重複する素性を持つNグラム言語モデル

松原 勇介[†]　宮尾 祐介[†]　辻井 潤一[†‡]
[†]東京大学大学院情報理工学系研究科
[‡]School of Computer Science, University of Manchester
National Centre for Text Mining, UK

1　はじめに

統計的言語モデルは、音声処理や自然言語処理の分野で古くから注目されている研究対象である。統計的言語モデルは任意の単語列に対して確率値を与えることができ、その確率値が単語列の統語的・意味的な文らしさを近似していると考えられる。したがって、音声認識や機械翻訳など、最終的に文を出力する自然言語処理アプリケーションでは重要な部分を占めている。

統計的言語モデルの研究において、確率値のスムージングは重要な課題のひとつである。対象の単語が未知であった場合、その単語に適切な確率値を与えることは一般に困難であり、統計的言語モデルの性能を低下させる大きな要因となる。また、単語が未知でない場合でも、統計的言語モデルにおいてはパラメータ数に対して訓練データが不足していることが頻繁に起こるため、信頼できるパラメータを推定するためにもスムージングが必須となる。これらの問題を解決するための方法として、統計的言語モデルのスムージング手法が数多く提案されてきた [1]。

既存のほとんどのスムージング手法は、単語の頻度分布に由来する情報のみを用いている。しかし、単語の接頭辞や接尾辞、単語に用いられている文字種、外部辞書が与える意味クラスなども、単語の確率を求める際に大きな手がかりになりうるのは明らかである。しかしながら、このような単語自身の特性に関する多様な情報源を用いた統計的言語モデルはごく少数しか見られない。その理由のひとつに、Nグラム言語モデルのような単純な確率モデルでは、独立性の仮定があるために複雑な情報を確率モデルに取り込むことができないということが挙げられる。

本研究は、最大エントロピーモデル [2] を統計的言語モデルに適用することにより、接頭辞や接尾辞、文字種の情報など、単語自身の特性を用いたNグラム言語モデルを構築する。最大エントロピーモデルは、機械学習手法の一種で、さまざまな自然言語処理タスクに応用され、大きな成功を収めている。最大エントロピーモデルの利点として、独立性を仮定することなく任意の素性を定義できるということが挙げられる。したがって、上記のようなさまざまな単語の特性を確率モデルに用いることができる。

2　背景

統計的言語モデルで広く用いられているNグラム言語モデルの表現能力が小さいことは従来から指摘されている [3]。その一つの原因として、互いに独立でない素性を定義できないという点がある。これに対して現在までに、複数のモデルを線形補完する手法 [4] や、最大エントロピー法 [2] に基づいて互いに独立でない素性を定義するモデル [3] が提案されてきた。

線形補間の手法は、独立に訓練した言語モデルによって得られる確率値の重みつき平均を取る手法である。平均化によってスムージングの効果が得られる反面、モデル１つにつき１つの重みを定義するため、一つ一つの素性の重要さを表現できないという問題がある。

一方、最大エントロピーモデルは任意の素性を定義できるため高い表現能力を持つモデルであるが、現在までに研究されているのは、トリガーモデル [5] など長距離の依存関係を利用するものがほとんどである。長距離の依存関係を採り入れる場合、可能な素性の数が極めて大きくなる。最大エントロピーモデルの学習は高コストなため、大規模な学習データが必要な統計的言語モデルの学習では、学習コストが大きな問題と

素性の種類	定義
接頭辞 (N) (N=1)	長さ N の文字列に対応して定義され、単語の先頭とマッチするとき発火
接尾辞 (N) (N=1,2)	長さ N の文字列に対応して定義され、単語の末尾とマッチするとき発火
文字種 (X) (X=平仮名, 片仮名, 漢字, 英数字)	単語が文字種 X だけを含んでいるとき発火
単語	訓練コーパスにある単語に対応して定義され、 対象の単語全体とマッチするとき発火

表 1: 基本素性のテンプレート

なる。Rosenfeld [3] は、計算時間を実用的な範囲に抑えるために素性の数を制限し、N グラムモデルとの線形補完を行うことで、精度向上を達成した。

互いに独立でない素性を定義できるという点で最大エントロピーモデルは有用だが、トリガーモデルのような長距離の依存関係に注目するモデルは、必要な計算時間に対して十分な精度が得られているとは言い難い。我々は距離を長くすることよりもむしろ、取り込む素性を応用タスクに応じて設計することが重要だと考える。

クラスベース言語モデル [6] は、品詞等の単語クラスを用いる N グラム言語モデルである。単語クラスから単語が生成されるというモデルに基づいており、異なる素性を定義しているモデルと捉えることができる。しかし、クラスベースの言語モデルでは異なる種類の互いに重複しうるクラスを定義することはできず、またその他の素性、例えば接頭辞、接尾辞などの素性と同時に定義することはできない。最大エントロピーモデルではこのような問題はないため、より汎用的であるといえる。

3 提案手法

本研究で提案する手法は、最大エントロピー N グラム言語モデル [7] をベースにしている。本手法では、以下のように長さ $N-1$ の単語列 h の後の単語 w の確率を定義する。

$$p(w|h) = \frac{1}{Z(h)} \exp\left(\sum_i w_i f_i(w,h)\right)$$

ここで、f_i は素性関数、w_i は素性の重み、すなわちパラメータである。訓練データおよび f_i が与えられたとき、w_i は訓練データの尤度を最大化するように推定される。

本研究では、文脈 h によらず対象の単語ごとに得られる基本素性と、対象単語の直前の単語との組み合わせで作られる 2 グラム組み合わせ素性を用いる。可能な 2 グラム組み合わせ素性の数は大きいため、素性テンプレートを用いて抽出するものを制限する。また、訓練コーパスに現れない組み合わせは有効に訓練することができないため、あらかじめ除外する。

3.1 基本素性

文脈 h によらずに単語に対して与えられる基本素性を表 1 に示す。表 1 は素性のテンプレートを表し、実際には各文字列や文字種で具体化されたものが素性となる。本研究では、単語の綴りに関する素性（接頭辞、接尾辞、文字種）と、従来の言語モデルで用いられていた単語の出現の素性を用いた。将来的には、外部辞書を用いた単語の意味クラスの素性などを導入する予定である。

ひとつひとつの基本素性は複数の単語をまとめた単語クラスとして働くため、訓練コーパスで頻度が少ない単語の確率値を正確に推定できることが期待され、スムージングに貢献すると考えられる。

図 1 に基本素性の例を示す。この例は、単語「パーセント」に与えられる基本素性を示している。たとえば「接頭辞/ント」は基本素性テンプレート接頭辞によって得られた素性の一つである。

(接頭辞, 接頭辞), (接頭辞, 接尾辞), (接頭辞, 文字種), (接頭辞, 単語), (接尾辞, 接頭辞), (接尾辞, 接尾辞), (接尾辞, 文字種), (接尾辞, 単語), (単語, 接頭辞), (単語, 接尾辞), (単語, 文字種), (単語, 単語), (文字種, 単語)

表 2: 2 グラム組み合わせ素性のテンプレート

“**パーセント**” → 接尾辞/ント 接尾辞/ト 接頭辞/パ 字種/片仮名 単語/パーセント

図 1: 基本素性の例

(“**で**”,” **最適**”) → (接頭辞/で, 接頭辞/最) (接頭辞/で, 接頭辞/最適) (接頭辞/で, 文字種/漢字) (接頭辞/で, 単語/最適) (接尾辞/で, 接頭辞/最) (接尾辞/で, 接頭辞/最適) (接尾辞/で, 文字種/漢字) (接尾辞/で, 単語/最適) (単語/で, 接頭辞/最) (単語/で, 接頭辞/最適) (単語/で, 文字種/漢字) (単語/で, 単語/最適)

図 2: 2 グラム組み合わせ素性の例

3.2 *N* グラム組み合わせ素性

N グラム言語モデルにおける *N* グラムの範囲内の単語列に関する素性を、*N* グラム組み合わせ素性と呼ぶ。この素性は基本素性と位置の組み合わせとして定義される。3 グラムの場合、「現在位置の単語が X であり、1 つ前の単語の接尾辞が Y、2 つ前の単語の意味クラスが Z」などが *N* グラム組み合わせ素性の例である。

本研究では、2 グラムの組み合わせ素性を用いた。表 2 に 2 グラム組み合わせ素性のテンプレートを示す。ここで、(X, Y) は、1 つ前の素性 X と対象単語の素性 Y の組み合わせを表す。

図 2 に 2 グラム組み合わせ素性の例を示す。この例は、表 2 のテンプレートによって単語 2 グラム「で 最適」に与えられる組み合わせ素性を示している。カンマで区切られた対が、一つの組み合わせ素性を表している。たとえば (単語/で, 文字種/漢字) は (単語, 文字種) のテンプレートから得られる組み合わせ素性である。

4 実験

4.1 実験設定

訓練および評価のためのコーパスとして、日本語話し言葉コーパス [8] を用いた。日本語話し言葉コーパスは、講演や自由対話の音声と書き起こし文からなるコーパスである。本稿では、書き起こし文のうち、A01M0101 から A01M0196 を訓練に、A08M0528 を評価に用いた。訓練データは 78436 単語、評価データは 2542 単語である。

評価指標として、単語当たりテストセットパープレキシティ[9] を用いた。テストセットパープレキシティは、モデルが評価コーパスで単語を予測する際の候補数の期待値である。

頻度の少ない組み合わせ素性による過学習を防ぐため、頻度 10 未満の素性は無視した。

4.2 実験結果

用いた素性の種類（テンプレート）、パープレキシティ、および素性数を表 3 に示す。ここで、“*” は任意の基本素性を表す。例えば、“(*, 接尾辞)” は、表 2 に挙げた 2 グラム組み合わせ素性のうち、(接頭辞, 接尾辞)、(接尾辞, 接尾辞)、(文字種, 接尾辞)、(単語, 接尾辞) の 4 種類の素性を表す。

表 3 の結果より、多くの組み合わせ素性を用いることにより、パープレキシティを低減させることができることが分かった。例えば、“T1 T2 T3 T4 T5” の結果と “T1 T4 T5” の結果を比較すると、接尾辞・接頭辞の素性が有効に働いていることが分かる。

5 おわりに

本稿では、最大エントロピーモデルを適用することで、様々な重複する素性を用いることができる *N* グラム言語モデルを提案した。実験では、既存の言語モデ

略称	素性
T1	(単語, 単語)
T2	(*, 接尾辞)
T3	(*, 接頭辞)
T4	(単語,*)
T5	(*, 単語)

素性	パープレキシティ	素性数
T1 T2 T3 T4 T5	23.1819709	43406084
T1 T4 T5	25.8564496	23600944
T1 T2 T3	109.3714858	32392600
T1 T2	206.8334103	22257412
T1 T3	214.1990345	17087640

表 3: 実験結果

ルで用いられる単語自身の素性に加えて、単語の接頭辞、接尾辞、文字種の素性を用いた 2 グラムモデルを構築した。

実験では、日本語話し言葉コーパスを用いてテストセットパープレキシティを測定した。実験結果より、組み合わせ素性を用いることによってパープレキシティを削減できることを示した。しかし計算コストの問題により小規模なデータで訓練する実験しか行うことができなかった。今後手法の改良によって、より大きなデータでの実験を行う予定である。

参考文献

[1] Stanley Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[2] Adam L. Berger, Stephen D. Della Pietra, and Vincent J. D. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71, 1996.

[3] Ronald Rosenfeld. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. PhD thesis, School of Computer Science, Carnegie Mellon University, 1994.

[4] M.; Gish H. Iyer, R.; Ostendorf. Using out-of-domain data to improve in-domain language models. *Signal Processing Letters, IEEE*, Vol. 4, No. 8, pp. 221–223, August 1997.

[5] R. Lau, R. Rosenfeld, and S. Roukos. Trigger-based language models: a maximum entropy approach. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 45–48 vol.2, 1993.

[6] Peter F. Brown, Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479, 1992.

[7] Stanley F. Chen and Ronald Rosenfeld. A gaussian prior for smoothing maximum entropy models. Technical Report CMU-CS-99-108, School of Computer Science, Carnegie Mellon University, February 1999.

[8] 国立国語研究所, 情報通信研究機構 [編]. 日本語話し言葉コーパス, 2004.

[9] Peter E. and Brown. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, Vol. 18, No. 1, 1992.