教師あり学習による連語抽出

内山 将夫　　井佐原 均
通信総合研究所

1 はじめに

その構成要素である単語の意味を単に合わせたのでは，それ自体の意味が十分に予測できないような単語列を連語と呼ぶ．連語には，「desktop publishing」「integrated circuit」などがある．連語の抽出が役に立つような応用としては，機械翻訳，情報検索，自然言語生成などがある (Smadja 1993; Manning and Schütze 1999).

連語の文法的な属性には色々なタイプがある．たとえば，「desktop publishing」は名詞句であり，「get ahead」は動詞句であり，「on purpose」は前置詞句である．このように様々なタイプがあるにもかかわらず，従来の連語抽出は，主に名詞句を対象としてきた．それは，従来研究が，主に，専門用語の抽出を対象としていて，専門用語は名詞句であることが大半であるからである (Kageura and Umino 1996)．ただし，動詞句の抽出を対象とした研究 (Dias, Kaalep, and Muischneck 2001) や，任意の連語を抽出しようという研究 (Schone and Jurafsky 2001) もある．

本稿では，任意のタイプの連語を抽出する．その理由は，我々が，連語抽出の結果を英語学習者のサポートに利用したいと考えていて，かつ，そのためには，任意のタイプの連語を抽出する必要があるからである．

従来の連語抽出もしくは専門用語の抽出は，主に，統計的な尺度 (Church and Hanks 1990; Dunning 1993) を利用するか，人手で書いた連語構成規則を利用するか，それらを組合せていた (Kageura and Umino 1996)．それらを組合せるときには，連語構成規則に一致しないような連語もしくは専門用語を排除し，そのあとで，残りを統計的尺度によりソートするものであった．これらは教師なし学習である．

それに対して本稿では，WordNet を教師データとして利用することにより，教師あり学習により，統計的尺度と連語構成パターンとを組み合わせて一般の任意のタイプの連語を抽出する方法を提案する．そして，その方法の精度が，教師なし学習の精度よりも高いことを示す．以下，統計的モデル，モデルに利用される素性，実験結果などについて順に述べる．

2 統計的モデル

n-gram として x が与えられたとき，$s(x)$ を，x の統計的な尺度値，$t(x)$ を，x を構成する単語の品詞列とし，u により，x が連語であるという事象を表す．$s(x)$ と $t(x)$ を条件とするとき，x が連語である確率は

$$\begin{aligned}&\Pr(u|s(x),t(x))\\ &= \frac{\Pr(s(x),t(x)|u)\Pr(u)}{\Pr(s(x),t(x))}\\ &= \Pr(u)\frac{\Pr(s(x)|u)}{\Pr(s(x))}\frac{\Pr(t(x)|u)}{\Pr(t(x))}\end{aligned}$$

である．ここで，統計的尺度と品詞列とは互いに確率的に独立であるとした．更に，$\Pr(u)$ がどの連語でも一定であるとして，x のスコアを

$$\mathrm{Score}(x) = \alpha \log \frac{\Pr(s(x)|u)}{\Pr(s(x))} + (1-\alpha)\log\frac{\Pr(t(x)|u)}{\Pr(t(x))} \quad (1)$$

とする．ここで，α は実験により定める定数であり，$0 \le \alpha \le 1$ である．

ここで，以下で述べる，スコア中の確率の定義に使われる記号を示す．まず，N により全ての異なり n-gram，M により N 中の全連語を示す．M は M_0 と M_1 に分け，M_0 を訓練，M_1 をテストデータとする．我々は，N と M_0 を利用して確率を推定し，$N - M_0$ 中の n-gram をソートし，その上位に M_1 があるかどうかにより，連語抽出システムの性能を評価する．

さて，$\Pr(t(x))$ と $\Pr(t(x)|u)$ の定義は，

$$\begin{aligned}\Pr(t(x)) &= \frac{\sum_{y\in N}[t(x)=t(y)]}{|N|}\\ \Pr(t(x)|u) &= \beta g(t(x)|M_0)+(1-\beta)\Pr(t(x))\\ g(t(x)|M_0) &= \frac{\sum_{y\in M_0}[t(x)=t(y)]}{|M_0|}\end{aligned}$$

である．なお，[命題] は，命題が成立するとき 1，そうでなければ 0 である．これらは，基本的には，連語の品詞列の割合であるが，もし，$\Pr(t(x)|u)=g(t(x)|M_0)$ とすると，$N-M_0$ 中の n-gram に確率 0 のものがでるため補完をする．なお，本稿では，$\beta=0.99$ を使用した．

次に，尺度値の確率については，$p(s(x)|u)$ と $p(s(x))$ を統計的尺度の密度関数とすると以下が成立する．

$$\frac{\Pr(s(x)|u)}{\Pr(s(x))}=\frac{p(s(x)|u)}{p(s(x))}$$

$p(s(x))$ と $p(s(x)|u)$ は，それぞれ，N と M_0 に属する n-gram の尺度値から推定できる．なお，密度推定には，混合正規分布により密度を推定するソフトウェア[1]を利用した．

3 素性

品詞としては，Penn Treebank Project[2]の品詞を利用した．たとえば，「chain store」の品詞列は「NN NN」である．これらの n-gram は「chain/NN store/NN」のように表現する．

尺度としては，三つの尺度，Freq, LLR(対数尤度比), MI(相互情報量 (Church and Hanks 1990)) を比較した．それらの尺度を定義するために，$F(x)$ を n-gram $x=x_1x_2...x_n$ の頻度とし，$f_{11}=F(x_1x_2)$, $f_{12}=F(x_1)-f_{11}$, $f_{21}=F(x_2)-f_{11}$, $f_{22}=F-f_{11}-f_{12}-f_{21}$, $f_{i.}=f_{i1}+f_{i2}$, $f_{.j}=f_{1j}+f_{2j}$ とする．ただし，F はコーパス中での全 2-gram 数である．これらの尺度は，正規分布になるべく近くなるように，必要に応じて対数を取っている．

まず，$\mathrm{Freq}(x)=\log(F(x))$ である．つぎに，LLR を定義するが，まず，(Dunning 1993) による定義は以下である．

$$\mathrm{LLR}_0(x_1,x_2)=2\sum_{i,j=1,2}f_{ij}\{\log\frac{f_{ij}}{F}-\log\frac{f_{i.}f_{.j}}{F^2}\}.$$

LLR_0 は連想の強さは示せるが，その方向が正か負かは示せない (Kageura 1997) ので，それを考慮して以下のように定義した．

$$\mathrm{LLR}=\begin{cases}\log(|\mathrm{LLR}_0|+1) & \text{if } f_{11}f_{22}>f_{12}f_{21}\\ -\log(|\mathrm{LLR}_0|+1) & \text{otherwise.}\end{cases}$$

MI は以下の通りである．

$$\mathrm{MI}(x_1,x_2)=\log\frac{\frac{f_{11}}{F}}{\frac{f_{1.}}{F}\frac{f_{.1}}{F}}.$$

MI と LLR は 2-gram について元々定義されている量である．それを n-gram に拡張する式は以下の通りである (Ries, Buϕ, and Wang 1995)．LLR も MI と同様である．

$$\mathrm{MI}(x)=\min_{1\le i\le n-1}\mathrm{MI}(x_1\ldots x_i,x_{i+1}\ldots x_n)$$

ここでは，$x_1\ldots x_i$ と $x_{i+1}\ldots x_n$ をそれぞれ一単語として尺度値を計算する．

4 実験

4.1 実験材料

連語データとしては，WordNet 1.7 から抽出した 59142 個を用いた．コーパスは British National Corpus (BNC)[3]を利用した．BNC の書き言葉の部分について，それを OAK system [4]によりタグ付けした．その結果から，頻度 10 以上の n-gram ($n\le 5$) を抽出し，それらの n-gram について，もし，二つの n-gram の単語が小文字化して考えたときに同じ字面となる場合には，頻度の少ない方を除去した．たとえば，"White/NNP House/NNP" と "white/JJ house/NN" は，単語を小文字化したときには同じになるので頻度の少ない "white/JJ house/NN" を除去した．こうすることにより，品詞や大文字小文字の揺れを吸収することを意図した．これ以外の正規化，たとえば，基本形を得る，などはしていない．

[1] http://www.nswc.navy.mil/compstat/density.html
[2] http://www.cis.upenn.edu/~treebank/
[3] http://www.hcu.ox.ac.uk/BNC/
[4] http://www.cs.nyu.edu/cs/projects/proteus/oak/

4.2 基本統計量

4.2.1 n-gram に含まれる連語

抽出した n-gram と，それに含まれる連語の数，および，連語のパーセンテージを表1に示す．この表より，nごとに連語の割合が大きく違い，かつ，4,5-gram における連語が非常に少ないことがわかる．そのため，実験では，2,3-gram のみについて別々に連語を抽出した．

	連語	n-gram	パーセント
2-gram	8442	674186	1.25
3-gram	1030	794424	0.13
4-gram	160	391954	0.04
5-gram	10	138402	0.007

表 1: n-gram に含まれる連語

4.2.2 品詞列パターン

2,3-gram の連語について，品詞列のパーセンテージを上位 10 位について表2に示す．ここで，2-gram 連語の異なり品詞列数は，178 であり，3-gram では 181 である．また，それぞれについて，一個の単語列にしか対応しない品詞列は，2-gram 連語については 58 であり，3-gram については 106 である．これらのことから，品詞パターンは多様であり，したがって，単に品詞パターンによりフィルタリングした場合には，多くの n-gram が残ると言える．実際，連語として実際に出現した品詞列に該当する n-gram のみを考慮したとしても，その数は，2-gram では 382687，3-gram では 150187 であり，このときの連語のパーセンテージは 2-gram では 2.21%，3-gram では 0.69%である．これらは，表1での数値よりは多いが，それでも，その中から統計的尺度のみにより連語を抽出するには低い数字である．

NN NN	30.1	NN IN NN	22.2
JJ NN	22.8	NNP NNP NNP	10.6
NNP NNP	13.3	IN DT NN	8.0
VB RP	4.7	NNP IN NNP	6.3
VB IN	3.6	JJ NN NN	3.3
IN NN	2.3	NN IN NNS	2.9
VBG NN	2.0	VB DT NN	2.3
NN IN	1.7	NN CC NN	2.2
VB RB	1.1	IN IN NN	2.1
VBN NN	0.9	IN JJ NN	1.7

表 2: 品詞列パターンとパーセンテージ

4.3 実験手法

10 回の実験を行った．各々の実験では，2,3-gram の連語について，それぞれを半分に無作為に分け，一方を訓練，他方をテストデータとした．評価の尺度は平均精度 (Baeza-Yates and Ribeiro-Neto 1999) である．平均精度とは，ソートされた連語を上から見ていき，連語がみつかるごとに，そこまでの精度を調べて，それらの精度の平均をとったものである．このとき，上位 10000 個の連語を見た．なお，ソートにおいて同一スコアを持つものの順位は無作為に決定した．また，平均精度は各実験ごとに求め，それらを平均した値を得た．以下では，それを報告する．

4.4 実験結果

表3と 4には，2,3-gram についての平均精度を示す．「フィルタリング」の列の数字は，訓練データ中に現われた品詞パターンに該当する n-gram のみについて，それぞれの統計的尺度によりソートしたときのものである．一方，「提案手法」の列の数字は，式 (1) によりソートしたときのものである．なお，$\alpha = 0.5$ は式 (1) において，α を 0.5 としたということであり，$\alpha = \max$ は，α を 0 から 1 まで 0.1 刻みに変化させたときの平均精度の最大値が列に示されていることを示す．

尺度	フィルタリング	提案手法	
		$\alpha = 0.5$	$\alpha = \max$
LLR	.068	.307	.328
Freq	.035	.271	.274
MI	.052	.239	.277

表 3: 2-gram についての平均精度

これらの結果より，提案手法の方が，フィルタリングのあとで統計的尺度によりソートす

尺度	フィルタリング	提案手法	
		$\alpha = 0.5$	$\alpha = \max$
LLR	.047	.096	.107
Freq	.026	.089	.098
MI	.025	.042	.043

表 4: 3-gram についての平均精度

るよりも顕著に精度が高いことが分かる．なお，式 (1) において，$\alpha = 0$，すなわち，統計的尺度を使わない場合について，既に，2-gram で平均精度が 0.24，3-gram では 0.033 であるので，品詞列の情報 (重み) が，連語を抽出するにあたって重要であることがわかる．しかし，フィルタリングでは，あるパターンに一致する全ての品詞列を全て同一視するため，この情報を利用できない．これは確率に基づく式 (1) の有効性を示している．また，これらの表から，LLR が Freq や MI よりも平均精度が高いことがわかる．

5 関連研究

(Schone and Jurafsky 2001) は，複数の統計的尺度について，WordNet における連語を抽出する能力を比較している．彼らの方法は，ヒューリスティクスは用いても，形態素解析器などの言語資源はなるべく使わないというものである．彼らの方法は教師なしの方法と言える．一方，我々の手法は，教師データを用いることにより，複数の情報を有機的に統合し，連語を抽出することを試みている．彼らの抽出手法と我々の抽出手法の精度を比較することは，連語を抽出したコーパスが違うため困難であるが，彼らの手法での最高精度は，任意の n-gram を抽出対象としたとき，0.282 であり，我々の手法の最高精度は，2-gram を抽出対象としたとき，0.328 である．このことは我々の手法の有効性を示していると考える．

6 おわりに

本稿では，連語の抽出に教師あり学習を適用することにより，教師なし学習を利用する場合に比べて高精度に連語が抽出できることを示した．一般の連語の抽出に教師あり学習を適用したのは，本稿が最初である．

本稿で利用した教師あり学習の方法は，非常に単純なものである．それは，統計的尺度と品詞パターンという二つの素性しか利用していないし，また，それらの素性間の確率的独立性を仮定している．教師あり学習の方法として，もっと洗練されたものには，サポートベクトルマシンや最大エントロピー法などがあり，素性として，もっと洗練されたものには，たとえば，構文情報がある．これらの学習手法や素性を利用することにより，より精度の高い連語抽出ができるものと考えている．

参考文献

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, chap. 3. Addison-Wesley.

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

Dias, G., Kaalep, H., and Muischneck, K. (2001). "Automatic Extraction of Verb Phrases from Annotated Corpora: A Linguistic Evaluation for Estonian." In *ACL-2001 Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, pp. 47–53.

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

Kageura, K. (1997). "Type-based and Token-based Learning of Kanji Morphemes." In *3rd International Conference on Quantitative Linguistics*, pp. 146–151.

Kageura, K. and Umino, B. (1996). "Methods of Automatic Term Recognition." *Terminology*, **3** (2), 259–289.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Ries, K., Buφ, F. D., and Wang, Y.-Y. (1995). "Improved Language Modeling by Unsupervised Acquisition of Structure." In *ICASSP-95*.

Schone, P. and Jurafsky, D. (2001). "Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?." In *EMNLP-2001*, pp. 100–108.

Smadja, F. (1993). "Retrieving Collocations from Text: Xtract." *Computational Linguistics*, **19** (1), 143–177.