述語の活用情報を用いたニューラル日英翻訳

黒澤 道希　　松村 雪桜　　山岸 駿秀　　小町 守
首都大学東京

{kurosawa-michiki, matsumura-yukio, yamagishi-hayahide}@ed.tmu.ac.jp,
komachi@tmu.ac.jp

1 はじめに

機械翻訳の研究において，以前は統計的機械翻訳（SMT）が主流であったが，近年ではニューラルネットを用いるニューラル機械翻訳（NMT）[1][3] が盛んに研究されている．

しかしながら，NMT においては出力単語の選択に膨大な計算を必要とするため，実用上では語彙サイズを制限する必要がある．このとき，一般には学習用コーパスにおける頻度が高い単語のみを語彙として使用するため，学習用コーパスに表層形が存在しない本来の**未知語**だけではなく，統計的機械翻訳では考慮できていた低頻度語も Out-of-vocabulary（**OOV**）として NMT における未知語となってしまう[1]．したがって，これらへの対応が必要である．

現在，これらへの対策として，辞書バックオフ [4] や Byte pair encoding（BPE）[8] などが提案されている．しかしながら，辞書バックオフは後処理のみで解決する手法であるため OOV は削減されず根本的な解決にはなっていない．BPE では単語を部分文字列に分割して全ての語をカバーすることにより未知語をなくすことができるが，語義などの文が持つ言語学的情報を欠落させてしまう可能性がある．

そこで，本研究では言語学的情報を保持した単位で語を扱うことを前提とし，述語の活用情報を用いた手法を提案する．この手法は，OOV の削減だけではなく未知語への対応も行うことができる．なお，活用情報の導入方法として，トークンとして扱う方法（**Conjugation token**）と埋め込みベクトル（**Conjugation feature**）として扱う方法の 2 種類を検討した．

日英および英日翻訳における実験により，本手法を適用することによって，語彙サイズを最大で 86.1%（Tanaka Corpus）まで圧縮することができ，また，未知語の出力が可能であることも確認した．結果として，トークンとして活用情報を用いる手法（Conjugation token）によって，BLEU スコアが日英翻訳で最大 0.86 ポイント，平均 0.44 ポイント，英日翻訳で最大 1.60 ポイント，平均 0.61 ポイント向上した．

[1]本論文では，学習用コーパスに出現しない語を “未知語”，低頻度（1 回以上）で未知語として扱われる語を “OOV” と表現する．

2 関連研究

辞書バックオフ　Luong ら [4] は，未知語トークンとして出力された語を辞書を用いて適切な語に書き換える手法を提案した．この手法は，入力文と出力文とのアライメント情報を用いて対応する語を特定し辞書を用いて翻訳を行う手法であり，様々な手法と同時に適応可能な手法ではあるが，OOV に対する直接的な解決にはなっていない．なお，この手法は提案手法と合わせて用いることができ，さらなる未知語の削減を期待することができる．

Byte pair encoding（BPE）　Sennrich ら [8] は，単語の全てを文字に分割し，頻度が高いものを結合することにより，文字ベースで語彙を決定する手法を提案した．そのため，最低でも文字ベースでは既知の語となるため，OOV が無くなるという利点があるが，結合は頻度に依存するため文法や語義が考慮されないという問題点が存在する．なお，日本語に BPE を用いた場合，漢字など学習用コーパス中に存在しない文字も出現する可能性があり，未知語は一定数存在する．

言語学的素性の追加　Sennrich ら [7] は，NMT に構文情報を明示的に学習させることを目的として，品詞や依存構造などを素性として追加する手法を提案した．この手法は BPE とともに用いることが可能であり，素性を追加することにより翻訳精度が向上されることが述べられている．しかしながら，原言語側のみにしか導入することができず，また，言語学的情報の量など言語に依存する部分が大きい．

3 活用情報の導入

日本語における活用語は語幹と活用形で成り立っているが，従来の単語分割で得られた語彙集合では語幹が同一であっても活用が異なるものは異なる語として扱われるため，語彙を圧迫しOOVを増加させる一因になっている．そこで，提案手法では活用語を語幹と活用形に分割して扱う．これにより複数の活用形を1つに集約することができ，OOVを減少させることができる[2]．また，語幹と活用形をそれぞれ独立したものとして扱うことにより，学習用コーパス中に存在しない語（活用）であったとしても，学習用コーパス中に存在する語幹と活用形を組み合わせることにより表現することができ，未知の活用語にも対応することができる．

なお，本研究では解析・分割にMeCab[3]（IPADic）を用いたため，語幹や活用形の情報についてはIPADicの基準を採用している．具体的には“表層形”・“品詞”・“品詞細分類1”・“活用型”・“活用形”・“見出し語”を用いており，語幹の代わりとして見出し語を使用した．以下活用のある語とは“動詞”・“形容詞”・“助動詞”の3品詞を指すものとする．

本研究では語幹と活用形の導入手法として，それぞれをトークンとして扱う“Conjugation token”と各情報の埋め込みベクトルを結合することによって1つのベクトルで表現する“Conjugation feature”の2手法を提案する．

Conjugation token　本手法では語幹と活用形をそれぞれのトークンとして扱う．活用形は3品詞の区別が可能な特殊トークンとして導入し，他の語とも区別する．この手法では，活用形の特殊トークンも語彙の一部を占めることになる．しかしIPADicを基準にすると，この特殊トークンは最大57トークンであるため，統一することで削減できる語彙サイズと比較すると影響はごくわずかである．また，この手法は語幹と活用形をトークンとして扱うため，出力が日本語の場合でも語幹と活用形の組み合わせによって復元が可能となるため，入力側・出力側のどちらに対しても適応可能な手法である．

具体的には次のような2トークンに変換される．

走る　→　走る　<&動詞・基本形&>
走れ　→　走る　<&動詞・命令形&>
だ　→　だ　<&助動詞・体言接続&>

[2]派生文法を用いた場合も複数の活用形を集約することができるが，派生文法では表層が同じ終止形と連体形の区別ができない．
[3]https://github.com/taku910/mecab

表1: 各コーパスの文数と最大文長．

コーパス	学習用	開発用	評価用	学習用の最大文長
NTCIR	1,638,742	2,741	2,300	60
ASPEC	827,503	1,790	1,812	40
Tanaka	50,000	500	500	16

Conjugation feature　本手法では，入力側の素性として活用形を用いる．具体的には，表層形以外に“品詞”・“品詞細分類”・“活用形”を用い，活用語以外の語についても品詞情報等を追加している．これらの素性は，関係性を考慮して“品詞”・“品詞&品詞細分類”・“品詞&活用形”の3つの組み合わせにし，それぞれの埋め込みベクトルを結合して用いる．なお，活用語については表層形の代わりに語幹を使用する．

この手法では，各素性の埋め込みが独立しているため語彙サイズを圧迫することはなく，語幹に統一したことによって減少した語彙サイズを全てOOVの削減に使用することができる．

ただし，この手法はベクトルから単語への復元ができないため原言語側にのみ適応できる手法である．

4 複数のコーパスにおける翻訳実験

4.1 実験設定

本研究では，2つの比較手法（ベースライン，BPE）と2つの提案手法（Conjugation token，Conjugation feature）について実験を行い，BLEU [6]で評価した．なお，ニューラルネットワークにおける重みの初期値による誤差を考慮するため，各手法につき4回ずつ実験を行い，その平均値を用いた．

実験にはNTCIR PatentMT Parallel Corpus - 10（NTCIR）[2]，ASPEC [5]，Tanaka Corpus（一部抜粋，前処理済み）[4]（Tanaka）の3コーパスを使用した．表1に各コーパスの特徴について示す．なお，ASPECに関しては，学習用データ約300万文のうち，文アライメントの確信度が高い100万文を用いた．また，コーパスの前処理として，日本語側に関してはMeCab（IPADic）を用いて形態素解析を行い，英語側に関してはMoses[5]のTokenizerおよびTruecaserを用いて処理を行い，学習用コーパスに関して，表1に示した最大文長を超える文については削除した．Tanakaのみリファレンス元の英語文がすでに小文字化されていたためTruecaserは用いずに実験を行った．

[4]http://github.com/odashi/small_parallel_enja
[5]http://www.statmt.org/moses/

表 2: 日英翻訳実験結果．数字は 4 回実験した BLEU スコアの平均値．

言語対	コーパス	Baseline	BPE 日本語のみ	Conjugation token	Conjugation feature	BPE 日英両方	BPE 英語のみ
日英	NTCIR	33.87	**34.17**	33.96	33.84	N/A	N/A
	ASPEC	21.08	21.10	**21.46**	21.33	21.43	20.55
	Tanaka	30.17	30.43	**31.03**	30.35	30.45	30.13
英日	NTCIR	36.41	35.96	**36.48**	N/A	N/A	N/A
	ASPEC	29.72	28.96	**29.89**	N/A	30.93	30.59
	Tanaka	28.86	28.66	**30.46**	N/A	29.27	29.15

ベースラインとしては Luong ら [3] のニューラル機械翻訳を実装したもの[6]を使用し，実験の設定を次に示す．なお，コーパスによって異なる設定については [NTCIR/ASPEC/Tanaka] で記載する．

最大エポック : [15/15/30] ，最適化手法 : AdaGrad，
初期学習率 : 0.01，埋め込み層の次元数 : 512，
隠れ層の次元数 : 1024，バッチサイズ : 128，
語彙サイズ : [30,000/30,000/5,000]，
出力語数制限 : [100/100/40]

以下にベースライン以外の各実験ごとの設定を示す．なお，特に言及しない場合，ベースラインと同じ設定を用いている．

Byte pair encoding 比較手法として BPE を用いた実験を行った．提案手法は日本語のみに対する手法であるため，BPE についても日本語のみに行い条件を統一した．

NTCIR，ASPEC に対しては BPE の結合操作回数を 16,000 回で行い，Tanaka は語彙サイズが小さいため 2,000 回で行った．この結果，日本語に関する語彙は語彙サイズを下回ったので OOV は存在しない．なお英日翻訳については，BPE の出力に対して後処理を行った上で評価する．

Conjugation token 英日翻訳の出力は活用形トークンを含むため，IPADic を用いてルールベースで復元した結果を用いて評価する．なお，この時の復元精度は 100%であるが，語幹のみを出力した場合は基本形として出力し，活用形のみを出力した場合は出力から削除する．

Conjugation feature 本手法は原言語側のみに適応できる手法のため，日英翻訳のみを行った．追加する各素性の埋め込みを結合したものを入力として用いた．この時，埋め込み層の次元数を 512 に統一するために，入力側の各素性の次元数は “品詞 : 4，品詞&品詞細分類 : 8，品詞&活用形 : 8” としたため，入力側の単語埋め込みベクトルの次元数は 492 である．

[6] http://github.com/yukio326/nmt-chainer

4.2 結果

実験の結果を表 2 左に示す．

結果として，Conjugation token とベースラインを比較すると，全てのコーパス，両翻訳方向で BLEU スコアが向上した．

また，日本語のみに BPE を用いた手法との比較においては，NTCIR の日英翻訳において BLEU スコアが若干低下したものの，全体的に同等か向上する結果を得ることができた．

5 考察

5.1 活用情報の効果に関する考察

実験結果より Conjugation token がベースラインと比較して BLEU スコアが向上した．また，Conjugation feature においてもベースラインと比較してスコアが概ね向上している．これらより活用情報を利用することが有用であり，特殊トークンとして用いた手法の方がより有用であることが示された．また，BPE と比較してもスコアが概ね向上することより，言語学的情報を保持することが重要であると考えられる．

Conjugation token では全体的にスコアが上がっており，コーパスのサイズが小さいほどスコアが向上する傾向にある．特に Tanaka において精度の向上が見られているが，これはコーパスのサイズが小さいことと，コーパス中に含まれる活用語の種類と量が多かったためだと考えられる．逆に NTCIR においては大きな効果はなく BPE を用いた英日翻訳では劣る結果となった．これはコーパスのサイズが大きいため十分に学習されており，今回の手法を適用することによる効果はほとんどなかったと考えられる．

5.2 BPE の結果に対する考察

今回，BPE については日本語側のみに適応した手法で比較した．しかしながら，今回の結果において BPE を日本語側のみに適応した手法ではベースラインを下回る結果となったものが複数存在した．そこで，日英

表 3: 英日翻訳の出力例.

入力文	where should i sit ?
参照訳	どこ に 座っ たら いい です か 。
ベースライン	どこ で <unk> ば いい です か 。
BPE	どこ で 座ら ない か 。
Conjugation token	どこ で 座れ ば いい でしょ う か 。

表 4: 語彙カバー率.

コーパス	ベースライン	Conjugation token	Conjugation feature
NTCIR	26.48%	27.43%	27.46%
ASPEC	18.56%	18.96%	18.96%
Tanaka	46.46%	53.95%	54.41%

両方に BPE を適応した結果と英語側のみに BPE を適応した結果をそれぞれ表 2 右に示す．なお，実験の都合上 ASPEC と Tanaka についてのみ行い，それぞれ 1 回のみ実験した結果を示している．

この結果より，日英両方に BPE を適用した場合では先行研究と同様ベースラインを上回る結果を得ることができ，英語のみ適用した手法では日英両方に比較するとスコアが悪くなった．また，日英，英日どちらの翻訳方向においても入力言語側に BPE を適用したほうが高いスコアを出すことがわかった．このことより，BPE は両言語に適用する必要があるが，今回の実験設定における結果は適切であると考えられる．

5.3 出力文

英日翻訳の出力例を表 3 に示す．Conjugation token では “座れ” という語が出力されているが，これは学習用コーパス中には存在しない未知語である．しかし，語幹（座る）と活用形（仮定形）がそれぞれ学習用コーパス中に存在しているため，その組み合わせによって出力することができた．このように，Conjugation token を用いることで未知語であっても出力することができる．

また，複数の活用形を一つにまとめることで，OOV が削減されて考慮できる語が増えることはもちろん，活用語の語幹としての頻度が高くなることで，適切な文や流暢な文が出力される傾向にあると考えられる．

一方で，語幹と活用情報を分離することにより，活用情報が適切に出力できなくなるため，流暢性が損なわれる文がいくつか存在した．また，語幹の品詞情報についても明確に保持することになるため，語幹の選択による後続の語への影響が大きい傾向にあった．

5.4 語彙カバー率

本提案手法において語彙サイズの削減を行うことができた．各コーパスごとの学習用コーパス中におけるタイプの語彙カバー率を表 4 に示す．なお，Conjugation feature においては見出し語のみの数を評価している．

この表より，どのコーパスにおいても OOV が削減されていることがわかる．特に Tanaka において大幅に改善されており，これは短文が中心のコーパスであるため活用語が多く含まれており，また，表現が固定されないことにより活用形が複数使われることが多かったため，本手法が有効であったと考えられる．

6 おわりに

本研究では，日本語活用情報を用いた分割を行い，単語が持つ言語学的情報を維持しながら翻訳を行うことで，語彙カバー率および翻訳精度の向上が示された．また，OOV だけにとどまらず未知語に対しても有効であることが確認された．加えて，活用形は素性として追加するよりもトークンとして追加する方が効果的であることを確認することができた．本研究では日英の言語対に関してのみ実験を行ったが，今後は他言語においても活用情報が有用であるかを調査したい．

参考文献

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

[2] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the patent machine translation task at the NTCIR-10 workshop. pp. 260–286.

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

[4] Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. In *Proc. ACL*, pp. 11–19, 2015.

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

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pp. 311–318, 2002.

[7] Rico Sennrich and Barry Haddow. Linguistic input features improve neural machine translation. In *Proc. WMT*, pp. 83–91, 2016.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proc. ACL*, pp. 1715–1725, 2016.