論　文

Technical Papers

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

An Example-Based Approach to Japanese-to-English Translation of Tense, Aspect, and Modality

村田 真樹　郵政省 通信総合研究所 〒 619-0289 京都府相楽郡精華町光台 2-2-2
Masaki Murata　Communications Research Laboratory, MPT, 2-2-2, Hikaridai, Seika-tyou, Soraku-gun, Kyoto, 619-0289, Japan
murata@crl.go.jp

馬 青　郵政省 通信総合研究所 〒 619-0289 京都府相楽郡精華町光台 2-2-2
Qing Ma　Communications Research Laboratory, MPT, 2-2-2, Hikaridai, Seika-tyou, Soraku-gun, Kyoto, 619-0289, Japan
qma@crl.go.jp

内元 清貴　郵政省 通信総合研究所 〒 619-0289 京都府相楽郡精華町光台 2-2-2
Kiyotaka Uchimoto　Communications Research Laboratory, MPT, 2-2-2, Hikaridai, Seika-tyou, Soraku-gun, Kyoto, 619-0289, Japan
uchimoto@crl.go.jp

井佐原 均　郵政省 通信総合研究所 〒 619-0289 京都府相楽郡精華町光台 2-2-2
Hitoshi Isahara　Communications Research Laboratory, MPT, 2-2-2, Hikaridai, Seika-tyou, Soraku-gun, Kyoto, 619-0289, Japan
isahara@crl.go.jp

keywords: modality, Japanese-to-English translation, example-based approach

Summary

We have developed a new method for Japanese-to-English translation of tense, aspect, and modality that uses an example-based method. In this method the similarity between input and example sentences is defined as the degree of semantic matching between the expressions at the ends of the sentences. Our method also uses the k-nearest neighbor method in order to exclude the effects of noise; for example, wrongly tagged data in the bilingual corpora. Experiments show that our method can translate tenses, aspects, and modalities more accurately than the top-level MT software currently available on the market can. Moreover, it does not require hand-craft rules.

1. は じ め に

近年，WWW などのインターネットの発展とともに機械翻訳システムの必要性が日に日に高まり続けている．本研究では，この機械翻訳において特に困難な問題の一つである，日本語のテンス・アスペクト・モダリティ表現の翻訳の研究を行なう．従来の手法ではテンス・アスペクト・モダリティ表現は表層の手がかり表現などを用いた規則を人手で作成することで翻訳を行なってきた [久米 90, 白井 90]．しかし，日本語のテンス・アスペクト・モダリティ表現は多様であり人手で細かなところまで逐一，規則を作成していくのは非常に難しい．そこで本研究では，対訳のデータベースから今解析しているテンス・アスペクト・モダリティ表現によく似た対訳例 (用例) を取り出し，これを参照して翻訳結果を出力することを試みる．

本研究で用いる，解析する表現によく似た対訳例を利用するという用例ベースの考え方は 1984 年に長尾により機械翻訳の問題において提案されたものである [Nagao 84]．この手法は，これまで名詞句「A の B」の訳し分け [Sumita 90] など数多くの研究 [Sato 90, 佐藤 91a, 佐藤 91b, Sumita 91, Watanabe 92, 古瀬 94] に用いられている，有用な手法である．本論文の目的であるテンス・アスペクト・モダリティの翻訳の問題に，用例ベースの手法を適用した研究はまだなく，この問題を用例ベース手法で精度よく解析できれば，用例ベース手法の有用性，有効性のさらなる確認となる．

本研究で強調したいことをあらかじめまとめておくと以下のようになる．

(1) 本研究は，テンス・アスペクト・モダリティの翻訳の問題に用例ベースの手法を初めて適用した研究である．
(2) 本研究で採用する用例間の類似度の定義は，文末からの一致文字列 (もしくは分類語彙表の分類番号も含めた文字列での一致) であり，非常に簡明なものである．
(3) 対訳例の代わりに日本語文に正解のテンス・アスペクト・モダリティの情報を付与したデータを用いることで，本研究の考え方は翻訳の問題に限らず単

言語におけるテンス・アスペクト・モダリティの推定の研究にも利用できる．

われわれは，本来機械翻訳の問題というものは意味や文脈などの理解が必要な難しい問題であること，また，人間でも行なうことが難しい問題であることから，言語処理の研究としては解析・生成の両面の研究がほぼ完璧に完成してからやっと本格的な研究が行なえるものであると考えている．ところが，機械翻訳というものは，単なるパズル的な置き換えだけでも正しく翻訳できる場合もあり，機械翻訳の需要を考えると簡単な技術であっても解析できるところが少しでもあればすぐにでも用いたいという側面もある．本研究は，概ね後者の考え方に立ち，意味処理などの深い処理を行なわず，人手の介入がほとんど必要とされない用例を利用するという簡単な方法で，現在のトップクラスの市販ソフト程度以上のテンス・アスペクト・モダリティの翻訳を可能とする手法について述べているものである．

2. 用例ベースによるテンス・アスペクト・モダリティの解析方法

2・1 文末一致文字列を類似度の定義に利用する考え方

われわれは以前に文末の省略の補完の研究を用例ベースの考え方で解いたことがある [村田 98]．例えば，以下の例では「ちょっとお願いが」の文で文末の動詞が省略されている．

(例)「ちょっとお願いが」

用例：あのう，お願いが あるのですが．

一致部分　後続部分

これを用例で解析するときに，まず「ちょっとお願いが」の文末の表現を最長に含む文を用例から取り出し，「あのう，お願いがあるのですが」という文を得て，一致部分の後続部分である「あるのですが」を補完すべき表現として抽出するということを行なっていた．用例ベースの手法では，類似した用例を取ってくるために入力と用例の間の類似度というものを設定する必要がある．この類似度というものをどのように設定するのかということが最も重要な問題となる．この類似度の定義の仕方によっては解析結果は良くも悪くもなる．この文末表現の省略の補完の研究では，文末からの文字列の一致の具合を類似度としており，類似度としては簡明なものであり，かつ，この問題に適切なものであった．

日本語文ではテンス・アスペクト・モダリティは文末にあるので，テンス・アスペクト・モダリティの翻訳にもこの考え方は利用できると考えられる．つまり，類似度を省略の問題と同様，文末からの一致文字列としてテンス・アスペクト・モダリティの翻訳を試みるのである．実際の解析は，解析する日本語文に対して，文末からの一致文字列が最も長い用例を取り出し，その用例 (対訳対) の英語文のテンス・アスペクト・モダリティをその日本語文のテンス・アスペクト・モダリティの英訳部分であるとすればよい．例えば，以下の文を翻訳することを考える．

(例)「彼は大望をいだいている」

用例：彼はふるさとへの激しい慕情をいだいている

一致部分

英訳：He has a great longing for home.

該当部分

まず，この文の文末部分を最長に含む用例を取り出す．ここでは上にあげた用例が文末の一致文字列が最も長かったとする．その用例の英訳側の動詞部分を見て，それが現在形であるのをみて，テンス・アスペクト・モダリティ部分は現在形であると解析し，それにあうように訳出すればよい．この問題は日本語文末表現を見た感じでは「ている」という表現から「進行形」を思い浮かべそうになるが，正しく「現在形」であると解析できる．

この文末一致文字列を利用するという考え方は，用例ベースに基づく手法が適用された「名詞 A の名詞 B」の訳し分けの問題における類似度よりも簡明で使い勝手のよいものである．「名詞 A の名詞 B」の問題では，類似度を決める際に名詞 A を優先した方がよい場合や名詞 B を優先した方がよい場合があり単純に名詞 A から求まる類似度と名詞 B から求まる類似度を平均すればよいというものではなく，どちらかに重みを与える必要があったり，さらにその重みを名詞ごとにかえていく必要もあったりして，複雑なものになっていた．それに対し，文末一致文字列というものでは，文末から順番に見ていくだけでよく，重みを考える必要性もないし，A と B のどちらを優先すべきかなどと悩む必要もない．文末一致文字列の利用というものは，そういう明解さを持っているのである．

2・2 文末一致を計る二つの指標

近年，様々な形態素解析システムが公開されるなど言語処理の技術はめざましい発展をとげている．このため，類似用例を取り出す際の文末の一致具合の計算を，単純に文字列で行なうのではなく，形態素解析をして単語の認定を行なったデータに基づいて一致具合を計算した方がよいとも考えられる．また，単語同士の類似度も文字列一致により計算するよりは，シソーラスでの近さにより求めた方がよいとも考えられる．そのように考え，文末の一致具合の調べ方としては，文字列を利用するものの他に，言語解析の情報を利用する方法も加え，以下の二つの方法を試してみることにした．

- 方法 1　単なる文字列の利用 (略称「文字列」)
  これは，前節でも述べた方法で，単なる文字列のまま文末から一致を調べ，その一致した文字列の長さを類似度とする方法である．
- 方法 2　言語解析結果の情報の利用 (略称「意味等」)
  これは，形態素解析システムや単語のシソーラス

表 1　言語解析結果により得られる情報

形態素	分類語彙表の分類番号	変化形
彼	1200003012	
は	1195038023	
大望	1304207024	
を		
いだいて	2153417012	タ系連用テ形
いる	2120002012	基本形

を利用してより高精度に一致の具合を調べることをめざした方法である．まず，形態素解析 [黒橋 97] を行ない形態素の認識を行なう．次に各形態素に分類語彙表 [国立国語研究所 64] の「分類番号」*1を付与する*2．また，変化する形態素の場合は形態素解析システムの出力から得られる「変化形」を付与する．

例えば，「彼は大望をいだいている」の場合は，表 1 のような情報が用いられる*3．「彼」「は」などの形態素ごとに分割でき，それぞれに分類語彙表の分類番号や変化形が付与されている．分類番号は 10 桁の数字から構成されている．この 10 桁の分類番号はシソーラスの 7 レベルの階層構造を示しており，上位 5 レベルは分類番号の最初の 5 桁で表現され 6 レベル目は次の 2 桁，最下層のレベルは最後の 3 桁で表現されている．

次にこれらの情報を用いて文末の一致具合を調べることになるが，方法 2 でも文字列を用いる方法 1 と同様に文末からの一致文字列を調べるだけでよいように，表 1 の情報を下記のように連結して用いることとする．

“彼 03 ０００２１：は 38 ０５９１１：大望 07 ２４０３１：を：いだいて 17 ４３５ １２タ系連用テ形：いる 02 ００２１２基本形”

ここで分類番号の数字は逆順にしている．このことにより文末から一致を調べる際に，ちょうど分類番号の上位桁から一致を調べることになり通常のシソーラスを用いて意味の類似度を調べる方法と同様の効果が得られる．また，分類語彙表の数字の 8 から 10 桁は用いていないが，これらは形態素自体を意味する最下層のレベルであり，形態素自体の情報は文字列として別に用いるので用いる必要はない．また，6,7 桁目の数字はその二つの組でシソーラスの 6 レベル目を意味するが，本研究では一致を 2 バイト (全角) ごとに数えるので，6,7 桁目の数字を半角にすることで (例えば「彼」の ‘03’)，シソーラスの 6 レベル目も一つのまとまったものとして適切にチェックできるようになっている．

方法 2 では上記の形に変形してから文末から 2 バイトごとに一致文字列の長さを調べ，その長さを用例ベースの手法で用いる類似度とする．上記の形であれば，文末からの一致文字列を調べることで，すべての形態素について「形態素の変化形」「形態素のシソーラスでの意味的近さ」「形態素の文字列としての近さ」を順に調べることができる．

2･3　ノイズ対策としての k 近傍法の利用

用例を用いる方法を包含する手法として k 近傍法 (k-nearest neighbor method) と呼ばれる方法がある [Fukunaga 72, Stanfill 86, 富浦 96, 岡本 97]．この方法は 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*4．たった 1 個の用例で判断する場合はその 1 個がノイズであるかもしれずあまり信頼性の高い解析にならないが，k 個の用例を利用することで少々ノイズがあったとしても全体として安定した解析を実現することができる．

本研究では k 近傍法の k として 1 から 9 までの奇数，つまり 1,3,5,7,9 の 5 種類の k を用いることにした．文献 [岡本 97] によると k を偶数にした場合は多数決で引き分けになることが多くなり，引き分けの場合は選択がランダムになり精度の低下を招くらしい．このため本研究では k としては奇数のみを選んでいる．

また，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．しかし，本研究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にした．このとき，用いられる用例は処理で偶然先に得られた用例になる．また，多数決の際に引き分けになった場合はその引き分けになっている用例において最も最初にあがった用例の分類を解であるとした．

以上が具体的にどのようになっているかを，表 2 の解析例を使って考察してみよう．表 2 は「彼は私の知り合いだ」の文のテンス・アスペクト・モダリティを解析し

*1　ここでは KNP[黒橋 98] に付属でインストールする分類語彙表の辞書を利用している．そこで用いられている分類語彙表は最新のものであって 10 桁より多く長い値の分類番号がふられているが，KNP ではそれをうまく 10 桁に変換しているようだ．

*2　分類語彙表では多義語には複数の番号をふる場合がある．本研究では簡単のため，複数の番号があった場合でも最初の番号しか用いないようにしている．しかし，本来は多義性の解消を行なうか，それが難しいならすべての番号をうまく総合的に判断して用いられるようにした方がいいだろう．

*3　表 1 で助詞「は」と助詞「を」で同じ助詞なのに，「は」には分類番号がついていて助詞「を」にはついていないことが一見奇異に思えるかもしれない．これは，分類語彙表で助詞「は」に分類番号がふられているわけではなく，「は」には助詞「は」の他に数量関係を意味する「は」という単語があり，その単語の分類番号を使ってしまったためである．本研究では単語の意味解析を行なっていないので，表層の表現が等しい単語の区別ができないため，このような誤った分類番号がふられてしまうことがある．

*4　k 個の用例の多数決の際に k 個の用例に類似度に応じて重みを加える方法もあるが，本研究では実験の節でも述べるとおり実験は人手による解析で行なっているため，重みを考慮するなどの負荷があがることはさけた．しかし，この重みを考慮した方が精度は向上するだろう．

表 **2** テンス・アスペクト・モダリティの解析例

		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	彼は膚触りのいい 人だ	現在	He is a gentle－mannered person.
10	10	なんと男振りのいい 人だ	現在	What a handsomelooking man he is!

表 **3** 文末からの一致による類似度の算出

入力文		彼03 0 0 0 2 1 :は38 0 5 9 1 1 :私01 0 0 0 2 1 :の07 0 0 0 1 1 :知り合い03 0 1 2 2 1 :だ基本形
番号	類似度	25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
1	25	彼03 0 0 0 2 1 :と :は38 0 5 9 1 1 :長年07 2 4 6 1 1 :の**07** 0 0 0 1 1 :知り合い**03** 0 1 2 2 1 :だ基本形
2	24	ふた03 0 5 9 1 1 :り :は38 0 5 9 1 1 :長い間 :の**07** 0 0 0 1 1 :知り合い**03** 0 1 2 2 1 :だ基本形
..	..	..
4	11	彼ら03 0 0 0 2 1 :は38 0 5 9 1 1 :多年07 0 1 6 1 1 :の07 0 0 0 1 1 :知己**03** 0 1 2 2 1 :だ基本形
..	..	..

ているものである．ここでは類似度の定義としては前節の「意味等」の情報を用いる方法 2 を使っている．方法 2 を用いた類似度の算出過程を表 3 に示す．表 3 では文末の方から太字になっている部分が入力文と一致した部分で，この一致部分の 2 バイトごとの文字数が類似度となる．例えば，1 番の用例は文末から 25 個一致しているので類似度は 25 となる．このようにして類似度を計算し類似度の高いものから順に用例を 10 個取り出した*5ものが表 2 のデータである．表 2 ではシソーラスの情報なども使っているので「顔見知りだ」「知己だ」という例文も上位にあがっている．本研究では先にも述べたように簡単のため多くてもこの 10 個の用例しか解析に用いない．また，表 2 の「現在」などのテンス・アスペクト・モダリティの分類は用例の英語文のテンス・アスペクト・モダリティ表現から得られるものである．

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．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．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．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．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．

*5 類似した k 個の用例の取り出しは，文末でソートしたデータに対して二分探索を行なうことで実現している．この方法だと容量もあまり使用せずに高速に用例を取り出すことができる．

3. 実 験 と 考 察

3·1 実 験

2 章で説明した方法でどのくらいテンス・アスペクト・モダリティを正しく解析できるかを確かめるために実験を行なった．この実験では講談社和英辞典 [清水 76] に含まれる日英の対訳データ (約 3 万 6 千文) を用例のデータベースとした*6．また，実験の入力文としてその対訳データから 300 文を任意に取り出した．本手法の有効性を客観的に調べるためにトップレベルの市販のソフトでも実験を行なうことにした．取り出した 300 文をそのソフトにかけると 11 文では動詞部分の訳出に失敗しテンス・アスペクト・モダリティの分類を取り出すことができなかった．このため，この 11 文はすべての実験を通じて省いて実験を行なうことにした．つまり，残る 289 文を用いて実験を行なうことにした．また，用例ベースの手法による解析では，クロスバリデーションのように各文の解析では今実際に解析している例文一つはその用例

*6 用例には一つの日本文に対して複数の英訳がふられているものもあったが，実験では最初の英訳の方しか用いなかった．また，辞書の例文中の見出しを意味する横棒の見出し語への自動変換ではなるべく見出しについている漢字を用いるようにした．また，用例には文でなく名詞であるものや括弧などの不要な記号を含む文が混じっていたが，300 文の取り出しはこれらの文が含まれないように注意して行なった．

表 4 解析結果

	合計	現在形	過去形	その他
市販ソフト	80.6% (233/289)	91.1% (112/123)	96.3% (105/109)	28.1% (16/ 57)
文字列 (k=1)	76.8% (222/289)	89.4% (110/123)	89.9% (98/109)	24.6% (14/ 57)
文字列 (k=3)	82.0% (237/289)	93.5% (115/123)	97.2% (106/109)	28.1% (16/ 57)
文字列 (k=5)	83.0% (240/289)	95.1% (117/123)	97.2% (106/109)	29.8% (17/ 57)
文字列 (k=7)	82.4% (238/289)	94.3% (116/123)	96.3% (105/109)	29.8% (17/ 57)
文字列 (k=9)	82.4% (238/289)	94.3% (116/123)	96.3% (105/109)	29.8% (17/ 57)
意味等 (k=1)	78.5% (227/289)	88.6% (109/123)	89.9% (98/109)	35.1% (20/ 57)
意味等 (k=3)	81.7% (236/289)	91.1% (112/123)	95.4% (104/109)	35.1% (20/ 57)
意味等 (k=5)	81.3% (235/289)	92.7% (114/123)	93.6% (102/109)	33.3% (19/ 57)
意味等 (k=7)	81.7% (236/289)	92.7% (114/123)	93.6% (102/109)	35.1% (20/ 57)
意味等 (k=9)	81.7% (236/289)	92.7% (114/123)	93.6% (102/109)	35.1% (20/ 57)

ベースになかったがごとく扱って解析している*7．

また，テンス・アスペクト・モダリティの分類は以下の 27 種類のものとした．

(1) { 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 (8 種類)

(2) 命令形 (1 種類)

(3) 助動詞相当語句 (be able to の現在形と過去形, be going to の現在形と過去形, can, could, have to の現在形と過去形, let, may, might, must, need, ought, shall, should, will, would の 18 種類)

“must” と “have to” や “can” と “be able to” などはまとめてもよさそうだが，違った意味合いがあるかもしれないので，本研究では厳密に表層の表現によりテンス・アスペクト・モダリティの分類を設定することとし，これらの場合も異なるものとして取り扱った．また，テンス・アスペクト・モダリティの検出に用いる英語側のテンス・アスペクト・モダリティ表現としては，単文ならばその唯一の動詞のテンス・アスペクト・モダリティを，複文ならば日本語文の文末の動詞に対応する動詞のテンス・アスペクト・モダリティを用いた．意訳などで対応する動詞がない場合は，文頭よりの動詞のテンス・アスペクト・モダリティを用いた．

3 万 6 千文もの用例すべてに人手で上の分類をふるのは大変なので，実際の解析は解析する 289 文それぞれに対して，表 2 の「分類」の列だけ欠けたような表を作り，その表において人手で逐一テンス・アスペクト・モダリティの分類を調べ，前節の k 近傍法での多数決などを人手で行なって解析している．

表 5 市販ソフトとの誤りの重なり具合

	精度	両方誤り	市販ソフトのみ誤り	用例手法のみ誤り
文字列 (k=1)	76.8%	43 文	13 文	24 文
文字列 (k=3)	82.0%	41 文	15 文	11 文
文字列 (k=5)	83.0%	40 文	16 文	9 文
文字列 (k=7)	82.4%	40 文	16 文	11 文
文字列 (k=9)	82.4%	40 文	16 文	11 文
意味等 (k=1)	78.5%	39 文	17 文	23 文
意味等 (k=3)	81.7%	39 文	17 文	14 文
意味等 (k=5)	81.3%	42 文	14 文	12 文
意味等 (k=7)	81.7%	41 文	15 文	12 文
意味等 (k=9)	81.7%	41 文	15 文	12 文

以上の実験環境において実験を行なったところ，表 4 のような結果を得た．表の「市販ソフト」は本研究の実験結果と比較するために使った市販ソフトの精度を意味する．「文字列」「意味等」は 2 章で説明した方法 1 と方法 2 を意味する．“k=1”,”k=3” などは k 近傍法における k の値を示している．「合計」はすべてのデータでの精度，「現在形」「過去形」は正解のテンス・アスペクト・モダリティが現在形，過去形である場合の精度，「その他」はその他のテンス・アスペクト・モダリティが正解である場合の精度を示している．表の各数値は精度とその精度の算出に用いた分母分子を示している．

また，文字列や意味等を用いる提案手法と市販ソフトとの誤りの重なり具合も調べた．これを表 5 に示している．「両方誤り」は市販ソフト，提案手法の双方が誤った個数を意味し，「市販ソフトのみ誤り」「用例手法のみ誤り」はそれぞれ市販ソフトのみが用例手法のみが誤った個数を意味する．

また，実験で用いた 289 文では設定した 27 種類のテンス・アスペクト・モダリティの分類すべては出現せず 13 種類のテンス・アスペクト・モダリティの分類しか出現しなかった．その 13 種類の分類ごとの精度を表 6 に示す．表中の「現進」「過進」「完了」はそれぞれ現在進行形，過去進行形，現在完了を意味する．

*7 本研究の実験では，実験の入力文を，用例ベースと同種の対訳データより取り出している．一般に用例ベースの手法では入力と用例ベースが同種のデータであることが有利に働くが，これはその同種のデータが分野依存性の高いデータの場合に限って生じるものである．本研究で用いているデータは，講談社和英辞典にあった例文データであることから分野依存性は低くかつ一般的な例文であるので，同種のデータであっても有利に働いていないと思われる．また，これと関連した議論として，本研究の手法は，入力と用例ベースで分野依存性の高いデータの場合は有利であると思われるので，講談社和英辞典のような一般性のある用例ベースではなく分野依存性の高い個別分野の用例ベースを用意することで，各個別分野に特化したような解析を行なうことも可能である．人手で規則を作成するような手法では個別分野に特化した規則を作成するのは困難である．われわれの手法には人手で規則を作成する手法に比べて，このような利点も存在する．

表6　各テンス・アスペクト・モダリティごとの精度

	合計	現在	過去	現進	過進	完了	命令	can	could	let	may	must	will	would
文数	289文	123文	109文	7文	1文	15文	12文	3文	2文	1文	2文	4文	9文	1文
市販ソフト	81%	91%	96%	29%	0%	0%	67%	67%	100%	100%	0%	25%	0%	0%
文字列 (k=1)	77%	89%	90%	14%	0%	7%	58%	33%	50%	100%	0%	0%	22%	0%
文字列 (k=3)	82%	93%	97%	29%	0%	0%	75%	33%	0%	100%	0%	0%	33%	0%
文字列 (k=5)	83%	95%	97%	14%	0%	0%	83%	33%	0%	100%	0%	25%	33%	0%
文字列 (k=7)	82%	94%	96%	14%	0%	0%	83%	33%	0%	100%	50%	25%	22%	0%
文字列 (k=9)	82%	94%	96%	14%	0%	0%	83%	33%	0%	100%	50%	25%	22%	0%
意味等 (k=1)	79%	89%	90%	29%	0%	13%	75%	33%	100%	100%	0%	0%	33%	0%
意味等 (k=3)	82%	91%	95%	29%	0%	13%	83%	0%	50%	100%	0%	25%	33%	0%
意味等 (k=5)	81%	93%	94%	29%	0%	7%	92%	0%	0%	100%	0%	25%	33%	0%
意味等 (k=7)	82%	93%	94%	14%	0%	7%	92%	0%	0%	100%	50%	50%	33%	0%
意味等 (k=9)	82%	93%	94%	14%	0%	7%	92%	0%	0%	100%	50%	50%	33%	0%

3·2　考　　察

(1)　精度について考察すると下記のことがわかる．

a　表4のように文字列のk=5が83%で最も精度がよい．

b　表4の「合計」の欄のように全てのテンス・アスペクト・モダリティをあわせたときの精度ではk=1以外の提案手法はすべて市販ソフト以上の精度を得ている．

k=1の提案手法ではノイズの影響をうけやすく精度が低下することがわかる．k近傍法の利用の効果がうかがえる．

c　表4のように現在形と過去形以外のテンス・アスペクト・モダリティを意味する「その他」での精度では，「意味等」を用いる方法2が他の手法よりも高い精度を出している[*8]．

高品質の機械翻訳システムを作成するには，解析が難しいテンス・アスペクト・モダリティでの解析精度が重要となる．全体としての精度が良かったとしても解を「現在形」か「過去形」だけを選んでいるだけでは高精度の機械翻訳システムの構築は難しい．「意味等」を用いる方法2は，全体での精度では「文字列」を用いる方法1に比べわずかに低いが，「その他」での精度では高い精度を出しているので方法1よりも優れていると考えることもできる．

d　といっても，「文字列」を用いる方法1は結構精度がよく，単純な文字列一致の方法でも比較的よい精度が得られることがわかる．

e　また，実験結果では，「その他」の結果は「現在形」「過去形」に比べて精度が著しく低かった．これはもともと「その他」の分類の方が「現在形」「過去形」に比べて難しい問題であるからと思われる．表6の各分類での精度を見てほしい．極端に悪いのは「現進」「過進」「完了」「will」「would」という分類で，進行相，完了相，未来事象といったものが絡むものである．これらの分類は，日本語で「過去」と「完了」，「現在」と「未来」を明確に区別して表現しないという問題や，英語の進行相の表出が微妙なニュアンスの違いにより行なわれたり行なわれなかったりするという問題があるため，もともと解くのが難しいものである．このためにその部分の精度が低くなっていると思われる．これらの分類の精度を向上させるには次に述べる(2)の問題点を解決する必要があると思われる．

(2)　提案手法の問題点として以下のことがあった．

a　本研究で用いた形態素解析システムJUMANでは，変化形に可能形といったものがないため，下記の例文の「解ける」などの解析がうまくいかなかった．

和文:　小学生なら大抵はこの問題は<u>解ける</u>
英訳:　Most elementary schoolboys and schoolgirls <u>can</u> solve this problem.

実際に上記の意味での「解ける」という例文があれば解けていたのかもしれないが，下記のような用例しか集まらず「現在形」と誤った解を出力した．

和文:　池の氷は3月に <u>解ける</u>
英訳:　The pond <u>thaws</u> in March.

和文:　太陽が雪を <u>解かす</u>
英訳:　The sun <u>melts</u> snow.

本研究では，形態素の変化形も用いていたが，JUMANでは「解ける」は「基本形」となってしまう．これでは可能の意味を持つ例文との間に大きい類似度を持つことはない．「解ける」は「解く」の可能の

*8　方法2が方法1よりも現在形と過去形以外のテンス・アスペクト・モダリティの解析に優れているのではないかということについては，実験結果とは別に理論的にも以下のような説明をつけることが可能である．もともと，現在形と過去形以外のテンス・アスペクト・モダリティの事例数は少ないため，事例の多数決を利用するk近傍法では解が現在形と過去形になりやすいが，方法1のような文字列だけを類似度とする方法では，用例を弁別する情報が少なく同点が生じやすく，同点の類似度で多数の用例が利用されやすく，事例の多い現在形と過去形になりやすい傾向がある．また，方法2のような形態素・意味情報も用いて類似度を求める方法では，用例を弁別する情報が増えるために，方法1に比べて同点の類似度になる可能性が下がりそれほど多数の用例が利用されず，事例の多い現在形と過去形になる傾向が緩和されやすくなっている．このため，方法2では現在形と過去形の精度が若干下がり，他の分類の精度が上がることになると思われる．

意味を持った場合の表現で，これを意味するような「可能基本形」とかいう変化形を設定できればこのような問題もうまく解析できるようになる[*9]．

b　本研究では訳出される英語文の構造を考慮せずに，ただ単純に日本語の文末表現からテンス・アスペクト・モダリティを推定した．しかし，訳出される英語文の構造がかわると用いるべきテンス・アスペクト・モダリティも変化する場合がある．例えば，一つ目の例文の「繁盛している」のテンス・アスペクト・モダリティは現在形であるが，二つ目の例文の「繁盛している」のテンス・アスペクト・モダリティは進行形である．

和文:　あの医者は 繁盛している
英訳:　The doctor has a large practice.

和文:　事務所は 繁盛している
英訳:　The office is flourishing.

日本語文ではこれらの文の動詞句部はほとんど同じ意味であり，同じテンス・アスペクト・モダリティを持っていると考えてもよいものだが，訳出に用いる動詞句をかえることで英語文でのテンス・アスペクト・モダリティが異なってくる[*10]．本提案手法でこの問題を解決したい場合は，機械翻訳システムの構造解析部で訳出に用いる英語動詞をあらかじめ推定しておいて，用例の取り出しの際に日本語側の照合だけでなく英語側の照合も行なって，翻訳システムの構造解析部が想定する英語側の構造 (あるいは動詞) とよく似た用例のみを取り出すということが必要になると思われる．なお，この処理を行なうには，機械翻訳システムの処理の手順において，英語側の構造 (あるいは動詞) の訳出を，テンス・アスペクト・モダリティの推定より先に行なうか，同時に行なうかする必要があることに注意してほしい．

c　本研究での実験結果では気にならなかったが，文末表現だけではなく文頭付近にある副詞なども利用した方がいい場合もある [久米 90]．例えば，下記の例では文末表現は同じで「もう」と「昨日」の違いだけで「過去完了」と「過去」の違いが出ている．

和文:　もう 登録しました．
英訳:　I've already registered.

和文:　昨日 登録しました．
英訳:　I registered yesterday.

用例を増やせば文末からの一致が文頭までいってこのようなものも扱えるようになるかもしれないが，そこまで用例が増えるのを期待するのは無謀なことであろう．このような例にも対処できるように文末だけでなく文の中身の表現も利用できるようにする必要がある．

おおよそ考察は以上のとおりである．本提案手法は，人手で規則を作成する手間がないという利点と実装の容易さをもっており，と同時に，最高クラスの市販ソフトと同程度以上の精度を得ている．このことから本提案手法の優秀性がわかる．また，この結果は，「はじめに」のところで述べた，用例ベース手法の有用性および有効性のさらなる確認を行なったことにもなる．

4. お わ り に

本研究は用例ベースの手法でテンス・アスペクト・モダリティの日英翻訳を試みたものである．本研究で採用した用例間の類似度の定義は，文末からの一致文字列 (もしくは分類語彙表の分類番号も含めた文字列での一致) であり，非常に簡明なものであった．また，用例におけるノイズの対策として k 近傍法を利用していた．実験の結果，市販のソフトと同程度かそれ以上の解析精度を得た．本手法は人手で規則を作る必要もなくまた機械翻訳についての専門的な知識がなくてもできるものなので，この方法で市販のソフトと同程度以上の解析精度を得ただけでもその手法の有用性が大いに認識できるだろう．

また，本研究では類似度の尺度として「文字列」を用いる方法 1 と「意味等」を用いる方法 2 の二つを用いたが，全体での精度では「文字列」の方がわずかによかった．しかし，「現在形」「過去形」以外のテンス・アスペクト・モダリティを意味する「その他」での精度では「意味等」を用いる方法 2 の精度の方がかなりよく，高品質の機械翻訳システムを作成するには解析が難しいテンス・アスペクト・モダリティでの解析精度が重要となると思われるので，「その他」での精度のよい「意味等」を用いる方法 2 の方が有用であると考えられる．とはいえ，用途によっては「文字列」の方を用いる方がいいかもしれず，適切に使い分けていくことも肝要であろう．

本研究はテンス・アスペクト・モダリティの日英翻訳の研究であったが，英語部分で主語を除いて動詞句を先頭にして文頭からの文字列一致などを使うとテンス・アスペクト・モダリティの英日翻訳にも用いることができるだろう[*11]．また，本研究は専門的な知識がなくてもで

*9　この例は実はもっと難しい問題をはらんでいる．「可能基本形」という変化形を設定できれば「この問題は 解ける」は解けるようになるが，「池の氷は 3 月に 解ける」の「解ける」も「可能基本形」にしてしまうとこちらの「解ける」の方は現在形ではなく可能であると誤って解析してしまうことになる．両方の「解ける」を正しく解析するには「解ける」という単語の多義性の解消を正しく行なえる必要がある．

*10　一つ目の例文では，「医者」などが繁盛している場合に用いる成句の "have a large practice" が用いられて進行形にはならない．

*11　この場合，英語部分での主語や動詞句の認定が必要なために，構文解析などの深い解析が必要になると思われる．しかし，近年高精度の構文解析システムが作成されはじめている [Collins

きるものなので，実際日本人以外の人が本研究を行なうことも容易である．さらには本手法は文法などをわざわざ設定する必要がないので文法などが未発達な言語においては極めて有用な手法となろう．

また，本研究の手法はテンス・アスペクト・モダリティ自体の研究にも利用できると思われる．例えば，対訳例の代わりに日本語文に正解のテンス・アスペクト・モダリティの情報を付与したデータを用いることで，日本語文のテンス・アスペクト・モダリティ推定はすぐに実現できる．また，大量の用例に対して本研究の方法 2「意味等」を用いて日本語文を言語解析結果の情報も盛り込んだような文字列に変換してから文末からの文字列でソートしたようなデータを作成すれば，そのデータではよく似たテンス・アスペクト・モダリティを含む文が近くに集まってくるので日本語研究の考察のための資料としても有用なものが作成されるであろう*12．

◇ 参 考 文 献 ◇

[Collins 96] Michael Collins: A New Statistical Parser Based on Bigram Lexical Dependencies. In *34th Annual Meeting of the Assocítation of the Computational Linguistics*, 1996.

[Collins 99] Michael Collins: Head-Driven Statistical Models for Natural Language Parsing. PhD Dissertation, University of Pennsylvania, 1999.

[Fukunaga 72] Keinosuke Fukunaga.: *Introduction to Statistical Pattern Recognition*. Academic Press Inc., 1972.

[古瀬 94] 古瀬 蔵, 隅田 英一郎, 飯田 仁: 経験的知識を活用する変換主導型機械翻訳. 情報処理学会論文誌, Vol. 35, No. 3, pp.414-425, 1994.

[国立国語研究所 64] 国立国語研究所: 分類語彙表. 秀英出版, 1964.

[久米 90] 久米雅子, 豊島孝之, 永田昌明: 話し言葉翻訳のための日本語アスペクト処理. 情報処理学会第 40 回全国大会予稿集，1F-7, pp. 415–416, 1990.

[黒橋 97] 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 3.5. 京都大学大学院工学研究科, 1997.

[黒橋 98] 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.

[村田 98] 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.

[村田 00] 村田真樹, 神崎享子, 内元清貴, 馬青, 井佐原均: 意味ソート msort — 意味的並べかえ手法による辞書の構築例とタグつきコーパスの作成例と情報提示システム例 —. 言語処理学会誌, Vol. 7, No. 1, 2000.

[Nagao 84] Makoto Nagao. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. Artificial and Human Intelligence, pp. 173–180, 1984.

[岡本 97] 岡本青史, 太田唯子, 湯上伸弘: k-最小近傍法におけるノイズの影響. 人工知能学会全国大会予稿集, 1997.

[Sato 90] Satoshi Sato and Makoto Nagao: Toward Memory-based Translation. In *COLING '90*, pp.203–212. 1990.

[佐藤 91a] 佐藤 理史: MBT1:実例に基づく訳語選択．人工知能学会論文誌, Vol. 6, No. 4, pp.592-600. 1991.

[佐藤 91b] 佐藤 理史: MBT2:実例に基づく翻訳における複数翻訳例の組合せ利用．人工知能学会論文誌, Vol. 6, No. 6, pp.861-871. 1991.

[清水 76] 清水護, 成田成寿 (編): 講談社和英辞典. 講談社, 1976.

[白井 90] 白井諭, 横尾昭男, Francis Bond: 新聞記事翻訳における時制の訳出について. pp. D–69, 1990.

[Stanfill 86] Craig Stanfill and David Waltz: Toward Memory-based Reasoning. Communications of the ACM, Vol. 29, No. 12, pp.1213–1228. 1986.

[Sumita 90] Eiichiro Sumita, Hitoshi Iida, and Hideo Kohyama: Translating with examples : A new approach to machine translation. In *The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, pp.203–212. TMI, 6 1990.

[Sumita 91] Eiichiro Sumita and Hitoshi Iida: Experiments and prospects of example-based machine translation. In *29th Annual Meeting of the Associtation of the Computational Linguistics*, pp.185–192. 1991.

[富浦 96] 富浦洋一, 日高達. k-nn 推定法に基づく統語的曖昧さの解消法. 言語理解とコミュニケーション研究会 NLC96-7, pp. 39–45, 1996.

[Watanabe 92] Hideo Watanabe: A Similarity-Driven Transfer System. In *COLING '92*, pp.770–776. 1992.

〔担当委員：山本秀樹〕

1999 年 12 月 8 日 受理

著 者 紹 介

村田 真樹 (正会員)

1993 年京都大学工学部卒業．1995 年同大学院修士課程修了．1997 年同大学院博士課程修了，博士 (工学)．同年，京都大学にて日本学術振興会リサーチ・アソシエイト．1998 年郵政省通信総合研究所入所．研究官．自然言語処理，機械翻訳，情報検索の研究に従事．言語処理学会，情報処理学会，人工知能学会，ACL，各会員．

馬 青

1983 年北京航空航天大学自動制御学部卒業．1987 年筑波大学大学院理工学研究科修士課程修了．1990 年同大学院工学研究科博士課程修了．工学博士．1990 ～ 93 年株式会社小野測器勤務．1993 年郵政省通信総合研究所入所，主任研究官．人工神経回路網モデル，知識表現，自然言語処理の研究に従事．日本神経回路学会，言語処理学会，電子情報通信学会，各会員．

内元 清貴

1994 年京都大学工学部卒業．1996 年同大学院修士課程修了．同年郵政省通信総合研究所入所．研究官．自然言語処理の研究に従事．言語処理学会，情報処理学会，ACL，各会員．

井佐原 均 (正会員)

1978 年京都大学工学部卒業．1980 年同大学院工学研究科修士課程修了．博士 (工学)．同年通商産業省電子技術総合研究所入所．1995 年郵政省通信総合研究所入所．2000 年同けいはんな情報通信融合研究センター感性情報処理研究室室長．自然言語処理，機械翻訳の研究に従事．言語処理学会，情報処理学会，人工知能学会，日本認知科学会，ACL，各会員．

96, Collins 99] ので，このことはさほど大きな問題にならなくなると思われる．

*12 文献 [村田 00] ではわれわれは意味ソートという方法について述べている．ここでのソートは，本研究の方法 2「意味等」において「形態素」「変化形」を用いず「分類語彙表の分類番号」のみを用いる場合，意味ソートを文末から多段に用いることと全く等価になる．