ニューラル日英翻訳における出力文の態制御

山岸 駿秀　叶内 晨　佐藤 貴之　小町 守
首都大学東京

{yamagishi-hayahide, kanouchi-shin, sato-takayuki}@ed.tmu.ac.jp,
komachi@tmu.ac.jp

1 はじめに

翻訳では，より自然な文を生成するために，言語ごとの表現方法の違いを考慮する必要がある．例えば，日英翻訳では日本語側の文と英語側の文で態が異なることがある．これは，文構造の違いや，語彙ごとの用法によって生じる．また，文書の翻訳を行う場合には，文書としての一貫性を保つために文の流れに適合した態に変えることもある．

表 1 に，本研究で用いた日英対訳コーパスである ASPEC コーパス [5] での，主な高頻度動詞における各態ごとの出現回数を示す．これは，科学技術論文の概要の対訳データを論文ごとに集めて文アライメントをとることで作成されたコーパスである．この表から，show は能動態の文で，find は受動態の文で使われやすいことがわかる．しかし，describe，develop などは使われる態の傾向が見られない．したがって，語彙によって，態の出現分布や，文書の情報構造上の理由での態の変わりやすさなどには傾向があるため，これらをルール化することは難しい．

近年の機械翻訳では，RNN を用いた手法であるエンコーダデコーダモデル [7] が，従来の統計的機械翻訳と比べて簡潔なモデルながらも高い精度の翻訳を行えることから注目されている．エンコーダデコーダモデルは出力を制御することが容易ではないが，出力を制御する試みはいくつか提案されている．

Kikuchi らは，エンコーダデコーダモデルを用いた文要約において，出力文長の制御を行った [2]．この研究では，モデルの機構を出力文長を考慮できるように変更することで，それまでに報告されていた精度を保ちながら制御を行えることを示した．

Sennrich らは，英独機械翻訳において出力文の敬意表現の制御を行った [6]．入力の英文にドイツ語側の敬意表現の有無を単語として組み込んだ文をコーパスとして用いて学習させた．評価時にも同様に，任意の敬意情報を単語として入力文に付与することで，付与した情報に一致した表現の文を出力でき，参照訳の敬意に揃えた場合は BLEU が 3.2 ポイント向上した．

表 1: 高頻度動詞における態ごとの出現回数

動詞	能動態	受動態		合計
show	27,106	11,082	(29.0%)	38,188
describe	16,338	18,043	(52.5%)	34,381
investigate	5,515	18,588	(77.1%)	24,103
develop	9,638	12,368	(56.2%)	22,006
find	3,401	16,358	(82.8%)	19,759
全動詞	604,158	499,178	(45.2%)	1,103,336

本論文では後者の手法をもとに，日英翻訳での出力文の態の制御に取り組んだ．英文側の態の情報を日本語文に単語として付与した文を新たなコーパスとして用いてモデルの学習を行った．評価時にも任意の態情報を日本語文へ付与することで，モデルが付与した情報と同じ態の文を生成しているかを調べた．その結果，制御率は能動態にする場合は 73.5%，受動態にする場合は 94.5%となり，参照訳と態を揃えることができれば BLEU が 1.87 ポイント向上することがわかった．

2 態制御のためのデータ作成手法

2.1 学習データのラベリング

日本語文に対する態情報付与の流れを，図 1 に示す．はじめに，英語側の文を依存構造解析する．得られた結果から，ROOT となった動詞が過去分詞形であるか，ROOT の子に be 動詞があるかを確認した．両条件が満たされたとき受動態であると判定し，満たされないときは能動態とした．この条件では become，get などを用いた受動態については考慮できないため，今回は be 動詞を用いない受動態の文も能動態の文としている[1]．文が重文や複文である場合は，構造解析器が ROOT とした動詞を判定対象とした．英語側の文が能動態であれば<Active>，受動態であれば<Passive>をラベルとし

[1] したがって，ある文の態が be 動詞を用いた受動態であるか否かを判定しているのであって，文の態を分類しているのではない．

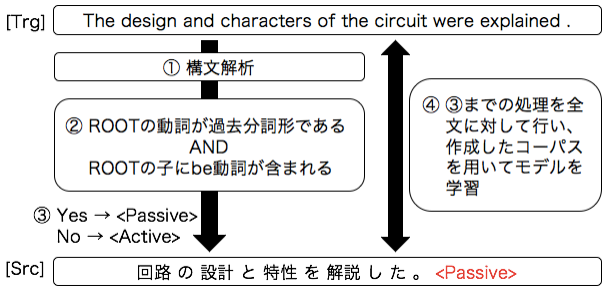

図 1: 学習データ作成のための自動ラベル付与の流れ

て取り出す．このラベルは，日本語側の文末に単語として付与する．このようにして，日本語の文に，対応する英文の態の情報が付与されたデータを作成し，そのデータを使ってモデルを学習した．

テスト時には学習時と同様に，英語側で生成したい態を表すラベルを末尾に付与した文を入力する．3.1 節ではラベルの貼り方を変えて実験を行い，付与したラベルの表す態の文が生成されたかどうかを調べた．

2.2 態予測

本論文で行う実験はすべての入力文に対してラベルを付与することを前提としている．しかし，どちらでもよい場合や，英語側の流暢性を保ちつつその文にあった態を決定することが使用者にとって困難である場合などが存在する[2]．このような場合は，入力文や過去の出力結果などをもとに英語側でのふさわしい態を選び，それをラベルとして付与する．本節では，態ラベルの予測手法を検討した．

2.1 節で得られた態ラベルを正解ラベルとし，各文ごとに得られた素性をもとにして正解ラベルを得ることを目標とした．素性として以下の 6 つを用いた．

SrcSubj 入力文の主語である文節の分散表現

SrcPred 入力文の述語である文節の分散表現

SrcPrevPred 1 文前の入力文の述語であった文節の分散表現

TrgPrevObj 1 文前の出力文の目的語の分散表現

PrevVoice 3 文前までの出力文の態情報

VoiceDist 入力文の述語ごとにまとめた，学習データ内の態分布での多数派の態情報

SrcSubj，SrcPred，SrcPrevPred は，日本語側の情報を使い，TrgPrevObj，PrevVoice は英語側の正解の情報を使った．文節の分散表現は，英数字と句読点などの記号を除いた文節中の全ての単語分散表現の平均を用いた．SrcPrevPred，TrgPrevObj，PrevVoice は，文書での情報の流れを用いることが目的である．VoiceDist は，学習データでの入力文の述語と出力文の態の対応関係を用いるものである．テスト文の述語が学習データ内で用いられている場合，出力の態として多く使われていた方の態を候補として得る．PrevVoice と VoiceDist は能動態，受動態，態情報なしの 3 値を与えた．素性は 1 つのベクトルとして結合し，ロジスティック回帰で学習を行った．予測の結果は 3.2 節で述べる．

3 実験

3.1 実験設定

本研究では，提案手法を態の制御率と翻訳結果の BLEU 値，Pairwise による人手評価の結果の 3 つで評価した．テスト文へのラベル付与方法を変えることで以下の 4 つの実験を行った．

ALL_ACTIVE 全文に能動態のラベルを付与する．

ALL_PASSIVE 全文に受動態のラベルを付与する．

REFERENCE 各文に参照訳と同じ態のラベルを付与する．

PREDICT 各文に 2.2 節で予測したラベルを付与する．

英語側の依存構造解析には Stanford Parser (Ver. 3.5.2) を用いた．得られた態ラベルを人手で評価したところ，95%の文には正しくラベルが付与されていた．PREDICT の実験では，日本語側の ROOT の文節を得るために CaboCha [3] (Ver. 0.68) [3]，MeCab [4] (Ver. 0.996, 辞書: IPADIC Ver. 2.7.0) を用いた．

ASPEC コーパス [5] の学習データは 3,008,500 文ある．これを論文ごとに分類し，文対の少なくとも一方が 50 単語以上の文を削除したのち，欠けた文がなく，かつ 2 文以上ある論文を集めた．この結果，学習データの文数は 1,103,336 文となった．評価には，ASPEC コーパスのテストデータ 1,812 文対から，参照訳の態が各 100 文対ずつになるように作成した計 200 文対の対訳データを用いた．参照訳の ROOT が自動詞である文対は受動態にならない可能性が高いので，作成したデータには加えていない．制御の評価は出力文の態のみを 1 人の評価者が調べた．翻訳性能は BLEU と人手評価で比較した．人手評価は，ベースラインと REFERENCE の結果を Pairwise によって評価した．これは態の評価者とは異なる 1 人の評価者が行った．

ベースラインは，エンコーダデコーダモデルにアテンション機構を付加したモデル [1][4] を，態ラベルのないデータで 15 epoch 学習させたものである．両言語の単語ベクトルの初期値として，ASPEC コーパス

[2] Sennrich らの実験では，ラベルを付与しないデータを混ぜることで制御しなくともよい場合について対応している．[6]

[3] https://taku910.github.io/cabocha/

[4] http://taku910.github.io/mecab/

表 2: 態予測の Ablation Test の結果

素性	使用した素性 (空欄である素性は使用していない)							
SrcSubj	○		○	○	○	○		
SrcPred	○	○		○	○	○	○	○
SrcPrevPred	○	○	○		○	○	○	
TrgPrevObj	○	○	○	○		○	○	
PrevVoice	○	○	○	○	○		○	○
VoiceDist	○	○	○	○	○	○		○
正解率 (%)	66.9	67.3	65.2	67.2	67.3	65.9	65.4	**67.7**

表 3: 各実験での出力の態分布と BLEU

実験	能動態	受動態	その他	正解率	制御率	BLEU
参照訳の態分布	100	100	—	—	—	—
ベースライン	31	163	6	60.5%	—	20.60
ALL_ACTIVE	147	44	9	57.5%	73.5%	20.22
ALL_PASSIVE	6	189	5	51.0%	94.5%	20.18
REFERENCE	82	113	5	89.0%	89.0%	**22.47**
PREDICT	74	118	8	64.0%	89.0%	21.05

全文を用いて Word2Vec [5] を学習させて得た分散表現を用いた．語彙数は 30,000，単語の埋め込み層と隠れ層の次元数は 512，バッチサイズは 64 である．学習時は初期学習率が 0.01 の Adagrad で最適化を行った．実装には，Chainer [6] (Ver. 1.18) [8] を使用した．2.2 節でのロジスティック回帰は，Python のライブラリである scikit-learn のものを用いた．分散表現を用いた素性は，コーパス整形後の 1,103,336 文で Word2Vec を学習させた次元数 100 の分散表現を使用した．

3.2 態予測の精度

態ラベル予測と Ablation Test の結果を表 2 に示す．ここでは前文の情報を用いる素性もあるため，1,812 文のテスト文に対してラベル推定を行い，得られたラベルと正解ラベルの正解率を調べた．SrcPred，PrevVoice，VoiceDist を用いたものが最も正解率が高く，67.7%であった．最も重要度の高い素性は SrcPred であった．VoiceDist は多数決の情報であるから，コーパス内に高頻度で現れ，かつ態分布の偏りが小さい述語に対しては悪影響を与える可能性がある．PrevVoice は文書の 1 文目には情報を与えないため，重要度が比較的低い．その他の素性は使うことで性能が下がった．SrcSubj，TrgPrevObj は，学習データで日本語側に主語のない文や英語側に目的語のない文が多く，うまく機能しなかった．SrcPrevPred は，人手の翻訳時は英語側の情報構造に沿って態が決定されるため，日本語側の過去の情報は必要なかったと考える．これによって得られた態ラベルを 200 文のテスト文に付与し，実験を行った．

4 実験の結果と考察

4.1 各スコアによる評価

得られた結果を表 3 に示す．表中の「その他」は，出力が非文であった場合や，動詞がなく態を判断できない場合などの文数を表す．表中の「正解率」は参照訳の態と出力文の態が同じであった文の割合を表し，「制御率」は入力文に付与したラベルの表す態と出力文の態が同じであった文の割合を表す．

[5]https://radimrehurek.com/gensim/models/word2vec.html
[6]http://chainer.org/

ALL_PASSIVE の制御率は ALL_ACTIVE に比べて 21%も高いため，能動文の生成が比較的難しいと言える．これは，本来受動態である方が自然な文を意図的に能動態にする命令をかけた場合，主語としてふさわしいものを選択できない傾向があることが主な原因である．ALL_ACTIVE の主語としては，we，this paper が頻出し，それぞれ 200 文中 41 文，33 文で用いられている．しかし，これらの主語を生成できたのは，高頻度動詞が用いられている場合が多かった．例えば，this paper を主語として生成した 33 文中，17 文では describe を用いている．describe は高頻度動詞であるため，コーパスから，能動文では主語に this paper を用いるという用例をうまく学習できたと言える．同様に，show は result を，report は we を主語として用いる傾向があった．一方で能動文を生成できない動詞は，コーパス内に能動文で用いられた例が比較的少なかった．したがって，能動態の生成には，何を主語として生成すべきかの用例が一定数必要であると言える．

ALL_ACTIVE と ALL_PASSIVE では，参照訳と異なる態の文の出力を試みたものが各 100 文存在するため，正解率が低い．これによって，BLEU はベースラインに比べて低くなった．BLEU が最も高かったものは，REFERENCE であった．この実験では BLEU を 1.87 ポイント改善できたので，参照訳と同じ態になるように出力することで翻訳性能の向上も認められる．PREDICT の実験では BLEU が 0.45 ポイント上がった．今回は予測器の精度が低かったためこのような結果となったが，REFERENCE が上限であるとすれば，予測器の精度を高めることで参照訳の情報を用いずに最大で 1.87 ポイントまで改善が可能であると言える．

人手評価の結果，200 文中 120 文が REFERENCE の方がよいとされた．参照訳の態に揃えることで，人手評価についても良い結果を得られた．英語側の出力として自然な態を選ぶことができれば，自然な文が生成されやすくなると言える．

4.2 出力結果の考察

表 4 に出力例を載せる．表中に赤字で示したものは能動態，青字で示したものは受動態を表す．例 1 は受

表 4: 出力例 (実際の入力文の末尾には，目的に合わせて<Active>または<Passive>のラベルを付与)

例 1	入力文	リサイクルに関する最近の話題を紹介した．
	参照訳	recent topics on recycling are introduced .
	能動態へ制御	this paper introduces recent topics on the recycling .
	受動態へ制御	recent topics on the recycling are presented .
例 2	入力文	Si ドットの密度及び直径分布は生成温度により変化させることができた．
	参照訳	the density and diameter distribution of Si dots could be changed through the production temperature .
	能動態へ制御	density and diameter distribution of the Si dots can be changed by the production temperature .
	受動態へ制御	density and diameter distribution of the Si dots can be changed by the production temperature .
例 3	入力文	自己組織化構造に分子の形と分子間相互作用が大きく影響する．
	参照訳	molecular shape and intermolecular interaction influence self-assembled structures greatly .
	能動態へ制御	the molecular structure and intermolecular interaction greatly affect the self-organization structure .
	受動態へ制御	the molecular structure and the intermolecular interaction are greatly influenced in the self-organization structure .
例 4	入力文	また，ドットの形状及び結晶性は温度に依存することも分かった．
	参照訳	it was also proven that the shape and crystallinity of the dots were dependent on temperatures .
	能動態へ制御	the morphology and the crystallinity of the dots depended on the temperature .
	受動態へ制御	it was also found that the shape and the crystallinity of the dots depend on the temperatures .

動態である文について制御を行い，成功した例である．introduce は，コーパス内では 6,217 文が能動態，8,789 文が受動態の形で用いられている．主語の用例が多いため，生成に成功したと言える．この例では，主語と目的語の位置も適切な場所に配置されている．

例 2 は，命令によらずに受動文を出力した例である．出現回数の低い動詞を用いる場合は失敗が目立った．今回の実験では，受動態にならずに能動態で出力されたという文が少なかったため，そのような場合の傾向はわからなかった．しかし，出力が能動文になった 6 文のうち 3 文は be 動詞+形容詞の形を出力しているため，こちらの誤りにも一定の規則があると想定される．より多くの出力を見ることで，こちらの傾向もつかめるはずである．

例 3 は，制御に成功しているが，主語と目的語の入れ替えが起こらず，結果的に意味が異なっている例である．本論文では，この入れ替えの評価を行っていない．しかし，本研究は談話構造の一貫性を意識した文書単位の翻訳を視野に入れたものであるため，入れ替えの有無についても検討したい．例 4 は，制御に成功しているが一部の情報が消えてしまった例である．この例で使われた prove や find などは，ALL_PASSIVE の実験で仮主語 it を伴って頻出していた．この場合は，主語と目的語を入れ替える代わりに，仮主語を用いた文にするなどによって対応したと考えられる．

5 おわりに

本論文では，エンコーダデコーダモデルにおける出力文の態を制御する取り組みについて報告した．前処理として，英語側の構文解析を行って ROOT の動詞の態を判別し，得られた態情報を日本語側の文の末尾に単語として付与したデータを作成した．ラベルを付与する位置や，複文や重文のときの対応，文単位以外のラベル付与などについては今後検討したい．

作成したデータを学習に用いて，テスト時に出力文の態制御を行った．受動態にする場合の制御率は 94.5%であり，能動態にする場合に比べて 21%高かった．参照訳と同じ態の文を出力できれば，BLEU は 1.87 ポイント上昇する．今後は態に限らず，文脈情報を用いた機械翻訳に取り組みたいと考えている．

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*, 2015.

[2] Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Controlling output length in neural encoder-decoders. In *EMNLP 2016*, pp. 1328–1338, 2016.

[3] Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002*, pp. 63–69, 2002.

[4] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP 2015*, pp. 1412–1421, 2015.

[5] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In *LREC 2016*, pp. 2204–2208, 2016.

[6] Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In *NAACL-HLT 2016*, pp. 35–40, 2016.

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS 27*, pp. 3104–3112, 2014.

[8] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. Chainer: a next-generation open source framework for deep learning. In *LearningSys 2015*, 2015.