統計的アプローチによる英語スラッシュ・リーディング教材の自動生成

行野顕正[†1] 田中省作[†2]
冨浦洋一[†3] 柴田雅博[†4]

スラッシュ・リーディングとは，意味のかたまりごとにスラッシュで区切られた英文を読むことにより，読解力の向上を目指す学習法である．多くのスラッシュ付き英文を読むことで，学習効果が上がると考えられるが，現在のところ十分な文書数のある学習教材が存在しないという問題がある．本稿では，統計的アプローチを用いて任意の英文にスラッシュを自動的に挿入する手法を提案する．英文中のスラッシュの位置を定める主な要因は，英文の部分的な構文構造・セグメント長のバランス・一部の単語であるという仮定に基づき，パラメトリックな確率モデルおよびSVMを構築する．既存の教材を学習データとしてモデルを学習することで，その教材のスラッシュ挿入規則を模倣したスラッシュ付き英文を作ることができる．3つの既存教材を対象とした実験では，提案手法が，様々な教材におけるスラッシュ挿入規則を，従来手法よりも高い適合率・再現率で模倣できるという結果が示されている．

Automatic Generation of Materials for Slash Reading Based on Statistical Models

Kensei Yukino,[†1] Shosaku Tanaka,[†2] Yoichi Tomiura[†3] and Masahiro Shibata[†4]

In Slash Reading, learners read English sentences separated into segments (sense groups) with slashes to improve their reading skills. The more texts for Slash Reading a learner read, the more effect of learning could be expected. However, there are not enough materials for Slash Reading. This paper proposes methods for transforming automatically a plain sentence into a slashed sentence based on statistical approaches. A parametric model and a SVM model are built on the assumption that the factors to decide where to insert slashes into a sentence are a portion of the syntactic structure of the sentence, the lengths of the segments and words around the slashes. The models are learned from an existing material for Slash Reading. The systems based on these models, therefore, can transform automatically a plain sentence into a slashed sentence by imitating positions of slashes in the material. The results of the experiments using existing materials for Slash Reading indicate that the proposed methods imitate positions of slashes of the materials with the higher precision and recall than the previous methods.

1. はじめに

スラッシュ・リーディング[☆]とは，通訳士訓練法を基にした英語リーディング学習法の1つである[1),2)]．学習者は意味のかたまり（セグメント[☆☆]）ごとにスラッシュで区切られた英文を，セグメント間の返り読みを極力行わずに読み，各セグメントの意味を順につなげるようにして文全体の意味を理解していく．

The massive devastation of Japan / during World War II / destroyed much of old Japan.

たとえば上の例文では，“The massive devastation of Japan” =「日本の大規模な荒廃は」，“during World

†1 九州大学大学院システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University
†2 立命館大学文学部
College of Letters, Ritsumeikan University
†3 九州大学大学院システム情報科学研究院
Faculty of Information Science and Electrical Engineering, Kyushu University
†4 九州大学ベンチャービジネスラボラトリー
Venture Business Laboratory, Kyushu University

☆ フレーズ・リーディング，チャンク・リーディング，区切り読みなどと呼ばれることもある．本稿ではスラッシュ・リーディングで統一する．

☆☆ フレーズ，センスグループ，チャンクとも呼ばれる．

War II” =「第二次世界大戦の間のですが」, “destroyed much of old Japan” =「旧来の日本のほとんどを破壊した」のように，順に文意を把握していく．このような学習を多く繰り返すことにより，学習者は「英語を英語そのものの語順で理解する」ことを習慣づけられる．また，ひとかたまりの意味としてとらえやすいセグメントを自然に把握する能力が訓練される．結果として，英文を素早く読解することができるようになるといわれている．

スラッシュ・リーディングは近年，ALC NetAcademy☆1や KUMON SPEED READING SYSTEM☆2といった CALL（Computer Assisted Language Learning）教材や，市販の英語教材で扱われるようになってきている．また，東京 SIM 外語研究所☆3などのリスニング学習においても，同様の英文分割提示が試みられている．

しかし，スラッシュ・リーディングを実際に学習に取り入れるためには，2 つの問題が残っている．第 1 に，既存のスラッシュ・リーディング用の教材の量が少ないことである．スラッシュ・リーディングでは，学習者が多くのスラッシュ付き英文を読むことで効果が上がると考えられる．しかし，教材生成を人手に頼っている現状では，十分な量の教材があるとはいい難い．たとえば上記の 3 つの教材では，それぞれ英文が 500 文程度しかなかった．

第 2 に，多くのスラッシュ・リーディング教材が，それぞれ独自のスラッシュ挿入規則☆4を提案していることである．著者らの調査した限りでは，挿入規則の良さを学習効率の観点から定量的に調査した研究はない．また，学習者の英語力などにより学習に適したセグメント長が異なるという意見もあり[1),2)]，本質的に決定的な規則というものはないとも考えられる．したがって，様々な挿入規則に基づいたスラッシュ付き教材を準備することが望ましい．しかし，多様なスラッシュ挿入規則に応じた教材を人手で作成していくことは現実的ではない．

教材不足を解消するには，英文中に自動的にスラッシュを挿入する手法があればよい．このとき，統計的アプローチ☆5ならば，多様な既存教材を学習データに使うことで，それぞれのスラッシュ挿入規則を模倣できる．

スラッシュ挿入問題に統計的アプローチを適用するには，スラッシュ挿入に関係する要因を，モデルに反映させなければならない．本稿では，既存教材および英語教育分野における既存研究の調査から，スラッシュ挿入に関係する主な要因を，英文の部分的な構文構造・セグメント長のバランス・一部の単語であると仮定した．また，各要因を変数化し，確率モデルのパラメータや，SVM の素性として導入する手法を考案した．既存教材を学習データとしてこれらのモデルを学習することで，その教材のスラッシュ挿入規則を模倣するシステムを構築できる．

本稿では，まず 2 章においてスラッシュ挿入に影響する要因と，それを活用した提案手法について説明し，3 章において従来研究との比較を行う．その後，4 章において提案手法の性能を実験により示し，5 章において残っている問題点について議論を行う．

2. 提 案 手 法

2.1 スラッシュ挿入に関係する要因

既存のスラッシュ・リーディング教材[3),4)] におけるスラッシュ挿入規則や，教育学におけるスラッシュ・リーディングに関する既存研究[1),2)] を調査したところ，スラッシュ挿入に関係する要因はどの教材でもおおむね共通していた．本稿では，次の 3 つの要因が，スラッシュ挿入に最も強く関係していると仮定する．その他の要因については，5 章で検討する．

部分的な構文構造

ある単語境界におけるスラッシュの有無は，ごく限られた範囲の構文構造から強く影響を受けると考えられる．たとえば，節や句の区切にはスラッシュが入りやすく，形容詞と名詞の間にはスラッシュが入りにくい．文献 1) では，スラッシュ挿入位置の原則として，節や句の直前や，5 文型における動詞の直後などをあげている教材が多いことが示されている．

セグメント長のバランス

一部の特徴的な構文構造を除くと，セグメント長はある程度の長さに揃えられると考えられる．なぜならば，長すぎるセグメントの意味は瞬時に把握することが難しく，短すぎるセグメントの意味は前のセグメントの意味とつなげていくことが難しいからである．文献 1) では，「学習者が一度にとらえられる単語数」や「一定の内容句数」を基準として，セグメント長をある程度の幅に揃えている教材が多いことが示されている．

一部の単語

特定の単語・単語列の存在が，スラッシュ挿入に影

☆1 http://www.alc.co.jp/netacademy/
☆2 http://www.kumon.ne.jp/srs/
☆3 http://www.tokyo-sim.com/
☆4 スラッシュを入れる位置を決めるための法則．節や句の直前に入れる，など．
☆5 サポートベクタマシン（SVM），最大エントロピー法（ME 法），条件付き確率場（CRF），確率モデルなど．

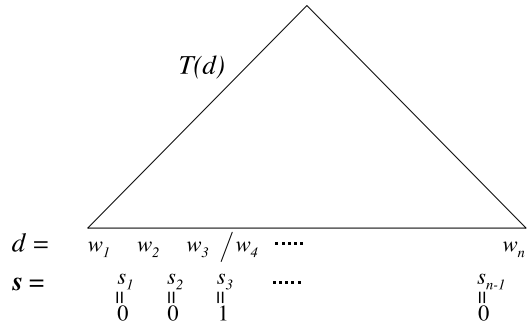

図 1 スラッシュ付き英文の表記法
Fig. 1 The notation system of slashed English documents.

響を与えることがあると考えられる．文献 1) では,「熟語☆の有無」や「単語間の結びつきの強さ☆☆」により，スラッシュ挿入位置が変化する教材が多いことが示されている．また，既存教材[3),4)] からは，文修飾副詞や that 節をともなう知覚動詞の直後にスラッシュが入りやすいといった傾向を読み取れる．

2.2 確率モデルを用いたスラッシュ挿入法

2.1 節であげた要因を表現するためのモデルの 1 つとして，確率モデルを採用する．確率モデルは，1 文全体に対する「スラッシュ挿入の良さ」を，尤度として表現できる．そのため，文中のスラッシュ間の相互関係を考慮しつつ，構築したモデルに対して最良のスラッシュ付けを選択できるという利点がある．一方，パラメータ数が増加すると，学習に時間がかかる，過学習に陥るといった問題があり，単語要因を導入しにくいという問題がある．本節では，確率モデルの表記および基本的な概念について説明する．

図 1 に，文 d にスラッシュが挿入された場合の表記法を示す．文 d は単語の列 $(w_1, w_2, ..., w_n)$ からなっており，その構文構造を $T(d)$ で表す．単語 w_i と w_{i+1} の単語境界 i のスラッシュ有無を $s_i \in \{1, 0\}$ で表す．$s_i = 1$ はスラッシュあり，$s_i = 0$ はスラッシュなしを表す．$\boldsymbol{s} = (s_1, s_2, ..., s_{n-1})$ は，d に対するスラッシュ付けを意味している．

文 d にスラッシュを挿入し，セグメントに分割する問題を考える．d にスラッシュ付け $\boldsymbol{s}$ が行われる確率は，$p(\boldsymbol{s} \mid d)$ で表される．d に最適なスラッシュ付け $\hat{\boldsymbol{s}}$ を与える問題は，

$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} p(\boldsymbol{s} \mid d) \qquad (1)$$

を求める問題と定義できる．ここで，文 d に対して構文構造 $T(d)$ が一意に定まるとすると，$p(\boldsymbol{s} \mid d)$ は式 (2) のように変形することができる．

$$\begin{aligned} p(\boldsymbol{s} \mid d) &= p(\boldsymbol{s} \mid T(d)) \\ &= p(s_1, ..., s_{n-1} \mid T(d)) \\ &= p(s_1 \mid T(d)) p(s_2 \mid T(d), s_1) \\ &\quad ...p(s_{n-1} \mid T(d), s_1, ..., s_{n-2}) \\ &= \prod_{i=1}^{n-1} p(s_i \mid T(d), s_1, ..., s_{i-1}). \end{aligned} \qquad (2)$$

したがって，文 d に最適なスラッシュ付けを行うには，各単語境界においてスラッシュが挿入される確率 $p(s_i \mid T(d), s_1, ..., s_{i-1})$ を与えるパラメトリックな確率モデルを設定し，式 (1) に基づいて $\hat{\boldsymbol{s}}$ を探索すればよい．

特定のスラッシュ付き教材におけるスラッシュ挿入規則を模倣するためには，その教材を学習データとして上記モデルのパラメータを推定すればよい．しかし，式 (2) は条件部（$T(d), s_1, ..., s_{i-1}$）の組合せが膨大であり，パラメータの学習に大量の学習データを必要とする．これは教材が少ないという現状に反する要求であり，このままではスラッシュ自動挿入システムに応用することはできない．

2.3 簡略モデル

ある単語境界 i のスラッシュ有無に関わる要因を 2.1 節で示した要因に絞ることで，2.2 節で示した確率モデルを簡略化する．提案手法では，上記要因を次の 3 つの変数で表現する．

- i 周辺の局所的構文構造 B_i
- i から直前のスラッシュまでの単語数 l_i
- i から文末までの単語数 r_i

また，これらの変数は互いに独立であると仮定する．単語要因は，確率モデルに導入するには様々な課題があるため，本モデルでは扱わない．これについては 5 章で考察を行う．提案モデルで用いる変数を，**図 2** に示す．以下，各変数について説明する．

局所的構文構造

B_i は i 番目の単語境界付近の構文構造についての情報を保持する変数であり，$B_i = (X, Y, Z, \text{head}, c)$ の 5 つ組で表現される．X，Y，Z は，i をまたぐ最下層の統語規則 $X \to \alpha Y Z \beta$ を構成する統語範疇である．ここで，α，β は統語範疇列を表す．たとえば,「動詞句→他動詞補文」が i をまたぐ統語規則であれば，X = 動詞句，Y = 他動詞，Z = 補文となる．head は，X から導出される句における主辞（head word）の位置を示す 3 値変数である（1：Y から導出，2：Z から導出，0：それ以外の統語範疇から導出）．c は，i

☆ 中学や高校で「熟語」として習う単語列を指す．
☆☆ 明確な定義は与えられておらず，感覚的なものである．

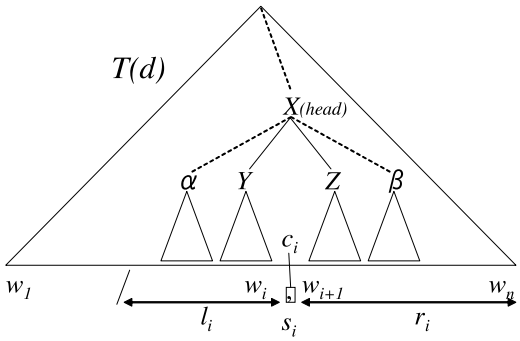

図 2 簡略化した確率モデル
Fig. 2 The simplified probabilistic model.

におけるカンマの有無を表す 2 値変数である．B_i は，2.1 節で示した要因のうち，部分的な構文構造を反映する．

前のスラッシュまでの単語数

l_i は i 番目の単語境界と，i 以前で最後にスラッシュが入った単語境界との間にある単語数である．2.1 節で示した要因のうち，セグメント長のバランスを反映する．

文末までの単語数

r_i は i 番目の単語境界と，文末との間にある単語数である．l_i のみでは最後尾のセグメントが短くなりすぎる可能性があるため，セグメント長に関する補助的な変数として r_i を導入する．

これらの変数を用いて，式 (2) を次のように簡略化する．

$$p(\boldsymbol{s} \mid d) = p(\boldsymbol{s} \mid T(d)) \simeq \prod_{i=1}^{n-1} p(s_i \mid B_i, l_i, r_i) \quad (3)$$

ここで，ある境界にスラッシュが入る確率とスラッシュが入らない確率の比率を，非負のパラメータ $\theta(B)$，$\alpha(l)$，$\beta(r)$ を用いて以下のように表現する．

$$\frac{p(1 \mid B, l, r)}{p(0 \mid B, l, r)} = \theta(B)\alpha(l)\beta(r)$$

この定義を用いて，$p(1 \mid B, l, r)$ は，以下のように表される．

$$p(1 \mid B, l, r) = \frac{\theta(B)\alpha(l)\beta(r)}{1 + \theta(B)\alpha(l)\beta(r)}$$

この式は，各パラメータ $\theta(B)$，$\alpha(l)$，$\beta(r)$ が大きい境界ほど，スラッシュが入りやすいことを意味している．

さらに，セグメント長がある程度の大きさに揃えられるという要因に従い，$\alpha(l)$，$\beta(r)$ に制限を加える．ある単語境界 i において，l_i が小さい（直前のスラッシュと近い）場合，i にスラッシュを入れると短いセグメントが生じてしまい，好ましくない．同様に，r_i が小さい（i が文末に近い）場合，i にスラッシュを入れると文末に短いセグメントが生じてしまう．l_i，r_i が大きくなるにつれ，スラッシュを入れても問題になりにくくなる．このことから，$\alpha(l)$，$\beta(r)$ を単調増加関数とし，非負の変数 a_k，b_m を用いて以下の関数で表す．

$$\alpha(l) = \sum_{k=0}^{l} a_k$$

$$\beta(r) = \sum_{m=0}^{r} b_m$$

これらの条件の下，各パラメータの値を既存教材から最尤推定する．ある教材に出現する文の集合を $\boldsymbol{D}$ とすると，尤度関数は次のように表現できる．

$$L = \prod_{d \in \boldsymbol{D}} p(\boldsymbol{s} \mid T(d)) \simeq \prod_{d \in \boldsymbol{D}} \prod_{i} p(s_i \mid B_i, l_i, r_i) \quad (4)$$

ある教材におけるスラッシュの入り方をシステムに模倣させるためには，式 (4) を最大にするパラメータ $\theta(B)$，$\alpha(l)$，$\beta(r)$ を求めればよい．本稿では山登り法を用いて，式 (4) を極大にするパラメータを推定する．

ある英文 d へのスラッシュ挿入に際しては，式 (4) を最大にしたパラメータを用いて，式 (3) を最大にするスラッシュ列 $\hat{\boldsymbol{s}}$ を探索すればよい．

2.4 SVM を用いたスラッシュ挿入法

2.1 節であげた要因を表現するためのもう 1 つのモデルとして，SVM を採用する．スラッシュ挿入問題を，各単語境界をスラッシュあり・スラッシュなしの 2 クラスに判別する 2 値判別問題ととらえることで，SVM の適用が可能になる．SVM はクラス間のマージンを最大化するという基準に基づいて判別面を構築する手法[5] である．汎化能力が高く，単語を素性として導入しても問題となりにくいことが実験的に知られており[6]，単語要因の導入に都合がよい．一方，SVM は 2 値判別器であるため，1 文全体に対する「スラッシュ挿入の良さ」を，総合的に判断することは困難であるという問題がある．

本稿で提案する SVM は，i 番目の単語境界付近の要因を表現する素性として，以下の 4 つの変数を用いる．3 つは 2.3 節の確率モデルで定義した $B_i = (X, Y, Z, \mathrm{head}, c)$，$l$，$r$ であり，それぞれ境界付近の構文構造，前のスラッシュとの単語数，文末までの単語数を表す．もう 1 つは単語要因を表す変数 $\boldsymbol{w}^{(t)}$ であ

り，境界前後 t 個ずつの単語からなる．

各変数を，2 値ベクトル☆として表現する．各変数は，そのとりうる値それぞれと 1 対 1 対応する次元を並べたベクトル（スロット）として表現される．ある単語境界は，各変数に対応するスロットを並べたベクトルとして表され，実際に出現した値に対応する次元のみが 1，それ以外の次元は 0 に設定される．

学習では，既存教材中の全境界を上記変数で表し，SVM の学習データとして与えればよい．ある英文 d へのスラッシュ挿入に際しては，文頭から順に各単語境界のスラッシュ有無の判別を行う．素性中に前のスラッシュとの単語数 l が含まれており，ある境界のスラッシュ有無を決めるためには，それより前方にある境界のスラッシュ有無が定まっている必要があるためである．l，r に若干修正を加えることで，後方から順にスラッシュの有無を決めることもできるが，前方から文を読むことが自然と考えられるため，本稿では考慮しない．

3. 関 連 研 究

3.1 人間の認知モデルに基づく手法

文献 7) では，心理言語学における「チャンキング」に着目し，文中の依存構造を利用したスラッシュ自動挿入システムを提案している．チャンキングとは，文読解において，それまでに読んだ文章をいくつかの適当な抽象的表象（チャンク）に変換する脳内作業のことである．人間が文を読解するとき，脳内のワーキングメモリにおいて，チャンクを作る，チャンクをまとめるといった作業を行っていると考えられている．チャンクがまとまった境界は何らかの意味がまとまった境界と考えることができるため，そこをスラッシュ挿入候補としている．

実装では，前後の依存関係が少ない境界ほどチャンクがまとまりやすいと仮定している．したがって，スラッシュによって断ち切られる依存関係（係り受け切断数）が少ない境界ほど，スラッシュが入りやすい．文献 7) では，さらにセグメント長に関する制限を加えて，スラッシュ挿入の制御を行っている．

文献 7) の手法には，学習データを用いずにスラッシュ挿入が可能であるという利点がある．しかし，逆にスラッシュ挿入規則はこのモデル独自のものに固定されてしまい，既存教材の様々なスラッシュ挿入規則を模倣することは難しい．

☆ 各次元が，1 か 0 かの 2 値からなるベクトル

3.2 短文との編集距離に基づく手法

文献 8) では，「文のセグメントへの分割」を「長文の短文への分割」ととらえ，スラッシュ挿入を行っている．「分割の良さ」は，各セグメントの短文としての発生確率と，各セグメントと対話コーパス中の短文との編集距離に基づいて定義される．発生確率は N-gram モデルに基づいて計算され，編集距離は，単語間のシソーラス上での距離に基づいて重み付けされている．

実験では，既存の学習教材のスラッシュをどれだけ再現できるかを調査しており，文献 7) の手法よりも高い精度が得られたことが報告されている．

文献 8) の手法は，比較的潤沢にある短文コーパスやシソーラスを利用でき，量の少ないスラッシュ・リーディング教材を学習データに使う必要がないという利点がある．また，構文解析器を使わないため，構文解析段階での誤りを考慮する必要がないという利点もある．しかし，収集した短文コーパス次第でスラッシュ挿入規則が変わってしまうという問題がある．また，得られたスラッシュ挿入規則と，既存の学習教材の挿入規則との関係も不明確である．短文コーパスの代わりに学習教材のセグメントそのものを与えることも考えられるが，単語間距離に基づくモデルの仕様上，大量のセグメントを必要としてしまう．そのため，文献 8) のモデルを既存の学習教材の模倣のために使うことは現実的ではない．

3.3 固定窓方式 SVM に基づく手法

SVM を自然言語処理に応用する際には，対象とする単語境界や単語の前後から，それぞれ一定個数の単語や品詞などを抜き出し，素性として使用することが多い[6]．この方法に基づいて素性を収集した手法を，提案した素性との比較対象として実装した．この手法と，提案した SVM とを比較することで，提案した要因とその表現手法の有効性を確認できる．

素性として，単語境界の前後 2 単語，前後 7 品詞，対象境界および前後 10 境界のカンマ有無，前方 15 境界のスラッシュ有無を使用した．提案手法と比較すると，品詞，カンマ有無が局所的構文構造要因，スラッシュ有無がセグメント長要因，単語が単語要因にそれぞれ対応する．文末からの単語数は，単語・品詞に文末を表す記号が出現することで，間接的に表現されている．それぞれ過不足はあるものの，提案手法と同等以上の情報量があると考えている．前方のみのスラッシュ有無を素性として利用するのは，提案した SVM と同様，スラッシュ挿入時に前方から順にスラッシュ有無判別を行う必要があるためである．

4. 実　　験

4.1 簡略確率モデルの実装

提案手法では，文書 d の構文構造 $T(d)$ が一意に定まると仮定している．実装では，理想的な $T(d)$ の近似として Charniak Parser[9] の第 1 位出力を用い，そこから局所的構文構造 B，単語数 l，r を抽出する．

局所的構文構造 B のうち，X，Y，Z には，Charniak Parser で用いられている統語範疇（Penn Treebank[☆]の Part of Speech Tags）をそのまま使用した．ただし,「カンマ」は変数 c として独立して B 中に取り入れてあるため，統語範疇としては扱っていない．

単語数 l，r には，Charniak Parser において単語として分割された項数を用いた．ただし，記号類（“$”，“%” など）や省略形（“I'm” における “'m” など）は数えていない．また，l，r のとりうる最大値をそれぞれ 15，8 に設定し，それ以上に長い l，r は，$l = 15$，$r = 8$ として扱った．

スラッシュ挿入問題では学習データが非常に限られているため，学習データ中に出現しなかった B が挿入対象英文中に出現する可能性がある．本手法では，スラッシュ・リーディングでは，挿入されていた方が読みやすいスラッシュが入っていない場合より，挿入されるべきでないスラッシュが入っている場合の方が学習者に与える違和感が大きいと仮定し，$\theta(B) = 0$ と設定した．これにより，このような構造に対してはスラッシュはいっさい入らなくなる．

4.2 SVM モデルの実装

B，l，r の取得・設定方法は，確率モデルと同様の手法を用いた．単語要因 $\boldsymbol{w}$ として，境界前後の 2 単語を用いた．単語には，Charniak Parser において単語として分割された項をそのまま利用し，記号や省略形もそれぞれ 1 単語として扱っている．

SVM の実装には，TinySVM[☆☆]を利用し，2 次の多項式カーネルを用いた．比較対象である固定窓方式の SVM も同じ設定を用いた．

4.3 実 験 手 法

提案手法が従来手法よりも精度良く既存教材を模倣できることを確かめるため，従来手法が実験に用いているのと同じ教材を対象に，比較実験を行った．また，提案した要因と，その表現手法の有効性を評価するため，2 種の提案手法と，比較用に作成した固定窓方式の SVM の間で，3 種類の教材に対して実験を行った．

☆ http://www.cis.upenn.edu/~treebank/

☆☆ http://chasen.org/~taku/software/TinySVM/

表 1 実験データ
Table 1 The experimental data.

	Rakuraku	NetAcademy	AllInOne
文数	480	455	598
単語境界数	7,419	8,569	11,282
スラッシュ数	840	824	2,476

数種類の教材に対して挿入実験を行うことにより，提案手法が様々な教材のスラッシュ挿入規則を模倣できるかどうかの調査も同時に行った．

実験では，各教材に対してクローズドテストおよび交差検定（データ分割数は 10）を行い，それぞれ適合率・再現率を調査した．適合率・再現率は以下の式で計算される．

$$\text{適合率} = \frac{\text{正解スラッシュ数}}{\text{提案システムの入れたスラッシュ数}}$$

$$\text{再現率} = \frac{\text{正解スラッシュ数}}{\text{オリジナル教材のスラッシュ数}}$$

実験で用いた教材は，以下の 3 種類である．各教材の文数などを**表 1** にまとめる．各教材のスラッシュ文の一部を，付録 A.1 に掲載する．

Rakuraku
七寶出版より出版されている教材「らっくらく英文解釈」[3] 中の英文．従来研究で実験に用いられている．

NetAcademy
ALC が販売している CALL 教材「NetAcademy」のリーディングセクションで用いられている英文．

AllInOne
Linkage Club 株式会社より出版されている英語教材「All In One Second Edition」[4] 中の英文．

4.4 実 験 結 果

Rakuraku を対象とした場合の，従来手法と提案手法の適合率・再現率を**表 2** に示す．表中の確率手法・提案 SVM は提案手法を，チャンク手法は文献 7) の手法[☆☆☆]を，コーパス手法は文献 8) の手法を表している．チャンク手法，コーパス手法は教師データを必要としないため完全なオープンテスト結果であり，提案手法は 10-fold 交差検定の結果である．

表 2 より，提案した 2 手法はそれぞれ従来手法よりも高い適合率・再現率を達成していることが分かる．提案手法の精度が高いのは，従来手法がそれぞれ独自

☆☆☆ チャンク手法は，構文解析器として Apple Pie Parser（APP）を利用している．提案手法で用いた Charniak Parser は APP よりも解析精度が高く，予備実験ではスラッシュ挿入精度も Charniak Parser を使用した方が高くなっている．このことから，チャンク手法も，構文解析器を変更することでより若干高い精度が得られる可能性がある．

表 3 様々な教材に対する適合率・再現率
Table 3 The precision and recall for some materials.

手法	Rakuraku		NetAcademy		AllInOne	
	適合率	再現率	適合率	再現率	適合率	再現率
確率手法	75.3	58.3	73.7	60.4	84.8	79.9
提案 SVM	76.6	68.2	80.1	69.3	87.5	86.6
固定窓 SVM	67.3	64.2	73.8	67.2	78.3	79.6

表 2 提案手法と従来手法の比較結果
Table 2 The comparison among the proposed method and the previous methods.

手法	適合率	再現率
確率手法	75.3	58.3
提案 SVM	76.6	68.2
チャンク手法	45.2	39.1
コーパス手法	47.2	51.1

のスラッシュ基準に基づいてスラッシュ挿入を試みるのに対し，提案手法は既存教材を学習データとして活用した結果である．

次に，各教材に対する提案手法 2 種と固定窓方式 SVM の適合率・再現率を**表 3** に示す．それぞれ，10-fold 交差検定の結果である．

表 3 より，確率手法と固定窓方式の SVM が同程度の性能であり，提案素性を用いた SVM がそれよりも高い精度を達成していることが分かる．確率手法は単語要因を利用していないにもかかわらず，単語素性を用いた固定窓方式の SVM と同程度の精度を達成しており，提案した局所的構文構造の表現手法が有効であったことが分かる．提案 SVM と，固定窓方式の SVM の差からも，同様のことが読み取れる．

提案 2 手法を比較すると，総じて SVM モデルが上回っており，文全体のスラッシュ挿入の良さを判断できることよりも，単語要因の導入の方が効果的であったことが分かる．確率モデルも単語要因を導入することで，精度が向上する可能性が高いと考えられる．

次に，提案手法の，多様な教材に対する適応可能性について見ていく．表 3 では，適合率は高いものの，NetAcademy，Rakuraku に対する再現率は十分には高くない．しかしながら，システムにより実際に挿入されたスラッシュを比較すると，各教材のスラッシュ挿入基準の相違が顕著に反映されていることが分かる（付録 A.2）．スラッシュ位置には若干の自由度があることを考えると，現精度においても，教材生成の支援システムとしては十分に活用可能なレベルに達していると思われる．

5. 議　　論

提案システムの出力をそのまま学習教材として利用するためには，いっそうの精度向上が不可欠である．そのためには，スラッシュ挿入に関係する要因のうち，現行モデルでは利用しなかった要因をモデルに反映していく必要がある．本章では，確率モデルにおいて，今の要因だけでは模倣ができなかった例を示しつつ，その他の要因について説明する．また，提案手法に残る問題点についても言及する．例文中では，システムの出力した誤ったスラッシュ付き英文の先頭に，“∗”を付ける．また，議論の対象としているスラッシュを，“ / ”で表す．

5.1 未利用の要因

広い範囲の構文構造の違いにより，スラッシュ挿入位置が変化する傾向が見られた．たとえば，いずれの教材でも，複雑な修飾部を持つ主格名詞句の直後にはスラッシュが入りやすく，単純な主格名詞句の直後にはスラッシュが入りにくいという傾向が見られた．提案手法は修飾部の有無を考慮しないため，これら 2 種類の構造を区別できない．

A stifling and restrictive environment / where individuality is not encouraged / is not an ideal place for anyone to learn.
∗ A stifling and restrictive environment / where individuality is not encouraged is not an ideal place / for anyone to learn.

上の例では，関係節に修飾された大きな主格名詞句“A stifling〜encouraged”の直後のスラッシュを，提案手法が再現できていないことが分かる．

また，Rakuraku や AllInOne では，that 節の内部にスラッシュが入りにくいという傾向が見られた．提案手法では that 節の内外を区別しないため，このような傾向を模倣できない．

I think / that the reliance on cram schools to “prepare” a student for education is detrimental.
∗ I think that the reliance on cram schools

/ to “prepare” a student for education is detrimental.

上の例では，教材において 1 セグメントとされた that 節が，提案手法では 2 つに分けられていることが分かる☆.

これらの傾向に対応するためには，句や節の末端位置の情報や統語範疇内部の複雑さ，X より上位の構造情報など，大域的構文構造情報をモデルに導入する必要がある．大域的構文構造の表現方法には様々な手法が考えられ，どのように変数化，モデル化するかが今後の課題となる．

5.2 単語要因利用に残る問題点

提案した確率モデルでは，単語要因を導入することができなかった．数多くの単語を無分別に導入すると，過学習に陥る可能性が高いためであった．既存のスラッシュ・リーディング教材や，中学・高校の英語教材を調査し，前後にスラッシュの入りやすい単語や，逆に入りにくい熟語を厳選する必要がある．または，大量のコーパスから単語の共起確率を計算しておき，それに基づくパラメータ設定を考案するなどの手法も考えられる．

提案した SVM では，出現したすべての単語を素性として導入した．構築された判別面から各素性の重み付けを調査したところ，スラッシュ挿入に強く影響する素性として「1 単語後に that」「1 単語前に say」などが出現しており，単語要因導入の利点が見られる．しかし，一方では「2 単語前が Hanako」「2 単語後が want」など，スラッシュ挿入要因としては不適当と思われる単語も強いスラッシュ挿入要因となっており，全単語を導入した弊害と考えられる．素性を，出現回数の多い単語に絞る，単語をシソーラスなどの利用によりベクトル化し，補完を行うなどの修正が必要と考えられる．

6. おわりに

本稿では，スラッシュ挿入問題に統計的アプローチを適用するため，既存教材などの調査を行い，スラッシュ挿入に影響を与える要因は，部分的な構文構造やセグメント長のバランス，一部の単語であると仮定した．また，各要因を変数化し，確率モデルのパラメータや SVM の素性として導入する手法を考案した．これらのモデルを既存のスラッシュ・リーディング教材から学習することで，その教材のスラッシュ挿入規則を模倣した英文を生成するシステムが構築できた．

3 種類の既存教材を用いた実験により，提案手法が従来手法よりも高い適合率・再現率で既存教材のスラッシュ挿入規則を模倣できることが分かった．また，一般的に SVM の入力として用いられる固定窓方式の素性よりも，提案した素性の方が有用であることが確認できた．

今後は，再現率の向上を目指すため，大域的構文構造をモデル内に取り入れるなどの改良を加えていく予定である．また，単語要因がスラッシュに与える影響をより詳しく調査し，導入する単語の厳選を行っていく予定である．

謝辞 本稿の執筆にあたり，複数の査読者から貴重なご助言をいただきました．また，本研究にあたり，自然言語研究室の山本祥平氏，馬場慎也氏，薬師寺亮太氏，飛松宏征氏らに多くの面でご協力をいただきました．ここに記して感謝の意を表します．本研究の一部は文部科学省の 21 世紀 COE プログラム「システム情報科学での社会基盤システム形成」および，日本学術振興会科学研究費補助金・基盤研究（C）（課題番号：17520383）の援助を受けて行われました．

参考文献

1) 寺島美紀子：英語「直読直解」への挑戦，あすなろ社 (2002).
2) 瀧澤正己：語学強化法としての通訳訓練法とその応用例，北陸大学紀要，Vol.26, pp.63–72 (2002).
3) 松本 茂：直読・直解・多聴式らっくらく英文解釈，七寶出版 (2001).
4) 高山英士：*ALL IN ONE SECOND EDITION*，株式会社 Linkage Club (2003).
5) 麻生英樹，津田宏治，村田 昇：パターン認識と学習の統計学，岩波書店 (2003).
6) 工藤 拓，松本裕治：Support Vector Machine による日本語係り受け解析，情報処理学会研究報告，Vol.2000-NL, No.138, pp.79–86 (2000).
7) 田中省作，冨浦洋一：スラッシュ・リーディング支援システムの構築，言語処理学会年次大会併設ワークショップ「e-Learning における自然言語処理」，pp.37–40 (2004).
8) 土居誉生，隅田英一郎：スラッシュ・リーディングのためのテキスト分割，情報処理学会研究報告，Vol.2004-CE, No.68, pp.25–32 (2004).
9) Charniak, E.: A Maximum-Entropy-Inspired Parser, Technical Report CS-99-12 (1999).

☆ 逆に，think の直後にスラッシュが入ると，文頭に短いセグメントができてしまうため，提案手法ではスラッシュが入っていない．

付　　録

A.1 各教材のスラッシュ文の例

実験で用いた各教材のスラッシュ付き英文を少数例掲載する．AllInOne のスラッシュ挿入基準が他とは大きく異なることが読み取れる．

Rakuraku

(1) When it happened, / I'd been living in Japan / for only two weeks. //
(2) Foreigners make mistakes / when they first come to Japan, / and I was no exception. //
(3) It happened / on a nice day, / the sky a perfect blue. //

NetAcademy

(1) The Chinese martial art of Kung Fu Wu Su / dates back over 6,000 years. //
(2) Though there is no written history of Kung Fu Wu Su, / there are plenty of myths and legends. //
(3) Of the various arts of Chinese self-defense, / Kung Fu Wu Su is the best known / as a method for not only fighting, / but also maintaining one's health. //

AllInOne

(1) He grinned / and said, / "I make lots of money. / On weekdays / I receive an average of 50 orders / a day / from all over the globe / via the Internet." //
(2) "What / does your son do?" //
(3) "Well, / he used to work / for a multinational corporation, / but / unfortunately / he is unemployed / at the moment." //

A.2 システムによるスラッシュ挿入例

各教材で学習したシステムを用いて，学習データとは異なる，同一の英文書☆にスラッシュを挿入した例を掲載する．A.1 で見られる AllInOne と他の教材との差が，システムでも再現されていることが分かる．

Rakuraku

(1) In the 5th century, / after the fall of the Roman Empire, / people from what is now Germany invaded Britain. //
(2) The islands were divided among various kingdoms, / and the Celts were pushed back to Scotland, / Wales, and Cornwall. //
(3) King Egbert of Wessex was the first king / to rule all of England. //
(4) In 1066, / England was invaded by the Normans, / who were from the western part of France / and spoke French, under William the Conqueror. //

NetAcademy

(1) In the 5th century, / after the fall of the Roman Empire, / people from what is now Germany invaded Britain. //
(2) The islands were divided among various kingdoms, / and the Celts were pushed back / to Scotland, Wales, and Cornwall. //
(3) King Egbert of Wessex was the first king / to rule all of England. //
(4) In 1066, / England was invaded by the Normans, / who were from the western part of France / and spoke French, / under William the Conqueror. //

AllInOne

(1) In the 5th century, after the fall / of the Roman Empire, / people / from / what is now Germany invaded Britain. //
(2) The islands were divided / among various kingdoms, / and / the Celts were pushed back / to Scotland, Wales, and Cornwall. //
(3) King Egbert / of Wessex was the first king / to rule all / of England. //
(4) In 1066, / England was invaded / by the Normans, / who were from the western part / of France / and spoke French, under William the Conqueror. //

(平成 18 年 1 月 3 日受付)
(平成 18 年 10 月 3 日採録)

☆ "History of the United Kingdom" written by Kenji Kitao, http://www.cis.doshisha.ac.jp/kkitao/library/student/reading/reading.htm

行野　顕正（学生会員）

2003 年九州大学大学院システム情報科学府知能システム学専攻修士課程修了．同年同大学院博士後期課程進学，在学中．自然言語処理に関する研究に従事．

田中　省作（正会員）

2000 年九州大学大学院システム情報科学研究科博士後期課程修了．九州大学情報基盤センター助手，同大学高等研究機構室員を経て，2005 年より立命館大学文学部助教授．博士（工学）．自然言語処理，言語教育への応用に関する研究に従事．言語処理学会，英語コーパス学会各会員．

冨浦　洋一（正会員）

1989 年九州大学大学院工学研究科電子工学専攻博士課程単位取得退学．同年九州大学工学部助手，1995 年同助教授，1996 年同大学大学院システム情報科学研究科助教授，2000 年同大学院システム情報科学研究院助教授，現在に至る．博士（工学）．自然言語処理，計算言語学，人工知能に関する研究に従事．言語処理学会，人工知能学会各会員．

柴田　雅博（正会員）

2005 年九州大学大学院システム情報科学研究科博士後期課程単位取得退学．2005 年九州システム情報技術研究所特別研究助手，2006 年九州大学ベンチャー・ビジネス・ラボラトリー PD，現在に至る．博士（工学）．自然言語処理，身体障害者支援に関する研究に従事．電子情報通信学会，言語処理学会各会員．