ゼロ主語補完のための評価関数の考察

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

1　はじめに

日本語の長文を複数の短文に自動的に分割する際、ゼロ主語補完の問題が生じる。我々はこの問題を解決するための一つの手段として統計的手法を用いた補完方式を採択している。この手法では、８つの特徴パラメータを基に、ある特定述語に対する各主語候補について「主語になりうる確率密度（ｐ）」と「なりえない確率密度（ｑ）」を別々に推定し、これらの主語候補の中で「ｐ／ｑ」の値が最大になるのを主語として認定する評価関数を採択している。

この研究では、現在主語認定に使われている特徴パラメータを変更せずに現評価関数より精度の高い方法の有無を調べた。その結果、現評価関数は最適に近い方式であることが分かった。しかし、現関数の弱点として、低頻度領域の候補に対して誤認定が起こる場合がある（全エラーの約１５～２０％）。この問題を改良するために現方式とスケール変換による方式を統合したハイブリット型の評価関数を検討した。

この報告では４種類の評価関数に対して性能比較実験を行い、その結果について述べる。

2　主語補完における評価関数の重要性

長い日本語ニュース文をより正確に英語へ機械翻訳するための手段として長文をいくつかの短文にあらかじめ分割して翻訳している［金］。しかし、分割によって主語のない短文が生成されるので、主語の自動補完が要求される。自動補完を行うために、まず、放送データベースから「述語－主語」と「述語－非主語」の形態の２種類の学習データ（Ｔ１とＴ２）を作成した。一般的に、１文から得られる「述語－非主語」のデータの数は「述語－主語」の数より多いので、Ｔ２のサイズはＴ１より大きい。ここで、異なる２種類の学習データを用いる理由は、極めて単純で、Ｔ１又はＴ２のどちらかを用いるよりは、Ｔ１とＴ２の両方を用いた方が制約が強く、認定率の向上につながるからである。この向上については、３章で述べる。

Ｔ１とＴ２のデータを８つの特徴パラメータに数量化した形式に変換する。これらの学習データに対する特徴パラメータ化及び統計モデルによる確率値の推定の細かいメカニズムについては文献［江原］を参照されたい。

この統計モデルは、ある特定の述語に属する各主語の候補（真の主語も含む）に対して、Ｔ１から「主語になりうる確率密度（ｐ）」とＴ２から「主語になりえない確率密度（ｑ）」両値を同時に出力する。

主語になりうる候補は、平均的にｐの値は高く、ｑの値は低い。主語になりえない場合はｐの値は低くて、ｑの値は高い。一方、ｐとｑの両値が共に高い（又は低い）場合もある。したがて、ｐとｑをどのように評価するかによって主語補完の精度が決定されるので、評価関数の選択は極めて重要である。

今回の研究では、現補完方式で用いる特徴パラメータの種類や数を変更せずに主語補完の精度を上げることが目的であるので、ｐとｑから主語認定に最も有効な情報を抽出することができる評価関数を推定することに重点を置く。

3　線形型評価関数

現在、我々が採択している補完方式は、主語候補の中で「ｐ／ｑ」の値が最大になるのを主語として認定する方式である。この方式は、２章で述べたようにｐの値が高くてｑの値が低いほど「ｐ／ｑ」の値が大きくなり、その候補が主語として認定されやすい。

この方式を可視化すると図１のようになる。例えば、ある述語「Ｖ」に対して、３つの主語候補「Ｃ１」、「Ｃ２」、「Ｃ３」があると仮定する。

現方式は、各候補のｐとｑの自然対数値を取ってプロットした際、４５度の線から垂直上方に最も離れて

An evaluation method for zero-subject resolution
Yeun-Bae Kim & Terumasa Ehara
NHK Science & Technical Research Laboratories

いる候補を主語として認定するのと同等である。この垂直距離（ｄ１、ｄ２、ｄ３）は"log(p)- log(q)" を取ることで簡単に計算できる。

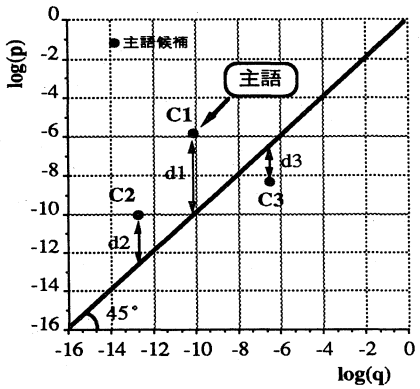

図１　ｐ／ｑによる評価関数

さて、４５度線より精度が高い線の有無を確認するために係数 α を次のように現評価関数に導入して性能を計った。

$$\alpha \log(p) - (1 - \alpha) \log(q)$$

where　$0 \leqq \alpha \leqq 1$　with step = 0.1

ここで、 $\alpha = 0$ の場合はｑ値のみを用いた評価であり、 $\alpha = 1$ の場合はｐ値のみによる評価になる。

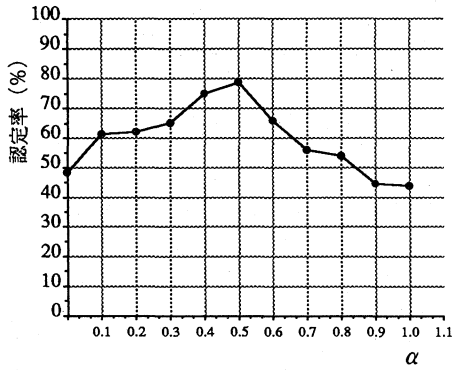

図２　α による主語認定率の変化

$\alpha = 0$ の場合の認定率（４８．４%）が $\alpha = 1$ 場合（４３．９%）より高い理由として、Ｔ２の学習データの数がＴ１より多いことが考えられる。

実験より、 $\alpha = 0.5$ の場合が最適であることが分かった。即ちＴ１とＴ２の両方を均等に用いる４５度線が最適である［図２］。ここでは、この現方式を「方式１」と呼ぶ。

線形型関数*1を用いた他の手法として、まず、各学習データＴ１とＴ２に対してｐとｑの自然対数値を推定する。次は、Ｔ１に対して回帰直線Ａを推定する。Ｔ２に対しても回帰直線Ｂを求める。各主語候補と直線Ａ間の距離を d_A、そしてＢとの距離を d_B とする際、「$D = d_A / d_B$」の値が最大になるのを主語として認定する［図３］。

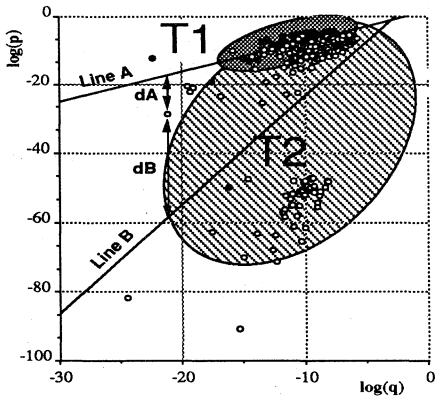

図３　回帰曲線による評価関数

この方式では、主語になりうる可能性が高い候補は、d_B は大きくて d_A は小さいと仮定するので、Ｄの値が最大になる候補を主語として認定する。これを「方式２」と呼ぶ。

4　スケール変換による評価方式

３章で述べた線形型関数の他、スケール変換による評価方式が存在する。対数によるスケール変換は、一般的に小さいまたは大きい数値の取り扱いが容易になるので便利である。スケール変換による方式は、ある候補のｐとｑの値が極端に小さいが（例えば e^{-10}）、ｐ／ｑの値が大きくなって１位に上がるのを防ぐために有効である。

実例として次のようなケースがある：述語「放つ」に対する３つの候補「今日、人、音」の「ｐとｑ」の値を表１の例１に示す。この例では、「放つ」の主語は「人」である。しかし、「今日」のｐとｑの値が小

＊１：「線形型」の意味は、ある曲線（ここでは直線）を基準に評価が行なわれることを示す。

さいにもかかわらずｐ／ｑの値が最も大きいので「今日」が主語として認定され、「人」は２位に下がる。一方、全候補のｐとｑの値を自然対数スケールに変換して、「log(q) / log(p)」の値で比較すると「人」が１位に上がり思わしい認定結果が得られる。方式１では４５度線からの距離で評価するが、この方式では、傾きで評価することになる。この方式を「方式３」とする。

表１　各主語候補のｐとｑの値

例１：　述語　=>　放つ 候補	p	q	log(p)	log(q)	p/q	log(q)/log(p)
今日	0.000012	0.00000015	-11.35	-15.68	75.61	1.38
人	0.002854	0.000041	-5.85	-10.10	69.62	1.72
音	0.000213	0.001469	-8.45	-6.52	0.14	0.77

例２：　述語　=>　戻る 候補	p	q	log(p)	log(q)	p/q	log(q)/log(p)
灯油	0.002107	0.000369	-6.16	-7.90	5.71	1.28
価格	0.000575	0.000086	-7.46	-9.35	6.65	1.25

スケール変換による逆転は、候補１：今日（ｐ１、ｑ１）と候補２：人（ｐ２、ｑ２）の間に次の条件が成立する場合生じるのが分かった。
まず、

$$p1 / q1 > p2 / q2 \quad (p1, p2, q1, q2 \neq 0)$$
$$\log(q1) / \log(p1) < \log(q2) / \log(p2)$$

として、この不等式を展開すると次の関係が得られる。

$$(\log(p2)/\log(q2)) < (\log x / \log y)$$
$$\text{where} \quad x = p1/p2 \quad y = q1/q2$$

この不等関係を可視化したのが図４である。これは、候補１と候補２を結ぶ直線の横軸からの角度（tan β）が原点と候補２を結ぶ直線の横軸からの角度（tan α）より小さければ「log(q2) / log(p2)」が「log(q1) / log(p1)」より大きくなる。

その結果、例上の候補「今日」と「人」のランクは逆転される。この結果を方式１と同じく４５度線からの距離の評価方法で示すことができる。即ち、距離は、"log(-log(q)) - log(-log(p)) " を計算することになる。図５で示すようにスケール変換後の４５度線からの距離は「人」の方が「今日」より大きい。スケール変換による方式は、この例のように低頻度の候補（図４の斜線部分）、即ち、ｐとｑの両値が小さくて信頼度が低い領域に有効である。しかし、表１の例２のように誤った認定が行なわれる場合もある。

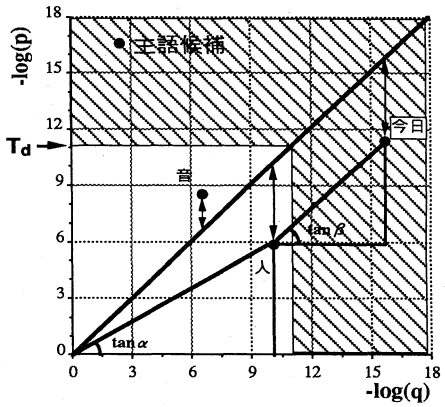

図４　スケール変換前の主語候補

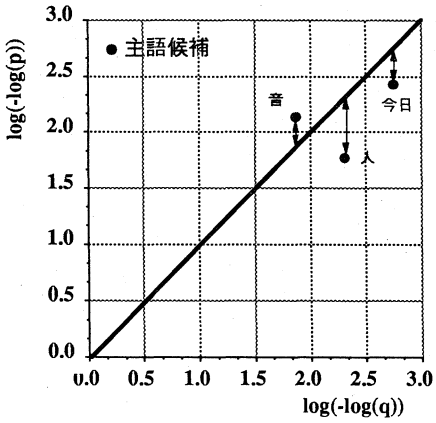

図５　スケール変換後の主語候補

５　統合型評価関数

４章で述べたスケール変換はｐとｑの値が極端に小さい場合に有効であり、線形関数は逆の場合に有効であるので、これらの２つの方式を統合してお互

いの良い所を利用することが可能である。この方法を方式４とする。統合方法のアルゴリズムは次である：

1）方式１と３を用いて各々の評価基準で１位と２位を決定する。

2）方式１と３による認定結果が一致した場合は、その結果をそのまま利用する。

3）方式１と３による認定結果が異なる場合は、方式１の１位の候補のｐとｑの両自然対数値があるスレッショールド値（t_d：図４を参照）より小さければ方式３の結果を優先する。

したがって、スレッショールド値を決める必要がある。今回の実験では、$t_d = -10$を用いた。

6　評価実験と性能比較

今回の精度実験では、ＮＨＫのニュースデータベースからランダムに選択された３８１文の中で自動分割によって主語補完が必要な１０８文を用いてＴ１（１１０個）とＴ２（２９８個）を作成した。４回のcross-validationによって各評価関数の精度比較を行なった［表２］。

表２　各方式による主語認定率

	方式1		方式2		方式3		方式4	
	1位	2位まで	1位	2位まで	1位	2位まで	1位	2位まで
1回目	84.1	96.3	70.7	90.2	84.2	96.3	84.2	96.3
2回目	82.9	96.3	73.2	94.0	80.5	96.3	81.7	96.3
3回目	86.6	96.3	81.7	97.6	86.6	96.3	87.8	96.3
4回目	88.9	96.3	65.4	92.6	88.9	93.8	88.9	96.3
平均値	85.6	96.3	72.8	93.6	85.1	95.7	85.7	96.3

（ａ）学習データに対する認定率

	方式1		方式2		方式3		方式4	
	1位	2位まで	1位	2位まで	1位	2位まで	1位	2位まで
1回目	74.1	88.9	70.4	88.9	81.5	92.6	81.5	92.6
2回目	81.5	92.6	77.8	92.6	74.1	92.6	81.5	92.6
3回目	77.8	96.3	70.4	88.9	74.1	96.3	77.8	96.3
4回目	82.1	96.4	78.6	92.9	85.7	96.4	82.1	96.4
平均値	78.9	93.6	74.3	90.8	78.9	94.5	80.7	94.5

（ｂ）試験データに対する認定率

学習データ［表２のａ］と試験データ［表２のｂ］の両データに対して、方式１と３を統合した方式４の認定率が方式１と３と比べて認定率が高い。表２ｂでは１．８％高いことが分かった。今回の実験では、逆転が起こりうるケースが少なく（２件）、正確にスケール変換の効果を計れなかった。

一方、方式２の認定率が最も低いが、これは主語と非主語のオーバラップが大きくて回帰曲線（直線）によって分離しにくいと考えられる。ここで、回帰曲線を直線から多項式に換える方法もあるが、実際に適用した結果、学習データに対して高い認定率を得ることができるが、試験データに対しては認定率が低下するので、オーバフィティングの問題が生じることがわかった。

7　おわりに

今回提案した４種類の評価方式とは異なる幾つかの方式が存在するが、現方式より遥に高い認定率は期待し難い。その理由は、今回の実験で得られた８０％という認定率は、現在用いる特徴パラメータを使って得られる認定率の限界値に近いと考えられるからである。

今後の課題として、１）スケール変換による低頻度候補への効果の確認、２）現在用いる特徴パラメータより優れた新パラメータ群の設定、３）新パラメータに群対する最適な評価関数の設定などがある。

参　考　文　献

［金］金、江原：日英機械翻訳のための日本語長文自動分割と主語の補完、情処学論、Vol.35, No.6,pp.1018-1028,1994

［江原］江原、金：統計的学習法を用いた主語の補完、信学会「自然言語処理における学習」シンポジウム論文集、pp25-32,1994