音声言語情報処理 33－8
(2000. 10. 27)

文字正解精度と相関の高い補正パープレキシティの算出法

永江尚義, 松浦博, 田中信一, 正井康之, 山中紀子

(株)東芝 研究開発センター
〒212-8582 川崎市幸区小向東芝町1
Email: {hisayoshi.nagae, hiroshi.matsuura, shinichi.tanaka, yasuyuki.masai, noriko.yamanaka}@toshiba.co.jp

あらまし　言語モデルの評価尺度として一般にパープレキシティが使用されている。また、未知語を考慮する場合には補正パープレキシティが使用されるが、未知語を扱う際に見出しの内容が考慮されていない。そのため、性能測定の単位として文字を使用した場合、認識性能とパープレキシティとの相関が弱くなる問題がある。提案する方法では、未知語の部分を文字単位に分割し、各文字間の出現確率の積を未知語に対する出現確率として使用することによって、未知語の見出しを考慮する。文字正解精度とパープレキシティとの相関を多項式回帰モデルを使って算出したところ、2次式モデルにおける決定係数の値が50.1%から71.1%に向上し、提案方法の有効性が示された。

A new method for computing perplexity which correlates with character accuracy

Hisayoshi Nagae, Hiroshi Matsuura, Shinichi Tanaka,
Yasuyuki Masai, Noriko Yamanaka

Corporate Research & Development Center, Toshiba Corporation,
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki 212-8582, JAPAN
Email: {hisayoshi.nagae, hiroshi.matsuura, shinichi.tanaka, yasuyuki.masai, noriko.yamanaka}@toshiba.co.jp

Abstract　This paper describes a method for computing perplexity which correlates with character accuracy. The proposed method of computing perplexity splits an unknown word into a sequence of characters and substitutes the product of conditional probability-based characters for conditional probability of an unknown word. In the case that the relationship between character accuracy and perplexity is computed using polynomial regression models, the value of a coefficient of determination for a quadratic model improved from 50.1% to 71.1% by using the proposed methoed for computing perplexity.

1　はじめに

言語モデルの評価尺度として広くパープレキシティ(perplexity)が使用されている。しかし、未知語を考慮しないで(無視して)パープレキシティを算出すると未知語の数が多いほどパープレキシティの値が小さくなってしまうという問題がある。この問題を解決するためには一般に補正パープレキシティ[1]が用いられている。

一方、日本語の場合、正解精度を測定する単位には形態素と文字が考えられる。文字を単位として性能を測定する方法には、比較対象のシステム同士の単語・品詞の体系が異なっていても良い。また、意味的に同一視して良い形態素を判定するために複雑な処理をする必要がないという利点を持っており、性能測定の処理を簡便にすることができる。

しかし、文字を採用した場合、補正パープレキシティと正解精度との相関が弱まると考えられる。一

般に文字単位で正解精度を算出する場合、未知語の見出しが長いと、未知語の影響で誤認識と判定される領域は大きくなる。したがって、未知語の見出しの内容が変化すると全体の正解精度に与える影響は変化することになる。しかし、従来の補正パープレキシティは認識対象中の未知語の領域について未知語という1つの単語としてひとくくりで扱っているため、この点が十分に考慮されていない。

そこで、本報告では未知語の見出しを考慮にいれたパープレキシティの算出方法を提案する。本方法では、未知語の見出しを文字単位に分割して、未知語の領域に対しては文字単位の出現確率を使用して確率推定を行い、その確率値の積を未知語に対する出現確率として扱う。2. では文字単位による正解精度算出法の利点を述べる。3. では、新しい補正パープレキシティの算出方法を提案する。4. では、認識評価実験を行った上で多項式回帰モデルを使用して正解精度とパープレキシティの相関を測定し、提案するパープレキシティが従来の補正パープレキシティよりも文字単位の正解精度との相関が強いことを示す。

2 正解精度算出の単位

今回の認識性能を測定する際の正解精度算出の単位は「形態素(単語)」ではなく「文字(見出し)」を使用することとした。

一般に日本語ディクテーションでは、認識性能を測定する際に形態素を単位として使用する場合が多い。しかし、今回は以下の二点(1), (2)を考慮して文字単位の正解精度を認識性能の指標として採用することにした。

(1) **単語の体系が異なる/不明なシステム同士の性能比較が容易に行える。**

一般に、単語や品詞体系が異なると形態素の単位が変化する。したがって、形態素の体系が異なるシステム同士の性能を比較する場合、認識結果中の形態素の区切り方が異なるために、どちらがどれだけ性能が優れているかを判定するのが難しくなる。

これに対し、文字の場合、判断基準が明確で普遍的なため、形態素の体系に依らず各種システムの認識結果を容易に比較できる。

(2) **評価の判定基準が明確で処理が簡単である。**

正解単語と認識結果単語の形態素の区切りが異なる場合に、厳密に一致(正解)しているかどうかを判定するのは難しい[2]。

次の例は『企画委員会は中長期的な観点から』という文章の正解単語と認識結果の例を示している。正解は8つの形態素で構成されている。これに対し、認識結果では「委員会」と「中長期的」の部分の区切り方が正解と一致していないために置換誤り(S)、挿入誤り(I)と判定されていることを示している。

正解：企画/委員会/は/中長期/的/な/観点/から
結果：企画/委員/会/は/中/長期的/な/観点/から
S I S S

一般に、形態素解析の解析結果は必ずしも1位候補として解析された候補だけが意味的に妥当で、それ以外はすべて誤りであるというわけではない。したがって、ディクテーションの性能を測定する際に、形態素解析の解析結果の1位候補とまったく同じ単語区切りに認識できていないからと言って、正しく認識されていないわけではない。しかし、正解判定を行う際にこのような形態素の意味的な同一性まで考慮すると判定処理は非常に複雑になってしまう問題がある。

一方、正解の判定基準を文字単位にした場合には、このような異なる形態素区切り同士の比較が必要なくなるため、非常に簡便に処理することができる。

ただし、文字単位の正解精度には次のような問題点(a),(b)もあるため、将来的にはより正確に性能を測定するための判断基準を検討していく必要があると考えられる。

(a) **意味を考慮した正解判定が難しい。**

形態素単位であれば、例えば同じ意味を表す同音語を同一視して評価を行うようなことも可能である。しかし、見出し情報だけに基づいて文字単位で処理を行う場合にはこのような意味を考慮した判定が非常に難しくなる。

(b) **たまたま表記が一致しているだけの場合でも正解になってしまう。**

例えば、「小野寺/田鶴子」という人名が「小野/寺田/鶴子」という誤った区切りの単語列と

して認識されていても文字単位での正解判定では誤りとして検出することができない。このように文字単位で正解かどうかを判定すると、たまたま見出しの表記が一致しただけの誤認識に対処することができない。

3 補正パープレキシティの算出法

3.1 従来の補正パープレキシティ

言語モデルの評価基準の一つであるテストセットパープレキシティPPは式(1)で定義される。

$$PP = P(w_1 \cdots w_n)^{-\frac{1}{n}} \quad (1)$$

一般にパープレキシティを算出する場合、未知語に対しては未知語部分をスキップするか、未知語を一つのカテゴリとして扱う方法が多い。しかし、この方法では、語彙サイズが小さいほどパープレキシティが小さくなる欠点がある。そこで、異なる未知語はそれぞれ等確率に生じると仮定した式(2)を用いる補正パープレキシティを使用する方法もある[1]。

$$\begin{aligned} P(unk|h) &\simeq P(unk|UNK)P(UNK|h) \\ &\simeq \frac{1}{m}P(UNK|h) \end{aligned} \quad (2)$$

上式は単語履歴hで未知語unkが現れた確率を表し、UNKは未知語のカテゴリ、mは未知語の種類数をそれぞれ表す。

したがって、補正パープレキシティAPPは次式で定義される。

$$APP = (P(w_1 \cdots w_n) \times m^{-o})^{-\frac{1}{n}} \quad (3)$$

ここで、oは未知語の総数、mは未知語の種類数である。

3.2 未知語の見出しを考慮した補正パープレキシティ

文字単位の正解精度を算出する場合、未知語の領域の長さに応じて正解精度への影響度が変化する。すなわち、見出しが長い未知語はその領域が誤認識することによって文字単位の誤りを多く発生させることになり、正解精度を大きく低下させる。そこで、ここでは未知語の見出しを考慮したパープレキシティPP'を以下のように算出する。

$$\begin{aligned} PP' &= P(w_1 \cdots w_n)^{-\frac{1}{n}} \\ &\simeq \prod_{i=2}^{n} P(w_{i-1}w_i) \end{aligned} \quad (4)$$

ここで、式(4)の単語間の出現確率$P(w_{i-1}w_i)$を次式から決定する。いま、単語w_iが未知語であったとする。そして、w_iが$< u_1^i, u_2^i, \cdots, u_m^i >$という構成要素の文字に分解されたとする。例えば、「宇宙開発事業団」、「TMD」という語が未知語であった場合、それぞれ「宇/宙/開/発/事/業/団」、「T/M/D」という構成要素に分解される。

この時、単語列$w_{i-1}w_i$の出現確率$P(w_{i-1}w_i)$を次式から算出する。

$$P(w_{i-1}w_i) = \begin{cases} \prod_{j=2}^{m} P(u_j^i|u_{j-1}^i) \times P(u_1^i|u_m^{i-1}) \\ \quad \textbf{if } w_{i-1}, w_i = \text{未知語} \\ \prod_{j=2}^{m} P(u_j^i|u_{j-1}^i) \times P(u_1^i|w_{i-1}) \\ \quad \textbf{if } w_{i-1} \neq \text{未知語}, w_i = \text{未知語} \\ P(w_i|u_m^{i-1}) \\ \quad \textbf{if } w_{i-1} = \text{未知語}, w_i \neq \text{未知語} \\ P(w_i|w_{i-1}) \\ \quad \textbf{if } w_{i-1}, w_i \neq \text{未知語} \end{cases} \quad (5)$$

式(5)を使用することにより、未知語の領域に対する出現確率を見出し文字の出現確率から推定することが可能になる。これにより、本パープレキシティは、従来のパープレキシティよりも文字単位の正解精度との相関が強くなり、より正確に文字単位の正解精度を表すと考えられる。

4 評価実験

4.1 実験方法

本研究で提案した補正パープレキシティの有効性を確認するために評価実験を行った。実験に用いた評価文は新聞記事からランダムに選択された50種類の文章を使用した。評価文は総数692文、16602

表 1: 各評価文に対する認識率、未知語数、パープレキシティ

文番号	正解精度 [%]	未知語数	従来の補正 perplexity	新しい補正 perplexity
1)	91.78	0	142.62	142.62
2)	93.89	0	104.95	104.95
3)	94.46	0	122.74	122.74
4)	89.49	3	103.14	146.19
5)	89.48	2	133.55	171.34
6)	89.08	4	168.88	269.86
7)	87.28	9	129.38	326.50
8)	88.36	7	164.06	295.63
9)	88.20	8	129.31	239.22
10)	84.39	14	208.89	548.54
11)	89.95	4	138.25	203.57
12)	93.90	1	125.02	147.98
13)	92.03	0	144.87	144.87
14)	89.45	1	184.02	268.55
15)	87.09	8	225.03	482.37
16)	90.10	5	140.58	539.38
17)	98.06	0	149.19	149.19
18)	93.37	0	144.58	144.58
19)	90.50	0	289.01	289.01
20)	91.54	0	151.64	151.64
21)	92.48	3	92.74	111.82
22)	94.40	1	76.16	96.41
23)	97.27	0	36.58	36.58
24)	91.73	0	83.40	83.40
25)	96.71	0	49.54	49.54
26)	89.25	0	155.81	155.81
27)	94.49	1	60.24	70.15
28)	96.10	0	92.96	92.96
29)	84.23	11	222.01	537.62
30)	85.27	1	221.98	254.45
31)	93.37	0	96.27	96.27
32)	83.09	9	140.54	439.83
33)	93.81	0	118.13	118.13
34)	88.85	0	99.11	99.11
35)	85.28	8	176.48	344.82
36)	81.84	13	200.77	591.01
37)	87.29	6	145.07	331.92
38)	90.94	3	92.08	121.55
39)	90.91	1	137.34	167.48
40)	86.80	0	220.78	220.78
41)	96.57	0	51.39	51.39
42)	85.79	2	163.42	334.00
43)	90.52	0	135.43	135.43
44)	96.13	0	68.20	68.20
45)	90.64	1	149.59	177.64
46)	88.31	2	106.19	181.76
47)	93.64	0	125.77	125.77
48)	89.74	0	104.85	104.85
49)	93.31	0	58.87	58.87
50)	93.85	1	66.98	84.97

形態素で構成され、未知語は延べ 129 語存在している (カバー率 99.2%)。

評価文全文を男性 3 名が発声した音声データを収集して評価データとした。データ収集は計算機雑音の存在する室内で接話型のマイク (ノウルズ社製 VR3310) を用いて、東芝製パソコン (DynaBook SS 3300) で録音することによって行った。なお、文章中の句読点や記号もすべて読み上げている。

認識実験に使用したディクテーションシステムの語彙サイズは 60000 語であり、言語モデルは 8 年分の新聞記事とメールやニュース文を学習コーパスとして作成した bigram, trigram データで構成されている。

4.2 実験結果

各評価文に対する認識率 (文字単位の正解精度) とパープレキシティは表 1 のようになった。

3.2 の計算式から明らかなように未知語が存在しない場合は算出されるパープレキシティの値は従来と同じであるため、未知語が存在する 26 文の時だけパープレキシティの値だけが異なっていることがわかる。また、未知語が存在する文に対するパープレキシティの値は本報告で提案した算出法の方が従来の補正パープレキシティの値よりも大きくなっており、パープレキシティの値の変動幅が大きくなる傾向がある。

4.3 正解精度とパープレキシティの相関

各文章に対するパープレキシティを入力、正解精度を出力とした時、多項式回帰モデルによって相関は次式のように算出された。

(1) 従来の補正パープレキシティと正解精度との相関

$$y = 97.2594 - 4.93 \times 10^{-2} x$$
$$y = 101.361 - 0.113962x + 2.20 \times 10^{-4} x^2$$

(2) 提案した補正パープレキシティと正解精度との相関

$$y = 95.1019 - 2.15 \times 10^{-2} x$$
$$y = 97.7927 - 4.87 \times 10^{-2} x + 4.63 \times 10^{-5} x^2$$

各文章に対するパープレキシティを横軸、正解精度 (accuracy) を縦軸にプロットし、相関を表す式を図示した図を図 1~4 に示す。図 1, 2 は相関を 1 次式モデルとして算出したもの、図 3, 4 は相関を 2 次式モデルとして算出したものである。

これらの結果から得られる、各パープレキシティに対する決定係数 (R^2、図中では R-sq と表記) は表2のようになった。この値は入力に対する出力が多項式回帰モデルの方程式によって説明される比率を表しており、値が大きいほど入力と出力の相関が高いことを意味する。

表 2: 各補正パープレキシティに対する決定係数 [%]

	1 次式モデル	2 次式モデル
従来の補正 perplexity	45.1	50.1
新しい補正 perplexity	64.4	71.1

したがって、従来の補正パープレキシティの R^2 が 1 次式モデルで 45.1%、2 次式モデルで 50.1%であるのに対し、本研究で提案した補正パープレキシティの R^2は 1 次式モデルで 64.4%、2 次式モデルで 71.1%と向上しており、今回提案した補正パープレキシティの方が文字単位の正解精度とより強い相関を持っていることが確認された。

5 おわりに

本報告では、文字単位の正解精度と相関の強い未知語を考慮したパープレキシティの算出方法を提案した。未知語に関する統計値を算出する際に未知語の領域の見出しを文字単位に分割し、文字単位の出現確率の積を使用して、未知語の出現確率を推定する。これにより文字単位で正解精度を算出する場合に、認識性能が未知語の見出しの長さに依存して変化する問題をパープレキシティ算出の際に取り扱うことができるようになる。

実際に認識実験を行い、文字正解精度とパープレキシティとの相関を多項式回帰モデルを使って算出し、相関の度合を表す決定係数 R^2の値を求めた。従来の補正パープレキシティの時、R^2は 1 次式モデルで 45.1%、2 次式モデルで 50.1%であったのに対し、提案したパープレキシティではそれぞれ 64.4%、71.1%に向上した。したがって、提案方法において従来方法と比較してパープレキシティが、文字正解精度とより強い相関関係を持つことが示された。

参考文献

[1] J. Ueberla, "Analysing a simple language model — some general conclusion for language models for speech recognition," Computer Speech and Language, Vol.8, No.2. pp.153–176, 1994.

[2] 山本俊一郎, 伊藤克亘, 鹿野清宏, 中村哲, "ディクテーションにおける日本語の特性を考慮した単語正解判定ツール," 日本音響学会講演論文集 3-Q-19, pp.155–156, 1999.

[3] 中川聖一, 赤松裕隆, "未知語を含む文集合のパープレキシティの算出法 — 新補正パープレキシティ —," 日本音響学会講演論文集 2-1-13, pp.95–96, 1998.

[4] 中川聖一, 伊田政樹, "連続音声認識のタスクの複雑さを表す新しい尺度," 電子情報通信学会 D-II, Vol.J81-D-II, No.7, pp.1491-1500, 1998.

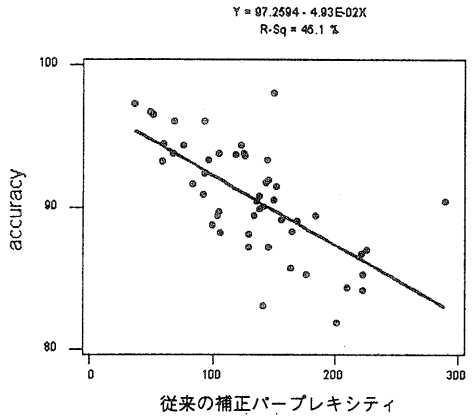

図 1: 正解精度と従来の補正パープレキシティの相関 (1 次式モデル)

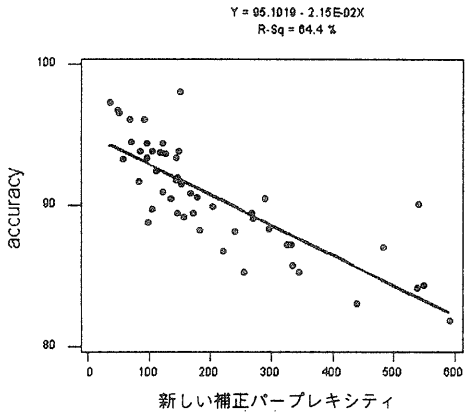

図 2: 正解精度と新しい補正パープレキシティの相関 (1 次式モデル)

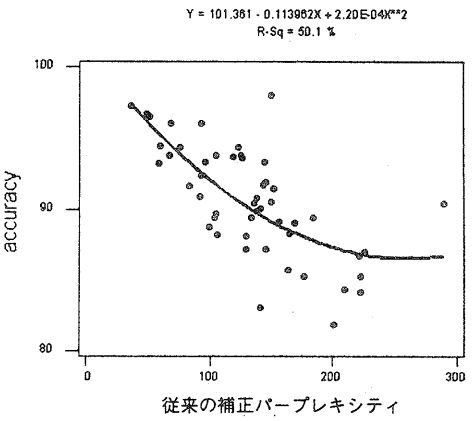

図 3: 正解精度と従来の補正パープレキシティの相関 (2 次式モデル)

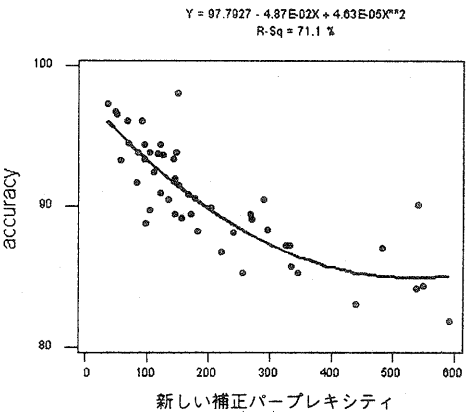

図 4: 正解精度と新しい補正パープレキシティの相関 (2 次式モデル)