2004－CE－75　(4)
2004／7／3

スラッシュ・リーディングのためのテキスト分割

土居 誉生, 隅田 英一郎
ATR 音声言語コミュニケーション研究所
{takao.doi,eiichiro.sumita}@atr.jp

外国語の学習法であるスラッシュ・リーディングを支援するために，テキストに自動的に区切記号を挿入する手法を提案する．本手法は，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対象文を分割する．英語学習参考書を使った実験では，本手法を使った場合，区切箇所の再現率，適合率，Ｆ値の面で従来手法を上回る結果が得られた．

Dividing Texts for Slash Reading

Takao Doi, Eiichiro Sumita
ATR Spoken Language Translation Research Laboratories
{takao.doi,eiichiro.sumita}@atr.jp

To assist foreign language learners in reading, we propose a method for placing slashes into reading texts. The method is based on a corpus and utilizes N-gram statistics and text similarity. In our experiments, the recall, precision and F-measure of slashes were evaluated. The proposed method achieved much better results than a previous method.

1　はじめに

スラッシュ・リーディングは，母国語と言語構造の異なる外国語を習得するために有効な学習法である．本稿では特に日本人の英語学習に焦点をあてることにする．英語と日本語は構造が大きく異なっている．日本人は英文を理解する際，英語の構文構造から日本語の構造を頭の中で再構築しがちである．この構造変換は英文理解過程の遅延を引き起こし，リーディングやリスニングに支障をきたす場合がある．この問題は，英文を元の語順のまま理解する能力を身に着けることができれば解決される．スラッシュ・リーディングはこの能力を獲得するための訓練法である．スラッシュ・リーディングでは，学習者は”/”（スラッシュ）によってチャンク（句や節などの意味的なかたまり）に区切られた英文を滞りなく読み進めるよう努める．その際，チャンクの順序を入れ替えて理解しようとしてはいけない．スラッシュにより区切られた英文の例を図1に示す．

I was passing / the small tobacco shop / when I heard a motor scooter racing down / from behind me.(松本, 2001)

図 1: スラッシュで区切られた文の例

この学習法はスラッシュの入った特殊なリーディング教材を必要とするが，あらかじめスラッシュの入った教材は限られている．(田中・

冨浦，2004) は，英文に自動的にスラッシュを挿入する機能を持つスラッシュ・リーディング支援システムを提案している．このようなシステムがあれば，学習者は自分の気に入った英文書を選び，それを教材としてスラッシュ・リーディングによる学習を進めることが可能となる．

本稿では，スラッシュ・リーディング支援システムにおいてスラッシュ挿入箇所を決めるための新たな手法を提案し，実験を通して従来の手法に対する優位性を確かめる．

以下，2章で従来法と提案法の概要，3章で提案法の詳細，4章で実験による両手法の評価について述べる．

2 スラッシュ挿入法

2.1 従来手法の概要

スラッシュ・リーディングのためのスラッシュの自動挿入法に関する先行研究 (田中・冨浦，2004) では，英文の依存構造に基づいた手法が使われている．この手法は，文読解プロセスに関する心理学的モデルに基づいている．このプロセスでは，文を構成する表現のチャンクが短期メモリ内に記憶される．文頭から 1 語ずつ読解処理を進めて行くにつれて，メモリ内のチャンクの個数は増減する．チャンクの個数が減るのは複数のチャンクが 1 つにまとめられる場合である．この手法では，依存構造解析に従ってチャンクをまとめる．ある単語まで処理を進めたとき，チャンクの個数が増えなければ，その単語の直後をスラッシュ挿入箇所の候補とする．得られた候補の中から，チャンク長の制約，依存関係によるまとめやすさ，記憶の負荷を考慮してスラッシュ挿入箇所を決定する．

2.2 提案手法の特徴

上記従来法では依存構造解析を利用しているが，我々は，依存構造解析や構文解析を用いないコーパスに基づく手法を提案する．本手法では，コーパスから作られた統計的言語モデル，

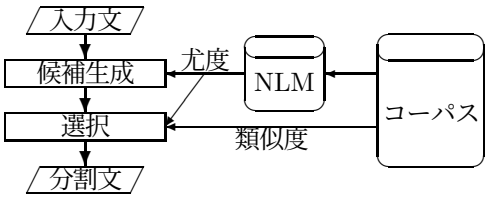

図 2: コーパスベース文分割の構成

および，コーパスを基準としたテキスト類似度を使って，対象文を分割する．分割のための知識はコーパスから自動的に獲得され，人手作成のルールは必要なく，依存構造解析器や構文解析器にも依存しない．

3 コーパスベース文分割

この章ではまず，提案手法をコーパスベース文分割手法，すなわち，文の集まりであるコーパスを利用して，与えられた文を複数の文に分割する手法として定義する．その後，スラッシュ挿入箇所決定問題への応用について説明する．

文分割に関する多くの研究では，N グラムなど単語連接の統計的な特徴に基づいて分割点を決めている (Berger et al., 1996; Lavie et al., 1996; 竹沢・森元，1999; 中嶋・山本，2001)．これらの研究と同様に本提案手法でも N グラム言語モデル（N-gram Language Model，以下 NLM と記す）(北，1999) を利用する．NLM は，注目点前後の数語の特徴を根拠に，その点で文を分割すべきかどうかの指標を提供する．その判断材料の局所性を補うため，我々は，対象文のより広い範囲を考慮に入れた類似度を新たな指標として導入する．以降の説明では，文を分割して得られる文のリストを分割文と呼ぶことにする．リスト中の文は元の文の部分文である．分割文は，まったく分割されなかった場合を含め，1 つまたは複数の文を要素として持つ．本手法の構成を図 2に示す．

3.1 分割文の尤度

コーパスから作られた NLM によって文の尤度を計算することができる．文の尤度を用い，

分割文の尤度 $Prob$ を式 (1) で定義する．ここで，P は NLM による文の尤度，N は N グラムの N，S は分割文，$|s|$ は文 s の単語数を表す．$Prob$ は分割文の要素である文の尤度の積を N グラムの個数で正規化した値である．N グラムは文頭，文末に加えられた擬似的な語を考慮している．

$$Prob(S) = \left(\prod_{s \in S} P(s) \right)^{\frac{1}{\sum_{s \in S} (|s| + (N-1))}} \quad (1)$$

ある点で分割すべきかどうか判断するため，その点で分割する場合としない場合の分割文の尤度を比較する．分割文の要素としての文を含め，文の尤度を計算する場合，文頭と文末に擬似語を加える．例えば，"I was passing" という文の尤度をトライグラム言語モデルで計算すると下のようになる．ここで，p(z | x y) は x y の 2 語が続いた後に z が出現する確率である．SOS と EOS はそれぞれ文頭と文末の擬似語を示す．

P(I was passing) = p(I | SOS SOS) ×
p(was | SOS I) × p(passing | I was) ×
p(EOS | was passing) ×
p(EOS | passing EOS)

3.2 分割文の類似度

2 つの文の類似度を単語列間の編集距離をもとに定義する．この編集距離には意味的要素が加味されている．編集距離は最小値 0，最大値 1 となるように正規化され，1 から編集距離を引いた値が類似度となる．2 文間の類似度の定義は式 (2) で与えられる．この式で I，D はそれぞれ，挿入，削除の個数を表している．Sem は置換された 2 語間の意味距離を表す．置換は同じ品詞の内容語の間でのみ許される．Sem はシソーラスを使って計算され 0 以上 1 以下の値をとる．Sem は，式 (3) のように，K（シソーラス中，2 語の共通の親となる最下層ノードのレベル）を N（シソーラスの階層構造の高さ）で割った値である (Sumita and Iida, 1991)．

$$Sim_0(s_1, s_2) = 1 - \frac{I + D + 2 \sum Sem}{|s_1| + |s_2|} \quad (2)$$

$$Sem = \frac{K}{N} \quad (3)$$

Sim_0 を使った分割文の類似度 Sim を式 (4) で定義する．この式において S は分割文であり，C は前提条件として与えられたコーパス，つまり文の集合である．Sim はコーパスに対する部分文の類似度を文長で重みを付けて平均した値である．コーパスに対する文の類似度は，この文とコーパス中の文との類似度の最大値である．

$$Sim(S) = \frac{\sum_{s \in S} |s| \cdot \max\{Sim_0(s, c) | c \in C\}}{\sum_{s \in S} |s|} \quad (4)$$

3.3 分割文候補の生成

Sim を計算するにはコーパスからの最類似文検索に相当する処理が必要となる．この検索処理はクラスタリング (Cranias et al., 1997) や単語グラフ上の A* アルゴリズム (土居ら, 2004) により効率的な実装が可能である．しかしながら，特にコーパスの規模が大きい場合，Sim は $Prob$ に比べより大きな計算コストを要する．本手法では，まず $Prob$ のみを使って分割文候補を生成し，その候補の中から $Prob$ と Sim を使って分割文を選択する．候補生成プロセスでは次の (i)，(ii)，(iii) の分割文を候補として生成する．

(i) 元の文のみからなる分割文

(ii) 要素数が 2 であり，$Prob$ が (i) よりも小さくない分割文

(iii) (ii) の 2 つの要素文それぞれについて分割文候補の生成処理を再帰的に適用した結果を組み合わせて得られる分割文

候補生成プロセスでは，$Prob$ の値が分割無しの場合より小さくなる分割文は候補から除外されることになる．

3.4 分割文の選択

分割文候補の中から，$Prob$ だけでなく Sim も使った基準で最適な候補を選択する．指標として式 (5) で定義する $Score$ を使用する．$Score$ は $Prob$ と Sim の積である．式中 λ は 0 以上 1 以下の定数であり，Sim の重みを表す．特に λ が 0 のときは，本方式は $Prob$ しか使わないことになる．

$$Score = Prob^{1-\lambda} \cdot Sim^{\lambda} \tag{5}$$

3.5 スラッシュ・リーディングへの応用

この文分割手法の効果は使用するコーパスの性質に依存する．もしコーパスが，完全な文ではなくチャンクから構成されていれば，本手法の導出する区切箇所がスラッシュ・リーディングに適した結果となることが期待できる．しかし，そのようなコーパスを新たに用意するには，チャンクを大量に集める必要がある．そのためには人手コストもしくは，テキストや文からチャンクを正しく抜き出す手法などが必要となる．特にチャンクを集めたコーパスを用意できない場合，短い文を多く含み，その中には句や従属節のみからなる文もあるようなコーパスを代替的に利用する方法がある．

3.6 例

図 3に生成された分割文候補の例を示す（英文は (松本, 2001) より引用）．5 つの分割文候補が生成されている．左側の数字は $Prob$ による順位を示す．この例では $Prob$ 基準で 2 位の候補が $Score$ 基準で 1 位 となり解として選択される．

4 実験

提案手法の評価として，スラッシュ挿入箇所の再現率，適合率，F 値を計算した．比較のため従来手法 (田中・冨浦, 2004) についても実験を行った．

4.1 テストセットおよび評価尺度

実験では，スラッシュ・リーディング用にあらかじめスラッシュ記号が入れられた教材をテストセットとして利用した．この教材は高校生向けの英文解釈の学習参考書である．内容は外国人によって書かれた日本文化についての 15 話のエッセイからなる．テストセットは 485 文からなり，その平均文長は 14.54（語／文）である．スラッシュの総数は 834 である．

評価尺度としては，手法の実行結果のスラッシュ挿入箇所が参考書と一致する程度，つまり，再現率（recall），適合率（precision），F 値（F-measure）を使った．ここで，F 値は以下の式 (6) で定義する．

$$\text{F-measure} = \frac{2 \cdot \text{recall} \cdot \text{presition}}{\text{recall} + \text{precision}} \tag{6}$$

4.2 従来手法の実装

従来手法の実装がウェブサイト[1]で実行可能な形で公開されている．ここでは，統語範疇の境界を考慮しない版と考慮する版の 2 つが用意されている．それぞれの版について，2 つのパラメータの値を決める必要がある．1 つのパラメータは，スラッシュで区切られたチャンクの単語数の上限を示し，5 から 9 までを値の範囲とする．もう 1 つのパラメータは語彙チャンクとして使う辞書セットを 4 種類の中から指定するものである．この手法中では語彙チャンクは 1 語として扱われている．実験では，2 種類の版それぞれについて，2 つのパラメータの全ての値の組み合わせでスラッシュ挿入処理を実行し結果を評価した．

4.3 提案手法の実装

4.3.1 学習コーパス

実験では ATR の旅行会話基本表現集とバイリンガル旅行対話データベースを使った (Takezawa and Kikui, 2003)．どちらも話し言葉で表現された文からなる対訳コーパスであ

[1] http://lengua.cc.kyushu-u.ac.jp/english/sr/

1.	He looked at me / as if I were some stranger spirit that had crawled / out of the mountains.
2.	He looked at me / as if I were some stranger spirit / that had crawled out of the mountains.
3.	He looked at me / as if I were some stranger spirit that / had crawled out of the mountains.
4.	He looked at me / as / if I were some stranger spirit that had crawled out of the mountains.
5.	He looked at me as if I were some stranger spirit / that had crawled / out of the mountains.

図 3: スラッシュで区切られた文の例

文数	224,535
語彙数	14,548
平均文長（語／文）	7.08
パープレキシティ	27.58

表 1: 学習コーパスの統計情報

	再現率	適合率	F 値
再現率最良	.4784	.2902	.3612
適合率最良	.3453	.4217	.3797
F 値最良	.3849	.3905	.3877

表 2: 従来手法（統語範疇の境界を考慮しない版）の結果

	再現率	適合率	F 値
再現率最良	.4868	.2944	.3669
適合率最良 F 値最良	.3909	.4515	.4190

表 3: 従来手法（統語範疇の境界を考慮する版）の結果

る．両者の英語部分を合わせて学習コーパスとして使用した．この学習コーパスからNLMを作るとともに，このコーパスに対する分割文の類似度を計算した．このコーパスの統計情報を表1に示す．ここでパープレキシティは単語トライグラム・パープレキシティである．コーパス中の文の長さは平均7（語／文）と短い．また，話し言葉コーパスであるため，句や従属節のみからなるような不完全な文も含まれている． 3.5節で述べたようにスラッシュ・リーディング用の分割処理のためには適当な学習コーパスだと考えられる．

4.3.2 その他の設定

Prob の計算に用いたNLMは単語トライグラムモデルであり，スムージングにはグッド・チューリング推定法を使った．

テストセットの1文あたりの生成される分割文候補数は30以内とした．また分割文の要素数，つまりスラッシュによって区切られたチャンク数，にも上限を設けた．指定された整数で文長を割った値以上の最小整数値を，その文のチャンク数の上限とした．ここで指定する整数は4から8までとした．式 (5) の λ は0または2/3とした．以後，λ が0の条件を「*Prob* のみを使った場合」，λ が2/3の条件を「*Score* を使った場合」と呼ぶ．

シソーラスは角川類語新辞典 (大野・浜西, 1984) のシソーラス構造に準拠したものを使った．その英語見出し語数は80,250である．

4.4 従来手法の結果

表2は従来手法の「統語範疇の境界を考慮しない版」について，パラメータ値の全組合せのうちで，再現率，適合率，F 値のいずれかが1番良かった結果を示している．表3は従来手法の「統語範疇の境界を考慮する版」についての同様の結果を示している．統語範疇の境界を考慮することで，より良い結果が得られている．

4.5 提案手法の結果

表4に提案手法の *Prob* のみを使った場合の結果を，表5に *Score* を使った場合の結果を示す．これらの表で x はチャンク数の制約（4.3.2項）を示す．チャンク数の上限は，文長を x で割った値以上最小の整数となる． *Score* を使った場合は *Prob* のみを使った場合よりも

x	再現率	適合率	F 値
4	.6367	.3590	.4591
5	.5576	.4122	.4740
6	.4736	.4379	.4551
7	.4209	.4795	.4483
8	.3825	.5104	.4373

表 4: 提案手法（*Prob* のみを使った場合）の結果

x	再現率	適合率	F 値
4	.6571	.3708	.4740
5	.5683	.4202	.4832
6	.5108	.4723	.4908
7	.4424	.5041	.4713
8	.4101	.5472	.4688

表 5: 提案手法（*Score* を使った場合）の結果

良い結果が得られている．いずれの場合も提案手法の結果は従来手法を上回っている．

4.6 類似度利用の効果

485 のテスト文中 442 文で複数の分割文候補が生成された．候補数を 30 に制限した実験条件下で，485 のテスト文について，分割文候補の平均個数は 23.7 であった．F 値の最良となる条件（表 5で $x = 6$ の場合）では，179 文において，*Score* による入れ替え，つまり *Prob* で 2 位以下の候補の *Score* での選択が発生した．この場合 *Score* 基準 1 位の候補の最も悪い *Prob* 順位は 28 である．表 6には，*Prob* による N ベストの N つまり候補数上限を変更したときの，*Score* による入れ替えが起る場合の数と評価指標の値を示している．この表から，入れ替えが起るとしても *Prob* 順位が上位の候補が選ばれる場合が多いことが分る．N が小さくても *Score* 利用による効果が評価値に現れ，N=10 で最大値に達している．候補数が小さくても効果があり，効果を得るための処理コストを抑えることが可能である．

N	入替	再現率	適合率	F 値
1	0	.4736	.4379	.4551
2	81	.5012	.4634	.4816
5	141	.5072	.4690	.4873
10	163	.5120	.4734	.4919
28	179	.5108	.4723	.4908

表 6: *Prob* による N ベストと性能

5 考察

5.1 評価方法

提案手法であるコーパスに基づくスラッシュ挿入は，学習参考書をテストデータとした実験で，再現率，適合率，F 値に関して，従来手法である依存構造解析に基づく手法を上回ることを示した．この評価指標は絶対的なものではないが，スラッシュ挿入箇所についての模範解答例との一致の度合いを示し，結果の良否を判定する判断材料となり得る．特に，両手法において性能向上のために有効であると考えられる方策，統語範疇の境界の考慮，文の類似度の利用のいずれもが評価指標値の向上につながっている点も，当指標と結果の良否との相関を支持する．このため評価指標で上回る提案手法が生成するスラッシュがより妥当なものであることが示唆される．

しかしながら，やはりこの評価指標には限界がある．次の例では，(a) は参考書 (松本, 2001) 記載のスラッシュ付き英文であり，(b) は提案手法の実行結果である．いずれが良いかを一概に言うことはできない．正解は一意には決められない．

(a) I lived in Japan almost two and a half years / without owning a car.

(b) I lived in Japan / almost two and a half years / without owning a car.

今後さらに，精度や有用度についての人間による主観的な評価も含め，何らかの別の評価方法を，効果とコストの兼ね合いを考慮しながら検討する必要がある．

5.2 学習コーパス

実験では，短い文からなる話し言葉コーパスを学習コーパスとして用いることにより，スラッシュ挿入のために適当な文分割を実現することができた．より分析を進めるため，他のコーパス，特に書き言葉コーパスを学習コーパスに使った実験も行う必要がある．

またスラッシュ・リーディング用の既存の教材は限られているとしても，少量のスラッシュ付きテキストの学習コーパスとしての利用も考えられる．これらのテキストは *Sim* の計算に利用できる． *Prob* の計算については，スラッシュ付きテキストから作られる NLM と，他のコーパスから作られる NLM を，線形補完等により混合して使う方法が精度向上に有望である．

5.3 関連研究

本稿では，提案手法としてコーパスに基づく文分割手法を，比較手法として，既存のスラッシュ・リーディング支援システムで使われている依存構造解析に基づく手法を取り上げた．他の関連する分野として節境界の検出の研究があげられる．この分野の研究は，(Leffa, 1998) のような人手で作成したルールを使う手法と，(Tjong Kim Sang and Dejean, 2001) のようなツリーバンクからの機械学習による手法に分けられる．ルールやツリーバンクの作成・保守にはコストがかかる．本稿の提案手法は，人手作成ルールもツリーバンクも必要としないためコストの点から有利である．本稿では英語を対象としたが，ツリーバンクや高精度パーザが利用できない言語に適用する場合，提案手法は有利になる．しかし，英語のように言語資源が豊富な言語を対象とする場合，節境界検出に関する先行研究との比較や提案手法との組合せは今後の課題として必要である．

6 おわりに

外国語の学習法であるスラッシュ・リーディングを支援するために，テキストに自動的に区切記号を挿入する手法を提案した．本手法は，コーパスに基づき，N グラム言語モデルと文の類似度を使って対象文を分割する．英語学習参考書を使った実験では，本手法を使った場合，区切箇所の再現率，適合率，F 値の面で従来手法を上回る結果が得られた．実験結果を踏まえて，評価方法，学習コーパス，他の節境界検出手法に関連した今後の課題について述べた．

謝辞

類語新辞典の使用許可をいただいた角川書店様に感謝いたします．なお，本研究は情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものです．

参考文献

A.L. Berger, S. A. Della Pietra, and V. J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):1–36.

L. Cranias, H. Papageorgiou, and S. Piperidis. 1997. Example retrieval from a translation memory. *Natural Language Engineering*, 3(4):255–277.

土居誉生，隅田英一郎，山本博史. 2004. 編集距離を使った用例翻訳の高速検索方式と翻訳性能評価. 情報処理学会論文誌, 45(6).

北　研二. 1999. 確率的言語モデル. 東京大学出版会.

A. Lavie, D. Gates, N. Coccaro, and L. Levin. 1996. Input segmentation of spontaneous speech in janus: a speech-to-speech translation system. *Proc. of ECAI-96 Workshop on Dialogue Processing in Spoken Language Systems*, pages 86–99.

V.J. Leffa. 1998. Clause processing in complex sentences. *Proc. of 1st International Conference on Language Resources and Evaluation*, pages 937–943.

松本　茂. 2001. 直読・速読・多聴式 らっくらく英文解釈. 七寶出版.

中嶋秀治，山本博史. 2001. 音声認識過程での発話分割のための統計的言語モデル. 情報処理学会論文誌, 42(11):2681–2688.

大野　晋，浜西正人. 1984. 類語新辞典. 角川書店.

E. Sumita and H. Iida. 1991. Experiments and prospects of example-based machine translation. *Proc. of 29th Annual Meeting of ACL*, pages 185–192.

竹沢寿幸，森元　逞. 1999. 発話単位の分割または接合による言語処理単位への変換手法. 自然言語処理, 6(2):83–95.

T. Takezawa and G. Kikui. 2003. Collecting machine-translation-aided bilingual dialogues for corpus-based speech translation. *Proc. of EUROSPEECH*, pages 2757–2760.

田中省作，冨浦洋一. 2004. 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. In 言語処理学会第 *10* 回年次大会ワークショップ「*e-Learning* における自然言語処理」, pages 37–40.

E.F. Tjong Kim Sang and H. Dejean. 2001. Introduction to the conll-2001 shared task: Clause identification. *Proc. of CoNLL-2001*, pages 53–57.