文字単位のbigram尺度に基づく複合漢字列の単位切り手法

影浦　峡
学術情報センター研究開発部
kyo@rd.nacsis.ac.jp

1　はじめに

日本語専門用語の多くは複合名詞であり、それを構成する基本単位を抽出することは、専門用語を巡る様々な研究の出発点として重要である。本発表では、複合専門用語を構成する要素のうち、語種別に見て最も比率の高い、複合漢字列の単位切りの一手法を提案する。

従来の単位切り研究の多くは、あらかじめ作成した要素語辞書を前提とするものであった。この場合、辞書の作成や更新に手間がかかる。武田・藤崎(1987)は、大量のデータから計算した漢字の遷移確率に基づき単位切りを行う手法を提案しているが、これは事前の学習を前提としている。

専門用語データの場合、分野毎に出現する漢字（列）が大きく異なるため、事前の辞書作成やトレーニングでは適応できない部分が大きい。従って、与えられた分野単位のデータに対して適応的に対応できる単位切り手法が望ましい。その場合、中小規模のデータに対応する必要が出てくる。また、現実問題として、単に規模が小さいだけでなく、専門用語辞書の見出しのように、異なり情報しか利用できない用語データも処理しなくてはならない。

森脇ら(1996)はこうした条件下での専門用語の自動分割手法を提案している。これは、漢字列のみでなくカタカナ列も扱えるという点で適用範囲は広いが、分割精度はそれほど高くない。我々が提案する手法は、上記の要件を満たしながら、漢字列についてかなりの分割精度を達成するものである。

2　基本アルゴリズム

基本的な考え方は単純である。まず、与えられたデータ（分野ごとの専門用語の異なり漢字列部分）を用いて、漢字bigramの結合強度を計算する。ここでの結合強度計算の考えは、定型表現抽出によく用いられる手法と共通のものである。

bigramの強度を計算する尺度として、相互情報量 (Church & Hanks 1990)、YuleのY (Delcourt 1992)、カイ二乗 (Fienberg 1977)等、色々なものがあるが、ここでは小頻度のデータに対しても比較的安全に適用できる、尤度比検定 (Dunning 1993; Fienberg 1977) を採用する[1]。

ここで用いた尤度比検定は、二項分布を前提としている。まず、注目する二文字の漢字AとBの組み合わせに対する以下のような分割表を考える。ここで、$\overline{A}$はA以外の漢字を意味し、$f(x)$はxの頻度を与えるものとする。

		第二文字 B	$\overline{B}$	合計
第一文字	A	$f_{11}=f(AB)$	$f_{12}=f(A\overline{B})$	$f_{1.}=f(A)$
	$\overline{A}$	$f_{21}=f(\overline{A}B)$	$f_{22}=f(\overline{A}\overline{B})$	$f_{2.}=f(\overline{A})$
	合計	$f_{.1}=f(B)$	$f_{.2}=f(\overline{B})$	$f_{..}$

この分割表に対する尤度比検定の値は、次の式で与えられる。

$$-2\log\lambda = 2[\sum_{c}\log L(f_{1c}/f_{.c}, f_{1c}, f_{.c}) - \sum_{c}\log L(f_{1.}/f_{..}, f_{1c}, f_{.c})]$$

[1] 実際の比較評価で、尤度比検定がこれら四尺度の中で最も良い結果を出した。これについてはKageura (1996a)を参照。

ただし、

$$\log L(p,n,k) = k\log p + (n-k)\log(1-p)$$

である。この式から算出される値は、二文字の組み合わせが偏って強い場合のみでなく、偏って弱い場合にも大きくなる。そこで、偏りの方向性を区別するために、Yule の Y の符号を用いる。

次に、これによって得られた結合強度の値を利用して、次の手続きで3文字以上からなる漢字列を分解する。これにより、国立国語研究所の β 単位に相当する単位への分割ができる。

```
decompose_string (string) {
   if (length of string <= 2) {
      return string;
   } else {
      divide string into head and tail,
         at the point where score is minimum;
      decompose_string (head);
      decompose_string (tail);
   }
}
```

武田・藤崎 (1987) のマルコフモデルに基づく手法が、学習データに基づき、接辞・語基からなるシンタグマティックな漢字列に対して遷移確率を与えるのに対し、この手法は、語基に相当する bigram 強度のみを用いるため、計算すべきパラメータ空間が小さい。逆に、この手法では接辞であることの強さは単位切りにおいて考慮されないことになる。

3 基本アルゴリズムの評価

評価実験のために、学術情報センターがサービスしている学会発表データベースのデータを用いた。このデータベースには、発表学会情報が含まれているので、学会を分野のまとまりとみなし、データ量の極端に小さい学会を除いた後に、学会単位の10%無作為抽出で取り出した3分野（林学会・人工知能学会・情報処理学会）のデータを用いた。

各分野の基本的なデータ量は次の通りである。

漢字	林学		人工知能		情報処理	
列長	異なり	延べ	異なり	延べ	異なり	延べ
1	579	9038	556	13030	1190	196032
2	2501	14250	1974	27394	6866	389042
3	2208	4574	1824	6286	12383	88815
4	2565	4094	2844	6303	22653	83150
5	1057	1427	995	1423	10721	21499
6 ≤	1054	1257	1078	1533	13940	22390
3 ≤	6884	11352	6741	15545	59697	215854
合計	9964	34640	9271	55969	67753	800928

評価のために、上記データを人手によって β 単位に分割した評価用データを別途準備した。

上記の3分野それぞれにつき、分野ごとに、異なり情報を用いて bigram 強度を計算し[2]、単位切りを行なった。

単位切り結果の評価は、漢字列単位の評価用データとの完全マッチングによった。従って、単位の区切りが評価用データと一箇所でも異なっている場合は、単位切り失敗とみなした。結果は以下の通りである[3]。なお、「学習 (字)」は、各分野で学習に用いた異なり漢字列の総文字数（延べ）を示す。

分野	学習 (字)	評価	漢字列数	成功数	成功率
林学	35072	延べ	11352	9580	84.39
		異なり	6884	5716	83.03
人知	35523	延べ	15545	14852	95.54
		異なり	6741	6286	93.25
情処	290367	延べ	215854	205647	95.27
		異なり	59697	55247	92.55
全体	(120321)	延べ	242751	230079	94.78
		異なり	73322	67249	91.72

結果は分野によってかなり異なっている。林学及び人工知能は概ね同じ大きさのデータであるが、人工知能分野の結果は、一桁大きい情報処理分野とほぼ同じであり、ともに延べ評価で95%の正解率を達成している一方、林学の結果はかなり低い。「全体」は、3分野の解析結果を漢字列のレベルでまとめた

[2]例えば、漢字列 ABCD が10回、ABD が2回データ中に現われるとすると、延べ情報に基づく計算では AB が12回、BC と CD が10回、BD が2回計数されるが、異なり情報に基づく計算の場合、AB が2回、BC、CD、BD はいずれも1回計数される。延べ計算による bigram 強度を用いた単位切りと異なり計算による単位切りとの比較実験では、一貫して後者の方が良い結果を得た。これについては Kageura (1996b) を参照。

[3]「評価」欄の「延べ」及び「異なり」は、それぞれ、延べ漢字列数に基づいて評価した値と異なり漢字列数に基づいて評価した値とをしめすもので、bigram 強度計算に延べ、異なり情報を用いたということではない。

ものであり、一応、この手法を異分野の同様データに適用したときに、全体としてどの程度の漢字列単位切り正解率を得るかの目安を示している[4]。概ね、平均12万字ほどの学習データで、延べに基づく評価で95%弱の正解率が得られることがわかる。

4 後処理の付加

基本アルゴリズムに基づく単位切りでは、単純なbigramの強度のみを考慮していた。前述したように、これはいわば、語基のみを考慮し、接辞は積極的には考慮していないことを意味している。

そこで、ある漢字が接辞であることの強さをも考慮するために、基本アルゴリズムによる単位切りの結果を再利用する、以下のような後処理ヒューリスティクスをつけ加えた。

(1) 基本アルゴリズムによる単位切りの結果から
・一文字漢字 x について：
漢字列の先頭に現れる頻度 $pre(x)$
漢字列の末尾に現れる頻度 $pst(x)$
総頻度 $ttl(x)$
・二文字漢字列 yz について：
総頻度 $ttl(yz)$
を計算する。

(2) 単位切りの対象となった3文字以上からなる漢字列について
(a) 漢字列長が偶数 (4、6 …) のとき：
・先頭4文字が/AB/C/D/で
$pre(A) > 0$ かつ $ttl(BC) > 0$ かつ
$pst(D) > 0$ かつ $ttl(AB) < 2$
ならば
/A/BC/D/に
$ttl(CD) > pst(C) + pst(D)$
ならば
/AB/CD/に
・先頭4文字が/A/B/CD/で
$pre(B) \times ttl(CD) < pst(D) \times ttl(BC)$
ならば
/A/BC/D/に
$ttl(AB) > 0$
ならば
/AB/CD/に
・先頭4文字が/A/BC/D/で
$pre(A) \times ttl(BC) < pst(C)$ かつ
$ttl(AB) > 0$ ならば
/AB/C/D/に
変換する
(b) 漢字列長が3以上のとき：
・先頭3文字が/A/BC/で
$pre(A) \times ttl(BC) < pst(C) \times ttl(AB)$
ならば
/AB/C/に
・先頭3文字が/AB/C/で
$pre(A) \times ttl(BC) > pst(C) \times ttl(AB)$
ならば
/A/BC/に
変換する
(c) 漢字列長が5以上のとき：
・部分文字列/A/B/C/がある場合
$pre(A) \times ttl(BC) > pst(C) \times ttl(AB)$
ならば
/A/BC/に
それ以外ならば
/AB/C/に
変換する

以上のヒューリスティクスは、今回実験に用いたデータに見られた切り誤りの主要なパターンについて、一般的に対処するものであるが、文字列長に応じた細かい対応を考慮する余地がある。

5 最終的な評価

この後処理を、前述の3分野に対し、分野別に適用した結果、最終的な単位切りの結果は以下のようになった。

分野	評価	漢字列数	成功数	成功率
林学	延べ	11352	10292	90.66
	異なり	6884	6156	89.42
人知	延べ	15545	15225	97.94
	異なり	6741	6513	96.62
情処	延べ	215854	209702	97.15
	異なり	59697	56748	95.06
全体	延べ	242751	235219	96.90
	異なり	73322	69417	94.67

基本アルゴリズムによる単位切り結果からの、後処理による正解率向上は、林学においてもっとも著しいが、他の2分野においても、延べ評価・異なり評価ともにかなりの正解率向上が見られる。平均12万字ほどの学習データで、延べに基づく評価で97%近い正解率が得られることがわかる。

[4] ただし、学習に用いた文字数は平均をとってある。

ただし、この3分野は無作為に選ばれたとはいえ、正解率の高い2分野の重なりが大きいため、全体の正解率が本当に適切かどうか、多少の疑問が残る。そこで、文部省学術用語集心理学編の見出しデータを用いて、単位切りの追実験を行なった。結果は以下のようになった。

漢字列数			学習	成功	成功
合計	≤ 2	≥ 3	文字数	数	率
5005	1159	3846	18597	3653	94.98

データが用語集の見出しであるため、延べ情報に相当するものはないから、上記の成功率評価は異なりに基づくものである。林学や人工知能における学習文字数の約半分で、異なり評価95%弱の正解率が出ている。厳密には、抄録から抽出した漢字列と全く同じに考えることはできないが、一応、延べで評価すると少なくとも96%程度の正解率が得られるものと考えて良いであろう。

6 おわりに

ここで紹介した手法は、冒頭で述べた要件、すなわち、(1) 与えられた分野単位のデータに対して適応的に対応できること、(2) 中小規模のデータに対応できること、(3) 異なり情報に基づいて良い結果が得られること、をすべて満たしている。数千から数万程度の辞書見出しに対して、異なり評価で95%程度の正解率を得ることが予測される。

現在、我々は、以下の方向で単位切り手法の改良及び拡張を行なっている。

- 同一の漢字を第一要素あるいは第二要素として持つ bigram の尤度比値を比較することにより、語基を構成しない bigram をあらかじめふるい落とす可能性の検討
- 漢字列長に応じた一般的な語構成パターンを考慮することによる、ヒューリスティクスの精緻化
- カタカナ列の単位切り手法の検討および、それの漢字列単位切りとの統合

また、別分野のデータを用いた事前学習の効果、分野における漢字の分布特性と単位切り正解率との関係等に関しても、検討を進めているところである。

補記

本研究の大部分は、発表者がシェフィールド大学計算機科学科自然言語処理グループ滞在中に行なったものである。同グループのメンバーに感謝致します。

参考文献

Church, K. W. & Hanks, P. (1990) "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics.* 16(1) p. 22–29.

Delcourt, C. (1992) "About the Statistical Analysis of Co-occurrences," *Computers and the Humanities.* 26(1) p. 21–29.

Dunning, T. (1993) "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics.* 19(1) p. 61–74.

Fienberg, S. E. (1977) *The Analysis of Cross-Classified Categorical Data.* MIT Press, Cambridge, MA.

Kageura, K. (1996a) "Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences," Internal Memo, Department of Computer Science, University of Sheffield.

Kageura, K. (1996b) "Type-based and Token-based Learning of Kanji Morphemes," Internal Memo, Department of Computer Science, University of Sheffield.

森脇敏、河辺恒、辻井潤一 (1996)「辞書を使わない日本語専門用語の自動分割」言語処理学会第2回年次大会.

武田浩一、藤崎哲之助 (1987)「統計的手法による漢字複合語の自動分割」情報処理学会論文誌. 28(9) p. 952–961.