固有ベクトル分析を用いた主語特徴の評価

5B−7

金　淵培　　江原　暉将
NHK放送技術研究所
email: kimyb@strl.nhk.or.jp　eharate@strl.nhk.or.jp

1　はじめに

日英機械翻訳における日本語ニュース文の翻訳精度を上げるため、自動短文分割を行なっている。しかし、分割を行なった結果、主語の無い文が発生することがある。このような文に対しては、主語補完を行なわねばならない。補完方法は主語になる名詞の特徴ベクトルの確率分布を推定し、各主語候補に対して確率を計算し、この値が最大となる候補を主語として採用している[1)]。

本報告では、主語になる名詞と主語にならない名詞の特徴の内、主語の認定に最も確率的に影響力が強い特徴を把握するため、特徴ベクトルの確率分布に対して主成分分析を行い、その結果について述べる。

2　特徴ベクトルの主成分分析

主語候補の名詞句（N）と主語補完の対象となる述語（V）の特徴を表す関係として次のような8個の変数[1)]を利用した。

x1）Nに付属する格助詞の種別
x2）Nに付属する意味マーカの存在
x3）NとVの意味的整合度
x4）NとVの間の連体節の存在
x5）NとVとの間にある「は」格要素の数
x6）NとVとの間にある「が」格要素の数
x7）NとVとの間にある「は」と「が」以外の格要素の数
x8）NとVとの間にある動詞の数

これらの変数を数量化し、あらかじめ人手で抽出した主語になる関係111件とならない関係297件に対して主成分分析を行なった。前者をケース1、後者をケース2と呼ぶ。

主成分分析は相関係数行列の固有値問題として定式化出来るので、まず各変数の分散が1となるように標準化し、そこから相関係数行列を求め、これに対して固有値と固有ベクトルをヤコビ法(Jacobi)で推定した。ケース1、2に対する固有値の値を図1に示す。

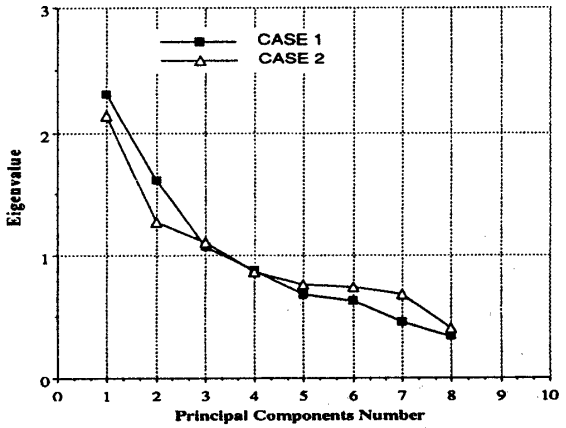

図1　各主成分と固有値

ここでは相関係数行列を使用しているので、総分散値(Total Variance)は変数の数（8）と一致する[2)]。各固有値を8で割ると、総分散に対する各主成分の寄与率が分かる。図2では、その累積寄与率を示す。

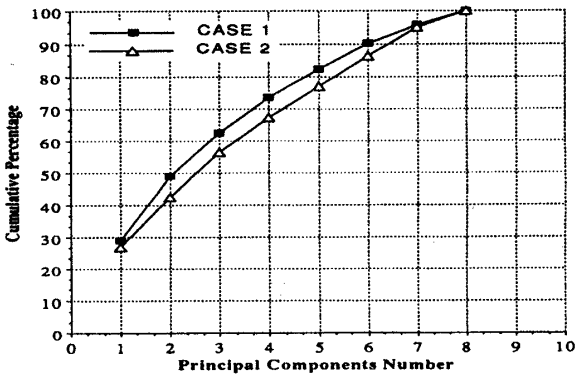

図2　各主成分と累積寄与率

一般的に、主成分分析はなるべく少数の主成分で全体の持つ変動を説明することが目的であるので、意味のある主成分の個数を定める必要がある。その定め方として累積寄与率が70〜80%程度の成分まで選ぶ

Principal components analysis on subject-predicate agreement by Yeun-Bae Kim and Terumasa Ehara
NHK Science & Technical Research Laboratories

のが普通である[3)]。

図２では、ケース１の場合、主成分の１から４までが総分散の約７３％を説明しているので４まで選択する。ケース２の場合は主成分５まで（７７％）選択する。

正規化された固有ベクトルの各要素（a_{ij}）と固有値λ_iを用いて第ｉ主成分（C_i）と変数（X_j）との相関係数（r_{ij}）は$r_{ij}=a_{ij}\lambda_i^{1/2}$となるので、例えば、相関係数が０．５より大きい$a_{ij}$は、次の値T(threshold)より大きくなければならない。

$$T = 0.5 / \lambda_i^{1/2}$$

表１と表２（全ての主成分１～８が表示）では、相関係数が０．５より大きa_{ij}は下線されてある。ここでは、ケース１とケース２の場合に分離して変数と成分の関係について述べる。

１）ケース１の場合（表１：主成分１～４）

ａ）主成分１は変数ｘ１、ｘ６、ｘ７、ｘ８と強い相関関係を持っている。これらの変数は格要素の種別や主語と述語の距離を計る幾つかの変数であるため、構文的情報(Syntactic Information)の重み付き平均(Weighted Average)と考えられる。

ｂ）主成分２は主語と述語の意味的整合度を計る変数ｘ２、ｘ３のみと強い相関関係を示している。この主成分は意味的情報を表していると思われる。主成分２が意味情報変数に対して高い値を持っていることから、その情報は主語の認定に重要であることが分かる。

ｃ）主成分３は連体節との関係（ｘ４）、主成分４は「は」格要素の数を計っていると考えられる。

２）ケース２の場合（表２：主成分１～５）

ａ）主成分１はｘ５、ｘ６、ｘ７、ｘ８の様に主語と述語の距離を計る変数と一致しているので距離が非主語の認定に重要であることが分かる。

ｂ）ここでも主成分２は意味情報変数と強い相関関係を持つので、意味情報は非主語の認定にも必要であると考えられる。

ｃ）主成分３は格助詞の種別と連体節との関係と相関が大きい。しかし、これらの情報は距離と意味情報ほど寄与率は高くない。

ｄ）主成分４ではＴより大きものはない。主成分５は距離の変数であるｘ６と相関関係を持つ。

表１　ケース１の主成分分析結果

Xj＼Ci	1	2	3	4	5	6	7	8
x1	0.403902	0.296086	-0.171534	-0.247825	-0.182878	0.711547	0.324602	-0.115090
x2	0.076206	0.624174	0.217917	0.066916	-0.174215	-0.489713	0.518614	0.116208
x3	-0.058377	0.618773	0.225663	0.332041	0.230692	0.211406	-0.540837	-0.249242
x4	0.079052	0.178853	-0.828555	0.242726	0.422766	-0.141579	0.073584	0.110241
x5	-0.278933	0.240237	-0.013743	-0.836002	0.384241	-0.090207	-0.095551	0.022050
x6	-0.423444	0.140941	-0.415003	-0.089826	-0.646738	-0.111976	-0.153049	-0.407801
x7	-0.517675	-0.107901	0.113455	0.201553	0.343131	0.202271	0.542227	-0.462830
x8	-0.544337	0.136568	-0.019321	0.131706	-0.143547	0.357149	0.034502	0.719705
T	0.329137	0.392631	0.481543	0.532547	0.604236	0.628880	0.740249	0.860240

表２　ケース２の主成分分析結果

Xj＼Ci	1	2	3	4	5	6	7	8
x1	0.327918	-0.049175	-0.512705	-0.480543	0.198500	0.269809	-0.511344	-0.150305
x2	0.060286	0.677750	-0.056818	-0.255289	-0.109100	-0.665277	-0.109321	-0.046603
x3	-0.114146	0.633036	-0.109226	0.495650	0.322052	0.427863	-0.084859	-0.186152
x4	0.014018	-0.164511	-0.793910	0.415110	-0.015795	-0.302742	0.210832	0.183791
x5	-0.411532	-0.293088	0.110310	0.277835	0.227160	-0.367109	-0.646733	-0.225269
x6	-0.429427	0.058794	-0.218284	-0.051701	-0.773394	0.225299	-0.100702	-0.320587
x7	-0.466697	-0.059111	-0.153531	-0.375563	0.427698	-0.062613	0.477478	-0.446374
x8	-0.551396	0.132615	-0.091076	-0.254968	0.109830	0.142973	-0.136144	0.744334
T	0.342082	0.442848	0.473487	0.536332	0.570279	0.580654	0.602758	0.784410

以上まとめると、ケース１、２共第４主成分までに全ての変数が含まれており、不必要な変数は存在しない（Ｔが０．５の場合）。また、上位の主成分との相関が極端に大きい変数も存在しない。

３　おわりに

本報告では、主語候補の特徴に対して主成分分析を行った。その結果、ケース１と２に対して、第４主成分まで考慮すれば十分であることが分かった。今後は分析の結果を利用して変数を増減させて主語補完率の変化を比較したい。又、より多くの特徴を用いて補完率の向上を図りたい。

参　考　文　献

１）金，江原：日英機械翻訳のための日本語ニュース文自動短文分割と主語の補完，情報処理学会自然言語処理研究会資料，９３－３（１９９３）

２）Afifi A.A., Virginia C.:Computer-Aided Multivariate Analysis,Van Nostrand Reinhold, (1990)

３）柳井，高木：多変量解析ハンドブック，現代数学社，（１９８９）