bigram 統計情報に基づくパージング

7L－5

佐藤 健吾, 中西 正和
慶應義塾大学理工学研究科計算機科学専攻

1. はじめに

現在の構文解析の手法は統語規則を用いるものが主流であるが，ドメインに最適な統語規則と巨大な辞書が必要である．一方，統計情報から得られる単語の関連性をもとにパーズする手法において必要なのは，大きなトレーニングコーパスのみである．

一般的な統計情報に基づくパーザは，構文解析済みのコーパスを用いて統語規則あるいは解析木の特徴を学習し，その結果を用いてパーズする [1, 2]．本稿では，生コーパスから得られる bigram 統計情報のみを用いたパージングを試みる．

2. 関連性に基づくパージング

関連性に基づくパージングは，隣接する単語間の関連性をもとにグルーピングを行なうことにより文の構造を発見する方法である [3]．図 1 の例では，他の組合せよりも a pen の関連性が高いため最初にグルーピングされ，同様の処理を繰り返すことにより文の構造を組み立てている。

```
this is a pen
this is <a pen>
this <is <a pen>>
<this <is <a pen>>>
```

図 1: 関連性に基づくパージング

2.1 Association Score

単語間の関連性を比較するために，bigram 統計情報を用いて Association Score を定義する．

グルーピングされる単語の組 w_1, w_2 に対して以下を仮定する．

1. w_1, w_2 は共起して出現することが多い．
2. w_1, w_2 の左右には多様な単語が出現する．

仮定 1 の尺度として単語 w_1, w_2 の間の相互情報量 $I(w_1, w_2)$ を用いる．$I(w_1, w_2)$ が大きくなる w_1, w_2 は共起関係が強いといえる．

$$I(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \quad (1)$$

(ここで $P(x)$ は x が出現する確率，$P(x, y)$ は x と y が連続で出現する確率)

仮定 2 の尺度として 単語 w_1, w_2 の左右に出現する単語の確率分布のエントロピーを用いる．$H_l(w_1, w_2), H_r(w_1, w_2)$ が大きくなる w_1, w_2 は左右に多様な単語が出現するといえる．

$$H_l(w_1, w_2) = \min(H_{d=-1}(w_1), H_{d=-2}(w_2)) \quad (2)$$

$$H_r(w_1, w_2) = \min(H_{d=2}(w_1), H_{d=1}(w_2)) \quad (3)$$

ただし，

$$H_{d=d'}(w) = \sum_{v \in \boldsymbol{W}} -P_{d=d'}(v|w) \log P_{d=d'}(v|w) \quad (4)$$

($P_{d=d'}(v|w)$ は w が出現した時に w の d' だけ右に離れて v が現れる確率，$\boldsymbol{W}$ はコーパス中の全単語の集合)

ここで，I, H_l, H_r を用いて単語間の Association Score $S(w_1, w_2)$ を定義する．$S(w_1, w_2)$ が大きくなる w_1, w_2 は関連性が高いといえる．

$$S(w_1, w_2) = \lambda_1 I(w_1, w_2) + \lambda_2 (H_l(w_1, w_2) + H_r(w_1, w_2)) \quad (5)$$

(λ_1, λ_2 は定数)

Parsing Based on Bigram Statistics
Kengo Sato, Masakazu Nakanishi
Department of Computer Science, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa Pref., 223, Japan

2.2 デリミタによる補正

at the や of the のような単語の組は，2.1 節で述べた仮定を満たしており，実際 Association Score S の値も他の単語の組に比べて極めて高くなることが実験によりわかった．このような単語の組がグルーピングされるのを避けるために新たに以下を仮定に加える．

3. グルーピングされるような単語の組 w_1, w_2 はコーパス中で「言いきりの形」(すなわち文) として存在し得る確率がある程度なければならない．

この仮定は，at the や of the のような単語の組はコーパス中で文の形では存在し得ず，the pen のような組は文として存在し得るので，関連性がより高いことを意味する．

この仮定を数値的に表す尺度として，デリミタ(文頭，“.”, “,”) との共起確率を用いる．

$$P(\boldsymbol{D}|w_1, w_2) = \sum_{v \in \boldsymbol{D}} (P_{d=-1,2}(v|w_1) + P_{d=-2,1}(v|w_2)) \tag{6}$$

(ここで $\boldsymbol{D}$ はデリミタの集合)

$P(\boldsymbol{D}|w_1, w_2)$ を用いて Association Score を新たに定義し直し，

$$S'(w_1, w_2) = S(w_1, w_2) \cdot P(\boldsymbol{D}|w_1, w_2) \tag{7}$$

とする．

3. 実験方法および結果

今回の実験では SUSANNE コーパス [4] の一部 (総単語数 77,726 語，語彙数 10,204 語，文数 3,128 文) を使用した．SUSSANE コーパスは構文解析情報が付加されているが，実験では単語のみを用い，評価に構文解析の情報を用いる．図 2 に実験結果の一部を示す．

```
<@@ <<<<<the <mayor's <present term>>>
of> office> expires <<jan. 1> .>>>>
```

図 2: 実験結果

コーパス中の構文解析情報に含まれる括弧の対応とこれまでに述べた手法により得られる括弧の対応を比較することにより評価を行なった．(a) 構文解析情報に現れる括弧が実験結果に現れる括弧の中に存在する割合，(b) 実験結果に現れる括弧が構文解析情報に現れる括弧に対して交差する割合を表 1 に示す．

(a) 存在した括弧	25.5 %
(b) 交差した括弧	56.0 %

表 1: 括弧の対応の比較

4. 今後の課題

今回の実験は品詞などの付加的な情報を用いず，単語のみの統計情報で Association Score を定義し，この値をもとにパーズを行なった．しかし，単語のみの統計情報ではデータが希薄になってしまうことは明らかである．このため，品詞情報を併用することで大幅に結果を改善することが可能であると思われる．

参考文献

[1] David M. Magerman. Statistical Decision-Tree Models for Parsing. *Proceedings of ACL-95*, 1995.

[2] Michael John Collins. A New Statistical Parser Based on Bigram Lexical Dependencies. *Proceedings of ACL-96*, 1996.

[3] David A. Evans, Chengxiang Zhai. Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. *Proceedings of ACL-96*, 1996.

[4] University of Sussex, School of Cognitive & Computing Sciences. *The SUSANNE Corpus*, Nov 1994. ftp://sable.ox.ac.uk/pub/ota/public/susanne.