

LE-002

テンス・アスペクト・モダリティの翻訳における機械翻訳システムの誤りの調査

Error Analysis of Translation of Tense, Aspect, and Modality in Machine Translation System

村田 真樹[†] 内元 清貴[†] 馬 青[‡] 金丸 敏幸^{*†} 井佐原 均[†]
 Masaki Murata Kiyotaka Uchimoto Qing Ma Toshiyuki Kanamaru Hitoshi Isahara

1. はじめに

テンス・アスペクト・モダリティは、翻訳が難しい問題の一つとして知られている。本稿では、テンス・アスペクト・モダリティの日英翻訳における機械翻訳システムの誤りを調査した。本稿では市販の6個の機械翻訳システムと、われわれが開発しているテンス・アスペクト・モダリティの翻訳システムについて誤りの調査をした。本稿の調査では、われわれのシステムが他のシステムよりも性能が高いことを確認し、われわれのシステムでほとんど誤らず市販システムでよく誤るパターンを抽出をした。これらの結果は市販システムの性能向上に役に立つ情報である。また、個々の市販の翻訳システムの特徴的な誤りも分析・報告した。これらの特徴的な誤りは他のシステムでは正しく処理できるため、個々のシステムでもこの誤りに着目することで比較的容易に性能を改善できる点と思われる。本稿ではこのようなシステムの改善に役に立つ情報を提供している。

2. 問題設定

本稿で扱うテンス・アスペクト・モダリティの翻訳は、与えられた日本語文から、日本語文を英語文に翻訳した際の英語文の主節の動詞句での表層的なテンス・アスペクト・モダリティの表現を求めることであり、この翻訳の性能および誤りを調査することが本研究の目的である。

このテンス・アスペクト・モダリティの表現として以下の分類の組み合わせを利用している。以降、この分類をテンス・アスペクト・モダリティの分類と呼ぶ。

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, be to の現在形と過去形, an, 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21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)

[†] 独立行政法人情報通信研究機構。

{murata,uchimoto,qma,kanamaru,isahara}@nict.go.jp
 National Institute of Information and Communications Technology.

[‡] 龍谷大学。
 qma@math.ryukoku.ac.jp
 Ryukoku University.

* 京都大学。
 kanamaru@hi.h.kyoto-u.ac.jp
 Kyoto University.

7. 間投詞, 挨拶文など (1種類)

翻訳の評価用のデータとしては、文献[1, 2]の40,198文のコーパスから取り出した800文を利用した。この800文において、市販の6個の機械翻訳システムとわれわれの翻訳システムの性能を調べ、また、誤りのパターンを調べることで、誤り分析をする。

評価する市販の6個の機械翻訳システムは、2003年10月における各社の最新のシステムで、またそれぞれすべて有名な機械翻訳ソフトのメーカーのものである。

また、評価するわれわれのテンス・アスペクト・モダリティの翻訳システムは、文献[3]でも示している機械学習手法のサポートベクトルマシン(SVM)[4, 5]を利用するものである。このシステムはテンス・アスペクト・モダリティの表現の日英翻訳のみを行なうもので、日本語文から前述のテンス・アスペクト・モダリティの分類を求めることで、テンス・アスペクト・モダリティの表現の日英翻訳を実現する。テンス・アスペクト・モダリティの分類を解とする分類問題としてとらえてSVMでこの問題を扱う。多分類問題の扱いには、ペアワイズ法を利用して対処した。学習データには前述の40,198文のデータから評価用の800文を取り除いた残りをを用いた。

機械学習の素性としては以下のものを用いた。

- 素性セット1 … 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。
 例: 「なかった。」 「かった。」 「った。」
- 素性セット2 … 入力された日本語文のすべての形態素。
 例: 「今日」 「は」 「走る」

実験では、素性セット1と素性セット2を利用するもの、それぞれ片一方の素性セットのみを利用するものの三種類の実験を行なった。

次に評価方法について述べる。日本語文が複数の英語のテンス・アスペクト・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に以下の手順で評価を行なった。この評価は外注して行なった。800文の評価用データにもともと対訳例文としてあった英語文の主節の動詞句のテンス・アスペクト・モダリティの分類をまず正解とした。(元の対訳例文は和英辞典の例文を利用している。)さらに、独立して作業してもらった三人のプロの翻訳家に、800文のデータについて日本語から英語に翻訳してもらって作成した英語文のテンス・アスペクト・モダリティの分類、6つの市販の翻訳ソフトの出力した英語文のテンス・アスペクト・モダリティの分類、われわれの翻訳システムの出力のテンス・アスペクト・モダリティの分類をテンス・アスペクト・モダリティの分類の正解の候補とし、この候補をさらに新しいプロの翻訳家二人に同時にチェックしてもらい、正解の候補のうち、正しいとされたものを正解とした。この二人の翻訳家の意見が分かれること

表 1. テンス・アスペクト・モダリティの分類の出現率

分類	出現率
現在形	0.65 (516/800)
過去形	0.45 (356/800)
現在完了形	0.32 (259/800)
can	0.11 (90/800)
will	0.11 (87/800)
現在進行形	0.10 (82/800)
命令形	0.09 (74/800)
should	0.07 (59/800)
must	0.05 (43/800)
would	0.05 (37/800)
過去進行形	0.04 (35/800)
現在完了進行形	0.04 (28/800)
ought to	0.04 (28/800)
could	0.03 (23/800)
may	0.02 (19/800)
be going to	0.02 (18/800)
had better	0.02 (13/800)
shall	0.01 (12/800)
have to	0.01 (11/800)
be to	0.01 (10/800)

表 2. テンス・アスペクト・モダリティの翻訳精度

方法	精度
ベースライン	94.50% (756/800)
SVM(全素性)	98.75% (790/800)
SVM(素性セット1のみ)	98.25% (786/800)
SVM(素性セット2のみ)	94.38% (755/800)
市販ソフト A	97.00% (776/800)
市販ソフト B	97.00% (776/800)
市販ソフト C	95.88% (767/800)
市販ソフト D	95.50% (764/800)
市販ソフト E	94.75% (758/800)
市販ソフト F	94.25% (754/800)

が少数のデータで存在したが、本稿の調査では明らかな誤りのみをまず調査しようと考へ、二人の翻訳家の意見が分かれた場合でも片一方の翻訳家が正しいとした正解の候補は正解として扱った。しかし、本稿の調査では特異な文脈を無理に想定しなければ正解にならない正解の候補については誤りとしている。

正解の分類の出現率を表 1 に示す。出現頻度が 10 未満の分類は表示していない。複数の分類が正解になる場合があるため、出現率の総和は 1 を超える。

3. 実験および調査

前節の方法で評価実験を行なった。その評価実験での精度を表 2 に示す。表 2 には、ベースライン手法の結果も示している。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英語文の動詞部分を英語文として出力しなかった場合はベースライン手法の出力を代りに利用して精度を求めている。6 つの翻訳システムについてはそれぞれ A,B,C,D,E,F と呼ぶ。

表 2 からわれわれのシステムである全素性を用いる SVM が最も性能が高いことがわかる。市販ソフト A,B もなかなかの精度を出しており、市販ソフトの中では最も良い精度を出している。市販ソフト E,F はベースラインの精度とほとんど変わらない精度である。

次に誤りの分析を行なった。これは上述の精度を求める際に誤りと判定された事例において、誤りパターンを調べることで行なった。ここでは誤りパターンとしては、正解の分類とシステムの誤った出力の分類の対を利用した。正解が複数の分類である場合ではその複数の分類のそれぞれの分類の場合を誤りパターンとした。(つまり、正解が「現在形」と「現在進行形」の二つある場合にシステムの出力が「過去形」の場合は、「現在形」と「過去形」の対と「現在進行形」と「過去形」の対の両方が誤りパターンとして取り出される。)市販の機械翻訳システムで英語文の動詞部分を英語文として出力しなかった場合は「出力なし」という分類を設けて集計したが、出現頻度が少なく本稿で示す表には現れていない。その集計結果を表 3 に示す。表では、合計の頻度が 10 以上、または個々のシステムでの誤りの頻度が 3 以上のものを示している。

また、双対尺度法を用いて誤りの分布の傾向を調べた。これは 6 個の市販の機械翻訳システムと全素性を利用した SVM のシステムの 7 個のシステムで調査した。(素性セット 1, 素性セット 2 のみを用いる SVM のシステムも混ぜて双対尺度法により図を求めた場合、考察しづらい複雑な図となったため、これらの二つの SVM のシステムを除いて双対尺度法を使った。)これらのシステムのうち一つでも頻度が 3 以上だった誤りパターンを取り出し、この誤りパターンと 7 個のシステムで共起頻度を求めてクロス表を作成し双対尺度法[6]を実行した。その結果を図 1 に示す。図から誤りパターンから見た各翻訳ソフトの類似性を見ることが出来る。また、図は各翻訳ソフトに偏った誤りパターンを大雑把に把握するのに役も立つ。例えば、市販ソフト F の近くには「過去形：現在完了形」「現在形：現在完了形」「過去進行形：現在完了形」の誤りパターンが現れているので、市販ソフト F は「過去形」「現在形」「過去進行形」を「現在完了形」に誤ることが他のシステムに比べて特に多いことがわかる。

まず表 3 において、われわれのシステムの SVM について考察する。全素性を用いるシステムに比べて素性セット 1 のみを用いるシステムはすべての誤りパターンで少しずつ性能が落ちている。また、素性セット 2 のみを用いるシステムは「現在形」と「過去形」の間での誤りがかなり多く、形態素情報だけではテンス・アスペクト・モダリティの分類は困難であることがわかる。文末の文字列の情報が特に重要であることがわかる。

次に表 3 において、市販ソフトについて考察する。まず市販ソフト A と B だが、翻訳結果はほとんど同じであり、テンス・アスペクト・モダリティの分類はすべてまったく

表3. 誤りパターン

誤りパターン		サポートベクトルマシン				市販ソフト						合計	
正解	システムの誤り	素性 全て	素性1 のみ	素性2 のみ	合計	A	B	C	D	E	F	合計	
現在形	現在進行形	1	1	2	4	7	7	9	10	4	8	45	49
現在形	過去形	1	2	19	22	2	2	1	2	2	3	12	34
過去形	現在形	1	1	9	11	2	2	5	5	4	1	19	30
will	現在形	3	3	2	8	3	3	3	3	4	4	20	28
現在完了形	現在形	1	1	5	7	3	3	3	3	4	1	17	24
現在完了形	現在進行形	1	0	1	2	3	3	2	2	2	4	16	18
命令形	現在形	2	2	1	5	3	3	0	0	5	2	13	18
現在形	現在完了形	0	0	2	2	0	0	2	2	2	8	14	16
現在形	命令形	1	3	4	8	1	1	1	1	2	1	7	15
現在進行形	過去形	1	2	2	5	2	2	1	1	2	2	10	15
現在完了形	過去形	1	2	0	3	2	2	1	1	2	2	10	13
can	現在形	2	2	2	6	1	1	1	1	2	1	7	13
should	現在形	1	1	1	3	2	2	0	0	4	1	9	12
would	現在形	1	1	0	2	2	2	1	1	2	2	10	12
過去形	現在完了形	0	0	0	0	0	0	1	2	3	5	11	11
過去形	過去完了形	0	0	0	0	1	1	4	4	1	0	11	11
must	現在形	1	1	1	3	2	2	0	0	4	0	8	11
will	過去形	0	0	8	8	0	0	0	0	0	0	0	8
現在形	will	0	0	2	2	0	0	3	2	0	0	5	7
過去進行形	過去完了形	0	0	0	0	0	0	2	2	3	0	7	7
can	過去形	0	0	6	6	0	0	0	1	0	0	1	7
過去進行形	現在完了形	0	0	0	0	0	0	0	0	1	4	5	5
命令形	過去形	0	0	4	4	0	0	0	0	0	1	1	5
現在形	can	0	0	0	0	0	0	0	1	0	3	4	4
現在形	might	0	0	0	0	0	0	0	0	3	0	3	3

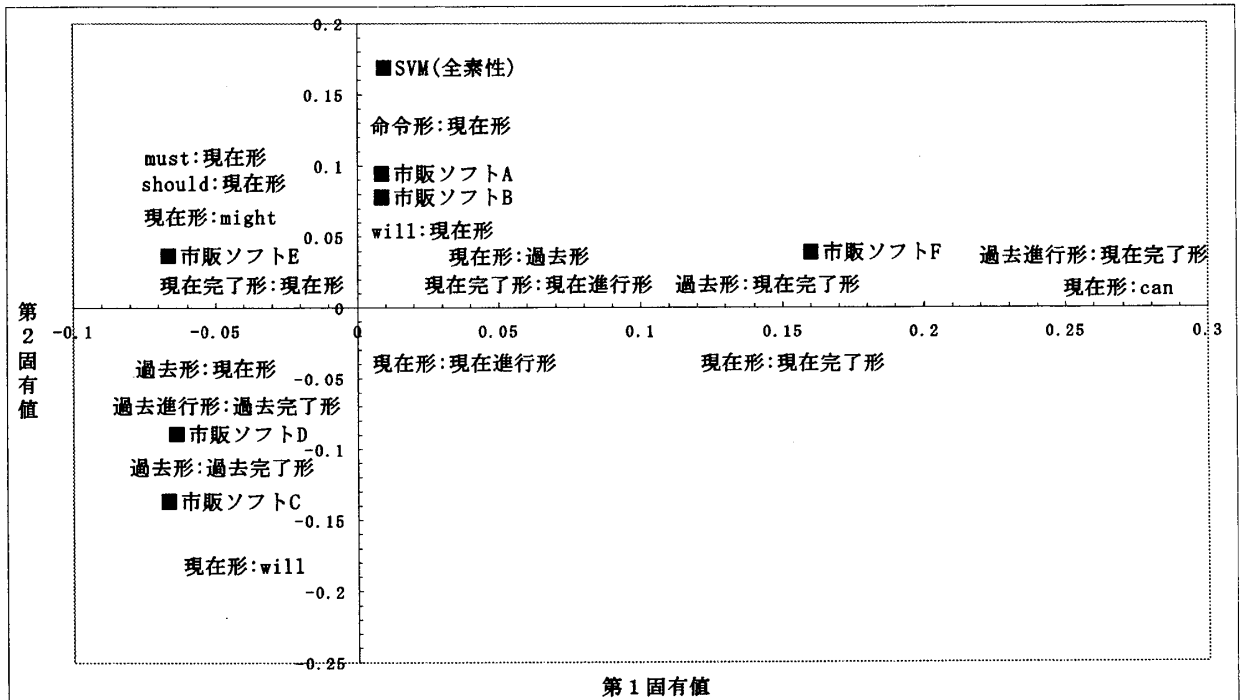


図1. 翻訳システムと誤りパターンの関係図

同じであり、誤りパターンの分布もまったく同じであった。市販ソフト A と B は、どちらも有名な翻訳メーカーであるが、出力がこれだけ酷似しているため、これらの会社で何らかの共同開発を行なっていることが予想される。

次に目立って気づくことは、どの機械翻訳システムでも高頻度に誤っていて SVM であまり誤っていない、「現在形」を「現在進行形」と誤る誤りパターンである。この誤りパターンの例文を以下に示す。

入力日本語文：この部屋には水道が取り付けられている。

翻訳結果：A water service is being installed on this room.

これはわざわざ進行形にしなくてもいいところを、進行形にしてしまう誤りである。機械学習法の SVM を用いるわれわれのシステムではこの誤りは市販ソフトに比べて極端に少なかった。機械学習手法による学習でこの問題を改善できることがわかった。このことは、今後の機械翻訳システムの開発に役に立つ知見である。その他にわれわれのシステムで誤りが少なく市販ソフトで誤りが多いものとして「過去形」を「現在形」また「現在完了形」を「現在形」や「現在進行形」に誤るものがある。特に「現在完了形」は翻訳が困難なことが予想されるものだが、これらに関する誤りは全素性を用いる SVM では誤りが少なくこの点の改良にもわれわれのシステムが役に立つ可能性がある。

市販システムとわれわれのシステムの両方でよく誤る誤りとしては「will」を「現在形」にする誤りである。日本語は未来のことも現在のことも同じ終止形で表現するため、「will」と「現在形」の訳し分けは困難であり、このためどのシステムでもこの誤りが多いと思われる。

次に、表3と双対尺度法の結果の図1をあわせて見ると、個々の市販の機械翻訳システムがどういった誤りを持徴的に示すかを調べてみた。市販ソフト C,D では「過去進行形」や「過去形」を「過去完了形」に誤ることが比較的多いことがわかる。以下に例を示す。

入力日本語文：風で帆がひらひらとひらめいていた。

翻訳結果：The sail had flashed flutteringly by the wind.

「過去進行形」か「過去形」でよいところ、わざわざ「過去完了形」にしている誤りである。これは市販ソフト C,D 特有の誤りであり、システムの調整を失敗して「過去完了形」を出力しやすくしていると思われる。市販ソフト C,D ではこれに注意すると性能を改善することができると思われる。また、市販ソフト C,D では「現在形」を「will」に誤ることが比較的多い。しかし、前述のように「will」と「現在形」の訳し分けは難しいので、この問題点を改善するのは難しいかもしれない。

また、市販ソフト E では、「現在形」を「might」に誤ることが比較的多い。例えば以下の例である。

入力日本語文：さて、相談したいことがあるのだ。

翻訳結果：Well, I might want to consult.

模範訳 1：Now, I have something to talk over with you.

模範訳 2：May I ask your advice?

「might」は少し不自然である。市販ソフト E ではこの誤りを考慮すると性能が向上すると思われる。

市販ソフト F では「現在形」「過去形」「過去進行形」を「現在完了形」に誤ることが比較的多い。例えば、以下の例である。

入力日本語文：それは私の好みに合っている。

翻訳結果：It has matched my taste.

模範訳：It suits my taste.

「現在形」で十分である。システムの調整を失敗して「現在完了形」を出力しやすくしていると思われる。また、市販ソフト F の翻訳結果を見るとわざと文を短く区切っているように見受けられる部分があった。以下に例を示す。

入力日本語文：彼は実際より若く見える。

翻訳結果：It is younger than practice and he can be seen.

模範訳：He looks younger than he really is.

なぜか二文に分割され、また「現在形」とすべきところが「can」になっていた。市販ソフト F でもこれらの誤りを考慮すると性能が向上すると思われる。

4. おわりに

本稿では、テンス・アスペクト・モダリティの日英翻訳における機械翻訳システムの誤りを調査した。まずわれわれの SVM を利用するシステムについて考察し、文末の文字列素性の重要性を確認した。次に市販の6つの翻訳システムについて考察し、市販のシステムはわれわれの SVM を利用するシステムに比べて特に「現在形」とすべきところを「現在進行形」に誤ることが多いことがわかった。われわれの SVM を利用するシステムは他のシステムよりも性能が高い上に、この種類の誤りが少なかったため、今後既存の翻訳システムにわれわれの SVM を利用するシステムの方法を組み込むことで既存の翻訳システムにおけるこの問題を改善し性能を向上させることができるとと思われる。また、市販の6つの翻訳システム特有の誤りについても調べて報告した。これらのシステム特有の誤りは他のシステムでは正しく処理できるため、個々のシステムでもこの誤りに着目することで比較的容易に性能を改善できる点と思われる。本稿で示した個々のシステム特有の誤りは、比較的容易に改善できる点として、各翻訳システムの性能向上の手がかりとして役に立つと思われる。

参考文献

- [1] 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均, 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, 言語処理学会第11回年次大会, (2005).
- [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 of Asian Language Information Processing*, (2005, to appear).
- [3] 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).
- [4] Nello Cristianini and John Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, (Cambridge University Press, 2000).
- [5] Taku Kudoh, TinySVM: Support Vector Machines, (<http://cl.aist-nara.ac.jp/taku-ku/software/TinySVM/index.html>, 2000).
- [6] 上田太郎, 刈田正雄, 本田和恵, 実践ワークショップ Excel 徹底活用多変量解析, (秀和システム, 2003).