ラティス構造を学習するニューラル単語分割

Neural Word Segmentation by Learning Lattice Structures

山口 修平　山根 丈亮　三輪 誠　佐々木 裕
Shuhei Yamaguchi　Josuke Yamane　Makoto Miwa　Yutaka Sasaki
豊田工業大学
Toyota Technological Institute
{sd13093, sd16432, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 はじめに

単語分割は自然言語処理においてしばしば最初に行われる処理であるため，単語分割を正確に行うことは重要である．日本語における単語分割の伝統的な手法（MeCab, JUMAN など）は，辞書を用いて図 1 のような単語のラティス構造を構築し，ラティス構造から単語の境界を決定する手法 [1, 2, 3] である．この手法は，辞書に存在する単語のみで構成された文については正しいラティス構造を構築できるため，高い精度で単語分割できる．しかし，辞書に存在しない単語（未知語）を含む文ではラティス構造を作成する段階で誤りを含んでしまう．一方で，文中の文字一つ一つに単語の境界に関するラベルを付与したコーパスを教師とし，文字ごとに単語の境界を予測する手法が提案されている [4]．この手法は辞書情報を用いていないため，辞書に依存しない単語分割が可能だが，辞書を利用した手法に比べて精度が低い．文字ごとに単語の境界を予測する手法として，ニューラルネットを用いたニューラル単語分割も提案されている [5]．そこで本研究では，辞書に依存しない単語分割の精度向上を目的として，ニューラル単語分割に辞書情報を教師として追加することで，辞書の情報を取り入れつつ，辞書に依存しない単語分割手法を提案する．

2 関連研究

本章では本研究の対象とする日本語単語分割と本研究の単語分割手法の基盤となる Long short-term memory（LSTM）について説明する．

2.1 日本語単語分割

日本語単語分割は系列ラベリング問題として扱われ，単語分割の可能性を辞書によってラティス構造で表現する手法 [1, 2, 3] と文字ごとの単語の境界に関するラベルを学習する手法 [4, 5] に分けられる．

2.1.1 ラティスを作成する単語分割

ラティス構造とは日本語文中に存在する辞書内の単語を列挙し，それらを繋げることで単語分割の可能性を表す．ラティス構造から単語の境界を決定する手法には，ルールに従って行うもの [2] や条件付き確率場 [1] を利用するものなどがある．これらの手法では辞書を用いてラティス構造を機械的に作成するため，未知語を含む文に対して正しくラティス構造を作成できない．

2.1.2 文字ごとの単語分割

ラティス構造を作成しない手法として，日本語文中の文字に対して単語の境界に関するラベルを学習する手法が提案されている [4, 5]. 文字ごとにラベルを学習する手法では辞書を用いておらず，ラティス構造を作成する手法と比べて，新語などの未知語に対しても自然な単語分割をできる可能

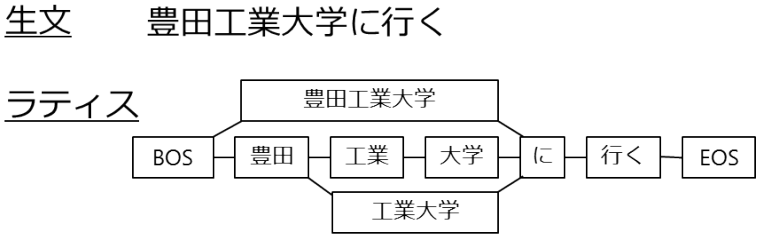

図 1　ラティスの例

性がある．北川らはこの文字ごとの単語分割を対象に，リカレントニューラルネットワークを用いたニューラル単語分割を提案している [5].

2.2　Long short-term memory (LSTM)

LSTM は系列を扱うことができるリカレントニューラルネットワークの 1 つで，長い系列のデータを考慮した出力が可能である．ある時点 t における入力を $\mathbf{x_t}$，出力を $\mathbf{h_t}$，LSTM のセルを $\mathbf{c_t}$ とすると，以下のように表される．

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i) \quad (1)$$
$$\mathbf{g}_t = \tanh(\mathbf{W}_{xg}\mathbf{x}_t + \mathbf{W}_{hg}\mathbf{h}_{t-1} + \mathbf{b}_g) \quad (2)$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f) \quad (3)$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o) \quad (4)$$
$$\mathbf{c}_t = \mathbf{c}_{t-1} \odot \mathbf{f}_t + \mathbf{i}_t \odot \mathbf{g}_t \quad (5)$$
$$\mathbf{h}_t = \tanh(\mathbf{c}_t) \odot \mathbf{o}_t \quad (6)$$

ここで，$\odot$ はアダマール積，$\sigma(x)$ は x に関するシグモイド関数，$\mathbf{i}_t, \mathbf{g}_t, \mathbf{f}_t, \mathbf{o}_t$ は，それぞれ時点 t における入力判断，入力，忘却判断，出力判断ゲートの値，$\mathbf{W}_{jk}$ は，j から k に対する重み行列，$\mathbf{b}_j$ は j のゲートにおけるバイアスを表す．

3　提案手法

本研究では未知語にも対応することができる文字ベースのニューラル単語分割に辞書情報を追加することで，未知語と辞書の両方を考慮した手法を提案する．単語分割はどの文字が単語の境界かを予測する問題であり，文字ごとに BMES ラベル（表 1）を予測する系列ラベリング問題とみなすことができる．本手法はラティス構造に対する BMES ラベルの学習（3.1 節）と単語の境界に対する BMES ラベルの学習（3.2 節）から成る．

3.1　ラティス構造に対する学習

ラティス構造に対する学習では，ラティス構造から限定される文字ごとのラベルを学習することで分かち書きコーパスにない単語の情報を取り入れる．ラティス構造から限定されるラベルとは，例えば図 2 に示すように，文字「豊」に対するラベルはラティス構造からラベル「$\mathbf{M}, \mathbf{E}$」にはなり得ず，ラベル「$\mathbf{B}, \mathbf{S}$」に限定されることである．この情報をラベル毎に 2 値分類をすることでラティス構造を学習する．図 2 のようなラティス構造学習のための BMES ラベルは，辞書と分かち書きされていない大量の日本語文を用いて生成される．文中の文字列が辞書に含まれていれば，それに対応する BMES のラベルを付与し，それを教師としてラティス構造を LSTM で学習する．ラティス構造に対する学習の目的関数は次式のように設定する．

$$\text{loss} = \sum_{\ell \in \{\mathbf{B},\mathbf{M},\mathbf{E},\mathbf{S}\}} \sum_{\text{c}} -\mathbf{t}_{\ell,\text{c}} \log \mathbf{y}_{\ell,\text{c}} \quad (7)$$

$\mathbf{t}_{\ell,c}$ はラベル ℓ に対する文字 c の正解分布，$\mathbf{y}_{\ell,c}$ は LSTM の出力に 2 値分類のソフトマックス関数を適用した予測分布である．

3.2　単語の境界に対する学習

単語の境界に対する学習では，図 3 のように分かち書きされたコーパスの各文字た対する BMES ラベルを LSTM によって学習する．目的関数は次式のように設定する．

$$\text{loss} = \sum_{\text{c}} -\mathbf{t}_{\text{c}} \log \mathbf{y}_{\text{c}} \quad (8)$$

$\mathbf{t}_c$ は文字 c の正解ラベル分布，$\mathbf{y}_c$ は文字 c について LSTM の出力に 4 値分類のソフトマックス関数を適用を適用した予測ラベル分布である．

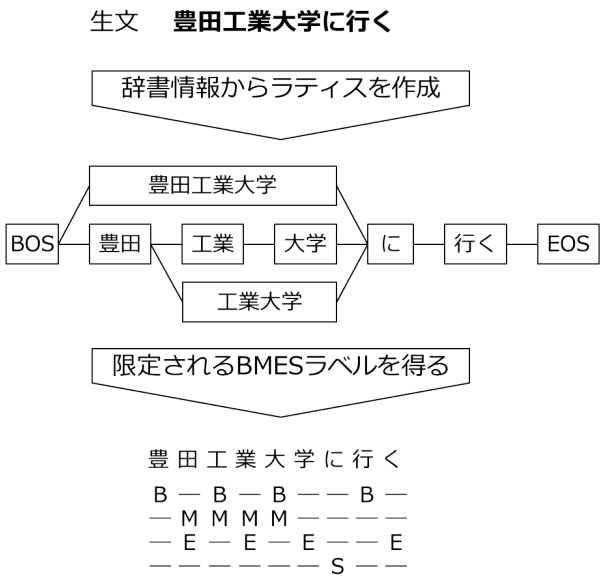

図 2　ラティス構造学習のためのラベル

3.3　ラティス構造を学習するニューラル単語分割

本手法では図 4 に示すように，3.1 節で述べたラティス構造に対する学習を行った後，3.2 節で述べた単語の境界に対する学習を行う．ラティス構造は辞書情報を含んでいるので，ラティス構造を学習した本手法は辞書にある単語に対して高精度に単語分割ができ，さらに単語の境界に対する学習を行うことで未知語にも対応することができると考えられる．

表 1　文字に対するラベルの意味

ラベル	意味
B	単語の始まりの文字
M	単語の途中の文字
E	単語の終わりの文字
S	一文字からなる単語

表 2　既存手法と提案手法の F 値での比較 [%]

モデル	テキスト	ウェブ
JUMAN	97.18	97.20
ラティス学習前	97.88	92.81
ラティス学習後	98.28	93.55

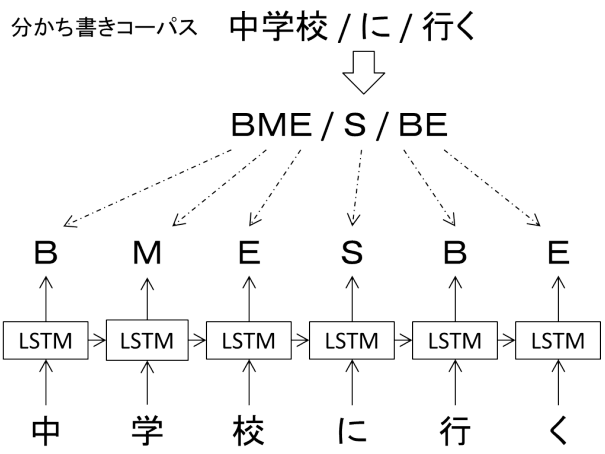

図 3　LSTM による単語境界ラベリング

4　実験設定

提案手法の辞書による学習が教師データに出現しない単語や表現にどのような影響を与えるかを調べるため，教師に用いたコーパスと同じコーパスでテストした場合とそうでない場合で比較する．単語の境界ラベル（分かち書き）を学習する際には京都大学テキストコーパスを使用し，テストに京都大学テキストコーパス（テキスト）と京都ウェブコーパス（ウェブ）をそれぞれ使用する．ラティスを学習しない手法は京都大学テキストコーパスのみを教師データとし，ラティスを学習する手法は事前にラティスの学習を行い，その後単語の境界を学習する．分かち書き教師データは約 35,000 文，テストデータは 2,000 文を用いた．ラティスを学習する手法の学習データには，辞書を mecab-ipadic と Neologd [6] として，ウィキペディア日本語版約 1,600,000 文を用いた．文字ベクトルの次元数は 100，LSTM 層の出力の次元数は 300 とした，実装には Tensorflow version 0.8.0 を用いた．

5　結果と考察

テストデータで評価した結果を表 2 に示す．ラティスを学習する後と前では学習後のほうが単語分割の F 値が大きくなった．既存手法に対して，分かち書き教師データとして利用していないウェブコーパスでは低い値となった．

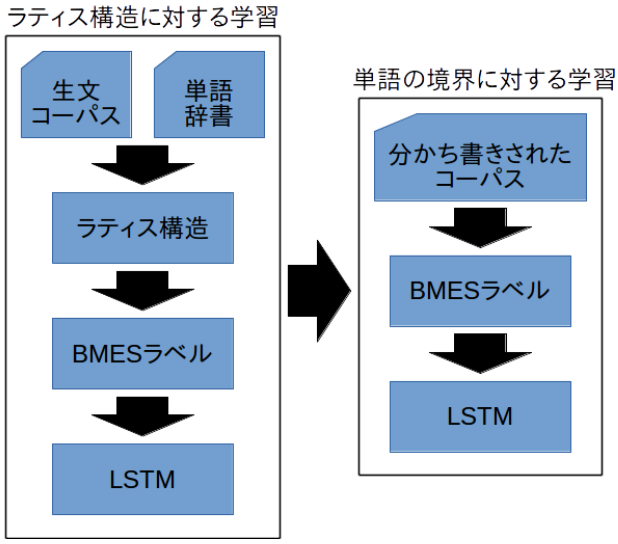

図 4　提案手法のモデル

表 2 からラティス構造を学習する手法はテキストコーパスとウェブコーパス両方で F 値が上がったため，ラティス構造を学習することで単語分割に有益な情報を得ることが出来ることがわかった．ウェブコーパスで評価した際 F 値が下がったが，ラティス学習前に比べ減少が少ないため辞書の情報を学習できていると考えられる．一方で，Juman に比べるとウェブコーパスの精度は低く，その要因としては利用したウィキペディア日本語版のコーパスが全体の 1 割ほどと小さく，辞書の情報を網羅出来なかったことが考えられる．

6　おわりに

辞書に依存しない単語分割の精度向上を目的として，ニューラル単語分割に辞書情報を教師として追加することで，辞書の情報を取り入れつつ，辞書に依存しない単語分割手法を提案した．結果として，辞書の情報を学習することで学習前より高い精度で単語分割をすることが出来た．今後はラティス構造を学習する別の方法を検討したい．

参考文献

[1] Taku Kudo et al. Applying conditional random fields to japanese morphological analysis. In *Proc. of EMNLP*, volume 4, pages 230–237, 2004.

[2] Sadao Kurohasi and Daisuke Kawahara. Japanese morphological analysis system juman 7.0 users manual, 2014. http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN (2016 年 8 月 4 日閲覧).

[3] Hajime Morita et al. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proc. of EMNLP*, pages 2292–2297, 2015.

[4] Graham Neubig et al. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proc. of ACL*, pages 529–533, 2011.

[5] 北川善彬ら．深層ニューラルネットワークを利用した日本語単語分割．言語処理学会第 22 回年次大会発表論文集, pages 933–936, 2015.

[6] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.