ニューラル日英翻訳における態の制御

首都大学東京　小町研究室　山岸駿秀　佐藤貴之　叶内晨　小町守　yamagishi-hayahide@ed.tmu.ac.jp

概要

日本語と英語では、表現の使われ方に違いがある。翻訳ではこの違いを意識することでより自然な訳にしなければならないが、これまでの機械翻訳では表現方法の違いという点はあまり考慮されていない。ニューラル英独翻訳での先行研究[Sennrich+ NAACL 2016]では、英文にはない敬語表現をドイツ語文から取得し、新たに作成したコーパスを用いて出力文の敬意制御を行った。
本研究では先行研究をニューラル日英翻訳に用いて、日本語の入力文に英語側の態情報を付けたコーパスを用いて学習させることで、英語の出力文の態制御を行った。

データ作成の手順

① 英語側の文を、能動態か受動態に分類
② 対応する日本語側の文末に、能動態か受動態かのラベルを、それぞれ<Active>、<Passive>として付与

翻訳の手順

① 作成した文を日本語側の学習データとしてアテンション型Encoder-Decoder[先行研究の]を学習
② テストの入力文の末尾に態ラベルをつける
③ 出力文でラベルに対応した態へ変更できたかを確認

ラベルは以下の4パターン

a. 全文に受動態のラベルをつけ、受動態へ変更
b. 全文に能動態のラベルをつけ、能動態へ変更
c. 参照訳の情報を用いて、参照訳と同じ態へ変更
d. 学習データで日本語の動詞ごとに調べた多数派の態へ変更

- 出力文の態は人手で各200文を確認
- BLEUは全文(1812文)を用いて評価

実験設定

コーパス: ASPECコーパス

- 初めの100万文のうち40単語以上の文を省いた827503文
- 300万文を埋め込み層のベクトルの初期値学習に使用

使用ツール

- Stanford Parser 3.5.2
- Word2Vec (gensim)、Chainer 1.12.
- MeCab (辞書: IPADIC ver. 2.7.0)

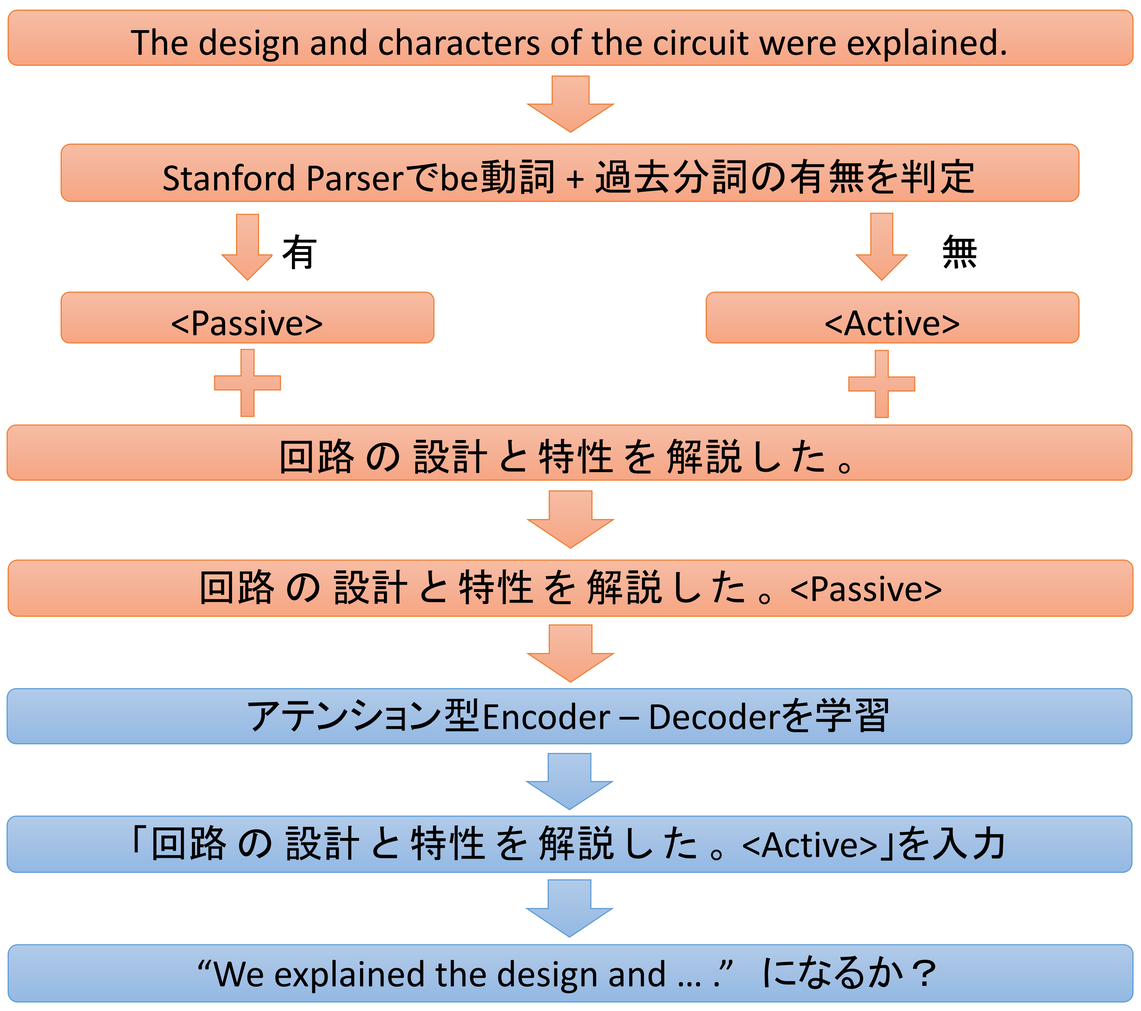
Encoder-Decoderのハイパーパラメータ

- 語彙数: 30000、epoch数: 10
- 埋め込み層次元: 512、隠れ層次元: 512、バッチサイズ:128
- optimizer: Adagrad (初期学習率0.01)

考察

1. do ⇄ be found to do、can be done ⇄ is able to be done のような場合や、be動詞を加えただけの場合など、主語と目的語を入れ替えない形での態変更が多い。
2. 受動態の用法でないbe動詞を用いた文は受動態へ変更不可。
3. be動詞以外の動詞ごとでも、正解率が変化する可能性

BLEUは、参照訳から態情報を取り出した場合のみ上がった。
ラベル予測は、高頻度かつ学習データで態の偏りが小さいと失敗



	能動態	受動態	エラー	Accuracy	BLEU
参照訳	100	100	0	-	-
変更なし	74	117	9	(72.0%)	20.53
受動態へ変更	17	175	8	87.5%	19.63 (-0.90)
能動態へ変更	151	36	13	75.5%	19.93 (-0.60)
参照訳と同じ態に変更	97	94	9	89.5%	21.26 (+0.73)
予測した態に変更 (参照訳と比較時)	72	121	7	69.5%	20.42 (-0.11)
(ラベルと比較時)	-	-	-	87.5%	-

展望

- 主語と目的語が変わるときと変わらないときの違いについて調べる。
- 学習データの動詞ごとの態分布を調べ、正解率にどう影響しているのかを見る。
- 今回の予測方法ではラベルなしの時より悪いので、参照訳の情報を使わなくとも制御を行える別の方法を考える。
- 時制など、他の表現についても同様の実験を行う。

正解例	参照訳が能動態のとき	参照訳が受動態のとき
入力文	熱戻り反応の機構を議論した	リサイクルに関する最近の話題を紹介した
参照訳	This paper discusses the mechanism of the heat return reaction.	Recent topics on recycling are introduced.
能動態へ変更	We discuss the mechanism of the thermal return reaction.	This paper introduces recent topics on recycling.
受動態へ変更	The mechanism of the thermal return reaction is discussed.	Recent topics on recycling are introduced.

失敗例	1. 態の変更が不完全な例	2. 態が変わっていない例
入力文	オゾン生成量が約2mg/h増大した。	テロメラーゼ活性は生殖細胞と癌細胞で高い
参照訳	Ozone formation increased about 2mg/h.	Telomerase activity is high in reproductive cells and cancer cells.
能動態へ変更	The ozone generation quantity was increased about 2mg/h.	The telomerase activity is high in the reproductive cell and cancer cells.
受動態へ変更	The ozone generation quantity increased about 2mg/h.	The telomerase activity is high in the reproductive cell and cancer cells.