自然言語処理 130−16
(1999. 3. 5)

用例ベースによるモダリティの日英翻訳

村田 真樹　馬 青　内元 清貴　井佐原 均

郵政省 通信総合研究所 関西先端研究センター 知的機能研究室

〒651-2401 神戸市西区岩岡町岩岡 588-2

TEL:078-969-2181 FAX:078-969-2189 http://www-karc.crl.go.jp/ips/murata

{murata,uchimoto,qma,isahara}@crl.go.jp

あらまし

本稿では用例ベースによるモダリティの日英翻訳の手法を提案する．本稿の用例の利用における類似度は，文末からの一致具合である．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る．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うかがえる．また本手法の考え方はモダリティ自体の解析にも利用できる．

キーワード　モダリティ，日英翻訳，用例ベースのアプローチ

The Example-Based Approach to Japanese-to-English Modality Translation

Masaki Murata　Qing Ma　Kiyotaka Uchimoto　Hitoshi Isahara

Intelligent Processing Section, Kansai Advanced Research Center,

Communications Research Laboratory, Ministry of Posts and Telecommunications

588-2, Iwaoka, Nishi-ku, Kobe, 651-2401, Japan

TEL:+81-78-969-2181 FAX:+81-78-969-2189 http://www-karc.crl.go.jp/ips/murata

{murata,uchimoto,qma,isahara}@crl.go.jp

Abstract

We propose a new method of Japanese-to-English modality translation by using the example-based method. We have defined the similarity of examples as the degree of the semantic match in an expression at the end of a sentence. We have also used the k-nearest neighbor method in order to prevent the problem of noise. We have carried out experiments on modality translation and have obtained a higher accuracy rate than the high-quality MT software currently available on the market. Our method has an advantage in that it does not need to be manually customized.

key words　Modality, Japanese-to-English Translation, Example-Based Approach

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，日本語のモダリティ表現の翻訳の研究を行なう．従来の手法ではモダリティ表現は時制や様相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[1][2]．しかし，日本語の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モダリティ表現に良く似た対訳例(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，解析する表現に良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だが[3]，その手法はその有用性のわりに名詞句「A の B」の訳し分け[4]に利用されて以来，目立った使われ方はしていない．本研究はそういう意味で用例ベースの手法の有効性の検証研究としての意味合いもある．

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，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．解析や生成の研究が現状のようなままで本格的な機械翻訳の研究を行なうのは本末転倒であり，土台を作らずに楼閣を建てるようなものである．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結構うまくできる場合もあり，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[5]．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だった．

日本語文ではモダリティは文末にあるので，モダリティの翻訳にもこの考え方は利用できるのではないだろうかと考えた．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)の英語文のモダリティをその日本語文の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，モダリティが現在形であるのをみて，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．この問題で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初めて適用された「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,B 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．本研究では類似した用例を取り出すために文末の一致具合を調べるが，これを単純に文字列で調べるのではなく，形態素解析をして単語の認定を行なったデータにおいて一致具合を調べた方がいいかもしれない．また，単語同士の類似度も文字列一致で見るよりは，シソーラスでの近さにより求めた方がいいかもしれない．そのように考え，文末一致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[6] 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[7] 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．

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

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 0 0 0 0 2 1 ：は 38 0 5 9 1 1 ：大望 07 2 4 0 3 1 ：を ：いだいて 17 4 3 5 1 2 タ系連用テ形 ：いる 02 0 0 2 1 2 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ここの部分はわざわざ用いる必要はない．また，5,6 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，5,6 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

[1] ここでは KNP[8] 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．

表 2: 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	彼は鷹触りのいい人だ	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03000021:は38059111:私01000021:の07000111:知り合い03012221:だ基本形
番号	類似度	252423222120191817161514131211109876543 21
1	25	彼03000021:と38059111:は38059111:長年07246111:の07000111:知り合い03012221:だ基本形
2	24	ふた03059111:り:は38059111:長い間:の07000111:知り合い03012221:だ基本形
..	..	..
4	11	彼ら03000021:は38059111:多年07016111:の07000111:知己03012221:だ基本形
..	..	..

る[9][10][11]．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[11]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」の文のモダリティを解析しているものである．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」などのモダリティの分類は用例の英語文の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

[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．

[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食わず高速に用例を取り出すことができる．

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)

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．

3 実験と考察

3.1 実験

2節で説明した方法でどのくらい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[12]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ベースになかったがごとく扱って解析している．

また，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，モダリティの検出に用いる英語側のモダリティ表現としては，単文ならばその唯一の動詞のモダリティを，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．

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節で説明した方法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

[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 6: 各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%

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

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 300 文では設定した 27 種類のモダリティすべては出現せず 13 種類のモダリティしか出現しなかった. その 13 種類のモダリティごとの精度を表 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.

3.2 考察

本節の考察では整理と理解しやすさのために, 二階層の箇条書によって記述することにする. 以下では, 1 精度, 2 傾向, 3 問題点, 4 注意事項の順に説明する.

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.
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.

2. 本実験を通じて得られた傾向として以下のものがある.
k 近傍法は, k の値が小さいときは少数の類似度の大きい用例によって解析することになっており, ある種, 賭けのような要素がある. k の値を大きくすると安定した解になり, 「現在形」や「過去形」などの頻度の多い分類になりやすく, 無難な解を選択するようになっている. (例えば表 4の「現在形」や「過去形」の欄の精度は概ね k の値を増やすと向上する傾向がある.) このことは, 今必要とされている問題が, 無難なものか, それとも, 少々勝負に出て細かいモダリティまで推定したいのかに応じて, k の値を調節することできるということを意味している.
以上のことは必ずしも精度の表からは読みとれないところもあるが, 実際の実験を行なった感覚から以上の

ことが生じているのではないかと思っている．

3. 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解ける
英訳: Most elementary schoolboys and schoolgirls can 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月に解ける
英訳: The pond thaws in March.

和文: 太陽が雪を解かす
英訳: The sun melts 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[7]．

(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「送っている」のモダリティは現在形であるが，二つ目の例文の「送っている」のモダリティは進行形である．

和文: 彼は質実な生活を送っている
英訳: He lives a sober and simple life.

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.

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．本提案手法を実際に用いて高品質な処理を行ないたい場合は，用例の取り出しの際に，日本語側だけの照合だけでなく，英語側の照合，つまり，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）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．

(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の頭あたりにある副詞なども利用した方がいい場合もある[(1)]．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．

- もう登録しました． I've already registered.
- 昨日登録しました． 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．

我々は以前に自由回答アンケートにおけるモダリティを文末表現だけでなく文全体のすべての箇所に存在する手がかり語を用いて自動推定する研究を行なっている[(13)]．そこでの方法は文中のあらゆる文字列を素性として扱って最大エントロピー法を用いて有用な文字列（手がかり語）を特定してモダリティを推定するものだった．本研究のモダリティの日英翻訳もこれと全く同じ方法で解けば，上記の問題は解消される．ただし，この手法は教師あり学習なので，すべての用例にモダリティの分類をふっておく必要がある．現状ではまだ分類をふらずに実験をしているので行なうことはできない．

4. 本研究の実験について注意しておいて欲しいこととして以下のものがある．

(a) 本実験での正解判定は極めて厳しいものである．もとの例文についている英訳のモダリティと寸分違わず一致したときのみ正解とするものである．もともと，日本語に対して英語はいろいろな訳が想定できるので，もとの例文についている英訳のように必ずしも訳す必要はない．実際に，それぞれのモダリティ表現が正しいかを日英の両方の言語に詳しい人にチェックしてもらうと，ほとんどの解析結果が正解であったということにもなるかもしれない．

また，市販ソフトとの誤りの重なり具合を示す表5を見てほしい．市販ソフトと提案手法の双方が誤っている結果が多いことに気づくだろう．これは本来なら双方正しい解を出力するところが，入力文についている英訳が特殊なものになっていて普通ならこう訳すというものとは違った訳がついているため，双方解析を誤るといった結果が多く出ていたためである．

[7] 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．

(b) 本研究はモダリティの解析を行なうものであったが，実験データのほとんどが「現在」か「過去」であった．つまり，適当にやってもうまく解析ができる可能性のある簡単なものである．偶然，正解していただけというものも多く存在すると思われ，実際の精度は実験を行なうたびに大きく揺れると思われる．(実際に本実験でも最初の100個を調べた途中段階では，提案手法の精度は市販ソフトの精度を大きく上回っていた．)

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，それでありながら最高クラスの市販ソフトと同程度以上の精度を簡単に得たことは，本手法の優秀性を物語っている．「はじめに」のところで用例ベースの手法の有効性の検証研究としての意味合いもあると述べていたが，この結果は大いに用例ベースの手法の有効性を検証したことになろう．

4 おわりに

本研究は用例ベースの手法で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モダリティの英日翻訳にも用いることができるだろう．また，本研究は専門的な知識がなくてもできるものなので，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マイナー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，日本文の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[8].

参考文献

(1) 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, (1990), pp. 415–416.

(2) 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, (1990), pp. D–69.

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

(4) Eiichiro Sumita, Hitoshi Iida, and Hideo Kohyama, Translating with examples : A new approach to machine translation, *The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, No. 3, (TMI, 1990), pp. 203–212.

(5) 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).

(6) 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).

(7) 国立国語研究所, 分類語彙表, (秀英出版, 1964).

(8) 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).

(9) Keinosuke Fukunaga, *Introduction to Statistical Pattern Recognition*, (ACADEMIC PRESS INC., 1972).

(10) 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, (1996), pp. 39–45.

(11) 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).

(12) 滑水護, 成田成寿(編), 講談社和英辞典, (講談社, 1976).

(13) 乾裕子, 内元清貴, 村田真樹, 井佐原均, 文末表現に着目した自由回答アンケートの分類, 情報処理学会 自然言語処理研究会 NL128-25, (1998).

(14) 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 —, 情報処理学会 自然言語処理研究会 130-12, (1999).

[8] 文献(14)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．