文字レベルの日中ニューラル機械翻訳における文字特徴情報の利用

張 津一　　松本 忠博
岐阜大学 大学院 工学研究科
{zhang, tad}@mat.info.gifu-u.ac.jp

1 はじめに

近年，ニューラル機械翻訳（NMT）は注目すべき成果をあげている [1, 2]．単語レベルの NMT における問題点として，語彙サイズが制限されることが挙げられる．日本語や中国語のように文中の単語の区切りが明示されない言語では，統一された正しい単語分割結果を得ることも容易ではない．文字レベルの NMT では，これらの問題を回避することができる．

一方，Sennrich ら [3] は，通常の単語レベルの NMT において，POS タグなどの単語の特徴情報が翻訳精度の向上に有効であることを示した．本研究では文字レベルの NMT においても何らかの文字特徴情報が有用ではないかと考え，漢字の部首および画数を入力特徴情報として加えて，文字レベルの NMT による日本語から中国語への機械翻訳を試みた．その結果，部首を特徴情報として加えることにより翻訳精度の向上が見られた．NMT システムは Luong ら [2] のものをベースとして用い，実験には WAT2017[4] の学術論文サブタスクでも用いられた ASPEC-JC コーパス [5] を文字ごとに分割して使用した．

本研究では文字の特徴情報の一つとして漢字の部首を用いる．六書では漢字の造字法・用字法を，象形・指事・形声（形聲）・会意・転注・仮借の 6 つに分類しているが，漢字の 80% 以上は，意符（意味成分，物事の類型を表す）と音符（発音を表す）を組み合わせて作られた形声文字であると言われている．例えば，「銅」の部首「釒」（かねへん）は金属という意味カテゴリを表し，「同」は音を表す．そこで，部首がもつ意味的な情報が翻訳精度の向上につながることを期待して，入力特徴情報に加えた．もう一つの文字特徴情報である画数は，その漢字の複雑さを表しており，その文字が表す概念の複雑さに関係している可能性がある．それが翻訳精度に良い影響が与えないかについても調べた．

2 関連研究

どの言語においても単語の数に比べて文字の数は遥かに少ないため，文字レベルの自然言語処理では語彙サイズの問題が解消される．文字レベルでの自然言語処理の利点は，言語モデル [6]， POS タグ付け [7]，固有表現抽出 [8]，構文解析 [9]，学習表現 [10] など，これまでにもいくつか示されてきた．

欧米の言語を対象とした NMT では，単語を文字ではなく部分文字列（サブワード）に分割することで語彙サイズの制約に対処する方法も提案されている [11]．しかし，単語が比較的多くの文字で表現される欧米の言語に比べ，表語文字である漢字を使用する日本語や中国語では単語の文字数が少なく，特に中国語では一文字の単語も多い．そのような単語をサブワードに分割することは困難である．

3 NMT と特徴情報の追加

NMT は原言語文に対する目的言語文の条件付き確率を計算する．ここでは本研究で使用する Luong ら [2] による NMT システムについて，文献 [3] を元に簡単に説明する．この NMT システムは，リカレントニューラルネットワークを用いたグローバルな注意機構付きのエンコーダ・デコーダモデルを実装したものであるが，本研究ではこれを文字レベルで利用する．エンコーダは，双方向 LSTM リカレントニューラルネットワークであり，入力系列 $\boldsymbol{x} = (x_1, \cdots, x_m)$ を読み取って，順方向の隠れ状態列 $(\overrightarrow{h_1}, \cdots, \overrightarrow{h_m})$ と逆方向の隠れ状態列 $(\overleftarrow{h_1}, \cdots, \overleftarrow{h_m})$ を求める．隠れ状態 $\overrightarrow{h_j}$ と $\overleftarrow{h_j}$ は連結され，アノテーションベクトルが作られる．

デコーダは，目的言語文 $\boldsymbol{y} = (y_1, \cdots, y_n)$ を予測する LSTM リカレントニューラルネットワークである．各単語（文字レベルの場合，各文字）y_i は，リカレント隠れ状態 s_i と，前回予測された単語（または文字）

y_{i-1}，文脈ベクトル c_i を元に予測される．文脈ベクトル c_i は，アノテーション h_j の加重和として計算される．各 h_j の重みは，y_i と x_j のアラインメントについての情報を表すモデル α_{ij} を通じて決められる．

エンコーダの順方向状態は以下のように表される．

$$\vec{h_j} = \tanh(\vec{W}Ex_j + \vec{U}\vec{h}_{j-1}) \quad (1)$$

ここで，$E \in \mathbb{R}^{p \times V_x}$ は単語埋め込み行列であり，$W \in \mathbb{R}^{q \times p}$ と $U \in \mathbb{R}^{q \times q}$ は重み行列である．p，q，V_x はそれぞれ，単語ベクトルのサイズ，隠れユニットの数，原言語の語彙サイズである．

入力特徴情報の数を $|F|$ とすると，式 (1) は次のように一般化することができる．

$$\vec{h_j} = \tanh(\vec{W}(\mathop{\Big\|}_{k=1}^{|F|} E_k x_{jk}) + \vec{U}\vec{h}_{j-1}) \quad (2)$$

ここで，演算子 $\|$ はベクトルの連結を表す．$E_k \in \mathbb{R}^{p_k \times V_k}$ は特徴埋め込み行列であり，$\sum_{k=1}^{|F|} p_k = p$ である．V_k は k 番目の特徴情報の語彙サイズ（種類数）である．特徴情報の埋め込みベクトルは，単語埋め込みベクトルと同じ方法で別々に求められ，最後に単語埋め込みベクトルと連結される．

4　日本語文字の特徴情報

本研究では，日本語から中国語への文字レベルのニューラル機械翻訳において，文字の埋め込みベクトルに，文字の部首，部首の画数，文字全体の画数を入力特徴情報として加える．

部首

部首は漢字の構成要素の一つであり，漢字を字画の構成で分類・配列する際に基準として用いられる．部首によって漢字を分類した辞典を（狭義の）字書というが，ある漢字がどの部首に分類されるかは字書による．意味カテゴリを表す意符（形符）と音声を表す音符（声符）で構成される形声文字では，意符が部首として用いられることが多い．例えば，「江」「河」の部首「氵」（さんずい）は水を表し，残り部分「工」「可」は音を表す．康熙字典では漢字が 214 の部首に分けられ，画数順に記載されている．康熙部首を図 1 に示す．康熙部首はすべて Unicode に収録されている（U+2F00〜2FD5）．

現代の日本語の標準的な文章には漢字と仮名，数字，英字，句点や括弧などの記号が混在する．漢字以外の文字には部首は存在しないが，本研究では日本語入力文中のすべての文字に特徴情報を付与するため，漢字以外の文字に対しても以下のように部首を設定した．

一	丨	丶	丿	乙	亅	二	亠	人	儿	入	八	冂	冖	冫	几
凵	刀	力	勹	匕	匚	匸	十	卜	卩	厂	厶	又	口	囗	土
士	夂	夊	夕	大	女	子	宀	寸	小	尢	尸	屮	山	巛	工
己	巾	干	幺	广	廴	廾	弋	弓	彐	彡	彳	心	戈	戸	手
支	攴	文	斗	斤	方	无	日	曰	月	木	欠	止	歹	殳	毋
比	毛	氏	气	水	火	爪	父	爻	爿	片	牙	牛	犬	玄	玉
瓜	瓦	甘	生	用	田	疋	疒	癶	白	皮	皿	目	矛	矢	石
示	禸	禾	穴	立	竹	米	糸	缶	网	羊	羽	老	而	耒	耳
聿	肉	臣	自	至	臼	舌	舛	舟	艮	色	艸	虍	虫	血	行
衣	襾	見	角	言	谷	豆	豕	豸	貝	赤	走	足	身	車	辛
辰	辵	邑	酉	釆	里	金	長	門	阜	隶	隹	雨	青	非	面
革	韋	韭	音	頁	風	飛	食	首	香	馬	骨	高	髟	鬥	鬯
鬲	鬼	魚	鳥	鹵	鹿	麥	麻	黃	黍	黑	黹	黽	鼎	鼓	鼠
鼻	齊	齒	龍	龜	龠										

図 1　214 康熙部首

仮名は，日本語を表記するために漢字の音を借用して用いられた万葉仮名（借字）が元になっており，平仮名は万葉仮名の草書化が進められて独立した字体になったもの（図 2），片仮名は漢文を和読するための訓点として万葉仮名の一部を省略して付記したものが始まりと考えられている．本研究では各仮名文字の元になった漢字の部首をその仮名の部首として使用する．

アラビア数字は，対応する漢数字の部首を使用する．英字には一律に「英」の部首（くさかんむり）を割り当て，記号には「符」の部首（たけかんむり）を割り当てる．

画数

漢字の字体は，点（てん）・横（よこ）・竪（たて）・提（はね）・捺（右はらい）・撇（左はらい）・鈎（かぎ）・折（おれ）・彎（左そり）・斜（右そり）などの筆画（ストローク．筆を紙面に下してから離すまでにできる線または点）の組み合わせで構成されている．現在の中国では約 30 ほどの筆画が設けられており，筆画の数を画数という．例えば，「銅」の部首「金」の筆画は撇・捺・横・横・竪・点・撇・提で，画数は 8 画となる．画数は漢字の複雑度を表し，親密度とも関連がある．

図 2　漢字から平仮名への変化（Wikipedia「平仮名」より）

表 1　日本語入力文字列と各文字の特徴情報の例

日本語文	溝幅は１０ｍｍ以上が必要と推定した．
康熙部首	水巾水一雨艹艹人一力心襾止手宀丿大灬
部首画数	３３３１８３３１１２４６４３３１３６

部首と画数の取得

表 1 は日本語の原文の一部と，その各文字に対応する康熙部首，部首の画数を示している．この文には，漢字，平仮名，数字，英字，記号が含まれている．

入力文中の各文字の部首や画数を取得するために，本研究の実験では cjklib[*1]を使用した．cjklib は中国，日本，韓国で使われる漢字の発音，部首，グリフの構成部品，筆画，異体字などの情報を得るための Python ライブラリである．現時点では Python3 に対応していないため，Python3 で使用できるように一部手を加えて使用した．

5　実験

実験には，ASPEC（Asian Scientific Paper Excerpt Corpus）[5] の日中学術論文抜粋コーパスを使用した．モデル中のパラメータは [− 0.1，0.1] を範囲とする一様分布の乱数により初期化を行い，バイアス項は 0 とした．各パラメータの学習には確率的勾配降下法 (初期学習率は 1.0) を用い，ミニバッチサイズを 10 とした．勾配ノルムは 1 でクリップした．また，単語ベクトル，隠れ層の次元は全て 512 とした．過学習を避けるため，dropout 確率は 0.8 に設定し，デコード時に行うビームサーチのビームサイズは 5 とした．

文字ベースでの翻訳のため，日本語テキスト・中国語テキストともに文字ごとに空白文字を挿入して分割するが，単語ベースの翻訳システムと同じ条件で BLEU スコアを計算するために，出力中国語テキスト中の空白文字をいったん取り除いた後，Python モジュール Jieba[*2]を使って単語に分割した．

翻訳システムの実装には OpenNMT[13] を用いた．訓練には NVIDIA 社の GeForce GTX 1080Ti を使用したところ，処理時間は 1 秒あたり約 3 千文で，モデルの訓練には 3〜4 日かかった．

実験の結果を表 2 に示す．表中の「ppl」は perplexity を表し，モデルが与えられた原文の参照翻訳をどの程度うまく予測できるかを示すのに有効な指標である．部首の画数，文字の画数を入力特徴として追加した場合はいずれも，特徴情報を何も追加しない文字レベルの翻訳結果よりも BLEU 値が下がった．画数は文字の複雑度を表すが，翻訳精度にはよくない影響しか与えなかった．一方，文字の特徴情報として部首のみを追加した文字レベルの NMT システムでは，テストデータで BLEU 値 39.65 を得た．これは同じコーパスを使用した WAT2017 の日中翻訳の評価結果の中で最も高かった BLEU スコアを約 4 ポイント上回っている．さらに，dropout を 0.3 に調整したとき，perplexity が 3.07，テストデータで BLEU 値が 40.61 となり，最も良い結果が得られた．この実験より，文字レベルの NMT システムは日中翻訳において非常に効果的であり，部首を特徴情報として加えることによりさらに良い結果が得られることがわかった．

翻訳結果の文を観察したところ，部首を特徴情報として追加した提案手法では，特徴情報を何も追加しない文字レベルの NMT と比較して，単語の翻訳精度が向上している例が見られた．例えば，「着生状況」，「界面活性剤」などの単語は提案手法により正しく翻訳できるようになった．ただし，「ヘキサデシルトリメチルアンモニウムブロミド」などは翻訳が困難であり，正しく翻訳されなかった．文字レベルの NMT により語彙サイズの制限は解消された．文字レベルの NMT 共通の問題点としては，単語レベルの NMT よりも入力列の長さが増加することがあげられる．

[*1] https://github.com/cburgmer/cjklib

[*2] http://github.com/fxsjy/jieba

表 2　実験結果

システム	ppl (↓)	BLEU (↑)	
	dev	devtest	test
Baseline (WAT2017 公開データ)	—	—	35.67
文字レベル（追加特徴情報なし）	3.73	39.03	39.25
文字レベル + 部首の画数	3.61	38.66	38.77
文字レベル + 文字の画数	3.62	38.61	38.97
文字レベル + 部首 + 部首の画数	3.61	38.93	38.94
文字レベル + 部首 + 部首の画数 + 文字の画数	3.61	37.06	38.78
文字レベル + 部首	3.64	**39.62**	**39.65**
（同上，dropout 調整時）	**3.07**	**40.58**	**40.61**

6　おわりに

本研究では，部首と画数を文字の特徴情報として追加することで，日本語から中国語への文字レベルのニューラル機械翻訳をさらに改善できないか，ASCPEC-JC コーパスを用いて実験した．その結果，漢字だけではなく，仮名やアラビア数字などの文字にも部首を設定し，文字の特徴情報として加えることにより，翻訳精度を向上させることができた．特徴情報を追加しない文字レベルの NMT と比較すると，パープレキシティは約 0.1，BLEU 値は devtest データと test データでそれぞれ約 0.5 および 0.4 向上した．一方，画数の情報は翻訳精度を下げる結果になった．中国語から日本語，あるいは，日本語から他の言語への文字レベルの翻訳においても，部首やその他の特徴情報が翻訳精度の向上に役立つ可能性があると考えられる．

参考文献

[1] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. ICLR2015*, 2015, pp. 1–15.

[2] M.T. Luong, H. Pham, C.D. Manning, “Effective approaches to attention-based neural machine translation,” in *Proc. 2015 Conf. on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.

[3] R. Sennrich and B. Haddow, “Linguistic input features improve neural machine translation,” in *Proc. First Conference on Machine Translation*, Vol.1: Research Papers. ACL, 2016, pp. 83–91.

[4] T. Nakazawa, S. Higashiyama, C. Ding, H. Mino, I. Goto, H. Kazawa, Y. Oda, G. Neubig, S. Kurohashi, “Overview of the 4th workshop on Asian translation,” in *Proc. 4th Workshop on Asian Translation (WAT2017)*, 2017, pp.1–54.

[5] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara, “Aspec: Asian scientific paper excerpt corpus,” in *Proc. LREC2016*, 2016, pp.2204–2208.

[6] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, “Character aware neural language models,” in *Proc. 30th AAAI Conf. on Artificial Intelligence*, 2016.

[7] C. D. Santos and B. Zadrozny, “Learning character-level representations for part-of-speech tagging,” In T. Jebara and E. P. Xing, editors, *Proc. ICML-14*, 2014, pp. 1818–1826.

[8] C. D. Santos and V. Guimar aes, “Boosting named entity recognition with neural character embeddings,” in *Proc. Fifth Named Entity Workshop*, ACL, 2015, pp. 25–33.

[9] M. Ballesteros, C. Dyer, and N. A. Smith, “Improved transition-based parsing by modeling characters instead of words with lstms,” in *Proc. 2015 Conf. on Empirical Methods in Natural Language Processing*, ACL, 2015, pp. 349–359.

[10] X. Chen, L. Xu, Z. Liu, M. Sun, and H. B. Luan, “Joint learning of character and word embeddings,” In Q. Yang and M. Wooldridge, editors, *IJCAI*, AAAI Press, 2015, pp. 1236–1242.

[11] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” *CoRR*, abs/1508.07909, 2015.

[12] Y. Li, W. Li, F. Sun, and S. Li, “Component-enhanced Chinese character Embeddings,” *EMNLP*, ACL, 2015, pp. 829–834.

[13] G. Klein, Y. Kim, Y. Deng, J. Senellart, A. M. Rush, “OpenNMT: open-source toolkit for neural machine translation,” *EACL (Software Demonstrations)*, ACL, 2017, pp. 65–68.