モダリティコーパスの修正：機械翻訳を利用した評価実験

村田 真樹 *1　内山 将夫 *1　内元 清貴 *1　馬 青 *2,*1　井佐原 均 *1

*1 情報通信研究機構　*2 龍谷大学　(murata @nict.go.jp)

1　はじめに

教師あり機械学習法を利用した言語処理研究の発展と共に，タグ付きコーパスの有用性が一段と高まってきた．また，研究分野の多様化に伴い作成されるコーパスの種類も増えてきた．しかし，一般にコーパスには人手でタグを付けるが，人手によりタグ付けされたタグには誤りが生じる場合がある．本稿ではこのコーパス中の誤ったタグを自動修正する研究について述べる．われわれのコーパス修正の方法は汎用的な手法である機械学習を用いる方法なので汎用性があり，多くの種類のコーパスの修正に利用できる．

本稿では特にわれわれが作成中であるモダリティコーパスに対して行なったコーパス修正の研究について述べる．このコーパスはモダリティ表現の機械翻訳の研究に役に立つものである．モダリティコーパスの修正については文献 [1] においても述べてきたが，本稿ではこの先行研究で行なったコーパス修正の概略とそのコーパス修正の効果を確かめるために実際に機械翻訳を利用して行なった評価実験について述べる[1]．

2　機械翻訳用モダリティコーパス

本研究の対象であるモダリティコーパスの一部を図1に示す．図のようにこのコーパスは日英の対訳文からなり，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．

また，日本語側の文の先頭に “c” や “d” といった記号が付けられているが，これらはこの対訳データの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[2])．モダリティの分類としては以下

[1]本稿は文献 [2] に掲載予定の内容のうち文献 [1] で述べていないことについて記述している．文献 [1] とともにご覧くだされば幸いである．なおコーパス修正に関して特許 [3] を取得している．

[2]図 1の一つ目のデータには “,” 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モダリティが右に<vj> で囲まれた動詞に対するモダリティが記述される．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．

, この子ども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 1: モダリティコーパスの一部

表 1: モダリティの分類の出現率

分類	出現率
現在形	0.41
過去形	0.35
命令形	0.05
現在完了形	0.04
助動詞 “will”	0.03
現在進行形	0.03
助動詞 “can”	0.02
その他	0.07

の 34 種類のものを用いた．これ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. { 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, had better, may, might, must, need, ought, shall, should, used to, will, would の 19 種類)
4. 名詞句 (1 種類)
5. 分詞構文 (1 種類)
6. 動詞省略 (1 種類)
7. 間投詞、挨拶文など (1 種類)
8. 日本語と英語で動詞の対応がとれない場合 (1 種類)
9. 作業不可 (1 種類)

この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のできあがりとなる．このため，機械学習に基づくモダリティ推定の研究では，このモダリティの分類を示すタグと日本語文のみが用いられる．

上記の仕様で，コーパスの作成を外注した．対訳データは，講談社和英辞典 [4] から取った約 40,000 文の例文を用いた．外注先の会社では，英語文への “<v>”

などのタグの付与およびその英語文の動詞を見てモダリティの分類のタグの付与を人手で行なった．また，複数回のチェック作業を行なってもらい，外注先の会社としては誤りは全くないという状態までこの作業を行なってもらった．各モダリティの分類の出現率を表1に示す．

3　コーパス修正の方法

本節では文献 [1] において良いとされているコーパス修正の方法を説明する．

まず，コーパス中の各タグが誤っているか否かの判定であるが，これは元のタグおよびそれ以外の場合のタグ (本稿では残る 33 種類のタグ) の生起確率を求め，元のタグの生起確率が最も大きい場合元のタグは正しいと判定し，元のタグ以外のタグの生起確率が最も大きい場合元のタグは誤っていると判定することによって行なわれる．次に誤っていると判断されたタグについては，そのとき生起確率が最も大きいタグを修正後のタグとして付け直すということを行なう (実際にはこの付け直しは人手で確認の上，行なう)．

各タグの生起確率の算出には，本研究では文献 [1] で良いとされている最大エントロピー法に基づく方法を利用した．詳細は文献 [1] を参照のこと．

最大エントロピー法により生起確率を推定する際には各タグの文脈となる素性を定める必要がある．本研究では以下のものを用いた．

- 英語文の最初の<v>/<vj> の前方の 1 ～ 5gram の文字.
- 英語文の最初の<v>/<vj> の後方の 1 ～ 10gram の文字.
- 英語文の最後の<v>/<vj> の前方の 1 ～ 10gram の文字.
- 英語文の文末の 1gram の文字.

上記の方法だけでもコーパスの修正はできるが，本稿のような確率値が求まる手法の場合は以下で述べるようにコーパス修正の確信度を定義できるので，コーパス修正の候補の各箇所をこの値によってソートして，コーパス修正の確信度が高いところから修正していった方が便利である．本稿では，このソートに用いるコーパス修正の確信度としては修正後のタグの生起確率を利用した．

機械学習の方法で確率の値を算出する際には，クローズで算出するかオープンで算出するかの二通りの方法がある．クローズによる確率値の算出は，修正すべきかを判断するタグの部分のデータも含めて確率値を算出する方法で，オープンによる確率値の算出は，修正すべきかを判断するタグの部分のデータを含めずに確率値を算出する方法である．本稿のオープンによる確率値の算出方法には 10 分割のクロスバリデーションを利用した．

文献 [1] での実験結果では，クローズでの確率推定をする方法はオープンでの確率推定をする方法に比べて精度は高いが誤りタグの検出数および修正数はオープンの方法の方が多かった．このため，コーパス修正の方法としては，まずクローズでの確率推定を用いてコーパス修正を高精度に行ない，次にオープンでの確率推定を用いてさらに多くのコーパス修正を行なうというのが良いと思われる．

4　コーパス修正の実験

前節で述べた方法で実際にコーパス修正の実験を行なった．講談社和英辞典のデータのうち “<v>” のタグのある総数 39,718 個のモダリティタグのデータに対して行なった．実験結果を表 2に示す．「総数での精度」はシステムの全修正箇所における精度を意味し，「上位 X 個」は確信度によるソートを行なったデータでの上位 X 個のシステム修正箇所での精度を意味する．「検出精度」はシステム修正箇所において誤りの検出を成功した割合を意味し，「修正精度」はシステム修正箇所においてタグの修正を成功した割合を意味する．「抽出総数」はシステムがタグ誤りと判定しタグを修正した箇所の総数を意味する．

実験結果の表 2から，検出精度と修正精度はほとんど同程度の精度であり，誤り検出だけでなく修正も同時に行なった方が良いことがわかる．精度以外の面でも実際に人手で確認しながら修正をする際には修正候補があるとどのように間違っているのかがわかって便利である．「総数での精度」と上位精度を比較すると圧倒的に上位精度の方が高い．確信度に基づいてソートすることが重要であるとわかる．精度ではクローズの方法が良いが抽出総数ではオープンの方法の方が良かった．クローズの方法とオープンの方法をあわせて，人手で確認も行なってコーパス修正を行なったところ，合計で 178 個の誤りを実際に修正できた．コーパス全体では 0.44%(=178/39718) の誤りを修正できたことになる．

コーパスに元々どのくらいの誤りが含まれているかを調べてみた．コーパスからランダムに 300 個取り出したところそのうち 4 個が誤りであった．このため，コーパスには元々 1.33%(=4/300) の誤りが含まれていると予想される．

オープンの方法で考えると約 700 個，すなわち，コー

表 2: コーパス修正の精度

	クローズで確率推定 (抽出総数 184 個)				オープンで確率推定 (抽出総数 694 個)			
	検出精度		修正精度		検出精度		修正精度	
上位 50 個	100%	(50/ 50)	100%	(50/ 50)	88%	(44/ 50)	88%	(44/ 50)
上位 100 個	92%	(92/100)	92%	(92/100)	88%	(88/100)	88%	(88/100)
上位 150 個	77%	(116/150)	77%	(116/150)	80%	(121/150)	80%	(120/150)
上位 200 個	69%	(127/184)	68%	(126/184)	67%	(135/200)	67%	(135/200)
上位 250 個	—	—	—	—	60%	(150/250)	60%	(150/250)
上位 300 個	—	—	—	—	52%	(158/300)	52%	(158/300)
総数での精度	69%	(127/184)	68%	(126/184)	25%	(178/694)	25%	(174/694)

表 3: タグ付け作業者による誤り

誤りのタグ	正しいタグ	頻度
現在形	過去形	53
過去形	現在形	28
現在進行形	現在形	15
命令形	現在形	6
現在完了形	現在形	6
現在形	命令形	6
現在形	現在完了形	6
助動詞 “can”	現在形	6
現在完了形	過去形	4
助動詞 “may”	助動詞 “must”	3
現在形	助動詞 “will”	3
現在形	現在進行形	3

パスの 1.76%(700/39718) を人手でチェックするだけで 178 個の誤りを修正でき，1.33% 程度含まれる誤りのうち 0.44%，すなわち，誤りの 1/3 を本手法で修正できたことになる．

最後にコーパスにはどのような誤りが多かったかを調べてみた．修正できた 178 個の誤りの内訳を表 3に示す．ただし頻度 3 以上のものだけを示している．「誤りのタグ」はタグ付け作業者が付けた誤ったタグで，「正しいタグ」は本研究のコーパス修正で修正したタグである．まず出現頻度の大きい「現在形」や「過去形」に関する誤りが多いことに気づく．次に，「現在形」を「現在進行形」に誤るパターンが多いことに気づく．タグ付け作業者はタグ付けの際には，日本語文と英語文の両方を見ていて，日本語文は「ている．」が文末にあるとたいていは「現在進行形」の意味になるため，日本語文が「ている．」で英語文が “-ing” を使わずに現在形の場合に，日本語文の影響を受けて誤って「現在進行形」とタグ付けしたようである．そういう種類の誤りが多く見られた．他の誤りのパターンとしては，助動詞 “must” を “may” と誤るものがあった．このコーパスでは “must” のためには “u” を “may” のためには “m” の記号をコーパスのタグ付けの際に利用した．このとき作業者は “must” の記号は “u” だと覚えるのが難しく誤って “may” の記号の “m” をタグ付けしたと思われる．タグ付けに利用する記号の定義も作業者の誤りの原因に影響するようである．

5 機械翻訳を利用した評価実験

前節で行なったコーパス修正の効果を調べるために，日本語のモダリティ表現を実際に英語に翻訳するシステムが，コーパスを修正したことによりどの程度性能があがるかを調べる実験を行なった．ここでは機械翻訳手法としては機械学習に基づくものを利用した．実験には 40,198 文のコーパスを利用した．(コーパス修正の研究で取り除いていたデータもこの実験では用いるのでデータ数はコーパス修正の実験よりも増える．) 入力は日本語文で，出力は主動詞のモダリティの分類である．コーパスから 800 文を抜き出してこれを評価用に用いる．残りの 39,398 文は機械学習の学習データとして用いる．最大エントロピー法とオープンの方法で抽出した 178 個の誤りを修正して修正済みコーパスを作成した．この修正済みコーパスで学習した結果と，元のコーパスで学習した結果の精度を比較することで，コーパス修正の効果を確かめる．

機械学習法としては，k 近傍法，決定リスト法，最大エントロピー法 (ME)，サポートベクトルマシン法 (SVM) を用いた．この実験はわれわれの先行文献 [5] とほぼ同じである．詳細は文献 [5] を参照のこと．

素性としては以下のものを用いた．

- 素性セット 1
  素性セット 1 は，入力された日本語文の文末の 1 から 10 文字までの文字列と，入力された日本語文のすべての形態素である．
- 素性セット 2
  素性セット 2 は，入力された日本語文の文末の 1 から 10 文字までの文字列である．

サポートベクトルマシンには素性セット 1 を利用し，他の機械学習法では素性セット 2 を利用した．それぞれの素性がそれぞれの機械学習法で有効であることは先行文献 [5] で確認済みである．

表 4: 機械学習を利用した実験結果

方法	修正済みコーパス	修正前コーパス
k 近傍法	93.50%(748/800)	93.38%(747/800)
決定リスト	92.25%(738/800)	91.88%(735/800)
ME	94.12%(753/800)	93.88%(751/800)
SVM	94.62%(757/800)	94.38%(755/800)

表 5: ベースライン手法と市販の機械翻訳システムを利用した実験結果

方法	精度	
ベースライン	87.00%	(696/800)
ソフトウェア 1	89.25%	(714/800)
ソフトウェア 2	88.50%	(708/800)
ソフトウェア 3	92.88%	(743/800)
ソフトウェア 4	88.25%	(706/800)
ソフトウェア 5	92.88%	(743/800)
ソフトウェア 6	88.62%	(709/800)

実験結果を表 4と表 5に示す．日本語文が複数の英語のモダリティ表現に翻訳可能な場合があるので，ここでは厳密な評価のための評価用データを作成した．この評価用データでは次のことを外注して行なった．各モダリティのタグの正解セットには，修正済みコーパスに付いていたタグ，独立して作業してもらった三人のプロの翻訳家が翻訳して作成した英語文のモダリティ表現に対応するモダリティのタグをまず加えた．次にさらに新しいもう一人のプロの翻訳家が入力の日本語文と翻訳された英語文を吟味してさらに追加可能なモダリティのタグがないかを調べてもらって追加可能な場合はそのモダリティのタグを正解セットに追加してもらった．この正解セットが評価用データである．このデータにあるモダリティのタグを出力すれば正解とし，なければ不正解とした．

表 4の「修正済みコーパス」は修正済みのコーパスを学習に利用した場合の精度を意味し，「修正前コーパス」は修正前のコーパスを学習に利用した場合の精度を意味する．

表 5には，ベースライン手法と 6 個の最新の市販の機械翻訳システムの結果を示している．ベースライン手法は，「た」(日本語の過去を意味する終助詞に「た」があり「た」で終わっている場合はたいていの場合過去である．) で終わっている文を過去形と判断しそれ以外の場合を現在形と判断する．市販の機械翻訳システムについては何も答えを出さない場合はベースライン手法の出力を代りに利用することにした．

表 4の実験結果では，修正済みコーパスを利用した手法で正解し修正前のコーパスを利用した手法で誤った事例が 8 個あり，修正済みコーパスを利用した手法で誤り修正前のコーパスを利用した手法で正解した事例は全くなかった．このことから，精度の向上が特に大きいというわけではなかったとしても，コーパス修正は機械翻訳の性能を安定して向上させることがわかった．サポートベクトルマシンの方法が最も高い精度を出しているが，この方法ではコーパス修正を利用することで 0.24%(= 94.62%−94.38%) の精度向上を実現している．元々誤っていた事例は 5.62%(=100%−94.38%) であるので，このうちの 0.24% が改善されたということで，改善率は 4.3% (=0.24/5.62) である．この改善率は無視できない数字である．われわれのコーパス修正では，39,718 個のデータのうち，178 個のデータを修正していた．このため，この修正の割合は 0.44% (=178/39718) である．精度向上の 0.24% は，コーパスを修正した割合の 0.44% の約半分であることがわかる．

われわれの機械学習に基づく方法と表 5の市販の機械翻訳システムの結果を比較すると，われわれの機械学習に基づく方法は市販の機械翻訳システムの性能と同等かそれ以上の性能を示していることがわかる．

6 おわりに

本稿ではわれわれが既に提案しているコーパス修正の方法について実際に機械翻訳を利用してその効果を確かめる実験を行なった．われわれのコーパス修正の方法により約 40,000 のデータのうちの 178 個の誤りを修正することができた．さらにこの誤りを正すことで機械翻訳システムの性能が実際に向上することを確認した．

参考文献

[1] 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).

[2] Masaki Murata, Masao Utiyama, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara, Correction of errors in a verb modality corpus used for machine translation with a machine-learning method, *ACM Transactions on Asian Language Information Processing*, (2005), (to appear).

[3] 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均, コーパス誤りの検出・修正システム，コーパス誤りの検出・修正処理方法およびそのプログラム記録媒体, 特許広報 (特許第 3396734 号), (2003).

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

[5] Masaki Murata, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara, Using a support-vector machine for Japanese-to-English translation of tense, aspect, and modality, *ACL Workshop on the Data-Driven Machine Translation*, (2001).