意味分類番号を用いた主語・述語の整合度の計算方法

5P－8

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

筆者らは、放送ニュースを対象にした日英機械翻訳の前処理として、長文の自動短文分割と、分割に伴って主語の無くなった文に対する主語の補完[1]を行っている。主語を補完するには、主語候補となる名詞と補完対象の動詞（形容詞、形容動詞、述語名詞を含む）の間の意味的な整合度が重要な情報であり、我々のシステムでは、「が」格関係の係り受けデータ[2]を利用して、整合度を計算している。実際には、係り受けを構成する名詞と動詞の双方に意味分類番号[3]を付与し、この分類番号を用いて、整合度を算出している。ここでは、同一の分類番号に属する語としては、その係り受けの性質が類似していると仮定している。分類番号の付与は形態素解析と分類番号辞書を用いて、自動的に行っているので、多義性が問題になる。つまり、現在は、分類番号が複数付与された語を含むデータに対して、その番号が均等に出現したと仮定して、データの度数を案分した案分データ（後述）を利用しているが、本来は、多義性を解消して、そのデータの意味に合った番号のみが出現したとするべきである。本論文では、分類番号の出現度数に基づいて、多義性を解消する２つの方法を提案し、その性能の比較評価実験について述べる。

2．多義性の解消法

係り受けデータは、名詞 i （Ni）が「が」を介して動詞 j （Vj）に係る関係とその度数（nij）から成っている。そして、Ni および Vjに、それぞれ、分類番号が付与されている。ここで、Ni、Vj に付与された分類番号の集合を、それぞれ、Ri、Sjと書く。これらの集合の要素数が２以上であれば、多義である。このとき、Ri、Sjの中から、「Ni がVjする」の意味として適切な番号の組を選択するのが多義性の解消である。その手順を示す。

まず、以下の様にして、度数nijを案分したデータを作成する。Ri、Sjの要素数をDri、Dsjとするとき、Riの各要素 rk （k=1, 2, ..., Dri）とSjの各要素 Sl （l=1, 2, ... Dsj）に対して、Niが rkの意味で、「が」を介して、Vjの Slの意味に係るデータの度数をnij /(Dri ×Dsj)とする。これを案分データと呼ぶ。

次に、案分データに基づいて、分類番号に関する係り受けデータを作成する。つまり、案分データから語形を無視して、分類番号rが「が」を介して、分類番号sに係る度数を足し上げ、mrsとする。（r, s, mrs）の全体を分類番号係り受けデータと呼ぶ。このデータに基づいて、Ri、Sjの中から、適切な組を取り出す方法を考えよう。

１つの方法として、r がsに係る度数mrsを考慮するとき、評価関数として

M1(r, s) = log(mrs)　(1)

を定義し、M1(rk, Sl)が最も大きい rkとSlを選択するやり方がある（相互情報量と合わせるためlogを取る）。これは、分類番号間の共起しやすさを用いているわけであるが、各分類番号にはそれに属する語の数に差があり、これが無視されることになる。すなわち、語を多く含む分類番号が選択されやすくなる。この点を改善する方法として、相互情報量を評価関数として用いる方法がある。これは、次のようにして計算される。案分データから、分類番号 r が係り元に出現する度数mrを求める。同様に分類番号 sが係り先に出現する度数msも求める。また、データの総度数をmとする。このとき、r が sに係る事象の相互情報量 M2(r, s) は

M2(r, s) = log {(mrs×m) /(mr×ms)}　(2)

である。このとき、M2(rk, Sl)の最も大きい rkとSlを選択する方法である。しかし、この方法は、第１の方法とは逆な問題がある。それを例で示そう。案分データが

名詞（番号）			動詞（番号）		度数
N1	(r1)	が	V1	(s1)	1
N1	(r2)	が	V1	(s1)	1
N2	(r1)	が	V2	(s1)	10

の場合を考える。このとき、r2よりも r1の方がs1に係りやすいと考えるのが自然であるが、実際に式（２）

A matching score estimation for Sub-Predicate agreement by using semantic categories: Y.B. Kim & T. Ehara
NHK Science & Technical Research Labs.

を用いて計算してみると、

$M2(r1, s1) = \log\{(11\times12)/(11\times12)\} = 0$

$M2(r2, s1) = \log\{(1\times12)/(1\times12)\} = 0$

となり、等しくなってしまう。

そこで、α（$0 \leqq \alpha \leqq 1$）をパラメータとして

$M3(r, s) = \log(mrs) + \alpha\times\log(m/(mr\times ms))$ (3)

を評価関数と定義し、αに対してM1とM2を比較する。式(3)では、αが0の場合はM1に一致し、αが1の場合はM2に一致する（αの増加幅：0．1）。

3．評価実験

［文献2］の係り受けデータ（174816個）を元に形態素解析と分類番号辞書（3桁）を用いて、案分データ（465678個）を作成した後、分類番号間の係り受けデータ8680個を番号の2桁のみを用いて作成した。このとき、度数が100以上の案分データ（上位89個）に関しては、人手で、多義の解消を行なった。次に、元の係り受けデータから、分類番号が複数ふられているもので度数が11以上の中から無作為に76データ抽出し、テストデータとした。テストデータに対する、分類番号の組の数は1～24の範囲であり、平均は6．13個である。このテストデータに対し、人間が多義を解消し、正解とした。このとき、人間によっても、分類番号の組が1つに決められなかったデータは、複数の正解があるとした（被験者1名）。このようなものは、15データあり、最多の正解数は4であった。つぎに、各αに対して式（3）の値が最も大きい組を選択し、正解と比較した。図1に、αに対する正しく選択された正選択率の変化を示す。

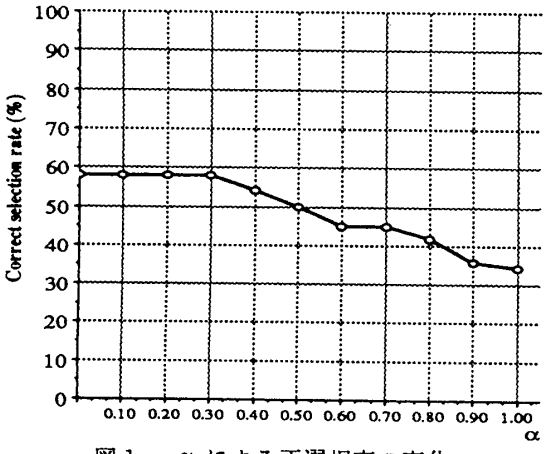

図1　α による正選択率の変化

なお、共起を考えず、名詞と動詞を独立に高頻度のものを選択する方法、つまり、$\log(mr\times ms)$の最も大きいrとsの組を選択する方法でも、正選択率は48．6％であったので、この値よりも大きい方法でなければ、共起を利用する意味がない。

図1から分かるように、相互情報量（α＝1）に近づくほど正選択率が低く、α＞0．5からは、独立に選択する方法よりも低い。$0 \leqq \alpha \leqq 0.3$の範囲は、今回の実験では一定値58．7％を取った。

M1とM2の相関図（図2）を参照すると、相関係数（R）は0．32で、互いにあまり相関が高くないことが分かる。

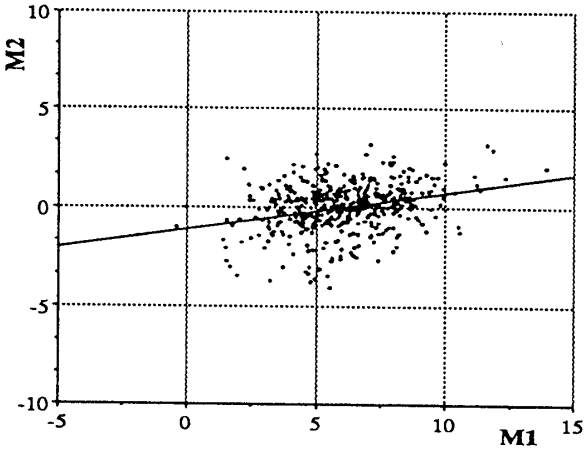

図2　M1とM2の相関図

4．おわりに

我々は、分類番号の出現度数に基づいて、多義性を解消する2つの方式の比較実験を行った。実験結果としてM1がM2より正選択率が高く、M1の方が人間による評価基準とより合致しやすいのが分かった。しかし、まだ正選択率が低く改善が必要である。また人間による評価基準は主観的であるので、それに一致することが多義性解消の意味で適切であるかを検討する必要がある。

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）田中：語と語の関係解析用資料（朝日新聞記事データ1年分）“が”を中心とした，（1989）

3）大野，浜西：類語新辞典，角川書店，（1984）