもそれほど精度低下を招かない傾向がある。

【００８１】以上のように，全手法通じて最も精度が高かったのは，ｄ＝１，素性集合Ｆ１のときのサポートベクトルマシン法であった。

【００８２】上記の結果のうち，サポートベクトルマシン法において，素性集合Ｆ１を用いる方が素性集合Ｆ２を用いるよりも良かった，すなわち，形態素の情報の追加が効果があった，という結果が有意なことなのかを調

16

べるために符合検定を行なった。これは，ｄ＝１の方が精度がよかったのでｄ＝１で行なった。全事例３９，６６０個のうち，素性集合Ｆ１で正しく素性集合Ｆ２で誤った事例は６４８個であり，素性集合Ｆ２で正しく素性集合Ｆ１で誤った事例は４２７個であったが，これを符合検定にかけると０．０００００００１％（計算では８桁で切っていたため，実際の値はこの値よりももっと小さい可能性がある。）以下の危険率で有意な差があると判定された。このことにより，サポートベクトルマシン法において，形態素の情報を追加する効果があったことは，ほぼ間違いないと考えてよい。

【００８３】次に形態素の情報といっても，実際にどのような素性が有効に働いているかを調べることにした。これは，素性集合Ｆ１で正しく素性集合Ｆ２で誤った事例は，６４８個に偏って出現している素性を調べることによって行なうことにした。ここでは二項検定を利用して，全事例３９，６６０個での出現確率よりも有意水準１％で大きいと判断されたものを偏って出現しているものとした。

【００８４】この有効に働いたと思われる形態素素性の頻度の大きいもの上位２０個を図８に示す。図８では，「もう」「最近」「だろう」「まだ」「なければ」「ましょう」「あす」など，テンス・アスペクト・モダリティの推定に役立ちそうな形態素素性が得られており，実際にこういった素性によって精度が向上したものと推測される。

【００８５】〔第２の例〕次に白書データを用いて処理を行った。この場合には，精度の良かったサポートベクトルマシン法を用いて行なった。本例でも１０分割のクロスバリデーションを行なうことでオープンでの精度を求めている。処理結果の精度を図９に示す。

【００８６】この処理結果より以下のことがわかる。

・白書データの精度は，最大で６４．６７％であった。

・白書データでも文末文字列のみを用いる素性集合Ｆ２よりも一文全体の形態素情報も加えて用いる素性集合Ｆ１の方が高い精度を得ている。また，白書データでは素性集合Ｆ２よりも，一文全体の形態素情報を用いる素性集合Ｆ３の方が精度が高い。これらの結果はさらに，一文全体の形態素情報の素性としての有効性を確かめるものとなっている。

【００８７】〔第３の例〕次に，K社データを学習データとして，白書データをテストデータとしたような場合，すなわち異分野のデータを教師データとした処理を行なった。本例により，異なる分野のデータを用いると精度がどのようにかわるのかを調べることができた。この処理では精度の良かったｄ＝１およびｄ＝２のデータを対象として，素性集合Ｆ１のサポートベクトルマシン法を用いて行なった。この処理でも学習データとテストデータが重なる場合は重なった部分において１０分割のクロスバリデーションを行なうことでオープンでの精度

を求めている。処理結果の精度を図１０に示す。

【００８８】この処理結果により以下のことがわかる。

・異なる分野のデータを用いると精度が非常に下がった。(白書データを学習データとしてＫ社データを解析したり，Ｋ社データを学習データとして白書データを解析したりすると，精度は１０％～２０％程度に落ちた。）このことから，入力されるデータと同分野の学習データを用いることが有効であることがわかる。

【００８９】人手で書いた規則を用いる手法では異分野に適応したシステムを作るのが難しい。これに対して本発明のような機械学習を用いる方法であれば，学習データをかえて学習し直すことにより，分野ごとに適応したシステムを作るのが容易となる。

・Ｋ社と白書の両方を学習データとして用いた場合は，精度はほとんどかわらないか，もしくは少し下がる程度であった。このことから，学習データは多ければよいというものではなく，異分野のデータの場合は，混在させて学習データを用いてもそれほど効果がないことがわかる。

【００９０】本例において，テンス・アスペクト・モダリティの翻訳の処理を，ｋ近傍法も含めて様々な機械学習手法を用いて行なった。また検証のため，機械学習法のうちどの方法がもっともよいかを調べた。

【００９１】従来の手法（参考文献１等）では，テンス・アスペクト・モダリティの翻訳の際に，素性として文末の文字列しか用いていなかった。本発明では文末の文字列以外にその一文中の形態素情報を追加して用いた。

【００９２】その結果，従来では用いていなかった一文中の形態素情報の利用が，処理精度を向上させる効果があることを検定を用いて確認した。

【００９３】また，機械学習手法として，決定リスト法，最大エントロピー法，またはサポートベクトルマシン法のいずれか一の手法を用いても，従来のｋ近傍法よりも高い精度でテンス・アスペクト・モダリティの翻訳を行なうことができた。

【００９４】特に，サポートベクトルマシン法による方法が最も高い精度を得て，従来手法のｋ近傍法による手法よりも高い精度で，テンス・アスペクト・モダリティの翻訳を行なうことができた。

【００９５】また，異なる分野(本例では，Ｋ社英和辞典データと白書データ）のコーパスを用いた処理の例を行なった。この処理では，異なる分野のデータを用いると精度が格段に落ちることを確認し，異なる分野ごとにテンス・アスペクト・モダリティの翻訳システムを構築する必要があることを確認した。このことは，異分野に適応するシステムを人手で作成することが困難であることを考えれば，機械学習手法を用いる本発明の有用性を示すことになる。

【００９６】以上，本発明をその実施の態様により説明したが，本発明はその主旨の範囲において種々の変形が可能である。例えば，本発明の実施の形態では，本発明に係るテンス・アスペクト・モダリティ翻訳システム１は独立して構成されるものとして説明してきたが，他の機械翻訳システムの一部として構成されることも可能である。

【００９７】また，機械学習部１３で用いる機械学習法は，決定リスト法，最大エントロピー法，サポートベクトルマシン法に限らず，異なる形式の素性を組み合わせて扱うことができる方法であればどのような手法であってもよく，また，テンス－素性組抽出部１２または素性抽出部１５で抽出する素性は，対象となる事例もしくは入力文から抽出可能な素性であれば種類は限定されないことは当然である。

【００９８】

【発明の効果】以上説明したように，本発明は以下のような格別の効果を奏する。

・本発明では，テンス・アスペクト・モダリティの翻訳の問題で，サポートベクトルマシン法に代表される，複数の形式の解析情報（素性）を取り扱うことができるような機械学習法を用いる。これにより，従来の類似度を用いるｋ近傍法よりも高い精度でテンス・アスペクト・モダリティの翻訳を行なうことができる，テンス・アスペクト・モダリティ翻訳処理方法および翻訳システムを提供することができる。

・本発明では，テンス・アスペクト・モダリティの翻訳の際に，文末の文字列以外にその一文中の形態素情報を新たに追加して用いている。これにより，素性として文末の文字列しか用いていなかった従来の手法に比べて精度の高い翻訳を行うことができる。

・本発明では，機械学習部１３で用いるテンス・アスペクト・モダリティと素性の集合の組を抽出するテンス・アスペクト・モダリティ・データベースとして，種々の分野のコーパスを利用することができ，さらにそのコーパスにもとづいて人手によらずに学習結果を取得することができる。これにより，異分野ごとに適応するテンス・アスペクト・モダリティ翻訳システムを容易に実現することができる。

【図面の簡単な説明】

【図１】本発明にかかるシステムの構成例を示す図である。

【図２】本発明にかかるシステムの処理フローチャートである。

【図３】テンス・アスペクト・モダリティと素性の集合の組の例を示す図である。

【図４】サポートベクトルマシン法におけるマージン最大化を説明するための図である。

【図５】入力文からの素性の集合の抽出の例を示す図である。

【図６】実施例におけるテンス・アスペクト・モダリティ・データベースの一部の例を示す図である。

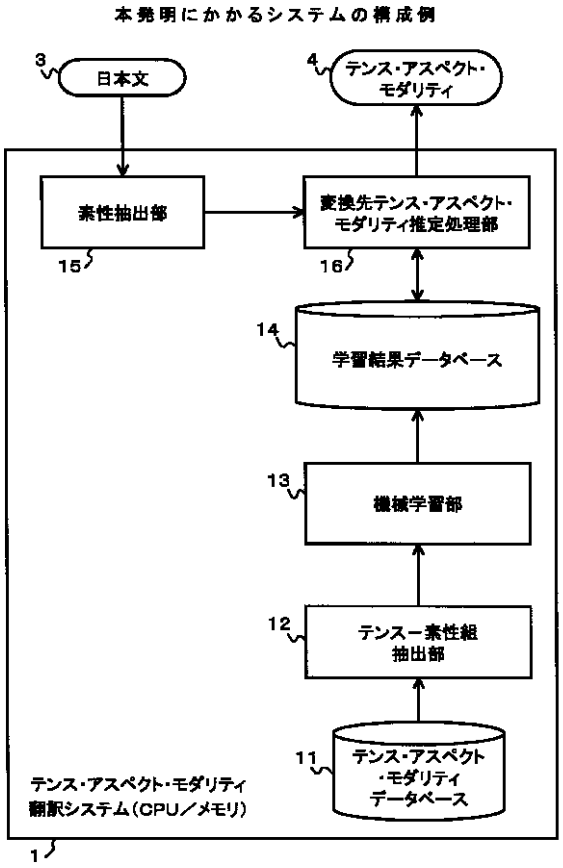
19

【図７】第１の例におけるテンス・アスペクト・モダリティの翻訳の精度を比較するための図である。
【図８】有効に働いたと思われる形態素素性の例を示す図である。
【図９】第２の例におけるテンス・アスペクト・モダリティの翻訳処理の精度を比較するための図である。
【図１０】第３の例におけるテンス・アスペクト・モダリティの翻訳処理の精度を比較するための図である。
【符号の説明】

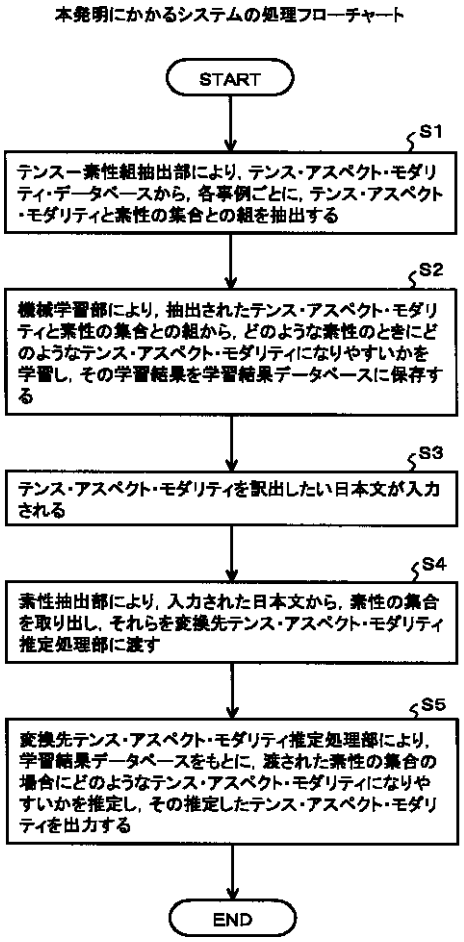
20

１　テンス・アスペクト・モダリティ翻訳システム
１１　テンス・アスペクト・モダリティ・データベース
１２　テンス－素性組抽出部
１３　機械学習部
１４　学習結果データベース
１５　素性抽出部
１６　変換先テンス・アスペクト・モダリティ推定処理部

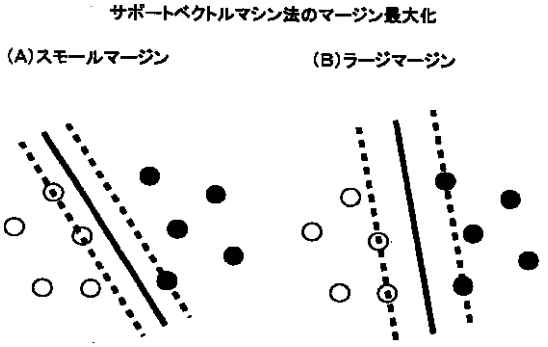
【図１】



【図２】



【図４】



【図９】

第2の例におけるテンス・アスペクト・モダリティの翻訳の精度

機械学習の手法	素性集合F1	素性集合F2	素性集合F3
サポートベクトル(d＝1)	60.10%　(99.81%)	56.61%　(89.87%)	56.14%　(96.67%)
サポートベクトル(d＝2)	64.67%　(99.81%)	56.74%　(89.87%)	62.07%　(99.83%)

＜括弧()内の数字はクローズの場合＞

【図３】

テンス・アスペクト・モダリティと素性の集合との組の例

11

現在完了	もう登録しました。 （対訳データは略）

⇩

文字列素性	「もう登録しました＄」 「う登録しました＄」 「登録しました＄」 「録しました＄」 「しました＄」 「ました＄」 「した＄」 「た＄」
単語素性	「もう」「登録」「し」「まし」「た」
テンス等	現在完了

【図５】

素性の集合の抽出の例

入力される日本文　もう行きました。　3

⇩

文字列素性	「もう行きました＄」 「う行きました＄」 「行きました＄＄」 「きました＄」 「ました＄」 「した＄」 「た＄」
単語素性	「もう」「行き」「まし」「た」

【図６】

テンス・アスペクト・モダリティ・データベースの一部の例

'	この子供はああ言えばこう言うから小憎らしい This child always talks back to me, and this <v> is </v> why I <vj> hate </vj> him.
d	彼がああおくびょうだとは思わなかった I <v> did not think </v> he was so timid.
c	ああ忙しくては休む暇もないはずだ Such a busy man as he <v> cannot have </v> any spare time.

【図７】

第1の例におけるテンス・アスペクト・モダリティの翻訳の精度

機械学習の手法	素性集合F1		素性集合F2		素性集合F3	
K近傍法（k＝1）	－	－	79.36%	(98.50%)	－	－
K近傍法（k＝3）	－	－	80.35%	(83.94%)	－	－
K近傍法（k＝5）	－	－	80.43%	(82.39%)	－	－
K近傍法（k＝7）	－	－	80.39%	(81.71%)	－	－
K近傍法（k＝9）	－	－	80.22%	(81.30%)	－	－
決定リスト	74.19%	(98.21%)	80.23%	(98.18%)	67.90%	(86.58%)
最大エントロピー	80.37%	(88.87%)	81.16%	(83.85%)	75.35%	(84.15%)
サポートベクトル（d＝1）	82.48%	(98.70%)	81.93%	(98.50%)	78.68%	(96.68%)
サポートベクトル（d＝2）	82.28%	(98.48%)	81.37%	(98.48%)	79.01%	(98.74%)

＜括弧（）内の数字はクローズの場合＞

【図１０】

第3の例におけるテンス・アスペクト・モダリティの翻訳の精度

処理対象のデータ	学習に用いたデータ					
	K社と白書		K社だけ		白書だけ	
K社データ（d＝1）	82.44%	(99.71%)	82.48%	(98.70%)	65.31%	－
K社データ（d＝2）	82.31%	(98.74%)	82.28%	(98.48%)	51.92%	－
白書データ（d＝1）	60.02%	(99.79%)	47.65%	－	60.10%	(99.81%)
白書データ（d＝2）	64.01%	(99.83%)	49.53%	－	64.67%	(99.81%)

＜括弧（）内の数字はクローズの場合＞

【図８】

有効に働いたと思われる形態素素性の例

頻度	形態素素性
221	が
121	いる
89	ない
28	ます
23	なら
23	きた
22	もう
19	中
18	最近
16	だろう
16	ました
14	まだ
13	れる
12	なければ
11	ましょう
11	すっかり
10	られ
9	っ
8	あす
8	なんて

【手続補正書】
【提出日】平成１４年7月２９日（２００２．７．２９）
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】特許請求の範囲
【補正方法】変更
【補正内容】
【特許請求の範囲】
【請求項１】　コンピュータにより一の言語から他の言語へ翻訳処理をする際に，変換元言語から変換先言語のテンス・アスペクト・モダリティを翻訳する方法であって，予め備えられた変換元言語の事例と当該事例の変換先言語におけるテンス・アスペクト・モダリティとの組を記憶するデータベースであるテンス・アスペクト・モダリティ・データベースにアクセスする過程と，前記テンス・アスペクト・モダリティ・データベースの各事例ごとに，テンス・アスペクト・モダリティと当該テンス・アスペクト・モダリティに対応する事例から抽出した単語素性および文字列素性を含む複数の形式の素性の集合とからなるテンス－素性組を抽出する過程と，前記テンス－素性組を教師データとして用いて複数の素性の出現パターンについて，それぞれのパターンのときになりやすいテンス・アスペクト・モダリティを，決定リスト法，最大エントロピー法，またはサポートベクトルマシン法のいずれか一の機械学習法により学習する過程と，前記機械学習する過程における学習結果を，入力文の変換先言語のテンス・アスペクト・モダリティを判定するための学習データとして学習結果データベースに保存する過程と，変換元言語の入力文から，当該入力文の素性の集合を抽出する過程と，前記入力文の素性の集合をもとに，前記機械学習法により，前記学習データベースに保存された学習データを参照して前記素性の集合の素性の出現のパターンについて，なりやすいテンス・アスペクト・モダリティを特定し，前記入力文のテンス・アスペクト・モダリティの推定解として出力する過程とを備えることを特徴とするテンス・アスペクト・モダリティ翻訳処理方法。
【請求項２】　コンピュータにより一の言語から他の言語へ翻訳処理をする際に，変換元言語から変換先言語のテンス・アスペクト・モダリティを翻訳するシステムであって，予め備えられた変換元言語の事例と当該事例の変換先言語におけるテンス・アスペクト・モダリティとの組を記憶するテンス・アスペクト・モダリティ・データベースと，前記テンス・アスペクト・モダリティ・データベースの各事例ごとに，テンス・アスペクト・モダリティと当該テンス・アスペクト・モダリティに対応す

る事例から抽出した単語素性および文字列素性を含む複数の形式の素性の集合とからなるテンス－素性組を抽出するテンス－素性抽出手段と，前記テンス－素性組を教師データとして用いて複数の素性の出現パターンについて，それぞれのパターンのときになりやすいテンス・アスペクト・モダリティを，決定リスト法，最大エントロピー法，またはサポートベクトルマシン法のいずれか一の機械学習法により学習する機械学習手段と，前記機械学習手段における学習結果を，入力文の変換先言語のテンス・アスペクト・モダリティを判定するための学習データとして学習結果データベースに保存する学習結果データベースと，変換元言語の入力文から，当該入力文の素性の集合を抽出する素性抽出手段と，前記入力文の素性の集合をもとに，前記機械学習法により，前記学習データベースに保存された学習データを参照して前記素性の集合の素性の出現のパターンについて，なりやすいテンス・アスペクト・モダリティを特定し，前記入力文のテンス・アスペクト・モダリティの推定解として出力する変換先テンス・アスペクト・モダリティ推定処理手段とを備えることを特徴とするテンス・アスペクト・モダリティ翻訳システム。

【手続補正２】

【補正対象書類名】明細書

【補正対象項目名】００１１

【補正方法】変更

【補正内容】

【００１１】

【課題を解決するための手段】上記課題を解決するため，本発明に係る方法は，コンピュータにより一の言語から他の言語へ翻訳処理をする際に，変換元言語から変換先言語のテンス・アスペクト・モダリティを翻訳する方法であって，予め備えられた変換元言語の事例と当該事例の変換先言語におけるテンス・アスペクト・モダリティとの組を記憶するデータベースであるテンス・アスペクト・モダリティ・データベースにアクセスする過程と，前記テンス・アスペクト・モダリティ・データベースの各事例ごとに，テンス・アスペクト・モダリティと当該テンス・アスペクト・モダリティに対応する事例から抽出した単語素性および文字列素性を含む複数の形式の素性の集合とからなるテンス－素性組を抽出する過程と，前記テンス－素性組を教師データとして用いて複数の素性の出現パターンについて，それぞれのパターンのときになりやすいテンス・アスペクト・モダリティを，決定リスト法，最大エントロピー法，またはサポートベクトルマシン法のいずれか一の機械学習法により学習する過程と，前記機械学習する過程における学習結果を，入力文の変換先言語のテンス・アスペクト・モダリティを判定するための学習データとして学習結果データベースに保存する過程と，変換元言語の入力文から，当該入力文の素性の集合を抽出する過程と，前記入力文の素性の集合をもとに，前記機械学習法により，前記学習データベースに保存された学習データを参照して前記素性の集合の素性の出現のパターンについて，なりやすいテンス・アスペクト・モダリティを特定し，前記入力文のテンス・アスペクト・モダリティの推定解として出力する過程とを備えることを特徴としている。

フロントページの続き

(72)発明者　内元　清貴
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内

(72)発明者　井佐原　均
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内

Ｆターム(参考)　5B091 AA15 AB13 BA03 CC03 CC16 EA01