クラスに基づいた単語間の依存関係の推定手法

W. R. ホヘンハウト　　松本裕治

奈良先端科学技術大学院大学 情報科学研究科

{marc-h,matsu}@is.aist-nara.ac.jp

自然言語の基礎的な性質として、単語と単語の間の係り受け関係があげられる。単語間の係り受け関係の統計的分布は、bigramと同様に、言語のもっとも基礎的な性質の一つであると考えられる。本論文では、単語の統語的振舞いに基づいたクラスタリング手法を提案する。もっとも粗いクラス分割は、品詞分類と等しいが、任意の数のクラスを得ることができ、最適な場合、3000個のクラスへの分割が得られる。獲得したクラスを用いて、単語間の係り受け関係の確率分布の推定精度を向上させることができる。コーパスを利用した実験の結果を報告し、我々の手法の有効性を検証し、得られたクラスの妥当性を示す。

A Class Based Method for Estimation of Dependency Relations

Wide R. HOGENHOUT　　Yuji MATSUMOTO

Graduate School of Information Science, Nara Institute of Science and Technology.

{marc-h,matsu}@is.aist-nara.ac.jp

The distribution of the grammatical relation between words is of one of the fundamental properties of language, much like bigram probabilities. We propose a method for clustering words on the basis of their grammatical behavior. In the extreme case this clustering is equal to the parts of speech, but in the optimal case we use around 3000 classes. In experiments, we applied the obtained classes to the problem of estimating the distribution of the grammatical relation between words. We show that this can significantly improve the estimation, which shows the viability of our method and the meaningfulness of the classes.

1 はじめに

係り受け文法に関して、統計的構文解析手法によって、有望な結果が報告されている。Collins [3] や藤尾ら [6] の研究では、単語と単語の間の係り受け関係の統計的分布が重要な役割を果たし、自然言語に対し、基礎的な性質の一つであると考えられる。さらに、表層語を用いた確率モデルは品詞確率モデルより優れている結果を示すとCollinsは、報告している。

しかしながら、Collinsの条件付確率のモデルにおいては、data sparsenessの問題があげられる。表層語を用いた係り受け関係の確率分布のパラメータの数は膨大であり、構文解析済みコーパスから得られるデータのみでは不十分であることはあきらかである。この問題の解決法として、smoothingが標準的に適用される。Collinsの研究では品詞を用いた確率モデルを利用したsmoothingが適用されているが、それは最適なモデルではないことが知られている。藤尾ら [6] の研究では分類語彙表の意味クラスを用いたsmoothingが行なわれているが、その手法では意味的曖昧性やクラスの妥当性の問題は避けられない。

本論文は、単語の統語的振舞いに基づいたクラスタリング手法を提案する。構文解析済みコーパスから得られるデータに基づいた表層語の係り受け関係を調べることによって、特定の係り受け関係の抽出が可能になることを示す。得られたクラスは言語の特徴的な性質の一つであると共に、smoothingに有効なクラスタリングであることを実験的に示す。もっとも粗いクラス分割は、品詞分類と等しいが、任意の数のクラスを得ることができ、最適な場合、3000個のクラスへの分割が得られる。

クラスタリングの手法においては、統語的振舞いの類似度を測定する基準が必要となる。一般には確率分布の距離を測定するdivergenceが用いられるが、本論文ではそれより妥当な距離を提案する。

実験の結果、3000個のクラスを用いることにより、係り受け関係の確率分布の推定結果が向上し、手法の妥当性が確認できた。実験の評価手法はcross entropyであるが、統計的構文解析の精度の向上も期待できると考えられる。

2 係り受け関係

本研究で、コーパスからの係り受け関係を獲得する手法は主辞の概念に基づいており、[3, 6, 7, 8, 9] と同様の

手順を用いている。まず、構文解析済みコーパスの構文木の葉ノードの主辞を選択してから、全ての上位ノードは下位ノードの主辞を選択するという条件の元で、すべてのノードに主辞 が与えられる。

Collins[3] が提案した手順を用いてコーパスの構文木を係り受け関係に書き直すと、構文木を係り受け関係で記述することが可能になる。文における単語は、その単語を含む最も小さい句の主辞に係る。

さらに、単語間の関係は係る単語が支配する句の非終端記号、主辞が支配する句の非終端記号及び二つの句を統治する句の非終端記号という、三つの非終端記号で記述する。表 1はこの記述の例を示す。

この関係を用いた確率モデルを次のように表す。

$$p(R|w_d t_d, w_h t_h). \quad (1)$$

ここで、Rが関係を表す三つの句名、w_dが係る単語、w_hが主辞、t_d, t_hがそれぞれの品詞を表す。

この分布は [3] が適用している分布とほぼ同様だが、本研究では、Collins が用いている文における距離と reduced sentence を無視し、より単純な関係を対象とする。構文解析システムに関しては、これは有用な情報であるが、クラスタリングに関しては、重要な情報ではないと考えられる。ただし、これらの情報を利用しているシステムに対しても、我々の手法の適用は容易である。

表 1: 例文 *John Smith works fast* の係り受け関係

係る単語		主辞		関係
John	固有名詞	Smith	固有名詞	- , NP , -
Smith	固有名詞	works	動詞	NP , S , VP
fast	副詞	works	動詞	- , VP , -

3 単語間の関係の分布の推定

式 (1) の確率空間によって起こり得る事象の数に対し、一般に学習データは非常に不足することが多い。Wall Street Journal のコーパスの全体を調査したところ、事象の約半分は一回しか観察できなかった。

bigram における sparse data の問題に対し、クラスに基づいた様々な手法の研究が行なわれてきた [2, 4, 5, 10]。これらのうち、bigram を用いた研究においては、我々の研究とは、問題の設定が異なっており、さらに意味的な類似度の測定と意味クラスの作成を目的としている。

我々の研究では、意味を表すクラスより、むしろ統語的振舞いに基づいたクラスの方が係り受け関係の確率分布（式 1）の近さを表すという立場に立つ。統語的振舞いは式 (2) と式 (3) によって定義する。

$$p(R, t_h|w_d t_d) \quad (2)$$

$$p(R, t_d|w_h t_h) \quad (3)$$

ただし、ここでは各記号は 2節と同じものを表す。本論文では「単語」は表層語と品詞のタグの組を指し、$w_d t_d$ と書く。

この二つの分布はそれぞれ、係る場合の振舞い（式 (2)）と主辞の場合の振舞い（式 (3)）を表す。係り受け関係における相手の単語の表層語を無視することによって、文法的な用法に注目することができる。例えば、ある英語の動詞は他動詞での使用の頻度が多いことを検出することが可能になるが、「食べる」と「ご飯」の関係といった意味的な関連は検出できない。

4 統語的振舞いの例

2節に述べた手法により、コーパス中の構文木から式 (2) の分布と式 (3) の分布のパラメータを推定した。英語のコーパスである Wall Street Journal (WSJ) を用い、最尤推定法に従い、次式によってパラメータを推定した。

$$p(R, t_h|w_d t_d) = \frac{f(R, t_h, w_d t_d)}{f(w_d t_d)}$$

ここでは、fは頻度である。

単語 Nippon と Rep. がそれぞれ係り側の単語である場合の頻度分布を表 2に示す。これらは固有名詞で、係られる主辞も固有名詞しかなく、関係も二種類しかない。Nippon は英語では会社名に使われているが、主辞にはならない。Rep. は共和国の省略で、国名によく使われるが、主辞にはならない。Rep. の頻度は Nippon より高い。両方とも普通の固有名詞ではなく、特別な固有名詞である。

WSJ においては、company(会社) は主語になることが多いのに対して、hostage(人質) は逆に目的語になることが多い。一方、year(年) は前置詞に係ることが特に多い。これらの単語は、統語的振舞いが異なるので、別のクラスに分けるべきであると考えられる。

動詞においては、including(含んで) は普通の動詞と違って前置句を構成することが多い。WSJ の特徴としては、fell(落ちた) が文の主辞になることが非常に多い。

本研究では、同じ品詞を持った単語において、統語的振舞いが似ている単語と、統語的振舞いがまったく違う単語があることに注目し、その振舞いの違いによってクラスタリングを行なう。クラスタリングの手法によって、任意の数のクラスを獲得することができ、人間が認識できない区別が期待できる。

5 距離尺度

統語的振舞いは確率分布によって定義される。確率分布の距離尺度としては Kullback-Leibler 距離があげられる。bigram を用いた研究においては、この尺度を用いて意味的距離またはそれに基づいたクラス抽出を目標にした研究がいくつかある [5, 11]。

表 2: 係る場合における、固有名詞 *Nippon* と *Rep.* の頻度分布

係る単語	主辞品詞	関係	頻度
Nippon 固有名詞	固有名詞	- NP-SBJ -	3
	固有名詞	- NP -	6
Rep. 固有名詞	固有名詞	- NP-SBJ -	23
	固有名詞	- NP -	45

Kullback-Leibler 距離は対称的ではないため、ここに両側の距離を合わせた Jeffery's 情報量（divergence ともよばれる）の適用も可能である。

$$Div(p,q) = \sum_x p(x) \log \frac{p(x)}{q(x)} + q(x) \log \frac{q(x)}{p(x)} \quad (4)$$

しかしながら、これは妥当ではないと思われる。特に、単語の頻度が異なっている場合に Jeffery's 情報量では、望ましい結果が得られない場合がある。さらに、推定された分布の一方が 0 であれば無限の距離になってしまうので、距離を計算する前に smoothing を行なう必要がある。例えば、

$$\hat{p}(R, t_h|w_d t_d) = \lambda(R, t_h|w_d t_d) + (1-\lambda)(R, t_h|t_d) \quad (5)$$

という式を適用すればこの問題を回避することができるが、表層語の確率分布の推定において品詞だけに基づいた分布を使うことは望ましいことではないと思われる。

本研究では、Jeffery's 情報量より妥当な距離を提案する。まず、$i = 1, \ldots, n$ のパターン（単語間の関係）があるとする。それから、単語 $w_a t_a$ と $w_b t_b$ において頻度 $a_1, \ldots, a_n$ および $b_1, \ldots, b_n$ の観察データがあるとする。ここで $A = \sum_i a_i$、$B = \sum_i b_i$ とすると、最尤推定法では例えば $p_a = a_x/A$ のように計算できる。本研究で提案する情報量を次のように定義する。

$$M(w_a t_a, w_b t_b) \stackrel{\text{def}}{=} \sum_i \left(\frac{a_i}{A} \log \left(\frac{a_i}{A} \frac{(A+B)}{(a_i+b_i)} \right) + \frac{b_i}{B} \log \left(\frac{b_i}{B} \frac{(A+B)}{(a_i+b_i)} \right) \right)$$

c を a と b を合併したクラスターとすれば、この距離は c と a、それから c と b の間の Kullback-Leibler 距離を合わせた距離である。この距離においては、$a_i = b_i = 0$ に対して両側 0 とすることによって、距離を計算する前に smoothing を行なう必要性はない。

上式を少し変換すれば、次のようになる。

$$M(w_a t_a, w_b t_b) = \log \left(\frac{A+B}{A} \right) + \log \left(\frac{A+B}{B} \right) + \sum_i \left(\frac{a_i}{A} \log \left(\frac{a_i}{a_i+b_i} \right) + \frac{b_i}{B} \log \left(\frac{b_i}{a_i+b_i} \right) \right) \quad (6)$$

これによって、上の情報量の計算の効率をかなり向上することができる。$a_i = 0 \neq b_i$ または $b_i = 0 \neq a_i$ のパターンを総和の中で無視することができる。

6 アルゴリズム

コーパスにおける単語の分布の統計を与えられるとして、すべての単語を要素が一個のクラスターとする。まず、分布の統計が等しい単語を融合し、すべての距離が 0 以上になるのようにする。特に、一回しか現われていない単語の場合、分布の統計が等しいことが多い。

本研究ではこの方法によって、クラスターの数が約 50 %減少する。（頻度の低い単語においては式 (2) または式 (3) の分布を推定するのは、妥当ではないかもしれないが、クラスタリングでは重要な手がかりである。）

次に greedy algorithm によって、距離のもっとも小さい二つのクラスターを発見し、それを融合することを繰り返す。この処理は、あらかじめ決められた数のクラスターに達すれば終了する。ただし、融合を行なうには、つぎの二つの条件を満たさなければならない。第一の条件は、融合する単語の品詞は等しくなければならないことである。この条件によって、例えば動詞と名詞の間の距離を計算する必要はなくなる。これは、効率の観点から重要である。第二の条件は、頻度 1 回（ちなみに、1 単語からなる）クラスターがあれば、それを優先的にもっとも近いクラスターに融合することである。

二つの分布があるため、このアルゴリズムは二通り行なう必要があり、係る場合の分布の式 (2) を用いたクラスタリングと主辞の場合の分布の式 (3) を用いたクラスタリングをそれぞれ行なう。

分布の等しい単語を融合するという条件、または、品詞が等しいクラスターのみを融合するという条件のため、本手法の計算量は比較的小さい。Ultra Sparc 1 において 43,000 単語を一日ぐらいかけて任意の数のクラスターにクラスタリングすることができるが、Kullback-Leibler 距離を用いれば 2、3 倍遅くなる。

7 実験に使用されたモデル

本手法の有効性を確かめるために、WSJ の解析済みコーパス中の単語と係り受け関係を抽出し、実験を行なった。ベースラインモデルとして、表層語モデルに品詞に基づいた smoothing を加えたものを用いた。

$$\begin{aligned} \hat{p}_{lexical}(R|w_d t_d, w_h t_h) = \\ \lambda_1(w_d t_d, w_h t_h) p(R|w_d t_d, w_h t_h) + \\ \lambda_2(w_d t_d, w_h t_h) p(R|t_d, w_h t_h) + \\ \lambda_3(w_d t_d, w_h t_h) p(R|w_d t_d, t_h) + \\ \lambda_4(w_d t_d, w_h t_h) p(R|t_d, t_h) \end{aligned}$$

ベースラインモデルの実験結果は、実験結果において "lexical" で示される。この smoothing は Bahl ら [1] の

手法に従って、別のデータからλの推定を行なうが、全ての表層語に対して独立したλを推定することはできない。そこで、Bahl らが示唆する手法に従って、表層語の頻度に基づいてλをグルーピングし、頻度の等しい単語のλを等しくする。データが十分でない頻度においては、頻度の近い単語のλも等しくするようになっている。グルーピングされた頻度においては、$\lambda_1 ... \lambda_4$ をニュートン法によって最適化した。

これを、次の単語のクラスタリング結果を導入した "combined" というモデルと比較する。

$$\begin{aligned}\hat{p}_{comb}(R|w_d t_d, w_h t_h) = \\ \lambda_1(w_d t_d, w_h t_h) p(R|w_d t_d, w_h t_h) + \\ \lambda_2(w_d t_d, w_h t_h) p(R|c_d, c_h) + \\ \lambda_3(w_d t_d, w_h t_h) p(R|t_d, c_h) + \\ \lambda_4(w_d t_d, w_h t_h) p(R|c_d, t_h) + \\ \lambda_5(w_d t_d, w_h t_h) p(R|t_d, t_h)\end{aligned}$$

smoothing は "lexical" のモデルと同じように行なう。テスト・データが学習データと異なるため、テストにおいてクラスタリングされていない単語が現れることは避けられない。ここでは、一回も現れていない、確率が 0 となる仮想クラスを導入する。このクラスを条件とする確率は必ず 0 となる。

8 実験

ここで行なった実験は、WSJ コーパスから抽出された 800,000 係り受けパターンについて、その内の 740,000 パターンで学習・クラスタリング・smoothing を行ない、デストデータとして 60,000 パターンを使用して評価を行なった。この実験では係る場合のクラスタリングと主辞の場合のクラスタリングをそれぞれ行なって、両方 3000 個のクラスを求めた。

評価にはテストデータに対するクロスエントロピーを用いたが、これは表 3の第一欄で示されている。第二欄においては一回以上現れたパターン（すなわち $f(R, w_d t_d, w_h t_h) > 0$）に対するクロスエントロピー、第三欄には一回も現れていないパターン（すなわち $f(R, w_d t_d, w_h t_h) = 0$）に対するクロスエントロピーを示す。

表 3が示すように、"combined" モデルではクラスを使用したため、約 4.6%クロスエントロピーが下がり、$f > 0$ のパターンと $f = 0$ のパターンの両方に対して有効である。

表 3: テストデータに対するクロスエントロピー

method	$f \geq 0$	$f > 0$	$f = 0$
lexical	1.862	0.729	3.178
combined	1.811	0.707	3.093

9 おわりに

本研究では、コーパスから抽出した係り受け関係を用いた、統語的振舞いに基づく単語のクラスタリング手法について述べた。従来の研究とは異なり、意味的類似度ではなく、係り受けの分布を対象にした。クラスタリング手法においては Jeffery's 情報量の代わりに、新しい確率分布の距離の尺度を提案した。

実験の結果、抽出されたクラスが統語的振舞いの種類を表しており、単語と単語間の係り受け関係の推定の精度向上が期待できることがわかった。

参考文献

[1] Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-5, No. 2, pp. 179–190, 1983.

[2] P. F. Brown, S. A. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479, 1992.

[3] M. J. Collins. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 184–191, 1996.

[4] I. Dagan, S. Markus, and S. Markovitch. Contextual similarity and estimation from sparse data. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 164–171, 1993.

[5] I. Dagan, F. Pereira, and L. Lee. Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 272–278, 1994.

[6] 藤尾正和, 松本裕治. 統計的手法を用いた係り受け解析. 情報処理学会研究報告 97-NL-117, pp. 83–90, 1997.

[7] W. R. Hogenhout and Y. Matsumoto. Training stochastical grammars on semantical categories. In *Proceedings of the IJCAI Workshop on New Approaches to Learning for Natural Language Processing*, pp. 65–70, Aug. 1995.

[8] F. Jelinek, J. Lafferty, D. Magerman, R. Mercer, A. Ratnaparkhi, and S. Roukos. Decision tree parsing using a hidden derivation model. In *ARPA: Proceedings of the Human Language Technology Workshop*, pp. 272–277, 1994.

[9] D. M. Magerman. Statistical decision-tree models for parsing. In *Proceedings of the 33d Annual Meeting of the Association for Computational Linguistics*, pp. 276–283, 1995.

[10] F. Pereira and N. Tishby. Distributional similarity, phase transitions and hierarchical clustering. In *Working Notes, Fall Symposium Series, AAAI*, pp. 108–112, 1992.

[11] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 183–190, 1993.