ニューラル機械翻訳における共起情報を考慮した語彙選択

勝又 智　　松村 雪桜　　山岸 駿秀　　小町 守

首都大学東京

{katsumata-satoru, matsumura-yukio, yamagishi-hayahide}@ed.tmu.ac.jp,
komachi@tmu.ac.jp

1 はじめに

近年，機械翻訳の分野において，統計的機械翻訳と比較して流暢な出力をもたらすニューラル機械翻訳（Neural Machine Translation，NMT）が注目を集めている．しかし，NMT では出力単語の選択時に大きな計算コストがかかるため，実際には出力単語の候補となる語彙のサイズに制限をかける．この処理のことを**語彙選択**と呼ぶ．このとき，一般的には語彙サイズに上限を設けて学習コーパス内での出現頻度が高い単語から順に取り出し，使用しなかった低頻度語は未知語トークンに置き換えている．

しかしながら，頻度を用いた出力語彙選択では，本来語彙として使用する必要性の低い単語であるにも関わらず，学習コーパスに複数回出現するために語彙として使用されてしまう場合がある．実際に田中コーパスの学習コーパス（5 万文）内で出現した回数が 1，2 回違うだけで語彙として使用されない単語を表 1 に示す．この表のように，頻度を用いて出力語彙を選択する場合はコーパス内に単純に多く出現する固有名詞が多く選択されてしまい，コーパスを表現するために必要な名詞や動詞などが未知語トークンとして扱われてしまう．固有名詞は対訳が一意に定まることが多いため，動詞などの一意に定まらない可能性の高い語を出力語彙として用意し，流暢に翻訳を行えるようにすべきであると考えられる．

そこで，本研究では**共起情報を単語の重要度として考慮した出力語彙選択手法**を提案する．NMT では，以前出力した単語を考慮し，文脈を用いて次の単語の出力を行う．そのため，文脈に基づいて出力語彙を決定する提案手法では，文脈として学習すべき単語を出力語彙として選択できると考えられる．本研究では HITS アルゴリズムを単語間の共起解析に適用し，単語の重要度として用いる．

田中コーパスを用いた翻訳実験の結果，頻度のみを用いて構築した語彙と比べて提案手法では使用する単語の約 1/10 が置き換わり，従来の NMT モデルに比べて，提案手法では日英，英日の実験において BLEU スコアの向上を確認した．

表 1: 田中コーパスにおいて使用する語彙を 3,000 とした時に使用される単語，されない単語 [1]

使用される単語	使用されない単語
広島（6）	暗かっ（4）
ローマ（6）	笑わせ（4）
道具（6）	支払っ（4）
農場（5）	困惑（4）
お客様（5）	好ま（4）

2 関連研究

NMT における語彙制限問題への先行研究として，Sennrich ら [9] は Byte Pair Encoding (BPE) を文字列に適用し，学習データ中の単語を全て表現できるような部分文字列集合を作成することで，使用語彙に含まれる単語を増加させた．本研究の目的は，使用する語彙を増やすことによる NMT モデルの改善ではなく，学習するコーパスのドメインにふさわしい語彙を使用することでモデルの表現力を改善することである．そのため，本研究の貢献は BPE とは異なっている．

また，Luong ら [7] は未知語トークンを対訳辞書を用いて置換する手法を提案した．未知語トークンを出力する際に，対応した原言語の位置を学習させ，その位置情報を元に未知語トークンを事前に用意した辞書から翻訳を行なっている．今回の提案手法は前処理として行うため，この手法と同時に使用することが可能である．

単語間の共起を用いて単語のランキングを行うアルゴリズムは，自然言語処理においても多くの研究で用いられている．例えば，PageRank [1] を重要語抽出に用いた研究として Mihalcea and Tarau [8] が存在

[1] ただし，括弧内の数字は学習コーパス内の頻度を表す．

アルゴリズム 1 HITS

Require: ハブ度スコアベクトル $\boldsymbol{i_0}$
Require: 隣接行列 $\boldsymbol{A}$
Require: 繰り返し数 τ
Ensure: ハブ度スコアベクトル $\boldsymbol{i}$
Ensure: 権威度スコアベクトル $\boldsymbol{p}$

```
function HITS(i_0, A, τ)
    i ← i_0
    for t = 1, 2, ..., τ do
        p ← A^T i
        i ← Ap
        i and p の正規化
    return i and p
end function
```

する．彼らは節点を単語に，辺を固定幅の単語間の共起としてグラフを構築し，PageRank アルゴリズムを実行して重要語を抽出した．この手法は教師なし学習であるが，重要語抽出の研究で当時の最高精度であった教師あり学習の手法と同等の精度を達成した．また HITS [5] を自然言語処理におけるブートストラップ法のシード単語とストップリスト作成に用いた Kiso ら [4] の研究が存在する．HITS や PageRank は単純な頻度では捉えられない各単語間の関連性の抽出に効果をもたらしていると考えられる．今回の研究ではコーパスを 1 つの文書としてみて，そのコーパス内の様々な重要語と共起する語を取り出すために HITS を用いる．

3 共起を考慮した語彙選択手法

3.1 HITS によるハブ度と権威度

Kleinberg ら [5] の提案した Web ページのランキングアルゴリズムである HITS では Web ページのリンク遷移を表す隣接行列 $\boldsymbol{A}$ を用いて，各ページに対してハブ度と権威度と呼ばれるスコアを求めている．権威度スコアの高い Web ページは，多数のハブ度スコアの高いページからリンクを張られているページであり，ハブ度スコアの高い Web ページは，権威度スコアの高いページへのリンクを張っているページである．アルゴリズム 1 に HITS の疑似コードを記述する．繰り返し数 τ を十分大きい値にすることで，ハブ度スコアと権威度スコアは収束することが知られている．

3.2 HITS を用いた語彙選択

本研究では，HITS アルゴリズムにおける各 Web ページを学習コーパス内の単語（節点）とし，Web のリンクを単語間の共起（辺）として捉えて隣接行列 $\boldsymbol{A}$ を作成した．

誰 が 一番 に 着く か 私 に は 分かり ませ ん 。
十中八九 彼 は 成功 する だろ う 。
あなた の 銀行 口座 を 教え て いただけ ます か 。
私 の 願い を 聞い て いただけ ます か 。
彼女 は 若さ を 十分 に 持っ て いる 。
多く の 動物 が 人間 によって 滅ぼさ れ た 。
そうなる と かなり きつい 仕事 という こと に なる 。
なぜ 彼 は 計画 を 変え た の です か 。
これら の 本 は 私 の 本 です 。
私 の 靴 は 修理 する 必要 が ある 。

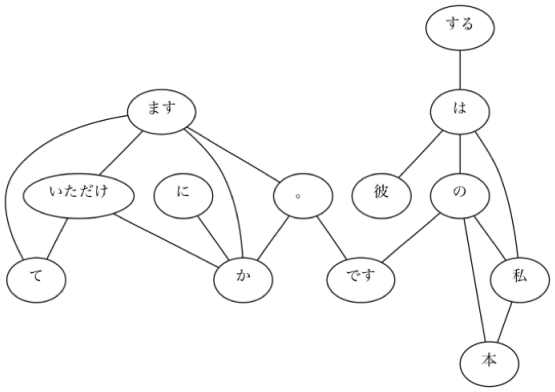


図 1: 単語グラフの例 [2]

Web ページとは違い，単語間の共起はリンクがあるかどうかの 2 値ではないので，辺の重みとしていくつかの共起尺度を使用した．用いた共起尺度と共起を測定した文脈については 3.3 節で述べる．このように作成した隣接行列 $\boldsymbol{A}$ を用いて HITS アルゴリズムを実行する．その結果として，学習コーパス内の各単語に対して重要度を表すスコアが獲得できる．

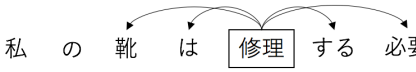
田中コーパスの学習コーパスから抽出した 10 文に対して作成したグラフ構造の例を図 1 に示す．HITS アルゴリズムを用いることで得られるスコアの高い単語は，様々な単語と共起する単語であると考えられる．また，図 1 から，様々な単語と共起する単語と共起（2 次の共起）する単語のスコアも高くなると考えられる．

本研究では，ハブ度スコアの高い単語が重要な内容語と共起する単語と考え，このスコアの高い単語を順に NMT の語彙として選択する．これにより単純に頻度のみを用いて選択した単語よりも学習にふさわしい単語が使用語彙に含まれると考えられる．

3.3 単語グラフの構築

本研究では共起関係を獲得する文脈として，ある単語に対してその周辺語との組み合わせを利用した．具体的には，注目している単語の窓幅 N 内の周辺語を，それぞれ注目している単語と組みわせて，出現頻度を数えていく．このように文脈を決めたとき，隣接行列

[2] 本研究では，学習コーパス内で出現頻度が 1 回の単語や，共起頻度が 1 回の単語の組み合わせはグラフから除いた．



得られた共起:（修理，する），（修理，必要），
（修理，は），（修理，靴）

図 2: 共起の文脈の例

$\boldsymbol{A}$ が対称行列になるためハブ度と権威度，どちらも同じものが得られる．ただし，文頭や文末の単語に対して注目している場合はそれぞれ右隣，左隣の単語を見ることになる．図 2 に窓幅 N を 2 とした例を示す．

今回作成した隣接行列の要素 A_{xy} として，単語間の共起の単純頻度（式 1）（Freq），Positive Pointwise Mutual Information（PPMI）をそれぞれ用いた．ただし，単純な PPMI では学習コーパス内の低頻度語に対して敏感に反応してしまうため，頻度を考慮するために PMI に対して共起回数の対数で重み付けしたものを求め，この重み付けした PMI を元にした PPMI を用いた（式 2）．

$$A_{xy}^{freq} = |x, y| \quad (1)$$

$$A_{xy}^{ppmi} = \max(0, \mathrm{pmi}(x, y) + \log_2 |x, y|) \quad (2)$$

以下に，ある単語 x と，その単語に対する共起語 y の PMI の式を示す．ここでの M は出現した組み合わせのトークン数を意味し，$|x, *|$ と $|*, y|$ はそれぞれ単語 x，共起語 y を固定した時の組み合わせのトークン数を意味している．

$$\mathrm{pmi}(x, y) = \log_2 \frac{M \cdot |x, y|}{|x, *||*, y|}$$

4 NMT における語彙選択実験

4.1 実験設定

本研究では，田中コーパス[3]を用いて英日翻訳実験と日英翻訳実験を行った．学習コーパスとして 50,000 文を使用し，開発セットとして 500 文，テストには 500 文を用いた．また日本語の単語分割は MeCab[4]（辞書に IPADic 2.7.0 を使用）で行った．

共起の窓幅 N は 2 に設定した．学習コーパス内で 1 回しか共起しない組み合わせについては，隣接行列の要素 A_{xy} の値を 0 に設定した．HITS アルゴリズムの繰り返し数 τ は 300 に設定した．提案手法によって選択した語彙を原言語側に対してのみ用いた場合と目的言語側のみに用いた場合，どちらにも用いた場合についてそれぞれ実験を行った．

[3]https://github.com/odashi/small_parallel_enja
[4]http://taku910.github.io/mecab/

表 2: 各手法の baseline に対する使用語彙の異なり数

	日本語側		英語側	
	Freq	PPMI	Freq	PPMI
異なり数	281	333	351	312

NMT のモデルには Luong ら [6] の global dot attention を使用した[5]．本研究は baseline として目的言語側の語彙を学習コーパス内における頻度で決定したものを用いる．語彙サイズは，baseline と提案手法どちらも，原言語側，目的言語側でそれぞれ 3,000 とした．隠れ層，埋め込み層の次元数はそれぞれ 512 次元，ミニバッチサイズは 150，最適化手法として AdaGrad を初期学習率 0.01 で用いた．また，dropout を 0.2 の確率で適用した．

また，実験では事前に辞書を用意し，モデルが未知語トークンを出力した際にアテンションスコアの最も高い原言語側の単語をクエリとして辞書に基づいて未知語トークンを置き換え，辞書に該当する単語がなかった場合は原言語側の単語をそのまま出力する処理を行った（unk_rep）[3]．この辞書は Dyer ら [2] による fast_align を学習コーパスに用いて得た単語アライメントを元に作成した．

評価は各手法について重み行列の初期値を変更し各 5 回ずつ実験を行い，得られた BLEU スコアの平均を用いた．

4.2 実験結果

英日，日英翻訳における baseline 及び各提案手法の結果を表 3 に示す．それぞれの手法について，翻訳する言語の方向に関わらず，baseline と比較して一貫した BLEU スコアの上昇が確認できる．

baseline に対して提案手法を用いて作成した語彙内のタイプの異なり数を表 2 に示す．baseline で使用した語彙と比べて提案手法ではそれぞれ使用する単語の約 1/10 が置き換わっていることがわかる．

5 分析

英日翻訳の出力例を表 4 に示す．baseline では未知語トークンを出力しており，辞書が誤っていたために未知語トークン処理がうまくできていなかった．一方で提案手法は baseline と比較して太字箇所がより流暢に翻訳ができていることがわかる．これは提案手法に

[5]https://github.com/yukio326/nmt-chainer

表 3: 田中コーパスを用いた英日，日英翻訳の BLEU スコア（5 回の実験の平均）

言語対	unk_rep	baseline	HITS（原言語側） Freq	HITS（原言語側） PPMI	HITS（目的言語側） Freq	HITS（目的言語側） PPMI	HITS（両方） Freq	HITS（両方） PPMI
英日	未使用	29.89	29.99	29.92	29.97	**30.23**	29.97	29.97
	使用	30.08	30.18	**30.30**	30.12	30.22	30.25	30.29
日英	未使用	28.95	29.26	29.38	29.43	29.11	29.57	**29.66**
	使用	29.19	29.50	29.61	29.65	29.51	**29.80**	29.69

表 4: 英日翻訳において提案手法を原言語側と目的言語側のどちらにも用いた際の出力例 [6]

原言語文	i think it 's worth a try .
baseline	それ は て† の 価値 が ある よ 。
HITS+Freq	**やっ て みる** 価値 が ある と 思う 。
HITS+PPMI	それ は **やっ て みる** 価値 が ある 。
参照訳	それ は **やっ て みる** 価値 は ある と 思う 。

よってモデルの表現能力が向上しているからだと考えられる．

次に英日翻訳の baseline と，提案手法を原言語と目的言語どちらにも用いた出力について，4.1 節で述べた未知語トークン処理を行う前の出力された未知語トークンの数を調べたものを表 5 に示す．このことから baseline と比べて提案手法で学習したモデルは未知語トークンを出力しやすいモデルであることがわかる．そのため，未知語トークン処理を施すことにより baseline と比較して BLEU スコアがより高くなると考えられるが，表 3 で示したように，unk_rep 後もそれほど baseline と差は開かなかった．これは辞書を田中コーパスの学習コーパス（5 万文）で作成したため，あまり辞書の精度が高くなかったからだと考えられる．

6 おわりに

本研究ではニューラル機械翻訳において，単語間の共起を用いた語彙選択手法を提案した．この手法は従来の語彙制限問題への研究とは違い，より学習にふさわしい語彙を選択する．この手法を用いることにより従来の頻度順に使用する語彙を決定する手法と比べて，出力未知語トークン数は上昇しているにも関わらず BLEU は一貫して向上した．

本研究では単語間の関係性を表現する際に対称行列を用いたが，次の課題としては，非対称な行列を設計して HITS アルゴリズムを行うことで得られた語彙を用いて翻訳を行うことが挙げられる．提案手法は NMT モデルだけでなく，ニューラル生成モデル全般に対して有効であると考えられるので，他の対話生成や画像

[6]未知語トークン処理が行われた単語を † で示す．

表 5: 出力された未知語トークン数の平均 [7]

手法	トークン数	未知語を含む文数
baseline	215.8	196.8
HITS+Freq	224.6	204.0
HITS+PPMI	237.8	208.4

の説明文生成などの他のタスクに対しても用いることが可能であると考えられる．

参考文献

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, Vol. 30, No. 1-7, pp. 107–117, April 1998.

[2] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In *Proc. of NAACL-HLT*, pp. 644–648, 2013.

[3] Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. Montreal neural machine translation systems for WMT15. In *Proc. of WMT*, pp. 134–140, 2015.

[4] Tetsuo Kiso, Masashi Shimbo, Mamoru Komachi, and Yuji Matsumoto. HITS-based seed selection and stop list construction for bootstrapping. In *Proc. of ACL*, pp. 30–36, 2011.

[5] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, Vol. 46, No. 5, pp. 604–632, September 1999.

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

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

[8] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into texts. In *Proc. of EMNLP*, pp. 404–411, 2004.

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

[7]未知語トークン数と未知語トークンを含む文数はどちらも 5 つのモデルの出力の平均である．