英語速読学習のための統計的手法による英文へのスラッシュの挿入

飛松宏征* 行野顕正 * 田中省作† 冨浦洋一‡

1 はじめに

英語 (外国語) 学習法の一つにスラッシュ・リーディングがある。これは、以下の英文のように適当な意味のかたまり (これを本稿ではセグメントと呼ぶ) ごとに“/”(スラッシュ) が挿入された英語文書を学習者がセグメント単位に素早く読み進める学習法である。

People used to think / that the sun and other stars / move around the earth.

日本人は、英文を理解する際、日本語の構造に並び替えて理解しがちである。これでは読み返しが多くなり素早く理解することが出来ず、リスニングでも致命的になる。スラッシュ・リーディングでは、学習者が、出現した順番にセグメントの意味を素早く捉え、文意をつかむ。こうすることで、英語本来の語順で英文を理解することを習慣付けでき、副次的にリスニング力の向上にも効果がある。

このような効果を得るには、多量の英文を読む、つまり多読の必要があるが、現在十分な量の教材は整備されていない。また、教材作成に複数の人間が携わると、各作成者のスラッシュ挿入の基準が異なるため、スラッシュ挿入に関する一貫性が不十分なものとなる可能性が高い。従って、一貫性のあるスラッシュ・リーディング教材を大量に得るためには、自動的に英文にスラッシュを挿入するシステムがあると望ましい。

そこで、部分的な統語構造を考慮した確率モデルに基づき英文にスラッシュを挿入する手法を提案する。この手法により、少量のスラッシュ付き英文から学習することが可能となる。

* 九州大学大学院システム情報科学府
† 九州大学情報基盤センター
‡ 九州大学大学院システム情報科学研究院

2 提案手法

2.1 統計的スラッシュ挿入法

まず、文 s へのスラッシュの挿入についての素直な確率モデルを考える。文 s に対して、一意にその統語構造 $T(s)$ が求まるものと仮定する。文 s の各単語を $w_0, w_1, ..., w_n$ とする。単語 w_{i-1} と w_i の間 (i 番目の境界) のスラッシュの有無を $b_i(\in \{1, 0\})$ と表すと、$\boldsymbol{b} = (b_1, b_2, ..., b_{n-1})$ は文 s へのスラッシュ付けを意味する (図 1)。統語構造 $T(s)$ の文 s へ $\boldsymbol{b}$ というスラッシュ付けがなされる確率が $p(\boldsymbol{b}|T(s))$ のとき、最適なスラッシュ付けはこの確率を最大にするような $\boldsymbol{b}$ として推定できる。つまり次のように表される。

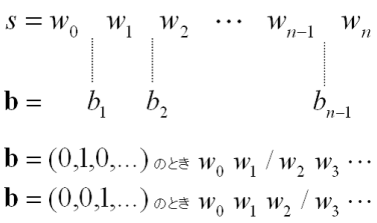

図 1 文 s へのスラッシュ付け

$$\arg\max_{\boldsymbol{b}} p(\boldsymbol{b}|T(s))$$

また、$p(\boldsymbol{b}|T(s))$ は次のように表される。

$$\begin{aligned} p(\boldsymbol{b}|T(s)) &= p(b_1, b_2, ..., b_{n-1}|T(s)) \\ &= p(b_1|T(s)) \cdot p(b_2|T(s), b_1) \cdot p(b_3|T(s), b_1 b_2) \\ &\quad \cdots p(b_{n-1}|T(s), b_1 \cdots b_{n-2}) \\ &= \prod_i p(b_i|T(s), b_1 \cdots b_{i-1}) \end{aligned}$$

上記確率モデルに基づいて最適な $\boldsymbol{b}$ を推定するには、スラッシュが挿入された統語構造付きの英文が学習データとして必要となる。スラッシュ付き英文自体が過疎、なおかつ統語構造のバリエーションは膨大であり、こ

のような確率モデルは現実的ではない。そこで、スラッシュ挿入の有無には統語構造全体ではなく、境界周辺の部分的な統語構造が強く影響すると仮定し、確率モデルの簡単化を行う。

スラッシュ挿入の有無を決める大きな要因は以下の三つと仮定する。

1. 境界周辺の部分的な統語構造
2. 前にスラッシュを入れた位置からの語数
3. 文末までの残りの語数

まず 1 について。文全体ではなく境界の周辺に限定しても、次の例のようにその境界へのスラッシュ挿入の有無をある程度推測することが出来る。

1. The * visitors (冠詞*名詞)
2. said * that they were ... (動詞*補文)

*へのスラッシュ挿入の有無を考えると、1 の例ではまず挿入は有り得ず、2 の例ではスラッシュの入る大きな候補になる。

ここで、境界周辺の統語構造を以下のように定義する。文 s がある文法によって生成されるとき、i 番目の境界の周辺の統語構造とは、その境界を分離する生成規則の左辺 X、右辺の中で境界に隣接する Y、Z、及び head とする (図 2)。head は、例えば図 2 においては、Y が主辞のとき 1、Z が主辞のとき 2、Y,Z 以外が主辞のとき 0 を値としてとるものとする。

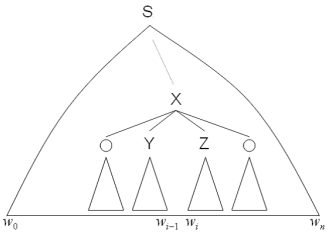

図 2　境界周辺の統語構造

Y と Z が必要なのは、先程の例で説明不要だろう。Y から導出される句と Z から導出される句の間にスラッシュを入れるか否かは、それらが結合してどのような句になるのか (つまり X) にも依存すると考えられる。head は、右辺の長さが 3 以上のときに意味を成す。長さが 2 のときは、Y と Z のどちらが主辞かは必然的に決まる。長さが 3 以上のとき、主辞が Y か Z かそれ以外かによっては、スラッシュ挿入の有無の傾向が変わることが考えられる。よって head を境界周辺の統語構造に加える。

2 と 3 は、セグメントの長さに関するものである。スラッシュ・リーディングでは、セグメント内の英文を素早く理解して読み進めなくてはならないため、セグメントが長くなりすぎると、理解しづらくなる。また、短すぎるセグメントも誤った読解を生むおそれがある。そのため、セグメントの長さに制約が必要となる。2 は、セグメントの長さが長くなるとスラッシュを入れやすくなる、3 は、逆に文末までの語数が少ないとスラッシュを入れにくくなるという要因である。

文 s の i 番目の境界の周辺の統語構造を $B(i;T(s))$、前にスラッシュを入れた位置からの語数を l、文末までの残りの語数を r として前述の $p(\boldsymbol{b}|T(s))$ を次のように近似する。

$$p(\boldsymbol{b}|T(s)) = \prod_i p(b_i|B(i;T(s)),l,r) \qquad (1)$$

なお、確率の条件部、(B(i;T(s)),l,r) のことを境界情報と呼ぶことにする。

2.2　確率モデルの学習

ある任意の B=(X,Y,Z,head),l,r に対して、$p(1|B,l,r)$、$p(0|B,l,r)$ を、非負のパラメタ $\theta(B)$,$\alpha(l)$,$\beta(r)$ を用いて次のように定義する。

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

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

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

$\theta(B)$、$\alpha(l)$、$\beta(r)$ それぞれが大きな値を取ると、スラッシュを入れる確率が大きくなる。前述したように、$\alpha(l)$ は前のスラッシュ位置からの語数 l に関する関数で、l が大きくなると $\alpha(l)$ の値は大きくなる。$\beta(r)$ は文末までの語数 l に関する関数で、r が小さくなると $\beta(l)$ の値は小さくなる。このような性質を満たす $\alpha(l)$、$\beta(r)$ は色々と考えられるが、今回は以下のように階段関数で与えた。

$$\alpha(l) = \begin{cases} 0 & l=0 \\ \alpha_1 & l=1 \\ \alpha_1+\alpha_2 & l=2 \\ \vdots & \end{cases} \qquad (\alpha_i \geq 0) \qquad (2)$$

$$\beta(r) = \begin{cases} 0 & r = 0 \\ \beta_1 & r = 1 \\ \beta_1 + \beta_2 & r = 2 \\ \vdots & (\beta_i \geq 0) \end{cases} \quad (3)$$

学習データからパラメタ $\theta(B)$、$\alpha(l)$、$\beta(r)$ の値を最尤推定により求める。尤度関数は、

$$L = \prod_s \prod_i p(b_i|B(i;T(s)),l,r) \quad (4)$$

である。

2.3 関連研究

土居ら [1] は、コーパスに基づき、統計的言語モデルとテキスト類似度を使って英文を分割するという手法を提案している。スラッシュ挿入を、長文を意味の通る短文に分割する過程と捉えた手法である。N グラムに基づく分割文 (セグメント全体を指す) の尤度と、分割文の類似度 (コーパスの文とセグメントとの類似度から成る) を組み合わせた値を指標としている。コーパスは旅行会話文という比較的短い文が収録されたものを使っている。

3 実験

3.1 実験データ

データに用いたのは、 [2] に収録されているスラッシュ挿入済み英文 472 文である (表 1)。また、英文統語構造解析プログラムとして Apple Pie Parser [3] を用いた。この解析器で用いられているタグ (品詞や句や節に付ける統語範疇) は詳細に分かれているため[*1]、独自に 28 種類にまとめた。

文総数	472
1 文平均語数	14.8
1 文平均境界数	16.3
1 文平均スラッシュ数	1.74

表 1　学習データの情報

3.2 実験方法

実験ではまず、学習データから境界情報を取得する。語数を数える際、カンマやピリオド、引用符などの記号類は含んでいない。しかし、こういった記号類は、スラッシュ付けに大きな影響を与えるので境界情報には含んでいる。また、Apple Pie Parser では “I’m” などの省略形や “5%” は、複数語扱いとなるが、1 語として扱う。

山登り法を用いて、式 4 を最大にするパラメタの値を求めた。また、$\beta(r)$ の段階数を、学習データの 1 文平均の語数である 15 とする。$r \geq 15$ の r に対しては $\beta(r) = \beta(15)$ とする。$\alpha(l)$ の段階数も 15 とする。これも同様に、$l \geq 15$ の l に対しては $\alpha(l) = \alpha(15)$ とする。

スラッシュ付けを決める際、学習されていない統語構造が出てきた場合、その境界にはスラッシュは入れないことにする[*2]。

関連研究で取り上げた土居らの手法 [1] と比較する。その手法ではテストデータとして、同じ [2] を用いている。

また、提案手法のセグメントの長さを考慮した効果を見るため、提案手法から長さの要因を除いた手法 (以下単純手法と呼ぶ) の実験も行った。長さの要因を除くと式 1 を最大にする $\boldsymbol{b}$ は各 i に対して $p(b_i|B(i;T(s)))$ を最大にするものであるから、各 i ごとに $p(b_i|B(i;T(s))) \geq 0.5$ となるように $b_i = 0$ or 1 を定めるという単純なものとなる。

3.3 実験結果

オープンテストを 5 回行った。その回ごとに、半分の 236 文を学習データとして、残りの半分をテスト用データとしてランダムに選択した。提案手法、単純手法および土居らの手法でスラッシュ付けしたときの再現率、適合率は表 2 のようになった。土居らの手法の数値は、F 値の最良となるものを論文から引用した。なお、再現率と適合率は以下のように定義する。

$$\text{再現率} = \frac{\text{正しくスラッシュが入った境界}}{\text{スラッシュが入るべき境界数}}$$

$$\text{適合率} = \frac{\text{正しくスラッシュが入った境界}}{\text{スラッシュを入れた境界数}}$$

また、α, β のセグメント長調整機能の効果を見るため、学習データ、提案手法及び単純手法のセグメント長 (セグメントの語数) に関するデータを表 3 に示す。

3.4 考察

提案手法は土居らの手法と比べると、再現率ではわずかに優り、適合率で大きく上回っている。土居らの手法

[*1] 例えば名詞は、単数形、複数形、固有名詞の単数形 . . .、動詞は、原形、過去形、三人称 . . . と区別される。

[*2] スラッシュがあるべきところにスラッシュがないよりは、変なところにスラッシュが入っていた方が読解に与える悪い影響は大きい。

	再現率	適合率
提案手法	52.66%	69.51%
土居らの手法	51.08%	47.23%
単純手法	47.80%	73.77%

表 2　再現率、適合率

セグメント長	2 以下	11 以上	平均長
学習データ	88.2	32.6	5.42
提案手法	55.2	61.4	6.39
単純手法	82.6	95.4	6.96

表 3　セグメント長の各個数と平均長

では、F 値にこだわらなければ、再現率を上げることもできるが、その分適合率が下がってしまう。

また、単純手法と比べると、再現率では上回っているが、適合率では下回っている。この原因はいくつか考えられる。

まず、単純手法がその定義通り、学習データにおいてスラッシュが入る確率が 5 割以上という限られた境界にしかスラッシュを入れないことである。そのため適合率が高く、スラッシュを入れる総数が少ないために再現率が低くなっている。

また、提案手法の $\alpha(l)$ と $\beta(r)$ が効いていることが挙げられる。このパラメタにより、セグメントは短くなりにくいようになっている、つまりスラッシュが入りにくくなっている。実際、2 以下のセグメント長の個数では、他よりかなり少ない。

他にも、学習データのスラッシュ付けの傾向が影響していることが考えられる。この参考書は、章が進むうちにセグメント長を長くしているようである。(このことは、スラッシュの挿入に「一貫性」がないことを意味しており、学習データとしてはあまりふさわしくないことになる。) 若い章のスラッシュ付けされた英文を見てみると、再現率と適合率の数値が表すように、スラッシュの誤挿入よりも、スラッシュの抜けが目立つ。つまり、学習データは章が進むうちにスラッシュが減っていくので、スラッシュが少ない方向にパラメタが学習されたと考えられる。

学習データの英文と提案手法によるスラッシュ付け英文を比較してみると、提案手法が学習データと異なったスラッシュ付けをしていても、決して悪くないものが多かった。以下はその例である。L が学習データ、P が提案手法によるスラッシュ付け英文である。

L: I was walking / from the university down the hill / to the train station.
P: I was walking from the university / down the hill to the train station.

L: High schools and businesses alike openly admit that they don't approve / of alteration of appearance of any kind.
P: High schools and businesses alike openly admit / that they don't approve of alteration of appearance of any kind.

どちらが良いスラッシュ付けかを言うことは出来ない。このように、スラッシュ付けに「正解」はない。よって、再現率と適合率やセグメント長による評価には限界があり、スラッシュ付けされた英文を用いた学習効果による評価が必要である。

4　まとめ

本稿では、英語学習法の 1 つであるスラッシュ・リーディング用の教材を作成するために、英文にスラッシュを挿入する統計的手法を提案した。評価実験では、少ない学習データからでも質の良いスラッシュ付けをできるという提案手法の高い性能が示された。今後の課題は、よりふさわしい学習データを手に入れ、再び評価実験を行うこと、また、実際にシステムを作成して学習効果を測ることである。

参考文献

[1] 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」, 研究報告 コンピュータと教育 No.75(2004 年 7 月)

[2] 松本茂, 「らっくらく英文解釈」, 七寶出版 (2001)

[3] Apple Pie Parser, http://nlp.cs.nyu.edu/app/

[4] 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」, 言語処理学会第 10 回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」論文集 pages37-40(2004 年 3 月)