場面を単位とする言語モデルに向けて

小嶋 秀樹　　伊藤　昭
郵政省 通信総合研究所 関西先端研究センター

本論文では，場面を単位としてテキストを予測するモデルを提案する．場面は，局所的な文脈をもったセグメントであり，意味的な段落と考えてよい．ひとつの場面の内部では，同じ対象が同じ状況のもとで語られる．場面の内部は，このように自由度が小さく，十分に予測可能である．しかし，隣接する場面間の推移や全体的な場面展開は，より大きな自由度をもつため予測が難しい．そこで本研究では，(1) テキストを場面に分節することと，(2) 場面ごとの文脈を抽出すること，(3) この文脈にもとづいて後続単語を予測することを組み合わせたモデルを考察する．本モデルによって，個々の場面がもつ局所的な文脈に依存して後続単語を予測することができる．

1 はじめに

自然言語を統計的な予測モデルとして記述する研究が数多く行なわれている．たとえば n-gram による予測モデル [1] では，単語（または単語の意味クラス）の n 組ごとの出現確率をコーパスにもとづいて最尤推定している．一般に，予測精度を上げるためには，大きなコーパスを必要とする．

このようなコーパスにもとづく最尤推定から得られる情報は，コーパスの意味的な非一様性を考慮していない．たとえば，コーパスに含まれる各テキスト（記事・論文など）はジャンルごとに一定の特徴をもつ．ジャンルを特定できれば，テキストに現われる単語を精度よく予測できる．Sekine [2] は，ここでいうジャンルを sublanguage として統計的に定義し，sublanguage の違いを考慮した予測モデルを提案している．

本論文では，ジャンルのような大きな単位ではなく，テキストよりも小さな単位である「場面」に注目した予測モデルを提案する．場面は局所的な文脈をもったセグメント（テキスト上の部分区間）であり，テキストは場面列に分節される．ひとつの場面の内部では，同じ対象が同じ状況のもとで語られる．場面の内部は，このように自由度が小さく，そこに現われる単語を高い精度で予測できる．一方，隣接する場面間の推移やテキスト全体としての場面展開は，より大きな自由度をもつため，その予測は難しい．

本モデルでは，自由度の小さな場面内部と自由度の大きな場面間推移とを区別し，個々の場面ごとにその局所的な文脈にもとづいて，漸進的 (incremental) な単語予測を行なう．まずテキストの先頭に場面開始マーカを置き，テキストを読み進みながら後続単語をつぎのように予測する．

- マーカから現在位置までをひとつの場面として，後続単語を予測する．
- もし場面境界に達したならば，その位置にマーカを移す．

この予測手続きは，人間にとって自然なものといえる．我々がテキストを読むとき，いま読んでいる場面の文脈にもとづいて後続テキストを予測する．もし新しい場面に入ったならば，そこから新しい文脈を読み取りはじめる．

以下，第2節では，場面の性質とテキストとの関係について説明する．第3節では，場面のもつ局所的な文脈を利用した単語予測モデルについて説明する．このモデルは，場面境界を認識することと，場面からその局所的な文脈を抽出することを同時に行なう．最後の第4節では，全体をまとめ，今後の課題を述べる．

2 場面とテキスト

テキストは，単語や文の単なる列ではなく，何らかの目的のために組織化されたテキスト構造をもってい

る [3]．個々の単語や文は，それらをとりまく構造単位のなかで役割を与えられ，はじめて意味が定まる．テキスト構造は，個々の単語や文を解釈するための，また照応・省略・曖昧性などを解決するための文脈（制約・選好など）をもたらす [4]．

場面は，このような文脈に深く関係したテキスト構造単位であり，局所的ではあるが意味的な一貫性をもったセグメント[1]である．ひとつの場面では，同じ対象（人物・事物など）が同じ状況（場所・時刻・視点など）のもとで語られる [5,6]．このような対象と状況についての一貫性によって，それぞれの場面は特徴づけられている．

この一貫性の制約は，場面内部における単語出現の自由度が小さいことを意味する．場面内部では同じ対象が同じ状況のもとで語られるため，そこには出現する単語は偏った分布をとる．もちろん同じ単語が繰り返し出現する保証はないが，互いに意味的な関連をもった単語が出現しやすい [7]．このことは，場面の冒頭部分を読めば，後続部分に現われる単語をある程度予測できることを示唆する．

一方，場面の列として構成されるテキストは，場面の展開（すなわち文脈の変化）を表現したものと考えられる．たしかにテキストは何らかのテーマについて書かれたものであるが，場面内部に比べて単語出現の自由度は大きい．このため，テキストの一場面を読んでも，それに後続する場面を予測すること（出現単語を予測すること）は難しい．

以上から，場面をテキストから切り出し，場面ごとの局所的な文脈を捉えることによって，精度の高い単語予測が可能となることがわかる．次節では，この考えにもとづいた単語予測モデルについて述べる．

3 場面にもとづく単語予測モデル

ここで提案する単語予測モデルは，テキストを漸進的に読み進みながら，(a) いま読んでいる場面がどこから始まっているのかを認識し，(b) この場面（正確には，場面のすでに読んだ部分）にもとづいて後続単語を予測する．具体的には，テキスト T を単語列 $w_1, \cdots, w_N$ とするとき，つぎのような手続きによって後続単語を予測する (図1)．

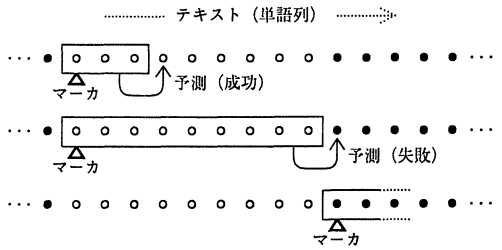

図1．場面にもとづく後続単語の予測

1. 現在位置を $i=0$ とし，ここにマーカ（場面開始位置を表わす）を置く．
2. マーカから現在位置までの部分テキストを場面として，後続単語 w_{i+1} を予測する．（場面が空のときは，予測は必ず失敗するものとする．）
3. この予測結果（成功・失敗）から，位置 $i+1$ が新しい場面の始まりであると認識された場合，マーカを位置 $i+1$ に移す．
4. i に1を加えて，2へ戻る．

以下では，ここで必要となる (a) 場面境界を認識することと，(b) 場面にもとづいて後続単語を予測することについて説明する．

3.1 場面境界の認識

テキストを局所的な文脈をもった場面に分節する手法がいくつか知られている．たとえば cue phrase [3] は場面変化を示唆するが，あくまで間接的な手がかりである．場面変化を直接的に捉える手法としては，テキストの各位置における新出単語の割合をプロットした VMP [8] や語彙的結束性（単語間の意味関係）の密度をプロットした LCP [5,6]，単語の出現頻度の偏りを利用した手法 [9]，このような統計的な指標に文法的な制約を加えた手法 [10] や表層的な手がかりを加えた手法 [11] などがあげられる．

ここでは，先行テキストにもとづいて後続単語を予測し，このときの予測誤差（予測の成功・失敗）から，場面変化を捉える手法を提案する．図2に示すように，現在の場面（マーカから現在位置までの部分テ

[1] が100語程度の場面が観察された [5,6]．新聞記事や論文では，より長い場面が予想される．

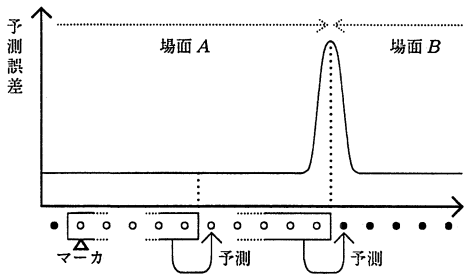

図2．予測誤差による場面境界の認識

キスト）にもとづいて後続単語を予測するとき，つぎのことが成り立つ．

- 場面内部では，局所的な文脈が一定であるため，予測誤差が小さくなる．
- 場面境界では，文脈は不連続に変化するため，予測誤差が大きくなる．

この原理にもとづいて，予測誤差が極大となる位置（またはある閾値を超えた位置）から場面境界を認識することができる．実際の予測誤差の計算については次項で述べる．

3.2 後続単語の予測

場面（マーカから現在位置までの部分テキスト）が与えられたとき，後続する単語を「場面依存的な単語連想」によって予測する．ここでは，場面 S に依存して任意の2単語 w, w' 間の意味距離 $d(w, w'|S)$ を計算する手法 [12,13] を概説し，この意味距離にもとづく連想によって後続単語を予測することを説明する．

場面 S に依存した意味距離 $d(w, w'|S)$ は，つぎのような手続きをとおして計算される．

1. **意味ベクトルの生成：**
   語彙 V の各単語を意味ベクトルによって表現する．この意味ベクトルは，英語辞書から構成された意味ネットワーク上の活性伝播によって得られる2851次元のベクトルを，主成分分析によってより小さな次元数（実験では281次元）に変換したベクトルである．
2. **意味空間のスケール変換：**
   場面 S を単語集合（重複をゆるす）とみなし，意味ベクトル空間における S の分布にもとづいて，この空間の各次元を拡大・縮小する．このスケール変換は，意味空間が S の分布を最も効率的に表現する（各次元での S のバラツキが等しくなる）ように行なわれる．

表1．$S=\{\texttt{bus}, \texttt{car}, \texttt{railway}\}$ からの連想

w	$\bar{d}(w, S)$
car_1	0.1039
railway_1	0.1131
bus_1	0.1141
carriage_1	0.1439
motor_1	0.1649
motor_2	0.1949
track_2	0.1995
track_1	0.2024
road_1	0.2038
passenger_1	0.2185
vehicle_1	0.2274
engine_1	0.2469

このスケール変換によって，意味ベクトル空間は場面 S の意味的な分布（広がり）に適応する．2単語 w, w' 間の意味距離 $d(w, w'|S)$ は，スケール変換された空間における意味ベクトル間のユークリッド距離として計算される．

この場面依存的な意味距離によって，与えられた場面 S からその関連単語を連想する．ここでは，意味空間における S の重心からみた各単語 w の意味距離 $\bar{d}(w, S)$ を計算し，この意味距離が小さい単語ほど連想されやすいとした．たとえば，場面 S として $\{\texttt{bus}, \texttt{car}, \texttt{railway}\}$ を与えたとき，$\bar{d}(w, S)$ が最も小さくなる12単語を表1にあげる．与えられた場面から連想すべき方向性（つまり文脈）を抽出し，それにもとづいて関連単語を連想できることがわかる．

このような場面依存的な単語連想にもとづいて，後続単語を予測する．ここでは，意味距離 $\bar{d}(w, S)$ の昇順に語彙 V をソートすることをもって，後続単語の予測とする．V は場面 S との意味的な関連性のつよさによってソートされ，その上位の単語ほど後続単語として現われやすいと考えられる．そこで，ここでは後続単語の予測誤差 $e(w|S)$ を，

$$e(w|S) = r(w|S)/|V|$$

とする．ただし $r(w|S)$ は S を場面としてソートされ

たVにおけるwの順位とする．後続単語をうまく予測できれば$e(w|S)\approx 0$となる．

4 おわりに

本論文では，場面のもつ文脈を利用した単語予測モデルを提案した．場面とは局所的な文脈をもったセグメントであり，場面内部では同じ対象が同じ状況のもとで語られる．このため場面内部では単語出現の自由度が小さく，場面の一部分から他の部分を十分予測できる．

本予測モデルは，テキストを漸進的に読み進みながら，その時点での場面にもとづいて後続単語を予測する．場面の境界は，後続単語の予測が失敗した位置として捉えられ，もしあればこの位置で場面を切り替える．こうして得られた場面にもとづく単語予測は，場面からみた意味距離によって語彙をソートすることによって行なう．

本予測モデルは場面の切り出しと場面にもとづく単語予測を同時に行なう．これは人間がテキストを読むときの文脈の捉え方を自然な形で表現したものである．モデルの具体化を進める上で，現在つぎのような点が課題となっている．

- **場面分節の不安定性：**
  場面境界が認識されたとき，それ以降を新しい場面として単語予測を進める．この初期段階では場面が極端に短いため，後続単語をうまく予測できず，正しく場面を分節できないことがある．これを解決するには，場面の長さに下限を与える必要がある．
- **単語予測の不安定性：**
  現在の場面にもとづいて予測されるのは，後続する1単語だけである．しかし，後続単語の性質には揺らぎ（内容語・機能語の区別や場面との関連度の違い）を含んでいるため，その予測誤差は不連続的に変化する．これを解決するには，何らかの平滑化を行なう必要がある．

今後は，これら課題を解決し，モデルの実装と実際のテキストを使った評価を進めていく予定である．

参考文献

[1] Brown, P. F. *et al.*: Class-based *n*-gram models of natural language, *Computational Linguistics*, Vol.18, pp.467–479, 1992.

[2] Sekine, S.: A new direction for sublangauge N.L.P., 自然言語処理, Vol.2, pp.75–87, 1995.

[3] Grosz, B. J. & Sidner, C. L.: Attention, intentions, and the structure of discourse, *Computational Linguistics*, Vol.12, pp.175–204, 1986.

[4] 角田達彦・田中英彦：英語名詞の多義性解消における文脈としての場面情報の評価，自然言語処理, Vol.3, pp.3–27, 1996.

[5] Kozima, H.: Text segmentation based on similarity between words, in *Proceedings of ACL-93 (Ohio)*, pp.286–288, 1993.

[6] 小嶋 秀樹・古郡 廷治：単語の結束性にもとづいてテキストを場面に分割する試み，情報処理学会研究報告 NL95-7, pp.49–56, 1993.

[7] Morris, J. & Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text, *Computational Linguistics*, Vol.17, pp.21–48, 1991.

[8] Youmans, G.: A new tool for discourse analysis: the vocabulary-management profile, *Language*, Vol.67, pp.763–789, 1991.

[9] Hearst, M. & Plaunt, C.: Subtopic structuring for full-length document access, SIGIR-93 (Pittsburgh), 1993.

[10] Nomoto, T. & Nitta, Y.: A grammatico-statistical approach to discourse partitioning, in *Proceedings of COLING-94 (Kyoto)*, pp.1145–1150, 1994.

[11] 望月 源・本田 岳夫・奥村 学：複数の知識の組合せを用いたテキストセグメンテーション，情報処理学会研究報告 NL109-7, pp.47–54, 1995.

[12] Kozima, H. & Ito, A.: Context-sensitive measurement of word distance by adaptive scaling of a semantic space, in *Proceedings of RANLP-95 (Bulgaria)*, pp.161–168, 1995.

[13] 小嶋 秀樹・伊藤 昭：意味空間のスケール変換による動的シソーラスの実現，情報処理学会研究報告 NL108-13, pp.81–88, 1995.