セミマルコフ CRF 自己符号化器による教師なし単語分割

和田 有輝也　村脇 有吾　黒橋 禎夫
京都大学大学院情報学研究科
{wada, murawaki, kuro}@nlp.ist.i.kyoto-u.ac.jp

概要

従来のニューラル教師なし単語分割手法には，(1) 少量のテキストしか参照しない，(2) 言語モデルに基づくため，単語分割モデルとしては不必要な制限を課しているという 2 つの課題がある．そこで，我々は事前学習済み文字レベル BERT を用いることで大規模テキストを活用する．さらに，入力文全体を見て分割を予測するセミマルコフ CRF 自己符号化器モデルを提案する．日本語を対象とした実験の結果，提案手法によって概ね妥当な単語分割が得られることが確認された．

1　はじめに

日本語や中国語等の明示的な単語の境界をもたない言語における自然言語処理では，前処理として単語分割が重要な技術となっている．単語分割の訓練手法としては教師あり学習が広く用いられている．しかし，その学習には大量のアノテーションデータが必要であり，言語リソースの少ない言語やドメインに適用することが難しい．そこで，アノテーションデータセットのない言語やドメインにも適用可能な教師なし学習による単語分割手法が求められている．

教師なし単語分割は，かつてはベイズ推定に基づく手法 [1, 2] が盛んに研究されていた．近年ではニューラルネットワークに基づく手法 [3, 4, 5, 6] が研究されているが，その課題として次の 2 点が挙げられる．第 1 に，モデルが参照するテキストが少量である．従来手法では，訓練データセットが 1–2 万文程度と小さく，かつ，事前学習済みモデルも採用しておらず，ニューラルネットワークを用いた複雑なモデルの学習には不十分であると考えられる．第 2 に，いずれのモデルも単語分割モデルであると同時に言語モデルでもあるため，分割の予測をする際に入力文の一部を適宜隠す必要がある．この制約は言語モデルでは必要だが，単語分割モデルでは不要である．本研究では 1 つ目の課題を解決するために，事前学習済み文字レベル BERT による入力文の埋め込みを行う．また，2 つ目の課題を解決する教師なしニューラル単語分割モデルとして，セミマルコフ CRF 自己符号化器を提案する．

2　関連研究

ニューラルネットワークに基づく教師なし単語分割手法としては，Segmental Language Model (SLM) [3] がある．SLM は入力文をエンコードする文脈エンコーダと単語の生成確率を計算するセグメントデコーダの 2 つの LSTM からなるモデルである．入力文における全ての部分文字列の生成確率を計算し，それらから最適な分割を求める．他の教師なしニューラル単語分割の手法としては，SLM に語彙メモリモジュールや単語長さについての正則化項を追加した手法 [4] や LSTM を双方向に拡張した手法 [5]，エンコーダとして LSTM の代わりに Transformer を用いた Masked Segmental Language Model (MSLM) [6] が提案されている．しかし，従来手法はいずれも 1 章で述べた 2 つの課題を有している．本研究ではそれらの課題を解決する教師なし単語分割手法を提案する．

3　提案手法

本研究では事前学習済み文字レベル BERT による入力文の埋め込みを行い，入力文全体を見て分割を予測するセミマルコフ CRF 自己符号化器モデルを用いる教師なし単語分割手法を提案する．提案手法の概要を図 1 に示す．

3.1　文字レベル BERT による埋め込み

大規模なデータセットで事前学習した文字レベル言語モデルは潜在的に単語についての情報を有していると考えられる．実際に Wang らは中国語を対象とした文字レベル BERT が潜在的に単語についての情報を有していることを示している [7]．そこ

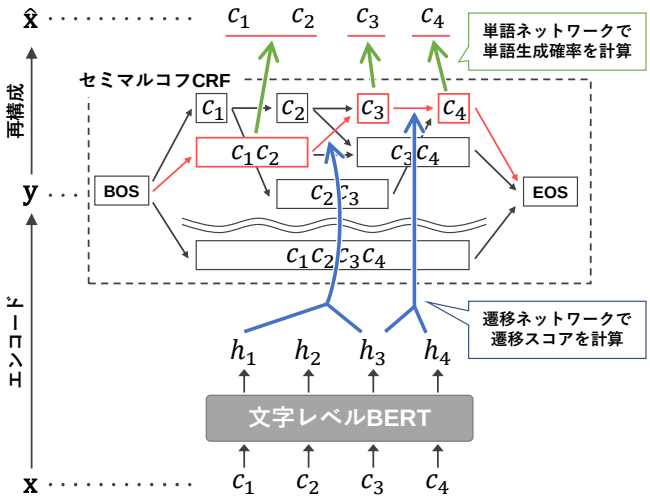

図 1　提案手法の概要

で，入力文 $\mathbf{x} = c_1, \cdots, c_n$ を事前学習済み文字レベル BERT を用いて埋め込み表現 $\mathbf{h} = h_1, \cdots, h_n$ に変換する．ここで，c_i は単一の文字，h_i は文字 c_i に対応する埋め込み表現であり，n は文 $\mathbf{x}$ の文字数である．この埋め込み表現を用いてモデルを学習することで，文字レベル BERT が潜在的に有している単語についての情報が単語分割の学習に活用されることが期待される．

3.2　セミマルコフ CRF 自己符号化器

セミマルコフ CRF 自己符号化器は CRF 自己符号化器モデル [8] における CRF をセミマルコフ CRF に拡張することで，単語分割タスクに適用したものである．訓練時には，入力文 $\mathbf{x}$ から可能な全ての分割 $\mathbf{y} = \mathbf{w}_1, \cdots, \mathbf{w}_m$ について分割の生成確率 $p(\mathbf{y}|\mathbf{x})$ を計算し，各 $\mathbf{y}$ から元の文 $\hat{\mathbf{x}}$ が生成される確率 $p(\hat{\mathbf{x}}|\mathbf{y})$ を計算する．予測時には，復元確率 $p(\hat{\mathbf{x}}, \mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})p(\hat{\mathbf{x}}|\mathbf{y})$ が最も高くなる $\mathbf{y}$ を選択する．ただし，m は分割 $\mathbf{y}$ における単語数である．

セミマルコフ CRF 自己符号化器は連続する 2 単語間の遷移スコアを計算する遷移ネットワークと単語の生成確率を計算する単語ネットワークの 2 つのモジュールによって構成される．ある分割 $\mathbf{y}$ の遷移スコア $\phi(\mathbf{x}, \mathbf{y})$ を $\mathbf{y}$ における全ての連続する 2 単語間の遷移スコアの総和

$$\phi(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{m} \tilde{\phi}(\mathbf{w}_i, \mathbf{w}_{i+1}) \tag{1}$$

で定義し，分割生成確率 $p(\mathbf{y}|\mathbf{x})$ は

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp(\phi(\mathbf{x}, \mathbf{y}))}{Z} \tag{2}$$

とする．ただし，$\mathbf{w}_0, \mathbf{w}_{m+1}$ はそれぞれ文の始端，終端を表すシンボル [BOS], [EOS]，Z は正規化定数である．また，$\mathbf{y}$ から元の文 $\hat{\mathbf{x}}$ が生成される確率 $p(\hat{\mathbf{x}}|\mathbf{y})$ を分割 $\mathbf{y}$ に含まれる単語の生成確率 $p_{\mathrm{w}}(\mathbf{w}_i|\mathbf{y})$ を用いて，

$$p(\hat{\mathbf{x}}|\mathbf{y}) = \prod_{i=1}^{m} p_{\mathrm{w}}(\mathbf{w}_i|\mathbf{y}) \tag{3}$$

で定義する．

遷移ネットワークは 3 層の MLP，単語ネットワークは前向き LSTM と後ろ向き LSTM の 2 つからなるネットワークでそれぞれ構成する．

3.2.1　遷移スコア

連続する 2 単語間の遷移スコア $\tilde{\phi}(\mathbf{w}_i, \mathbf{w}_{i+1})$ は $\mathbf{w}_i, \mathbf{w}_{i+1}$ のそれぞれの先頭の文字に対応する埋め込みベクトル $h_{i,1}, h_{i+1,1}$ を結合したベクトルを遷移ネットワークに入力して得られる値であるとする．なお，シンボル [BOS], [EOS] に対応する埋め込みはそれぞれモデルパラメータとして学習する．この遷移スコアの定義は，学習の結果として，真の単語について BERT のもつ潜在的情報がその先頭の文字に対応する埋め込みベクトルに現れるようになることを期待するものである．

3.2.2　単語生成確率

分割 $\mathbf{y}$ における単語 $\mathbf{w}_i = c_{i,1}, \cdots, c_{i,k}$ の生成確率 $p_{\mathrm{w}}(\mathbf{w}_i|\mathbf{y})$ を

$$p_{\mathrm{w}}(\mathbf{w}_i|\mathbf{y}) = \frac{p_{\mathrm{f}}(\mathbf{w}_i|\mathbf{y}) + p_{\mathrm{b}}(\mathbf{w}_i|\mathbf{y})}{2} \tag{4}$$

とする．ここで $p_{\mathrm{f}}(\mathbf{w}_i|\mathbf{y})$ は [BOW], $c_{i,1}, \cdots, c_{i,k-1}$ を入力，先頭の文字 $c_{i,1}$ に対応する埋め込みである $h_{i,1}$ を初期隠れ状態としたときの前向き LSTM の $\mathbf{w}_i$ の生成確率であり，$p_{\mathrm{b}}(\mathbf{w}_i|\mathbf{y})$ は [BOW], $c_{i,k-1}, \cdots, c_{i,1}$ を入力，末尾の文字 $c_{i,k-1}$ に対応する埋め込みである $h_{i,k-1}$ を初期隠れ状態としたときの後ろ向き LSTM の $\overleftarrow{\mathbf{w}_i}$ の生成確率である．なお，[BOW] は単語の始端を表すシンボルであり，$\overleftarrow{\mathbf{w}_i}$ は単語 $\mathbf{w}_i$ を反転させた文字列である．

3.2.3　目的関数

学習では入力文 $\mathbf{x}$ について可能な全ての分割に対する復元確率の和

$$p(\hat{\mathbf{x}}|\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})p(\hat{\mathbf{x}}|\mathbf{y}) \tag{5}$$

$$= \frac{\sum_{\mathbf{y}} \exp\left\{\phi(\mathbf{x}, \mathbf{y}) + \sum_{i=1}^{m} \log p_{\mathrm{w}}(\mathbf{w}_i|\mathbf{y})\right\}}{\sum_{\mathbf{y}} \exp\left\{\phi(\mathbf{x}, \mathbf{y})\right\}} \tag{6}$$

の負の対数尤度 $-\log p(\hat{\mathbf{x}}|\mathbf{x})$ を目的関数とし，これを最小化する．この目的関数の最小化は $\mathbf{x}$ の可能な全ての分割について，尤度を向上させるようにモデルパラメータを変化させることに相当する．この過程で単語ではない部分文字列に対する単語生成確率や遷移確率も向上する．しかし，単語である部分文字列はそうでない部分文字列に比べて頻出すると考えられるため，最終的には単語である部分文字列に対する単語生成確率や遷移確率が大きい状態で収束することが期待される．その結果として，正しい分割が予測できるようになると考えられる．

3.2.4 動的計画法による学習・推論の高速化

長さ n の文字列 $\mathbf{x}$ において可能な分割 $\mathbf{y}$ は 2^{n-1} 通りだけ存在する．そのため，全ての $\mathbf{y}$ について $p(\mathbf{y}|\mathbf{x})$ を求めてから $p(\hat{\mathbf{x}}|\mathbf{x})$ を計算する方法では，学習に指数時間を要する．そこで，動的計画法を用いることで多項式時間での学習を実現する．具体的には次のように $p(\hat{\mathbf{x}}|\mathbf{x})$ を求める．まず，文 $\mathbf{x}$ の i 文字目から j 文字目までの部分文字列を $\mathbf{x}_{i:j}$，文 $\mathbf{x}_{i:j}$ について可能な分割のうち，末尾の単語の長さが k のものを $\mathbf{y}^k_{i:j}$ とする．ここで，n 以下の自然数 t と末尾の単語の長さ k について次の関数を定義する．

$$f_D(t,k) = \sum_{\mathbf{y}^k_{1:t}} \exp\left\{\phi(\mathbf{x}_{1:t}, \mathbf{y}^k_{1:t}) + \sum_{\mathbf{w}\in\mathbf{y}^k_{1:t}} \log p_{\mathrm{w}}(\mathbf{w}|\mathbf{y}^k_{1:t})\right\} \quad (7)$$

$$f_N(t,k) = \sum_{\mathbf{y}^k_{1:t}} \exp\left\{\phi(\mathbf{x}_{1:t}, \mathbf{y}^k_{1:t})\right\} \quad (8)$$

関数 f_D, f_N は文 $\mathbf{x}$ の始めの t 文字である $\mathbf{x}_{1:t}$ についての可能な分割のうち，末尾の単語の長さが k である分割のみを対象とした場合の式 (6) の分子・分母に相当する．すなわち，

$$p(\hat{\mathbf{x}}|\mathbf{x}) = \frac{\sum_k f_D(n,k)}{\sum_k f_N(n,k)} \quad (9)$$

である．関数 f_D, f_N についての更新式をそれぞれ

$$f_D(t,k) = \sum_l \exp\left\{\log f_D(t-k,l) + \tilde{\phi}(\mathbf{w}^{\mathrm{last}}_{t-k,l}, \mathbf{w}^{\mathrm{last}}_{t,k}) + \log p_{\mathrm{w}}(\mathbf{w}^{\mathrm{last}}_{t,k}|\mathbf{y}^k_{1:t})\right\} \quad (10)$$

$$f_N(t,k) = \sum_l \exp\left\{\log f_N(t-k,l) + \tilde{\phi}(\mathbf{w}^{\mathrm{last}}_{t-k,l}, \mathbf{w}^{\mathrm{last}}_{t,k})\right\} \quad (11)$$

とする．ただし，$\mathbf{w}^{\mathrm{last}}_{t,k}$ は文 $\mathbf{x}_{1:t}$ における長さ k の末尾の単語であり，$f_D(0,0) = 1, f_N(0,0) = 1$ とする．これらの更新式に従って $f_D(n,k), f_N(n,k)$ を求め，$p(\hat{\mathbf{x}}|\mathbf{x})$ を計算することで多項式時間での学習を実現する．推論において $p(\hat{\mathbf{x}}|\mathbf{x})$ を最大化する $\mathbf{y}$ を求める際も以上の更新式で総和を取るところを最大値を取るように変更することで，同様に多項式時間で計算する．

3.2.5 モデルの学習

モデル全体を学習する前に，事前学習として BERT 部分を固定して 1 エポックだけ学習する．これは事前学習済み BERT のもつ情報が，事前学習されていない遷移ネットワーク及び単語ネットワークを含むモデル全体の最適化によって忘却されるのを抑止するための工夫である．

また，予測する単語の長さを最大 K 文字とする制限を設ける．これは任意の長さの単語を認めた場合に考えられる，訓練データセットに現れる文の一つ一つが単語であると学習するような局所解に陥ることを避けるためのものである．

4 評価実験

提案手法による単語分割の精度を評価するために日本語を対象として実験を行った．ここでは，実験設定について述べた後，実験結果及びそれに対する考察を述べる．

4.1 実験設定

データセットは京都大学ウェブ文書リードコーパス (KWDLC) [1] を用いた．KWDLC は日本語ウェブ文書を対象としたコーパスであり，その訓練データ，検証データ，テストデータはそれぞれ 12,271 件，1,585 件，2,195 件である．事前学習済み文字レベル BERT は日本語 Wilipedia 全体で学習した BERT-base-Japanese-char[2] を用いた．予測する単語の長さの最大値 K は 5 とした．KWDLC のテストデータにおける単語の約 99%が 5 文字以下であることを踏まえると，この制約のために分割精度が大きく損なわれる可能性は小さいと考えられる．

表 1 提案手法による単語分割の定量的評価．モデル全体の学習についてのエポック数を 5,10,15,20 としたそれぞれの場合における結果を示す．

エポック数	F	P	R
5	34.0	40.7	29.1
10	**63.5**	**61.8**	**65.3**
15	57.2	53.7	61.2
20	56.1	51.6	61.4

1) https://github.com/ku-nlp/KWDLC
2) https://huggingface.co/cl-tohoku/bert-base-japanese-char

表 2　提案手法（エポック数を 10 とした場合）で得られた予測の例

正解	いずれ / も / プロジェクト / は / 、 / 個人 / ベース / で / 行わ / れる / 。
予測	い / ず / れ / も / プロ / ジェクト / は、 / 個人 / ベース / で / 行わ / れる / 。
正解	ローズ / の / 花 / が / １０００ / 輪 / 分 / 、 / 入って / い / ます / 。
予測	ローズ / の / 花が / １０００ / 輪 / 分、 / 入って / います / 。

学習のための最適化アルゴリズムは AdamW を用いた．また，AdamW のハイパーパラメータとして $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}, \lambda = 0.01$ を用い，学習率は 2×10^{-5} とした．モデル全体の学習についてのエポック数を 5,10,15,20 としたそれぞれの場合について実験を行った．なお，いずれの場合においても，BERT 部分を固定した事前学習を 1 エポックだけ実施した．また，学習全体を通して BERT の入力側 8 層は固定した．

4.2　実験結果

表 1 にモデル全体の学習についてのエポック数を 5,10,15,20 としたそれぞれの場合におけるテストデータに対する分割結果の F 値，P 値，R 値を示す．エポック数が 10 の場合において F 値が最大 (63.5) となった．表 2 にエポック数 10 の場合のモデルで得られた予測の例を示す．いずれの例も，いくつかエラーが見られるものの概ね妥当な分割であるといえる．エラーの例としては読点（“、”）を直前の単語と結合しており，一つの単語として判別できていないことが挙げられる．後処理でモデルの予測に関わらず読点を一つの単語として見なすようにすると，テストデータに対する F 値は 68.6 に増加した．このことから，このモデルによる予測の多くで読点についてのエラーが生じていることがわかる．

エポック数を 15,20 とした場合では，エポック数を 10 とした場合に比べて F 値が減少している．それぞれの場合において予測された単語の長さの分布（図 2）から，学習が進むにつれて分割が細かくなっていくことが確認された．これがエポック数 15,20 の場合に分割精度が低下した大きな原因であると考えられる．ここで，妥当な分割が得られる状態で学習が収束せずに分割が過剰に細かくなっていく要因について考察する．提案手法では分割 $\mathbf{y}$ をボトルネックとすることで，BERT の埋め込みベクトル h_i が単語についての抽象的な表現を獲得することを期待している．しかし，実際にはモデルが十分に柔軟であるために h_i が入力文字についての具体的な表現を獲得，すなわち対応する文字 c_i の情報をそのまま有するように学習が進んでいる可能性がある．ベクトル h_i が c_i の情報をそのまま有していると，h_i を初期隠れ状態として与えられる前向き・後ろ向き単語ネットワークにとって，予測対象の文字列の先頭・末尾の文字である c_i の予測は他の文字の予測に比べて非常に容易であるというバイアスが生じる．この結果として，学習が進むにつれて長さ 1 の単語の生成確率が過大に評価されるようになり，分割が過剰に細かくなっていくのだと考えられる．

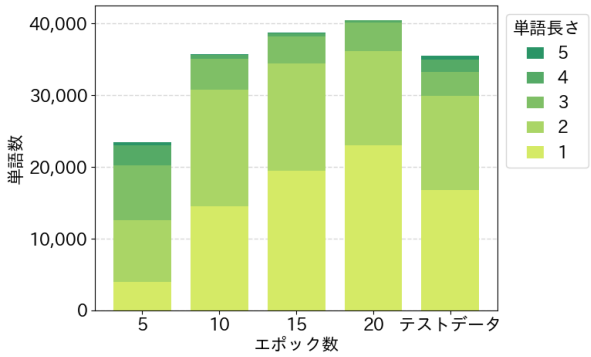

図 2　エポック数と予測単語長さの分布の関係．左から順にエポック数 5,10,15,20 の場合のモデルが予測した単語の長さの分布であり，右端はテストデータの真の単語の長さの分布である．なお，6 文字以上の単語については省略した．学習が進むほど長さ 1 の単語が占める割合が増加し，分割が過剰に細かくなっていることが確認できる．

5　おわりに

本研究ではモデルが参照するテキストの量を増加させるために事前学習モデルとして文字レベル BERT を導入し，また，入力文全体を見て分割を予測するためにセミマルコフ CRF 自己符号化器を設計した．提案手法により F 値 63.5 の概ね妥当な単語分割が得られた．

今後の課題としては，中国語等の他言語への適用，さらなる分割精度向上のための訓練データセットの大規模化がある．また，実験で明らかになった読点についてのエラーや分割が過剰になっていく問題について調査し，改善を図りたい．さらに，単語分割だけでなく品詞クラスタリングも同時に行うモデルの設計も検討したい．

参考文献

[1] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. A bayesian framework for word segmentation: Exploring the effects of context. **Cognition**, Vol. 112, No. 1, pp. 21–54, 2009.

[2] Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In **Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP**, pp. 100–108, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[3] Zhiqing Sun and Zhi-Hong Deng. Unsupervised neural word segmentation for Chinese via segmental language modeling. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, October-November 2018.

[4] Kazuya Kawakami, Chris Dyer, and Phil Blunsom. Learning to discover, ground and use words with segmental neural language models. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 6429–6441, July 2019.

[5] Lihao Wang, Zongyi Li, and Xiaoqing Zheng. Unsupervised word segmentation with bi-directional neural language model. **arXiv preprint arXiv:2103.01421**, 2021.

[6] CM Downey, Fei Xia, Gina-Anne Levow, and Shane Steinert-Threlkeld. A masked segmental language model for unsupervised natural language segmentation. **arXiv preprint arXiv:2104.07829**, 2021.

[7] Yile Wang, Leyang Cui, and Yue Zhang. Does Chinese BERT encode word structure? In **Proceedings of the 28th International Conference on Computational Linguistics**, December 2020.

[8] Waleed Ammar, Chris Dyer, and Noah A Smith. Conditional random field autoencoders for unsupervised structured prediction. In **Advances in Neural Information Processing Systems**, Vol. 27. Curran Associates, Inc., 2014.