シンボルグラウンディングによる分野特有の単語分割の精度向上

友利 涼[†]・亀甲 博貴[††]・二宮 崇[†††]・森 信介[††††]・鶴岡 慶雅[††]

本稿は, 自動単語分割における精度向上を実現するために, 非テキスト情報とその説明文に対するシンボルグラウンディングを用いた新しい単語分割法を提案する. 本手法は, 説明文が付与された非テキスト情報の存在を仮定しており, 説明文を擬似確率的単語分割コーパスとすることで, 非テキスト情報と分野固有の単語との関係をニューラルネットワークにより学習する. 学習されたニューラルネットワークから分野固有の辞書を獲得し, 得られた辞書を単語分割のための素性として用いることでより精度の高い自動単語分割を実現する. 将棋局面が対応付けされた将棋解説文から成る将棋解説コーパスを用いて実験を行い, シンボルグラウンディングにより得られた辞書を用いることで単語分割の精度が向上することが確認できた.

キーワード：シンボルグラウンディング, 単語分割, 辞書

Improvement in Domain Specific Word Segmentation by Symbol Grounding

SUZUSHI TOMORI[†], HIROTAKA KAMEKO[††], TAKASHI NINOMIYA[†††], SHINSUKE MORI[††††] and YOSHIMASA TSURUOKA[††]

We propose a novel framework for improving a word segmenter using information acquired from symbol grounding. The framework uses a dataset consisting of pairs of non-textual information and a commentary. We generate a pseudo-stochastically segmented corpus from the commentaries, and then build a neural network to predict relationships between non-textual information and the words. We generate a domain specific term dictionary by using the neural network for word segmenter. We applied our method to game records of Japanese chess with commentaries. The experimental results show that the accuracy of a word segmenter can be improved by incorporating the generated dictionary.

Key Words: *symbol grounding, word segmentation, dictionary*

[†] 京都大学 大学院情報学研究科, Graduate School of Informatics, Kyoto University
[††] 東京大学 工学系研究科, Graduate School of Engineering, The University of Tokyo
[†††] 愛媛大学 大学院理工学研究科 電子情報工学専攻, Graduate School of Science and Engineering, Ehime University
[††††] 京都大学 学術情報メディアセンター, Academic Center for Computing and Media Studies, Kyoto University

自然言語処理 Vol. 13 No. 2 April 2006

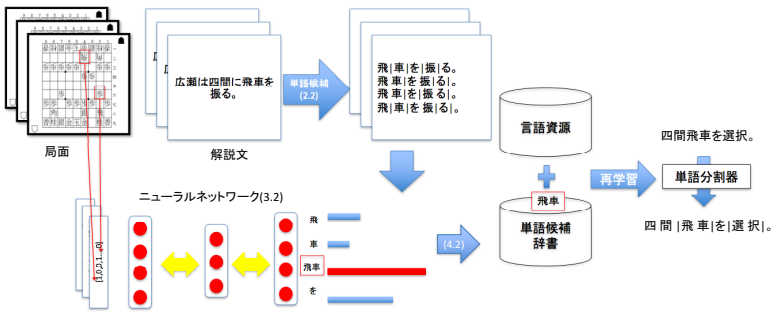

図 1 提案手法の概観

1 はじめに

近年, インターネットなどからテキストとそれに紐づけられた非テキスト情報を大量に得ることができ, 画像とそのキャプションや経済の解説記事とその株価チャートなどは web などから比較的容易に入手することができる. しかし, テキストと非テキスト情報を対応させる研究の多くは, 画像から自然言語を出力する手法 (Farhadi, Hejrati, Sadeghi, Young, Rashtchian, Hockenmaier, and Forsyth 2010)(Yang, Teo, Daumé, and Aloimonos 2011)(Rohrbach, Qiu, Titov, Thater, Pinkal, and Schiele 2013) のように非テキスト情報から自然言語を出力することを目的としている. Kiros らは非テキスト情報を用いることにより言語モデルの性能向上を示した (Kiros, Salakhutdinov, and Zemel 2014).

本稿では, 非テキスト情報を用いた自動単語分割について述べる. 本稿では, 日本語の単語分割を題材とする. 単語分割は単語の境界が曖昧な言語においてよく用いられる最初の処理であり, 英語では品詞推定と同等に重要な処理である. 情報源として非テキスト情報とテキストが対応したデータが大量に必要になるため, 本研究では将棋のプロの試合から作られた将棋の局面と将棋解説文がペアになったデータ (Mori, Richardson, Ushiku, Sasada, Kameko, and Tsuruoka 2016) を用いて実験を行う. 似た局面からは類似した解説文が生成されると仮定し, 非テキスト情報である将棋の局面からその局面に対応した解説文の部分文字列をニューラルネットワークモデルを用いて予測し, その局面から生成されやすい単語を列挙する. 列挙された単語を辞書に追加することで単語分割の精度を向上させる.

本手法は 3 つのステップから構成される (図 1). まず, 将棋の局面と単語候補を対応させるために生テキストから単語候補を生成する. 単語候補は将棋解説文を擬似確率的分割コーパスを用いて部分単語列に分割することで得られる. 次に, 生成した単語候補と将棋の局面をニューラル

ネットワークを用いて対応させることでシンボルグラウンディングを行う. 最後にシンボルグラウンディングの結果を用いて将棋解説文専用の辞書を生成し, 自動単語分割の手法に取り入れる.

本稿の構成は以下の通りである. まず 2 章で単語の候補を取り出すために確率的単語分割コーパスを用いる手法について述べる. 3 章で将棋解説文と局面が対応しているデータセットのゲーム解説コーパスについて触れ, シンボルグランディングとして単語候補と将棋局面を対応させる手法の説明を行う. 4 章ではベースラインとなる自動単語分割器について述べたあと, 非テキスト情報を用いた単語分割として, シンボルグランディングの結果を用いて辞書を生成し, 単語分割器を構築する手法を述べる. 5 章で実験設定と実験結果の評価と考察を行い, 6 章で本手法と他の単語分割の手法を比較する. 最後に 7 章で本稿をまとめる.

2 確率的単語分割コーパス

本研究では将棋の局面とその言語表現をシンボルグラウンディングの対象とし, 将棋解説文専用の辞書を獲得する. 本章では, 辞書獲得のために用いる確率的単語分割コーパス (Mori and Takuma 2004) について説明する. 確率的単語分割コーパスは文字列の分割境界が確率的に与えられたコーパスであり, 確率的単語分割コーパスを用いることでコーパスに出現する各単語の期待頻度を計算することができる. しかし, 辞書に追加するための単語候補を確率的単語分割コーパスから選択するためには, コーパスに出現するほとんどすべての文字列を単語候補として期待頻度を計算する必要があり, 語彙サイズが非常に大きくなり, 高い計算コストを要する. そのため, 本研究では擬似的な確率的単語分割コーパス (森，小田 2009) を用いる.

2.1 確率的単語分割コーパス

確率的単語分割コーパスは生テキストコーパス C_r (以下, 文字列 $x_1^{n_r}$ として表す) と境界の分割確率 P_i の組み合わせで定義される. ここで P_i はある文字 x_i と x_{i+1} の間に分割境界が存在する確率である. この分割確率は x_i と x_{i+1} の周辺の文字列を参照したロジスティック回帰モデル (Fan, Chang, Hsieh, Wang, and Lin 2008) により推定される. ただし, ここで用いるロジスティック回帰モデルは人手で単語分割したコーパスを用いて学習される. コーパスの最初の文字の前と最後の文字の後は分割確率を 1 とする ($P_0 = P_{n_r} = 1$). 確率的単語分割コーパスで推定される単語の期待頻度 $f_r(\boldsymbol{w})$ は以下で計算される.

$$
\begin{aligned}
f_r(\boldsymbol{w}) &= \sum_{i \in O} P_i \left\{ \prod_{j=1}^{k-1} (1 - P_{i+j}) \right\} P_{i+k} \qquad (1) \\
O &= \{ i \mid x_{i+1}^{i+k} = \boldsymbol{w} \}
\end{aligned}
$$

ここで, O はテキストの単語候補となりうる部分文字列 x_{i+1}^{i+k} における i の集合である.

2.2 擬似確率的単語分割コーパス

確率的単語分割コーパスを用いた単語の期待頻度の推定は非常に高い時間的・空間的計算コストを要する. そのため, 本研究では, 確率的単語分割コーパスから直接単語の期待頻度を推定するのではなく, 擬似確率的単語分割コーパス (森，小田 2009) と呼ばれる具体的に単語分割が付与されたコーパスを用いて単語の期待頻度を推定する.

疑似確率的単語分割コーパスは, 確率的単語分割コーパスにより定義される確率分布に従って単語分割を行うことにより得られる. 具体的には, 確率的単語分割コーパスの各文に対し, その文に与えられた確率分布に従って単語分割を行うことで, 単語分割文の生成を行う. ただし, 同じ文に対して 1 度だけ単語分割文を生成するのではなく, 複数回単語分割を行い, 複数の単語分割文を生成する. この手法はサンプリングの一種であり, より多くの単語分割文を生成することで, より良く確率的単語分割コーパスを近似する. 生成された疑似確率的単語分割コーパスは陽に単語分割がされているため容易に各単語の期待頻度を推定することができる. 次の手続きは, 確率的単語分割コーパスから擬似確率的単語分割コーパスを生成する具体的な手続きを表す.

- For $i = 1$ to $n_r - 1$
  - (1) x_i を出力
  - (2) $0 < p < 1$ となる p をランダムに生成
  - (3) if $p < P_i$: 単語境界を出力
    otherwise: 何も出力しない

上記のプロセスを m 回繰り返し, x_i と x_{i+1} の分割境界の数を m で割ることで P_i の推定値を得ることができる. ここで $m \to \infty$ とすると, 大数の法則より P_i と P_i の推定値の誤差が 0 になることが保証される.

3 シンボルグラウンディング

本稿では, 将棋の局面とその言語表現をシンボルグラウンディングの対象とする. 後述する実験では, ゲーム解説コーパス (Mori et al. 2016) を用いる. 本研究の手法は素性設計を除いて分野特有ではないので, 画像とその説明文の組み合わせ (Regneri, Rohrbach, Wetzel, Thater, Schiele, and Pinkal 2013) など他の種類のデータにも適用可能である.

3.1 ゲーム解説コーパス

将棋は 2 人で行うボードゲームで, お互いに自分の駒を動かしながら相手の玉の駒を取ることを目的とする. 将棋の大きな特徴として, 取った相手の駒は自分の持ち駒として使うことができることや一部の駒は相手の敵陣に入るなど特定の条件を満たした場合に駒を裏返して別の動きに変えることができることが挙げられる.

多くのプロ棋士間での対局は, 他のプロ棋士により指し手の評価やその局面の状況, 次の指し手の予想などが解説されている. ゲーム解説コーパスの各解説文には, 対象とする局面が対応しており, ほとんどの解説文は局面に対するコメントをしているが局面に関係のないコメント（対局者に関する情報など）が少量含まれる.

3.2 グラウンディング

将棋局面 S_i $(i = 1, ..., n)$ とその解説文 C_i の大量のペアを学習セットとし, 将棋局面 S_i は素性ベクトル $f(S_i)$ に変換して用いる. ここで n は学習セットに含まれる局面の数である. まず, C_i から擬似確率的単語分割コーパス C'_i を生成する. C'_i は m 個のコーパス C'_{ij} $(j = 1, ..., m)$ を含んでおり, それぞれのコーパスは同じ解説文から成るが, 異なる単語分割が与えられている (本実験では $m = 4$ とした). 次に将棋局面の素性ベクトル $f(S_i)$ を入力として用いて C'_i の単語を予測するモデルを 3 層のフィードフォワードニューラルネットワークで構築する. 隠れ層の次元数は 100 とし, 活性化関数には標準シグモイド関数を用いる. 出力は d 次元の実数値ベクトル (d は単語候補の総数) であり, 実数値ベクトルのそれぞれの要素はある特定の単語候補に対応しており, 解説文にその単語候補が含まれている確率を出力する. 学習の際には解説文にその単語候補が含まれていれば 1, 含まれていなければ 0 とし, 損失関数として 2 乗誤差を用いる. つまり, 将棋局面からその解説文の単語候補の Bag-of-Words を 3 層のニューラルネットワーク用いて予測することでグラウンディングを行う.

将棋局面の素性はコンピュータ将棋プログラムの激指 (Tsuruoka, Yokoyama, and Chikayama 2002) によるゲーム木探索の素性や結果, 評価の一部を用いた. 本実験では以下の素性を用いた.

a: 将棋の駒の位置

b: 持ち駒

c: **a** と **b** の組み合わせ

d: その他ヒューリスティックな素性

c は, 2 駒関係 (ある 2 つの駒の位置関係) や 3 駒関係などであり, **d** のヒューリスティックな素性の例として, 駒の価値に関する素性や玉の危険度に関する素性などがある. 将棋局面の素性の多くは **a**, **b**, **c** であり, 次元数では 94.7%, 発火数では 87.9% を占めている.

一般のシンボルグラウンディングとは違い, 解説文から作られた擬似確率的単語分割コーパスに出現する単語の候補は, 正しい単語文字列と正しく分割されていない文字列が含まれる. それらの正しく分割されていない文字列は正しい単語文字列よりも出現する確率が低いと推測できる. 似た局面からは同じ文字列が出現しやすいと仮定すると, ニューラルネットワークを用いたモデルでは, それらの局面と正しく分割されていない文字列は強い関係を結ぶことができず, 出力する値は正しい単語文字列よりも小さくなると推測される.

自然言語処理 Vol. 13 No. 2 April 2006

4 シンボルグラウンディングの結果を用いた単語分割

この章ではベースラインとなる自動単語分割とシンボルグラウンディングの結果を用いた単語分割について述べる.

4.1 ベースラインとなる単語分割

さまざまな日本語の単語分割手法や形態素解析手法があるが, 品詞情報なしで新しい単語を追加することができる唯一の単語分割手法である点予測に基づく手法 (Neubig, Nakata, and Mori 2011) を採用する.

点予測による単語分割の入力は分割されてない文字列 $x = x_1, ..., x_{n_r}$ である. この単語分割器はサポートベクターマシン (Fan et al. 2008) を用いて単語境界があれば $P_i = 1$ なければ $P_i = 0$ として境界を推定する. このときの素性は周辺の 6 文字から文字 n-gram と文字種 n-gram ($n = 1, 2, 3$) を用いる. また, もし周辺の文字 n-gram が辞書の文字列と一致した場合にはそれも素性として用いる.

4.2 非テキスト情報を用いた自動単語分割器の学習

非テキスト情報を自動単語分割に用いる最初の試みとして, 非テキスト情報と関連性の高い単語候補を加えた辞書を生成する手法を提案する. 非テキスト情報から単語候補を予測するニューラルネットワークを構築することでシンボルグラウンディングを行う. 構築されたニューラルネットワークを用いることで非テキスト情報と関連する単語候補を取得できる. 例えば将棋の場合, 局面と局面から生成される解説文の単語は強い関連があり, 似た局面からは同じ単語が生成される可能性が高いと考える. つまり, 非テキスト情報と強い関連の単語候補を選ぶことで良い辞書を作ることができる.

辞書生成のために, シンボルグラウンディングの結果として将棋局面 S_i から d 次元の実数値ベクトルを計算し単語候補のスコアを得る. 本稿では, 単語候補のスコアから以下の 3 つの方法で辞書を生成する.

sum すべての局面の d 次元の実数値ベクトルの和をとり, 上位 $R\%$ の単語を辞書に追加する.

max すべての局面の d 次元の実数値ベクトルの要素の最大値を取り, 上位 $R\%$ の単語を辞書に追加する.

each 各局面の d 次元の実数値ベクトルの上位 $R\%$ の単語を辞書に追加する.

例えば, 以下のように局面 S_1, S_2 から計算される単語候補 [四間, 間, 間飛, 飛, 飛車] の 5 次元の実数値ベクトルがあり, その上位 40% の単語候補を辞書に加えるとする.

- S_1 から計算される単語候補のベクトル $[1.4, 1.5, 0.2, 0.5, 3.8]$
- S_2 から計算される単語候補のベクトル $[4.9, 0.8, 0.1, 0.9, 3.2]$

表 1 コーパスの概要

	文数	単語数	文字数	局面数	単語分割
シンボルグラウンディング用コーパス					無
将棋 (局面)	33,151	-	-	33,151	無
訓練コーパス					
BCCWJ-train	56,753	1324,951	1,911,660	0	有
新聞記事	8,164	240,097	361,843	0	有
日常会話文	11,700	147,809	197,941	0	有
開発コーパス					
将棋 (局面)	253	3,898	4,961	137	有
テストコーパス					
BCCWJ-test	6,025	148,929	212,261	0	有
将棋 (局面なし)	3,000	21,261	26,767	0	有
将棋 (局面あり)	1,788	31,220	41,104	928	有

sum では S_1, S_2 から計算される単語候補のベクトルの要素を足しあわせた $[6.3, 2.3, 0.3, 1.4, 7.0]$ について上位 40% の単語候補である「四間」と「飛車」を辞書に加える. **max** ではそれぞれの要素の最大値からなるベクトル $[4.9, 1.5, 0.2, 0.9, 3.8]$ から「四間」と「飛車」を辞書に加える. **each** では $[1.4, 1.5, 0.2, 0.5, 3.8]$ と $[4.9, 0.8, 0.1, 0.9, 3.2]$ のそれぞれの上位 40% の単語候補「間」「飛車」と「四間」「飛車」を辞書に追加する. この時すでに辞書に登録されている単語候補 (この場合は「飛車」) は二重に登録しない.

最後に, それぞれの方法で生成された辞書を用いて自動単語分割器の再学習を行う.

5 評価

4 章で述べた提案手法の効果を検証するために自動単語分割の実験を行った. 提案手法の効果を検証するために, シンボルグラウンディングにより獲得された辞書を用いる場合 (提案手法) と用いない場合を比較した.

5.1 コーパス

表 1 は今回の実験で用いたコーパスの詳細を示している. コーパスは, シンボルグラウンディングのためのコーパス (シンボルグラウンディング用コーパス) と自動単語分割のための訓練/開発/テストコーパスから成る.

シンボルグラウンディング用コーパスは, 33,151 組の将棋局面と将棋解説文から成る. ただし, シンボルグラウンディング用コーパスの将棋解説文には単語分割が付与されていない. この将棋解説文から疑似確率的単語分割コーパスを生成し, シンボルグラウンディングの学習 (ニューラ

ルネットワークの学習) を行った.

自動単語分割のための訓練コーパスには, 現代日本語書き言葉均衡コーパス (BCCWJ) (Maekawa, Yamazaki, Ogiso, Maruyama, Ogura, Kashino, Koiso, Yamaguchi, Tanaka, and Den 2014) と, 日経新聞 (1990-2000) の一部からなる新聞記事のコーパス, 英語表現辞典からなる日常会話文のコーパスを用いた. BCCWJ の一部は学習には用いず, テストコーパスとして用いた. 将棋解説文からランダムに抽出した 5,041 文を人手で単語分割し, これを開発コーパス (253 文), 局面なしテストコーパス (3,000 文), 局面ありテストコーパス (1,788 文) の 3 つに分けた.[1] 将棋解説文のための辞書は, 局面ありテストコーパスに対し提案手法を適用することで獲得する. 局面なしテストコーパスは局面の情報を持たない将棋解説文だけから成るコーパスであり, 局面ありテストコーパスから得られた辞書の汎用性を評価するために用意した. 実験では, 局面ありテストコーパスから得られた辞書を用いて, 局面なしテストコーパスの単語分割精度を評価した.

5.2 単語分割システム

本実験では以下の 2 つの単語分割モデルを用いてその精度を評価した.

ベースライン: 単語分割のための訓練コーパスと UniDic (234,652 単語)[2] を用いて学習されたモデル.

+擬似確率的単語分割辞書: ベースラインで用いた言語資源に加え, 擬似確率的単語コーパスから出現頻度の高い単語候補を加えた辞書を用いて学習されたモデル.

+シンボルグラウンディング: ベースラインで用いた言語資源に加え, シンボルグラウンディングにより獲得された辞書を用いて学習されたモデル.

ベースラインおよび提案手法のいずれにおいても UniDic を辞書として用いた. 提案手法では UniDic に加えて, シンボルグラウンディング用コーパスから獲得される辞書を用いる. ベースラインの単語分割モデル構築と辞書獲得のために必要となる擬似確率的単語分割コーパスの生成にはロジスティック回帰を用いており, 表 1 に示した自動単語分割のための訓練コーパスを学習用に用いた. ロジスティック回帰は単語境界の確率値 P_i を出力し, ベースラインではこの P_i が 0.5 以上なら分割境界があるとし, 擬似確率的単語分割コーパスには P_i の出力値をそのまま用いて生成した. このとき $m = 4$ とし, 擬似確率的単語分割コーパスを 4 つ生成した.

シンボルグラウンディングの手法を評価するために, 擬似確率単語分割コーパスの単語を頻度順に並べ, その上位 $R'\%$ を追加した辞書を生成し, モデルを構築した.

辞書獲得において, シンボルグラウンディングの辞書生成の手法 (**sum**, **max**, **each**) と R', R の値には開発セットの単語分割精度 (F 値) を用いて最も高くなるパラメータを採用した. 擬似

[1] シンボルグラウンディング用コーパスと自動単語分割のための開発/テストコーパスはそれぞれ異なる将棋解説文から抽出して作成した.

[2] http://pj.ninjal.ac.jp/corpus_center/unidic/

表 2 BCCWJ (6,025 文) の単語分割結果

単語分割手法	適合率	再現率	F 値
ベースライン	99.36%	96.37%	99.37
+シンボルグラウンディング	99.34%	99.35%	99.34

表 3 将棋解説文 (4,788 文) の単語分割結果

単語分割手法	適合率	再現率	F 値
ベースライン	90.78%	91.03%	90.90
+擬似確率的単語辞書	90.84%	91.53%	91.19
+シンボルグラウンディング	90.92%	91.57%	91.24

確率的単語分割辞書では $R' = 0.074$ のときに最も精度が高くなり, 辞書には 110 単語が追加された. 提案手法では, 手法 **each** で $R = 0.074$ のときに最も精度が高くなり, 辞書には 110 単語が追加された.

5.3 結果と考察

単語分割精度の評価尺度には以下で表される, 適合率と再現率, F 値を用いた.

$$\begin{aligned} \text{適合率} &= \frac{\text{正解単語数}}{\text{システムの出力文の単語数}} \\ \text{再現率} &= \frac{\text{正解数単語数}}{\text{正解文の単語数}} \\ \text{F 値} &= \frac{2 \cdot \text{適合率} \cdot \text{再現率}}{\text{適合率} + \text{再現率}} \end{aligned}$$

表 2 は BCCWJ に対する単語分割の精度を示しており, 表 3 は局面なしの将棋解説文に対する単語分割精度と局面ありの将棋解説文に対する単語分割の精度を示している. このときの辞書は局面ありの解説文のみを用いて生成された. BCCWJ に対する単語分割精度 (表 2) と将棋解説文に対するの単語分割精度 (表 3) を比較すると, 将棋解説文の単語分割は一般ドメインの単語分割より難しいことが分かる. 将棋解説文には将棋特有の単語や表現が大量に含まれるため単語分割の精度が低くなったことが考えられる.

表 3 において, 提案手法はベースラインや擬似確率的単語辞書を追加した手法に比べて精度が向上しており, 再現率についてはマクネマー検定で 1% の統計的有意差があった. 本手法における辞書獲得は教師なし学習にもかかわらず自然注釈による手法 (Liu, Zhang, Che, Liu, and Wu 2014) と同程度のエラー削減率を実現できた. この結果よりシンボルグラウンディングによる単語分割は注釈付けと同様に有用であると言える.

表 3 において, ベースラインと辞書を追加した手法の適合率と再現率を詳しくみると, 再現率

表 4 局面なしの将棋解説文 (3,000 文) の単語分割結果

単語分割手法	適合率	再現率	F 値
ベースライン	90.90%	88.91%	89.89
+擬似確率的単語辞書	90.96%	89.51%	90.23
+シンボルグラウンディング	90.95%	89.44%	90.19

表 5 局面ありの将棋解説文 (1,788 文) の単語分割結果

単語分割手法	適合率	再現率	F 値
ベースライン	90.70%	92.47%	91.58
+擬似確率的単語辞書	90.76%	92.91%	91.83
+シンボルグラウンディング	90.89%	93.03%	91.95

は適合率よりも大きく向上していることが分かる. この結果より, 正しい単語と少量の間違った単語を学習していることが分かる. 実際に, 擬似確率的単語分割辞書とシンボルグラウンディングによる辞書の両方には「飛車」や「同歩」,「先手」などの将棋解説文に出現する頻度が高い単語が登録されていた. シンボルグラウンディングによる辞書には「休憩」や「成」など擬似確率的単語分割辞書には無い単語が追加されており, 擬似確率的単語分割辞書には「.」や「・」などのシンボルグラウンディングによる辞書には存在しなかった単語が追加されていた. また, 表 2 より本手法は一般のドメインに深刻な精度低下をもたらさないことが分かる.

表 4 と表 5 は局面なしの将棋解説文と局面ありの将棋解説文の単語分割精度を示している. 表 4 より, 辞書を追加する 2 つの手法において, 局面なしの将棋解説文の単語分割の精度が向上しており, 生成された辞書が将棋解説文の分野に対して有効であることがわかる. この精度向上は高頻度で出現する将棋用語によるものと考えられる. 局面なしの将棋解説文では, 擬似確率的単語分割辞書を追加する手法が最も精度が高かった. しかし, 局面ありの将棋解説文において, シンボルグランディングによる辞書を追加した手法が最も精度が高かった (表 5). また, 局面なしの将棋解説文よりも局面ありの将棋解説文の方が精度向上の割合が大きい. 以上より, 本稿で提案する手法は局面に対応した単語を効果的に学習できていると結論できる.

最後に, ニューラルネットワークを用いて将棋局面と解説文をグラウンディングする際のデータサイズを変更し, その学習曲線を図 2 に示す. 横軸が学習に用いる局面数であり, 縦軸が将棋解説文 (4,788 文) における単語分割の精度 (F 値) を表している. 局面数が 12,000 程度で学習が収束している.

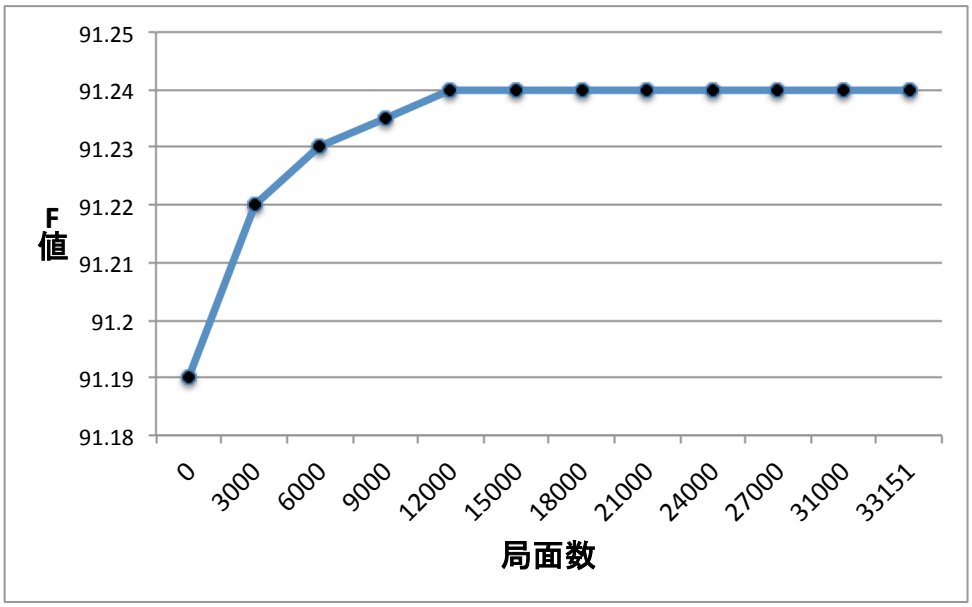

図 2 学習のデータサイズを変更したときの学習曲線

6 関連研究

本稿では日本語の単語分割を行った. 単語分割の代表的な手法は隠れマルコフモデル (Nagata 1994) である. また, Sproat らは類似した手法で中国語の単語分割を行った (Sproat, Gale, Shih, and Chang 1996). これらの手法は単語をモデルの単位として扱っている.

近年, Neubig らはそれぞれの文字の間に単語分割があるかどうかを点予測によって判定する手法 (Neubig et al. 2011) を提案しており, タグの制約のない, 文字への BI タグのタグ付けとして解くことができる. 中国語の単語分割では BIES タグをタグ付けし系列ラベリング問題 (Xue 2003) として解く手法がある. BIES はそれぞれ単語の始まり, その続き, 単語の終わり, 1 文字の単語を表している. 我々の予備的な実験で日本語の単語分割では BI タグを用いたサポートベクターマシンは BIES タグを用いた CRF よりもわずかに精度が高かった. これが本稿で点予測を用いた理由の 1 つである. しかし, 本手法は BIES タグと CRF の単語分割にも適用可能である.

本稿で述べた提案手法は教師なし学習でハイパーパラメーターを調整するための少量の注釈付きデータを必要とした. この観点では, この手法は自然注釈 (Yang and Vozila 2014)(Jiang, Sun, Lü, Yang, and Liu 2013)(Liu et al. 2014) に類似している. しかし, これらの研究ではハイパーテキストのタグは部分的な注釈と見なし, 部分的な注釈を含むデータを用いて学習された CRF で単語分割の性能を向上させた. また, Tsuboi らは大量の生のテキストから新しい単語を抽出する

手法を提案し (Tsuboi, Kashima, Mori, Oda, and Matsumoto 2008), Mori らは類似した設定でのオンライン手法を提案した (Mori and Nagao 1996).

グラウンディングに基づく教師なし単語分割には (Roy and Pentland 2002; Nguyen, Vogel, and Smith 2010) がある. Roy らは音声情報と画像情報をグラウンディングすることにより, 音声情報から単語を獲得する手法を提案した. これは, 音声信号と画像の物体の類似性を用いて, 物体とその名前を連続音声から獲得する. Nguyen らは機械翻訳のための教師なし単語分割を提案した. 分かち書きされている翻訳先の言語の単語と対応するように翻訳元の言語をノンパラメトリックな手法で単語分割した. これらの研究に対して, 本稿では言語以外のモダリティを扱い, 単一言語内でテキストの単語分割を行う最初の試みとして将棋局面を用いた.

7 終わりに

本稿では非テキスト情報を用いた教師なし学習により辞書を生成し, それを用いることによる自動単語分割の精度向上について述べた. 単語候補を生テキストから取り出すために, まず確率的に文を分割し, 単語の候補と元の解説文に対応する将棋局面をニューラルネットワークを用いて結びつけてシンボルグラウンディングを行った. 最後にシンボルグラウンディングの結果を参照しスコアの高い単語の候補を辞書に追加した. 実験結果より非テキスト情報を用いた手法はテキスト情報のみを用いた手法よりも精度が高く, 非テキスト情報を用いる手法の有効性が確認できた. 今後は, この手法を他の深層学習のモデルに適用することやシンボルグラウンディングの結果を分散表現として単語分割の手法 (Ma and Hinrichs 2015) に適用, 及び画像などの他の非テキスト情報を用いることが課題としてあげられる.

謝 辞

本研究は JSPS 科研費 26540190 及び 16K00293, 25280084 の助成を受けたものである. ここに敬意を表する.

参考文献

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). "LIBLINEAR: A Library for Large Linear Classification." *Journal of Machine Learning Research*, **9**, pp. 1871–1874.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). "Every Picture Tells a Story: Generating Sentences from Images." In

Proceedings of the 11th European Conference on Computer Vision, pp. 15–29.

Jiang, W., Sun, M., Lü, Y., Yang, Y., and Liu, Q. (2013). "Discriminative Learning with Natural Annotations: Word Segmentation as a Case Study." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 761–769.

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014). "Multimodal Neural Language Models." In *Proceedings of the 31st International Conference on Machine Learning*, pp. 595–603.

Liu, Y., Zhang, Y., Che, W., Liu, T., and Wu, F. (2014). "Domain Adaptation for CRF-based Chinese Word Segmentation using Free Annotations." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 864–874.

Ma, J. and Hinrichs, E. W. (2015). "Accurate Linear-Time Chinese Word Segmentation via Embedding Matching." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pp. 1733–1743.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). "Balanced corpus of contemporary written Japanese." *Language Resources and Evaluation*, **48**, pp. 345–371.

Mori, S. and Nagao, M. (1996). "Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis." In *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 1119–1122.

Mori, S., Richardson, J., Ushiku, A., Sasada, T., Kameko, H., and Tsuruoka, Y. (2016). "A Japanese Chess Commentary Corpus." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pp. 1415–1420.

Mori, S. and Takuma, D. (2004). "Word n-gram probability estimation from a Japanese raw corpus." In *Proceedings of the Eighth International Conference on Speech and Language Processing*, pp. 1037–1040.

Nagata, M. (1994). "A Stochastic Japanese Morphological Analyzer Using a forward-DP backward-A* N-best Search Algorithm." In *Proceedings of the 15th Conference on Computational Linguistics*, pp. 201–207.

Neubig, G., Nakata, Y., and Mori, S. (2011). "Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 529–533.

Nguyen, T., Vogel, S., and Smith, N. A. (2010). "Nonparametric Word Segmentation for Machine Translation." In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 815–823.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., and Pinkal, M. (2013). "Ground-

ing Action Descriptions in Videos." *Transactions of the Association for Computational Linguistics*, **1** (Mar), pp. 25–36.

Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., and Schiele, B. (2013). "Translating Video Content to Natural Language Descriptions." In *Proceedings of the 14th International Conference on Computer Vision*, pp. 433–440.

Roy, D. K. and Pentland, A. P. (2002). "Learning words from sights and sounds: a computational model." *Cognitive Science*, **26** (1), pp. 113 – 146.

Sproat, R., Gale, W., Shih, C., and Chang, N. (1996). "A Stochastic Finite-state Word-segmentation Algorithm for Chinese." *Computational Linguistics*, **22** (3), pp. 377–404.

Tsuboi, Y., Kashima, H., Mori, S., Oda, H., and Matsumoto, Y. (2008). "Training Conditional Random Fields Using Incomplete Annotations." In *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 897–904.

Tsuruoka, Y., Yokoyama, D., and Chikayama, T. (2002). "Game-Tree Search Algorithm Based On Realization Probability." *Journal of the International Computer Games Association*, **25**, p. 2002.

Xue, N. (2003). "Chinese Word Segmentation as Character Tagging." *The Association for Computational Linguistics and Chinese Language Processing*, **8**, pp. 29–48.

Yang, F. and Vozila, P. (2014). "Semi-Supervised Chinese Word Segmentation Using Partial-Label Learning With Conditional Random Fields." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 90–98.

Yang, Y., Teo, C. L., Daumé, III, H., and Aloimonos, Y. (2011). "Corpus-guided Sentence Generation of Natural Images." In *Proceedings of 2011 the Conference on Empirical Methods in Natural Language Processing*, pp. 444–454.

森信介，小田裕樹 (2009). 擬似確率的単語分割コーパスによる言語モデルの改良. 自然言語処理, **16** (5), pp. 7–21.

略歴

友利　涼：2016 年愛媛大学工学部卒業. 同年より京都大学大学院情報学研究科修士課程在学.

亀甲　博貴：2015 年東京大学大学院工学系研究科修士課程卒業. 同年より同大学院博士課程在学.

二宮　崇：2001 年東京大学大学院理学系研究科情報科学専攻博士課程修了. 同年より科学技術振興事業団研究員. 2006 年より東京大学情報基盤センター講師. 2010 年より愛媛大学大学院理工学研究科准教授, 2017 年同教授. 東京大

学博士 (理学). 言語処理学会, 情報処理学会, 人工知能学会, 電子情報通信学会, 日本データベース学会, ACL, ACM 各会員.

森　信介：1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了. 同年, 日本アイ・ビー・エム株式会社入社. 2007 年より京都大学学術情報メディアセンター准教授, 2016 年同教授. 京都大学博士 (工学). 1997 年情報処理学会山下記念研究賞受賞. 2010 年, 2013 年情報処理学会論文賞受賞. 2010 年第 58 回電気科学技術奨励賞受賞. 言語処理学会, 情報処理学会, 日本データベース学会, ACL 各会員.

鶴岡　慶雅：2002 年東京大学大学院工学系研究科博士課程修了. 同年より科学技術振興事業団研究員. 2005 年マンチェスター大学研究員. 2009 年北陸先端科学技術大学院大学准教授. 2011 年より東京大学大学院 准教授. 東京大学博士（工学）. 言語処理学会, 情報処理学会, 人工知能学会, 各会員.