E-037

入力予測手法における候補の尤度評価について
Credibility evaluation of candidate for input prediction method

烏　日娜† 荒木　健治† 栃内　香次‡
Wu Rina† Kenji Araki† Koji Tochinai‡

1. まえがき

我々は，帰納的学習による入力文予測を用いた中国語ピンイン入力手法を提案した[1]．入力文予測手法においては，効率的予測を行うために，予測候補の尤度評価が非常に重要である．理想的な尤度評価結果は，正解の尤度が最大であること．すなわち，正解が候補のトップであること．しかしながら，実際のシステムでは，すべての正解を尤度最大と評価することは困難である．如何に正解を予測候補の上位に持ってくるということは，効率的予測の重点になっている．本手法では，予測用ルールの尤度評価関数及び関連語の関連性を用いて予測候補の尤度評価を行うことにより，より効率的予測を図っている．ここでの効率的予測は，正解が予測候補の上位 5 位の中に含まれている場合の正しい予測のことである．関連語辞書は帰納的学習[2]により自動的に生成される．

2. 概要

本提案手法[1]の初期段階の目的は，文章の表層情報のみにより，帰納的学習を用いて予測用ルールを獲得して予測を行う場合の精度を確認することである．学習の持続に伴い，獲得されるルールの数が増加し，予測される候補の数も増加する．したがって，ルールの尤度評価がより重要になる．

獲得されたルールのフォーマットを表 1 に示す．フォーマットの括弧中の「文字列 1」は検索部分になり，辞書検索時使用される．「文字列 2」は予測される文字列になる．また，「#A；B；F；L#」はルールの尤度評価に用いられる情報が記述されている部分になる．

表 1 ：ルールのフォーマット

フォーマット	（文字列 1） 文字列 2 #A；B；F；L#		
各部分	文字列 1	文字列 2	#A～L#
用途	検索部分	予測文字列	尤度計算用

本稿では，ルールの利用された履歴情報（正予測頻度，誤予測頻度）とルールの特徴量（相互頻度，長さ）により決められる尤度評価関数 PEF (Priority Evaluation Function) 及び関連語を用いて予測候補の尤度評価を行う．

（1） ルールの尤度評価関数

尤度評価関数の定義を次の式（1）に示す．

$$PEF = \alpha \times A - \beta \times B + \gamma \times F + L \qquad (1)$$

A： 正予測頻度
B： 誤予測頻度
F： ルールの検索部分と本体部分の相互頻度
L： ルールの文字数（本体部分の文字数）
α，β，γ ：係数

式（1）は，精度が高く，誤りが少なく，頻繁に出現し，文字数が多いルールの尤度が高くなるという意味である．式（1）で正予測頻度 A については，予測候補の中，正解である候補の A が 1 増加する．A の初期値は 1 と設定されている；誤予測頻度 B については，予測候補の中，正解を除く候補の B が増加する．これは予測候補中正解以外の候補が全て誤って予測されたという意味である．本稿では，予測候補の上位 5 位の中に含まれている誤って予測されたルールの B が 1 増加するように設定している．B の初期値は 1 と設定されている．

同一文章において，同じ表現が重複して出現する可能性が高いという特徴を考え，次の規則及び尤度評価式を用いてルールの尤度を評価する．

規則：ユーザが入力している文章の既に入力された部分から獲得されたルールを最優先する．

したがって，予測候補の尤度評価を行う際：

（A）まず，ユーザが現在入力している文章から獲得された予測候補に対して，上述式（1）を用いて尤度評価を行い，尤度値の高い方が上位という順に順位付ける．

（B）次に，残りの候補に対して，同様に式（1）を用いて尤度を評価し，尤度値の高い方が上位というように（A）に続く順で順位付ける．

（2） 関連語

本手法では，帰納的学習を用いて既に入力された文章により関連語を獲得する．関連語の定義は：二つの文が二箇所で二文字以上マッチした場合，そのマッチした共通部分を関連語とする．

表 1 を用いて関連語を獲得する方法を説明する．表 1 で，「X」はそれぞれ異なる任意の文字を示す．「W1，W2，WA，WB，WC」はそれぞれ文字数が 2 文字以上の単語を示す． W1 と WA が同時に文 1 と文 2 に出現する．前述の定義によると，W1 と WA が関連語になる．同様に，W1 と WB が同時に文 3 と文 4 に出現するため，W1 と WB も関連語になる．WA と WB がそれぞれ W1 と関連語になるため，W1 (WA # f #, WB # f #)のように表記する．

関連語の関連性を CRW (Correlation of the Related Word) と呼ぶ．この関連性が関連語の出現頻度，関連語になる単語の距離など多数の要素により決められる．

† 北海道大学大学院情報科学研究科
‡ 北海学園大学大学院経営学研究科

表２：関連語獲得の例

文 1	xxx W1 xxxxxx WA xxxxxxx
文 2	xxxxx W1 xxxxxxx WA xxxxx
文 3	xxx W1 xxxxxx WB xxxxxx
文 4	xxxx W1 xxxxxx WB xxxxx
文 5	xxx W2xxxxxxx WB xxxxx
文 6	xxxxxxx W2 xxxxxxxx WB xxxxx
文 7	xxxxxxx W2 xxxxxxxx WC xxxxx
文 8	xxxxxxx W2 xxxxxxxx WC xxxxx
共通部分	(W1 WA) (W1 WB) (W2 WB) (W2 WC)
関連語	W1 (WA # f #, WB # f #) W2 (WB # f #, WC # f #)

本稿では関連語の出現頻度のみを用いて尤度評価を行う．ルールの尤度評価関数 PEF と関連語の関連性 CRW を考慮した場合の予測候補の尤度評価式を次の式（２）に示す．

$$V = PEF + \varepsilon \times CRW \qquad (2)$$

ε：係数

係数α，β，γとεの値はそれぞれ予備実験により求めた．

3. 実験及び考察

実験の目的は，前述の予測候補尤度評価方法の有効性を確認することである．

実験データとして，中国語で書かれた自然言語処理分野の論文 5 編約 51,800 文字を用いた．実験では，まず，実験データの約 33,170 文字を用いて学習させ，ルール辞書，セグメント辞書及び関連語辞書を作成する．次に，実験データの残りの約 18,630 文字を用いて予測候補の順位を調べ，尤度評価関数を利用する場合と利用しない場合の結果，関連語を利用すると利用しない場合の変化をそれぞれ調べた．

3.1 実験方法

本実験では，まず，尤度評価式の有効性を確認し，次に，候補の尤度評価における関連語の影響を調べた．ルールの尤度評価関数 PEF においては，予測候補の上位 20 個に対する影響を調べた；関連語の関連性 CRW においては，予測候補の上位 5 位に対する影響を調べた．

3.2 実験結果及び考察

実験結果をそれぞれ表３，表４に示す．表３は尤度評価関数のみを用いた場合の実験結果である．表３の「候補の位置」の列は，予測候補の上位 20 位をそれぞれ 1～５，６～１０，１１～２０と三つに分割したことを指す．ここでの正解率は，候補位置ごとに当たる正解の数と総正解数の割合になる．本手法では，正解率を用いて正解の候補中の分布状況を測定する．表４は関連語の関連性を考慮した場合と尤度評価関数のみを用いて評価した場合の結果を比較し，関連性を考慮した場合の正解位置の変化を示す．表４で，予測回数の 337 は，正解が含まれた予測と正解が含まれていない予測の総予測回数である．正解数の 145 は総予測回数 337 の中正解の数である．

表３：尤度評価式の影響

候補の位置	正解率 (%)	尤度評価なしの正解率 (%)	変化量 (%)
1-5	45.3	23.5	21.8
6-10	24.7	23.5	1.2
11-20	24.1	47.1	-23
総数	94.1	94.1	---

表４：関連語の影響

予測回数	正解数	順位上昇	順位下降	順位不変
337	145	83	40	22
割合%	43.0	57.2	27.5	15.1

表 3 から分かるように，PEF を用いて評価することにより，約 21%の正解が上位 5 位以内に入った．また，表 4 によると，関連語を考慮した場合，正解の約 57%が上位 5 位以内に移動された．

4. おわりに

本稿では，入力予測手法における予測候補の尤度評価について述べた．ここでは，予測用ルールの尤度評価関数 PEF 及び関連語の関連性 CRW を用いて予測候補の尤度評価を行い，それぞれの影響を調べた．実験結果から，入力予測システムの候補評価方法として，本評価方法は有効的な評価を行えることを確認した．

本稿では，関連語の出現頻度のみを用いて関連性 CRW を計算した．今後は，関連語の距離情報を用いて評価を行う予定である．

参考文献

[1] Wu Rina, K. Araki and K. Tochinai, A Method for Intelligent Association of Chinese Input Using Inductive Learning, *Proceedings of International Conference on Information Technology & applications Information Technology & Applications ICITA 2002), 234-17, 25-28 November 2002,* BATHURST, AUSTRALIA.

[2] 荒木健治，高橋祐治，桃内佳雄，栃内香次，“帰納的学習を用いたべた書き文のかな漢字変換，”信学論，Vol. J79-D-Ⅱ，No. 3，pp. 391-402，March 1996.

[3] M. Harada, S. Shimizu, a Simple Way of Guidance: Making Relevant Keyword From Anonymous User Behavior on WWW Search, NTT Software Labs.