D-6

深層学習モデルによる日本語動詞の双方向形態屈折の検証

深津 聡世，原田 宥都，関澤 瞭，田村 鴻希，大関洋平
東京大学総合文化研究科

要旨

人間の形態処理に必要な知識は規則と類推のどちらなのか、あるいはその両方なのか、といった議論は形態論の分野で今も続いており、過去時制論争（Past Tense Debate）と呼ばれる。特に日本語は、動詞の屈折が形態的に複雑である上に、そもそも過去形と現在形のどちらが基底形なのかも自明ではない。そのため日本語は、過去時制論争において重要な言語でありつつも、その形態処理のメカニズムは未だ明らかでない。そこで本研究では、日本語の動詞の屈折現象を対象にし、類推のモデルである深層学習モデルを用いて双方向形態屈折の学習を行い、モデルがどちらの時制方向により適しているのかを検証した。実験の結果、8 つのうち 6 つの実験設定において、現在形から過去形への方向でモデルの正解率が高くなった。その差は平均で 3.01%であり、時制方向によって大きな差があることが分かった。これは、現在形から過去形への方向では類推的な処理が適することを示唆している。

1 導入

形態論では、人間がもつ形態処理の知識には規則と類推のどちらが必要か今も議論が続いており、この議論は過去時制論争（Past Tense Debate）と呼ばれている。日本語を対象とした関連の研究では、日本語母語話者が実在語に見られる屈折パターンに則して非実在語を屈折させることができないことが広く観察されており、これが類推による形態処理を支持する証左として言及されている (Klafehn, 2003, 2013)。しかしながら、Oseki et al. (2019) は、獲得研究での観察をもとに、日本語においては現在形から過去形の方向ではなく過去形から現在形の方向に規則が派生しているという仮説を提案した。そして、双方向 wug テスト（現在→過去と過去→現在）と規則ベースのモデルを用いた検証を行ったところ、様々な評価尺度から過去形から現在形の方向で一般化による演算処理が行われていることが示唆されている。

この研究では、方向によって異なる性質の処理（過去→現在では規則による演算処理、現在→過去では類推による処理）が行われていることが示唆されたものの、類推ベースのモデルによる検証はまだ行われていない。そこで本研究では、類推に基づくモデルである深層学習モデルを用いて日本語動詞を対象に双方向に形態屈折のモデリングを行い、結果のエラー分析を行った。

2 先行研究

2.1 過去時制論争

過去時制論争 (Pinker & Ullman, 2002) は、 Rumelhart and McClelland (1986) によるコネクショニストモデル（以下、RM モデル）の提案を発端に始まった。彼らは、英語を獲得する子どもの言語獲得の過程でみられるような U 字型発達曲線と類似した学習曲線を RM モデルが表現できたことから、人間の言語知識には明示的な規則は必要ないと主張した。それに対し、Pinker and Prince (1988) は、RM モデルには、単語間の音韻的な類似性を捉えられない、モデリングの際に人手でインプット内の不規則形動詞の割合を変化させている、「過剰不規則化」のような人間に見られない産出をするなど、様々な問題点があると指摘した。

このような指摘がなされたのち、長いあいだコネクショニストから新しい提案はされていなかったが、近年のニューラルネットワークが発展したことに伴い過去時制論争は再燃している。Kirov and Cotterell (2018) は、機械翻訳のために開発された類推ベースのモデルである Encoder-Decoder モデル (以下、ED モデル）を英語の形態処理に応用し、最新のニューラルネットワークモデルが Pinker and Prince (1988) によって指摘された RM モデルの欠点を克服できたと主張した。

しかし、後続の研究では、Kirov and Cotterell (2018) のモデルの精度評価方法の問題点 (Corkery et al., 2019) や、モデルのパフォーマンスが頻度に影響を受ける点 (McCurdy et al., 2020) が指摘されている。このように、形態処理における明示的な規則の必要性については、規則派と類推派のあいだで今も議論が続いている。

2.2 日本語動詞における過去時制論争

一方で、日本語動詞の過去時制論争研究においては、日本語母語話者が実在語の音韻的なパターンに則って動詞を屈折することができないことから、彼らは規則を用いて形態処理をしていないといわれている (Klafehn, 2003, 2013; Vance, 1991)。これらの研究に共通しているのは、母語話者に、現在形の動詞をもとに過去形を答えさせる点であり、これは、英語を対象にした研究 (Berko, 1958) で用いられている、現在形から過去形へ非実在語を屈折させる実験パラダイムを踏襲したもとだと考えられる。しかしながら、Oseki et al. (2019) は、英語で仮定されている屈折の方向性を十分な検討をしないまま日本語に適用するべきではないと指摘し、その根拠として言語獲得研究において報告されている観察を挙げている。

1 つ目に、英語を母語として獲得する子どもの発話では、現在形が過去形に先行して出現する一方、日本語においては、過去形が現在形に先行する。2 つ目に、時制の形態素の獲得が不完全な英語母語話者の子どもは 2 歳ごろに現在形と語形が同じである原形不定詞を過剰に使用するのに対し、日本語母語話者の子どもは、過去形を原型不定詞の代替形として用いる。 3つ目に、英語のおいて、過剰規則化は現在形から過去形の方向におきるものの、日本語では過去形から現在形の方向に起きる。そして、最後に、日本語母語話者の子どもは、現在形よりも過去形を好む帰納バイアスを持っている。

これらの観察を踏まえて、日本語動詞における屈折の基底形は過去形であるという仮説を立てた Oseki et al. (2019) は、39 名の大人の日本人母語話者を対象とした双方向の wug テスト（現在→過去と過去→現在）と規則ベースのモデルである Minimal Generalization Learner (MGL; Albright & Hayes, 2002) を用いたモデリングを行った。結果、現在形から過去形への方向よりも過去形から現在形の方向へ屈折される方が、参加者が実在語の屈折パターンから予想される語形を産出する確率が高く、MGL の人間データへの適合度も高く、またモデルの複雑さも低いことが示された。これらの結果から、Oseki et al. (2019) は、過去→現在の方向では一般化によって動詞が規則的に処理されており、現在形→過去形の方向では類推的な処理が行われていることを提案している。

2.3 問題提起と課題設定

Oseki et al. (2019) の研究では、過去→現在の方向で規則的な処理が起きており、現在→過去の方向では類推的な処理が行われていることが示唆されたものの、類推に基づく言語モデルによる検証はまだ行われていない。そのため、ED モデルなどの深層学習モデルを用いて双方向に形態屈折のモデリングを行うことで、方向性によって形態処理のメカニズムが異なる可能性をさらに検討することができると考えられる。

そこで、本研究では、深層学習モデルを用いて双方向に形態屈折のモデリングを行い、得られた結果を人間データ (Oseki et al., 2019) と比較することで、類推ベースのモデルが現在→過去と過去→現在のどちらの方向で働いているのかを検討する。もし、Oseki et al. (2019) の提案するように方向性によっ

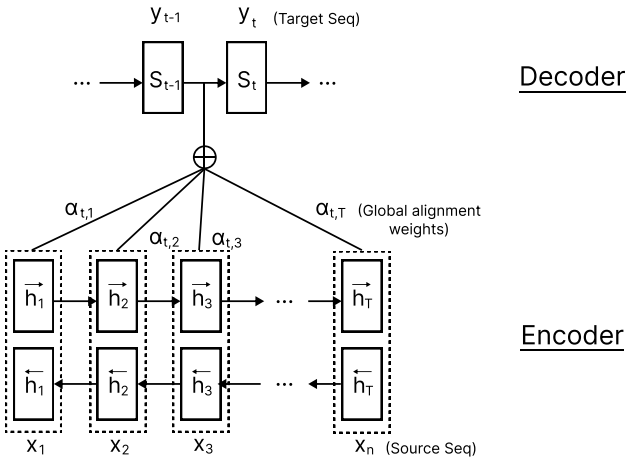

図 1: Attention 付き RNN モデルの概要

て処理の性質に違いがあるのであれば、類推ベースのモデルを用いて検証した場合、過去→現在に比べ現在→過去の方向においてより正答率が高くなることが予想される。

3 実験

3.1 Attention 付き RNN モデル

形態屈折を学習するモデルは、規則ベース、類推ベースの 2 種類に大別される。従来のアプローチである規則ベースのモデルは、有限状態オートマトンや辞書法などの技術によって実現されているが、作成コストの高さ、システムの壊れやすさなど多くの課題を残していた。これを解決するために、近年では機械学習を用いた類推ベースの学習モデルが提案された (Ahlberg et al., 2015; Hulden, 2014; Nicolai et al., 2015)。しかし、これらもコストの高い特徴量エンジニアリングを必要とする手法であり、その後の研究ではさらに作成が容易な深層学習モデルでの変換が提案された (Faruqui et al., 2016)。本研究では、深層学習モデルを英語における過去時制論争に用いた Kirov and Cotterell (2018) と同様のモデルである Attention 付き RNN モデルを実装し、実験を行った。図 1 にモデルの概要図を示す。

長さ n の入力シーケンス $\boldsymbol{x}$ と、長さ m のターゲットシーケンス $\boldsymbol{y}$ を以下のように定義する。

$$\boldsymbol{x} = [x_1, x_2, ..., x_n] \tag{1}$$

$$\boldsymbol{y} = [y_1, y_2, ..., y_m] \tag{2}$$

形態屈折の学習では、ラテンアルファベットや IPA などで表記された、1 つの入力単語が入力シーケンスに相当する。ターゲットシーケンスはその単語の屈折先に相当する。エンコーダには双方向 LSTM を用いており、前向きの隠れ状態 $\overrightarrow{\boldsymbol{h}_i}$ と後ろ向きの隠れ状態 $\overleftarrow{\boldsymbol{h}_i}$ を結合することでエンコーダ部の表現 $\boldsymbol{h}_i$ を得る。デコーダネットワークは、位置 t の出力語に対して隠れ状態 $\boldsymbol{s}_t$ を持つ。

$$\boldsymbol{s}_t = f(\boldsymbol{s}_{t-1}, y_{t-1}, \boldsymbol{c}_t) \tag{3}$$

ここで $t = 1, 2, ..., m$ であり、$\boldsymbol{c}_t$ はエンコーダの出力の重み付け和である。

$$\boldsymbol{c}_t = \sum_{i=1}^{n} \alpha_{t,i} \boldsymbol{h}_i \tag{4}$$

$$\alpha_{t,i} = \frac{\exp\left(\text{score}(\boldsymbol{s}_{t-1}, \boldsymbol{h}_i)\right)}{\sum_{i'-1}^{n} \exp\left(\text{score}(\boldsymbol{s}_{t-1}, \boldsymbol{h}'_i)\right)} \tag{5}$$

アラインメントスコア α は隠れ層が一つのフィードフォワードネットワークによってパラメータ化され、活性化関数には tanh を用いている。そのため、スコア関数は以下のようになる。

$$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{v}_a \tanh \boldsymbol{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i] \tag{6}$$

$\boldsymbol{v}_a$ と $\boldsymbol{W}_a$ がフィードフォワードネットワーク内で学習される重み行列である。

3.2 データセット

本プロジェクトでは 2 種類のデータセットをもとに動詞の現在形、過去形のデータを収集し、実験に使用した。以下に実験に使用したデータセットについて述べる。

京都大学テキストコーパス 京都大学テキストコーパスは、1995 年に出版された毎日新聞の記事と社説、それぞれ約 2 万文に対して、形態素・構文情報を人手で付与したテキストコーパスである (Kurohashi & Nagao, 2003)。このデータセットに含まれる単語のうち、動詞の過去形のみを取り出した。MeCab(Kudo et al., 2004) を用いてそれらの動詞の原形を取り出し、間違いのあったものについては人手で修正を加えた。これにより、日本語動詞の現在形と過去形のペア 1300 個が得られた。

IPA 辞書 IPA 辞書[1]、あるいは ipadic は、IPA コーパスに基づいて作成された日本語の形態素辞書である。この中から動詞の現在形と過去形のペアを取り出し、5300 個のデータが得られた。

日本語動詞の表記 作成したデータセットは PyKakasi[2]を用いて全てアルファベット（ラテンアルファベット）に変換し、この表記のデータセットのバージョンを”latin”とした。また、phonemizer(espeak)[3]を用いて国際音声記号（International Phonetic Alphabet; IPA）に変換し、こちらの表記のデータセットのバージョンを”IPA”とした。本実験では”latin”と”IPA”の異なる 2 種類の表記を用いてモデルの学習を行う。

3.3 実験設定

モデルのハイパーパラメータについては、Kirov and Cotterell (2018) を参照し、設定を合わせている。学習時の最適化アルゴリズムには AdaDelta を用い、単語の埋め込み次元数は 300、LSTM のユニットサイズは 100 に設定した。バッチサイズは 20 で実験を行なった。訓練データとテストデータを 8:2 の割合で分割している。

[1]https://taku910.github.io/mecab/
[2]https://codeberg.org/miurahr/pykakasi
[3]https://github.com/bootphon/phonemizer?ref=morioh.com&utm_source=morioh.com

訓練データ	単語表記	テストデータ	正解率 (%) past → present	正解率 (%) present → past
京大コーパス	IPA	京大コーパス (n=260)	87.69	**89.62**
京大コーパス	IPA	IPA 辞書 (n=4000)	85.68	**89.88**
京大コーパス	latin	京大コーパス (n=260)	75.00	**77.69**
京大コーパス	latin	IPA 辞書 (n=4000)	69.75	**74.38**
IPA 辞書	IPA	IPA 辞書 (n=800)	**96.50**	95.38
IPA 辞書	IPA	京大コーパス (n=1300)	88.54	**88.77**
IPA 辞書	latin	IPA 辞書 (n=800)	**97.00**	93.63
IPA 辞書	latin	京大コーパス (n=1300)	88.46	**92.85**

表 1: それぞれの実験設定ごとの実験結果。表中の n はテストに用いたデータ数を示す。太字表記はそれぞれの実験設定において精度の高かった屈折の時制方向を示している。

4 結果と考察

3.2 節で述べたように、本実験では 2 種類の訓練データ、2 種類の単語表記、2 種類のテストデータを用いて、計 8 通りの実験設定を試している。そのそれぞれにおいて過去→現在、現在→過去の 2 方向での形態屈折の学習を行った。その結果を表 1 に示している。

実験の結果、8 つのうち 6 つの実験設定において、現在形から過去形への方向（present → past）での正解率が高くなった。6 つの設定における正解率の改善幅の平均は 3.01%であり、類推ベースのモデルである Attention 付き RNN が、現在形から過去形への方向の学習により適していることが示唆された。これは、規則ベースのモデルで同様の実験を行った Oseki et al. (2019) とは時制方向について逆の結果である。

訓練データのサイズによる影響　訓練データに用いた 2 種類のデータセットを比較すると、京都大学コーパスからは 1300 個の動詞、IPA 辞書からは 5300 個の動詞が得られており、この 2 種類のデータセットで重複している動詞は 1169 個あったため、訓練の際もテストの際も、重複した単語は IPA 辞書から取り除いている。そのため、IPA 辞書のデータセットサイズは京都大学コーパスの約 3 倍である。訓練データによる結果の影響を確認すると、 IPA 辞書で学習したモデルの方が全体的に正解率が高いことがわかる。訓練データのサイズはこれ以降に述べるどの要素よりも精度への影響が大きく、データセットのサイズが重要な要因であることが伺える。IPA 辞書は多くの動詞を収録しているため、テキスト等での出現頻度が低いような珍しい動詞も含むが、これが正解率を下げる要因にはならなかった。

単語表記の種類による影響　実験では”IPA”と”latin”の 2 種類の単語表記を用いており、同じ単語に対しての情報の量では IPA 表記が優れている。そのため、IPA 表記を用いて学習したモデルの精度が高くなるという仮説を立てていた。サイズが小さい京都大学コーパスを訓練に用いたモデルでは、実際にそのような傾向が見られ、latin 表記のモデルは大きく精度を落としている。しかし、サイズの大きい IPA 辞書を訓練に用いたモデルでは、IPA 表記と latin 表記のどちらを用いても精度に大きな差が見られなかった。訓練に用いることができるデータサイズが限られている場合は IPA 表記の持つ豊富な情報は有用であるが、データサイズが十分に確保できる場合は latin 表記による学習でも十分である可能性がある。

テストデータのサイズによる影響　今回はそれぞれの実験で 2 通りのテストデータを試しており、表に示したようにテストデータごとにサイズが異なる。京都大学コーパスの IPA 表記で学習したモデルは、サイズの大きな IPA 辞書でテストした場合も精度が大きく落ちることがなく、高い汎化性能を獲得して

入力	正解	予測
tokihanatta	tokihanatsu	tokihanaru
tonda	tobu	tomu
kisokudatta	kisokudatsu	kisokudaru
zannengatta	zannengaru	zanengaru
futekusatta	futekusaru	futtekusaru

表 2: モデルの誤答例。IPA 辞書を latin 表記で訓練し、IPA 辞書でテストした、過去形→現在形の方向の屈折を学習したモデルについて調査した。

いると考えられる。反対に、IPA 辞書で学習したモデルは京都大学コーパスでテストした際に大きく精度落とす傾向にあるが、これは京都大学コーパスがより基礎的な動詞を含んでおり、IPA 辞書からはそれらが重複分として全て削除されていることが影響していると考えられる。

屈折の時制方向による影響　2.3 節で述べた通り、屈折の時制方向がモデルの振る舞いに影響するという仮説を立てていた。実験の結果、8 つのうち 6 つの実験設定で、過去形から現在形への方向で正解率がより高くなった。IPA 辞書で学習したモデルを IPA 辞書でテストした場合のみ、過去形→現在形の時制方向で正解率が高くなった。京都大学コーパスはより基礎的な動詞を含み、IPA 辞書からはそれらが重複分として削除されているため、訓練データの性質の差が何らかの影響を及ぼした可能性が考えられるが、更なる調査が必要である。

エラー分析　最も精度の高い結果を示したモデルに対して、簡易なエラー分析を行った。モデルは IPA 辞書を latin 表記で訓練し、IPA 辞書でテストした、過去形→現在形の方向の屈折を学習したものである。モデルが間違った答えを出したケースについて、いくつかの例を表 2 に示した。”tonda”を”tomu”とするなど、人間の産出でも見られる過剰規則化や、”zannengatta”を”zanengaru”としてしまうような、活用は正しいものの語幹を誤って変形してしまう例が見られた。

5　結論

本研究では、日本語の動詞の形態屈折を類推のモデルである深層学習モデルによって学習し、時制方向による正解率の差について検証した。実験の結果は、現在形から過去形への方向では類推的な処理が適していることを示唆するものであった。今後は、非実在語である「wug」語によるテストや、他の深層学習モデルや規則ベースのモデルによる検証など、更なる詳細な調査を行う。

参考文献

Ahlberg, M., Forsberg, M., & Hulden, M. (2015). Paradigm classification in supervised learning of morphology. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1024–1029. https://doi.org/10.3115/v1/N15-1107

Albright, A., & Hayes, B. (2002). Modeling English past tense intuitions with minimal generalization. *Morphological and Phonological Learning: Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, 58–69.

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2-3), 150–177.

Corkery, M., Matusevych, Y., & Goldwater, S. (2019). Are we there yet? encoder-decoder neural networks as cognitive models of english past tense inflection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3868–3877.

Faruqui, M., Tsvetkov, Y., Neubig, G., & Dyer, C. (2016). Morphological inflection generation using character sequence to sequence learning. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 634–643. https://doi.org/10.18653/v1/N16-1077

Hulden, M. (2014). Generalizing inflection tables into paradigms with finite state operations. *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM*, 29–36. https://doi.org/10.3115/v1/W14-2804

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 651–665.

Klafehn, T. (2003). *Emergent properties of japanese verbal inflection* (Doctoral dissertation). University of Howai'i.

Klafehn, T. (2013). Myth of the wug test: Japanese speakers can't pass it and English speaking children can't pass it either. *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*, 170–184.

Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 230–237. https://aclanthology.org/W04-3230

Kurohashi, S., & Nagao, M. (2003). Building a japanese parsed corpus, treebanks:building and using parsed corpora.

McCurdy, K., Goldwater, S., & Lopez, A. (2020). Inflecting when there's no majority: Limitations of Encoder-Decoder neural networks as cognitive models for German plurals. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1745–1756.

Nicolai, G., Cherry, C., & Kondrak, G. (2015). Inflection generation as discriminative string transduction. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 922–931. https://doi.org/10.3115/v1/N15-1093

Oseki, Y., Sudo, Y., Sakai, H., & Marantz, A. (2019). Inverting and modeling morphological inflection. *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 170–177.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73–193. https://doi.org/10.1016/0010-0277(88)90032-7

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Science*, *6*(11), 456–463. https://doi.org/10.1016/s1364-6613(02)01990-3

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In D. E. Rumelhart, J. L. McClelland, & C. PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2: Psychological and biological models* (pp. 216–271). MIT Press.

Vance, T. J. (1991). A new experimental study of Japanese verb morphology. *Journal of Japanese Linguistics*, *13*, 145–166.