文法を利用したN-gramモデルのタスク適応

伊東 伸泰* 荻野 紫穂* 新島 仁†
iton@trl.ibm.co.jp shiho@trl.ibm.co.jp niijima@m.u-tokyo.ac.jp

1 はじめに

N-gramに基づく言語モデルは、統計的言語モデルの代表として、言語処理分野、特に音声認識で広く用いられている。このモデルを作成するには、精度のよい統計パラメータを求めるのに十分な学習データが必要だが、そのような大量のテキストは入手が難しいことが多い。特に日本語テキストの場合、新聞など限られたドメイン以外の電子化テキストを大量に入手することは非常に困難である。この問題に対して、新聞などを用いて作成された言語モデルを、対象となる分野の少量のテキストを使ってその分野に適応させようとする試みが報告されている [Rudnicky95, 伊藤 96]。一方、ある対象分野において出現し得る単語や構文が非常に限られる場合は、文脈自由文法に基づく文法規則などで現象を記述して制御することが可能であるばかりか、より効果的なこともある [Periera92]。

しかし、実際の音声認識アプリケーションでは、現実に入力される文のほとんどが定型文で単語も限られているにも関わらず、文法では規定しきれない部分を無視できず、任意の単語列を受理する枠組が必要となることも多い。例えば、画像診断などの医療所見では、使用される文の90%以上が定型文だと言われるが、それに該当しないフリーコメントを排除することはできない。

本論文では、このような問題に対処するため、法医学レポートを対象に、定型表現を記述した文法規則から生成した文を*N*-gramモデルの学習に使用することを試みる。文法で記述された言語制約を*N*-gramと組み合わせてその枠組で表現する手法は、[Kilian95] や [Eckert96] でも提案されている。前者は、正規文法をbigramの枠組で表現し、その場合に辞書サイズを抑制するためのデータ構造を提案している。後者はtransition networkで記述された文法をclass-based *N*-gramモデルで表現している。どちらもその目的の主眼は文法の持つ制約を*N*-gramの枠組で利用したい、という点にある。

これに対し我々は、充分なコーパスが得られない分野において、文法から生成された文を*N*-gramの学習データまたはその一部として利用することにより、言語モデルを当該分野に適応化することを提案する。2章ではまず、対象分野である法医学レポートの概略と文法の記述方法、およびその*N*-gramへの適応を述べ、3章で評価実験の結果について触れる。

*日本IBM(株) 東京基礎研究所
†東京大学大学院医学系研究科法医学教室

2 法医学レポートとその文法記述

2.1 法医学レポートの概略

法医学レポートは検死解剖の結果を司法当局に報告するために作成されるもので、その多くは、解剖者の口述を書記が記録した後、専門のオペレータがタイプで書き起こす、という作業手順で作成される。最終的に提出されるのは印刷された紙だけで、電子化されたソフトコピーは保存されないのが常であり、大量の電子化テキストを手に入れるのは困難な状況にある。

法医学レポートに使用される文の多くは図1に挙げたような定型文である。

文の内容から分かるように、図1の(1-a)で"死体硬直"という単語は必ず使われるが、場所や程度を表す部分は他の適当な単語も入れ換えて使用できる。例えば、"肩関節"の代わりに"顎"を、"強度"の代わりに"中等度"を使った(1-b)のような文もよく使用される。(2)の文も同様で、場所や程度、または色に関する用語が少し違う文も法医学レポートには頻出する。こうした文は、簡単な文法規則で記述することができる。

一方、死体の傷・損傷の状態や内臓の変化などに関する記述は極めて多岐に渡っていてフリーコメントに近いため、全てのバリエーションを文法でカバーすることは不可能に近い。しかし、このような種類の文でも表現の長さを1文節か2文節程度に絞って見ると、ある一定の頻出する言い回しがあることが分かる。例えば、図1の(3-a)(3-b)などは、傷の記述によく見られる部分表現である。

2.2 文法の記述

図1の例文を見ると分かる通り、法医学レポートは専門用語や特殊な言い回しが非常に多く、部外者が文法を記述するのは困難である。このため、文法記述は法医学に携わる専門家に行なってもらう必要がある。法医学専門家は文法記述については必ずしも専門家ではない。従って文法記述形式は、文法記述にさほど慣れていない作業者にも直観的に理解しやすいものでなければなら

死体 硬直 は 肩関節 に 強度 発現 している	(1-a)
死体 硬直 は 顎 に 中等度 発現 している	(1-b)
死斑 は 背面 に 中等度 存し 色 紫赤色 指圧 により 消褪 しない	(2)
創底 は 骨膜 に 達し …	(3-a)
創底 は 筋肉内 に 留まり …	(3-b)

図 1: 法医学レポートの例文

```
＊死体:シタイ-硬直:コウチョク-は:ワ-(左右)-(場所)-(に)-(見える)
[左右]
左 ヒダリ
右 ミギ
左右 サユウ
[場所]
肩関節 カタカンセツ
背面 ハイメン
顎 ガク
手根 シュコン
その他-の-関節 ソノタ-ノ-カンセツ
[に]
に ニ
で デ
[見える]
緩解-消失 カンカイ-ショウシツ
検出-不能 ケンシュツ-フノウ
```

図 2: 文法記述例

ない。この点を考慮した上で、語彙を V としてある文Sに関する文法は

$$\begin{aligned} S &\rightarrow S_1 S_2, \ldots, S_n \quad (1 \leq n) \\ S_i &\rightarrow a \mid C \quad (1 \leq i \leq n) \\ C &\rightarrow D_1 \mid D_2 \mid \ldots \mid D_m \quad (1 < m) \\ D_j &\rightarrow b_1 b_2, \ldots, b_l \quad (1 \leq j \leq m, 1 \leq l) \end{aligned}$$

で表せるものとした（ただし $a, b_k (k = 1, \ldots, l) \in V$）。これは簡潔には

$$\begin{aligned} S &\rightarrow C_1 C_2, \ldots, C_n \quad (1 \leq n) \\ C_i &\rightarrow a \quad (1 \leq i \leq n, a \in V^+) \end{aligned}$$

と表せる。これは 3 型文法であり、文を生成する有限オートマトンを容易に構成することができる。

図 2にこの文法に従った文法記述の例を挙げる。

“＊” で始まる文パターン中の、“()” で括られたタグの位置で使用できる単語のリストが、パターンの直後に並んでいる。例えば、“(場所)” の部分では “肩関節” “手根” などの単語を使うことができる。

2.3 *N*-gram への統合

N-gram モデルは、文の生起確率を N 個の連続した単語列の生起確率から近似する手法である。文 S を単語列 $S = w_1 w_2, \ldots, w_n$ とすると、その生起確率 $P(S)$ は次の式から計算される。

$$P(S) \approx P(w_0 w_1) \times \prod_{2}^{n} P(w_i \mid w_{i-1} w_{i-2})$$

ただし、我々の実験では、N は通常用いられる 3、つまり trigram とした。また、w_0 は文頭に割り当てられた特殊な記号である。

[Eckert96] は全ての S_i をユニークなクラスとして、一般の *N*-gram と組み合わせている。従って、文法から生成される候補パスと一般の *N*-gram モデルから生成されるパスが交わることはなく、確率の推定も容易である。

しかし、我々の目的はコーパスの不足を文法で補うことなので、このアプローチをとることはできない。そこで、文法から得られる *N*-gram 確率 P_G と一般コーパスから学習することによって得られる確率 P_C、および、対象分野の実テキストから得られる確率 P_D を補間することを考える。上で述べたように、我々の実験では trigram を使用するので、

$$\begin{aligned} P(w_i \mid w_{i-1} w_{i-2}) = {} & \lambda_0 P_D(w_i \mid w_{i-1} w_{i-2}) + \\ & \lambda_1 P_G(w_i \mid w_{i-1} w_{i-2}) + \\ & \lambda_2 P_C(w_i \mid w_{i-1} w_{i-2}) \end{aligned}$$

ここで $\lambda_0 + \lambda_1 + \lambda_2 = 1$ である。

各 λ は対象分野の実テキストから推定することになるが、パラメータの数が少ないので、推定用の実テキストはそれほど大量に必要ではない。また、低頻度事象への対応はより低次のモデル（我々の実験では bigram など）で補間するのが一般的である。従って上の式は、

$$\begin{aligned} P(w_i \mid w_{i-1} w_{i-2}) = {} & \alpha \quad (\lambda_0 P_D(w_i \mid w_{i-1} w_{i-2}) + \\ & \lambda_1 P_G(w_i \mid w_{i-1} w_{i-2}) + \\ & \lambda_2 P_C(w_i \mid w_{i-1} w_{i-2})) + \end{aligned}$$

$$\beta \quad (\lambda_0 P_D(w_i \mid w_{i-1}) + \lambda_1 P_G(w_i \mid w_{i-1}) + \lambda_2 P_C(w_i \mid w_{i-1})) +$$

$$\gamma \quad (\lambda_0 P_D(w_i) + \lambda_1 P_G(w_i) + \lambda_2 P_C(w_i))$$

と書き換えられる。ただし$\alpha+\beta+\gamma=1$である。

3 実験

3.1 使用データと方針

以上で述べた手法の有効性を確かめるための実験で使用した文法とデータの詳細について述べる。

2.1で述べたように、文法は法医学の専門家が記述している。実験時の文法は、これらは完全な文を表す規則以外に、文節や部分的な言い回しに関する規則も含めて618の規則がある。単語セットは、我々が一般言語モデルを作成した際に採用したもの（約40,000語[西村98]）から法医学専門家が不要と考える単語を取り除いて法医学用語約800語を追加したセット（18,143語）を使用した。

一般コーパスとしては、EDRコーパスから未知語の割合が非常に高い文を除いた約3,566,000語のテキストを使用した。法医学レポートの実テキストは18件（46,286語）用意し、そのうち1件（170文、2,216語）をテスト文とし、残り17件（44,070語）を9:1に分け、前者を実テキストデータとして、後者を$\alpha,\beta,\gamma,\lambda$の推定（closed）に使用した。単語セットのテスト文に対するカバレージは94.8%である。

前に述べたように、我々のアプローチは文法で記述された各ノードをユニークなクラスとして表現するというものではない。また、法医学の専門家には、同一のコンテキストがどこかに出現するかどうかを意識させることなく、自由に文法を記述してもらっている。従って、単純にBranching factorから計算することができない。これを計算するには

1. Branching factorに基づいて、各ルールの各パスに初期確率を付与し、N-gramの初期生起確率を計算
2. すべてのルールから遷移前の状態が等しいN-gramを集め、その確率の総和が1になるように正規化
3. 得られたN-gramの確率から、再び各パスの生起確率を計算し、すべてのパスについて正規化
4. 2から繰り返し

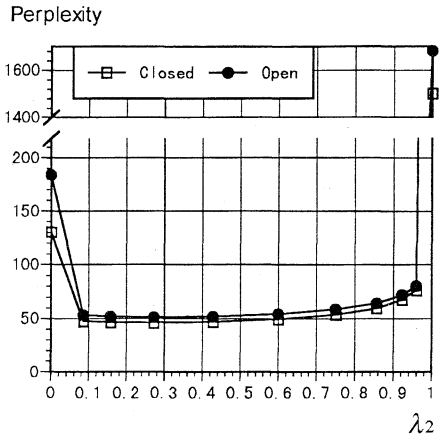

図3: 実テキストと一般文

という過程を経ることが考えられるが、実際には適切な初期確率を与えることは難しい。そこで各文法ルールから生成する単語列数の最大数(N_{max})と最低数(N_{min})を与え、当該ルールから生成されうるパターンの数がN_{max}を越える場合はランダムにN_{max}個を選択する。またN_{min}に満たない場合は同一単語列を複数回生成することにより生成される学習データの単語数を制御し、λを変化させた。

3.2 実験結果

まず、2.3の式のλ_1を0、つまり、文法から生成された文を全く使用せず、λ_2の値を変化させてperplexityとの関係を調べた結果を図3に示す。

図3を見ると、少量の法医学レポートの実データに、ほんの少しでも一般文を入れると、perplexityは飛躍的に下がる。そこからλ_2が0.6を越すあたりから少しずつperplexityはまた上がり始め、一般文だけの言語モデルでは、perplexityが最大になってしまうことがわかる。λ_2を0、つまり、一般文を全く使わずにλ_1の値を変化させる実験も行なったが、やはり同様に、文法から生成した文を少し入れるとperplexityは飛躍的に下がり、その後変化は緩やかになって、最小のperplexityは48.7だった。このことから、一般文も文法から生成された文も、単独では法医学レポート用のよい言語モデルのソースにはなり得ないが、法医学レポートの実データの不足を補うためにはよいデータとなり得ることがわかる。

次に、λ_1を0にした時に補間係数推定用データ（closed）に対して一番perplexityが改善されるλ_0とλ_2の割合を

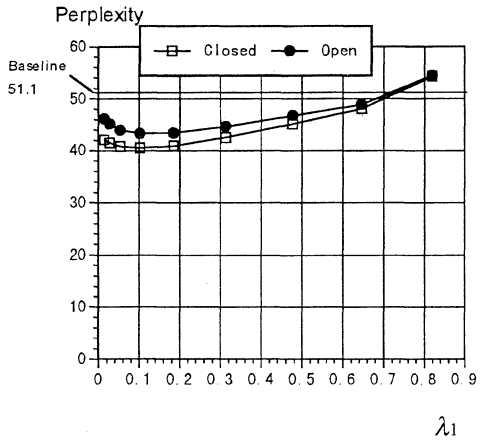

図 4: 実テキストと一般文と文法

言語モデル	医	医+文	医+般	医+般+文
Perplexity	184	48.7	51.1	43.3
誤認識率	5.7	4.4	4.9	3.9

医: 法医学レポート実テキスト
般: 一般文
文: 文法から生成した文
(語認識率は未知語を含まない)

表 1: 認識テスト

求めたところ、$\lambda_2/\lambda_0 = 0.37$ だった。このλ_0とλ_2の比率を変えないようにしたまま、λ_1の値を変えて、perplexityの変化を調べた結果を図 4に示す。図 4を見ると、文法から生成された文を補充することによって、perplexityが更に 15%ほど改善されることがわかる。

更に、それぞれのソースの組み合わせでperplexityが一番よかった割合の言語モデルを使って、実際に音声認識を行ない、単語の誤認識率を調べた。テスト文は上記の実験で使用した法医学レポートを使用し、法医学に携わる医者が離散発声で読み上げた。この際、話者適合は行なっていない。表 1 に実験結果を示す。

一般文と文法から生成された文との双方が認識率に貢献しているが、文法から生成された文を加えた言語モデルを使用したほうが、やや認識率が低い。言語モデルの大きさを N-gram の異なり数で比べると、法医学レポートに文法から生成した文を加えたものは、法医学レポートに一般文を加えたものの約 1/25 でしかない。このことから、一般文よりも文法から生成した文のほうが、実データの少なさを効率よく補うことができる、といえる。

更に、3 種類のデータをうまく混ぜると、更に誤認識率が下がることが表 1から読みとれる。

4 おわりに

文法から生成された文を使用して、言語モデルを法医学分野に適応する手法を示し、その有効性を実験的に明らかにした。この手法は、定型文の割合が多いが非定型文も排除できないさまざまな分野において利用可能であると思われる。今後は法律など他分野に本手法を適用する実験を行ないたい。

謝辞

本研究に有益なコメントを下さり、かつ認識タスク作成に御協力くださった日本アイ・ビー・エム西村雅史氏、田原義則氏に深謝する。

文献

[Rudnicky95] A. I. Rudnicky (1995) "Language modeling with limited domain data," *Proc. of ARPA Spoken Language Systems Technology Workshop*, pp. 66-69.

[伊藤 96] 伊藤, 好田 (1996)「対話音声認識のための事前タスク適応の検討」電子情報通信学会技術報告 NLC96-50, SP96-81.

[Periera92] F. Periera and Y. Schabes (1992) "Inside-outside Reestimation from Partially Bracketed Corpora," *Proc. of the Speech and Natural Language Workshop*, pp. 122-127.

[Kilian95] U. Kilian et al. (1995) "Representation of a finite state grammar as a bigram language model for continuous speech recognition," *Proc. of European Conf. on Speech Commun. and Tech.*, pp. 1241-1244.

[Eckert96] W. Eckert, F. Gallwitz and H. Niemann (1996) "Combining stochastic and linguistic models for recognition of spontaneous speech," *Proc. of ICASSP*, pp. 243-426.

[西村 98] 西村他 (1998)「単語を認識単位とした日本語の大語彙連続音声認識」情報処理学会音声言語情報処理研究会報告 20-3.