d-bigram を用いた形態素解析

2H－3

延澤 志保, 堤 純也, 孫 大江, 佐藤 健吾, 佐野 智久, 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

本研究の目的は、日本語のように単語等に切り分けられていない言語の文を統計情報を用いて形態素列に切り分けることである。ここで紹介するシステムでは、文法等の知識は全く利用せず、純粋に形態素間の統計情報のみを用いて日本語の切り分けを行ない、closed corpus を用いた実験では 99%、open corpus でも 80% 以上の正解率を得た。

2. 文の評価値

2.1 文の評価値

入力ひらがなべた書き文の複数個の漢字かな混じり候補文の中から正しい文を選び出すために、本システムでは文の評価値を導入している。

2.2 相互情報量

相互情報量 (MI: mutual information) とは、事象間の関係の情報量のことである [1]。自然言語処理では、事象として形態素や文字などをとることが多い。

2.3 文の評価値の計算法

本システムで導入した文の評価値は、与えられた文字列が統計的に見てどの程度文らしいかを示すものである。

ここでの文の評価値の計算には 2 つの事象がある距離をおいてどの程度同時に出やすいかを示す d-bigram 情報 [1][2] を用いている。

2.4 d-bigram

自然言語処理の分野では統計情報として bigram など n-gram がよく用いられる。本研究では新しい統計情報として d-bigram[1][2] を導入した。

d-bigram は bigram 同様 2 事象間の関係値であるので、trigram など他の n-gram 情報と比べてコーパスから取得できる情報量が濃く、しかも組み合わせが少なくなるため情報量も少なくなる。trigram はコーパス中の n 個の事象に対し $O(n^3)$ となるが、d-bigram ではたかだか $O(n^2)$ である。

2.5 文の評価値の計算

以下に 2 形態素間の情報の算出式を示す [2]。

$$MI_d(x_i, x_{i+d}, d) = \log_2 \frac{P(x_i, x_{i+d}, d)}{P(x_i)P(x_{i+d})} \quad (1)$$

x_i : 入力列中の i 番目の形態素
d : 2 形態素間の距離

文の評価値は、先に与えられた距離 d までのすべての形態素の組についてその間の d-bigram を計算しそれらに距離に応じた重み付けを施したものを足し合わせることで得ている [2]。

$$I(W) = \sum_{d=1}^{d_{max}} \sum_{i=0}^{n-d-1} \frac{MI_d(x_i, x_{i+d}, d)}{d^2} \quad (2)$$

W : 入力文
n : 入力文中の形態素数

3. MSS

ここで述べる MSS(Morphological Segmentation using Statistical information) は、日本語ひらがなべた書き文を入力としてとる。入力文字列を辞書に照らし合わせてその文字列中に含まれ得る形態素を抽出し、それらを並べて漢字かな混じり文の候補とする。その後それらの候補文それぞれについて評価値の計算を行ない、評価値の一番高いものを解として採用する。

Segmenting a Sentence into Morphemes Using D-bigram
Shiho NOBESAWA, Junya TSUTSUMI, Da Jiang SUN, Kengo SATO, Tomohisa SANO, Masakazu NAKANISHI
Department of Computer Science, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa Pref., 223, Japan

4. 結果

本研究では英語の教科書の日本語訳の一部[1]をtraining corpus として用いた。約 1500 語の見出し語から成る辞書[2]を手作業で用意した。入力は日本語ひらがなべた書き文各 100 文ずつで、評価値計算に用いる形態素間の距離の上限は 5 とした。

表 1: Experiment in Japanese

	best score	~ 2nd best	~ 3rd best
α	99 %	100 %	100 %
β	100 %	100 %	100 %
γ	100 %	100 %	100 %
δ	95 %	98 %	98 %
ϵ	80 %	90 %	95 %

α : the very sentences in the corpus
β : replaced one morpheme in the sentence (the buried morpheme is in the corpus)
γ : replaced one morpheme in the sentence (the buried morpheme is not in the corpus)
δ : sentences not in the corpus (the morphemes are all in the corpus)
ϵ : sentences not in the corpus (include morphemes not in the corpus)

5. 考察

5.1 実験結果

表 1 中の α、β、γ はコーパスにある文に相当するひらがな文を入力としてとったもので、β と γ はそれぞれ一つずつ文中の形態素をほかのものと置き換えられている。当然その形態素の周りでは d-bigram 情報は薄くなるが、他の形態素間の結び付きが強いので、その一つの影響は小さく抑えられて、結果はさほど悪くならない。

δ と ϵ には口語文など文法的に問題の有るものも含まれている。今回の実験のコーパスは小さいのでこのように形態素同士の関係が摑みにくい文では条件はあまり良くないように思われるが、実際には充分と言って良い結果が得られている (表 1)。MSS ではコーパスをそのまま用いるのではなく MI を利用することで情報を有効に活用している。この点において、MSS は期待どおりの結果を得ていると言ってよい。

[1]中学生用の英語の教科書 Horizon 中の約 630 文
[2]本研究では、training corpus に含まれる形態素と比較実験のためのコーパスに含まれない形態素から成る。

5.2 コーパス

本研究で用いたコーパスは実際の日本語文のモデルと呼ぶには小さ過ぎるが、それでも、実験結果を見ると MSS は充分な効果ををあげていることが判る。

本研究では日本語の形態素解析を行なったが、コーパスとそれに対応する辞書さえ用意すればシステムにはほとんど変更を加えずにさまざまな用途に用いることが可能である。中国語の発音記号 (Pinyin) からの漢字 (Hanzi) 文の生成 [3] などもその一つである。

5.3 文節数最少法との比較

候補文中の形態素がコーパスに出てこない場合、MSS ではこの形態素と他の形態素との MI として負の値を与える。したがって、ので、形態素の数が少ないものが選ばれやすくなってくる。これは、文節数最少法の考え方とよく似ている。実際、このような条件の元では MSS の結果と文節数最少法の結果は非常によく似てくる。これは文中の形態素間に関係がほぼ無い場合のことであり、形態素間の情報を用いることができる場合には当然結果は良くなってくる。つまり、MSS は最悪でも文節数最少法程度の有効性を示すと言える。

6. 結論

このように、d-bigram を用いた形態素解析システム MSS は日本語形態素解析において非常に有用であり、その候補文がコーパスに有るか無いかにかかわらず、全体的に良い結果を得ている。MSS は充分大きなコーパスを用いることでさらに良い結果が期待できる。

参考文献

[1] 堤 純也, 新田 朋晃, 小野 孝太郎, and 延澤 志保. 統計情報を用いた多言語間機械翻訳システム. 人工知能学会研究会, pages 7–12, 1993.
[2] Nobesawa, S., Tsutsumi, J., Nitta, T., Ono, K., Sun, D. J. and Nakanishi, M. Segmeting a Japanese Sentence into Morphemes Using Statistical Information between Words. *Coling*, pages 227–233, 1994.
[3] Sun, D. J., Tsutsumi, J., Nitta, T., Ono, K., Nobesawa, S. and Nakanishi, M. An intelligent Chinese input system using statistical information between words. *Qualico*, pages 102–107, 1994.