自 然 言 語 処 理 147－3
（２００２．１．２１）

機械翻訳の汎用的な前処理手法における係り受け関係の改善方法について

長島康人　　荒木健治　　栃内香次

北海道大学大学院工学研究科
〒060-8628　　札幌市北区北 13 条西 8 丁目
TEL：011-706-7389　FAX：011-709-6277
E-mail:{naga,araki,tochinai}@media.eng.hokudai.ac.jp

機械翻訳において翻訳精度を向上させるための方法の 1 つに，入力文に対してあらかじめ処理を施す前処理がある．我々はこれまで様々な機械翻訳システムに対して適用可能な汎用的な前処理手法を提案したが，その手法の問題点の 1 つに係り受け関係の改善があった．本研究ではその問題に対し，新たに文の広い範囲の情報を用いて学習を行い係り受け関係を改善する方法を提案する．本稿では実際にシステムを作成し，複数の種類のテキストを用いて実験を行いその有効性を確認した結果について述べる．

キーワード：前処理，機械翻訳，汎用性，係り受け関係

A Method to Improve Dependency Relations in General Preprocessing on Machine Translation

Yasuto Nagashima, Kenji Araki and Koji Tochinai

Graduate School of Engineering, Hokkaido University
Kita-13, Nishi-8, Kita-ku, Sapporo, 060-8628, JAPAN
TEL:+81-11-706-7389　Fax:+81-11-709-6277
E-mail: {naga, araki, tochinai}@media.eng.hokudai.ac.jp

Preprocessing is one of methods to improve translation accuracy. We have proposed a method for preprocessing that can be applied to various machine translation systems. However, the method was not so effective on errors of dependency relations. In order to solve this problem, we propose a method to correct dependency relations by using broad information in a sentence. To confirm the effectiveness of proposed method, we carried out experiments with several texts.

Keywords: preprocessing, machine translation, generality, dependency relation

1　はじめに

近年の世界規模でのネットワークの普及に伴い, 電子化された異言語文書に触れる機会が増大している. 一方, 機械翻訳ソフトは多くのものが市販されているが, それらの翻訳結果には訳語選択や係り受け関係の誤り, 不自然な表現などが多く含まれ, いずれも満足のいくものではない[1][2].

機械翻訳精度向上の方法として, 入力文に対してあらかじめ処理を施す, 前処理がある. 前処理は同言語間での変換であり, ある意味では言い換えである. しかし, 従来行われてきた言い換えの研究[3][4][5]には純粋に翻訳精度を向上させるためのものは少ない. また, 翻訳精度向上のために行われている前処理に関する研究には, あらかじめ規則を与えておく解析的な手法[6][7][8][9]や, 統計的な手法[10]などがある. しかしそれらの多くは個々の機械翻訳システムに付随して研究されているものであり, 現在, 既に多数存在する機械翻訳システムに対して汎用的に適用できるというものではない.

これらの問題を解決するために, 我々は様々な機械翻訳システムに対して適用可能な帰納的学習を用いた汎用的な前処理手法（IL-PP）[1]を提案している[11]. この手法は入力文と翻訳結果が正しくなるようにその入力文を変換した正変換結果から差異部分を抽出することにより変換規則を獲得し, その規則を用いて前処理を行う手法である. IL-PP ではあらかじめ規則を与えておくのではなく, 対象とする機械翻訳システムに適した規則を獲得していく. また, 特定の機械翻訳システム特有の情報を用いていないので様々な機械翻訳システムに対して適用可能である. しかし, 規則の獲得, 適用に際して差異部分とその前後の情報しか用いていないため, 訳語選択の誤り等には有効であったが係り受け関係の改善に対しては満足のいくものではなかった. そこで本研究では, IL-PP の改善方法として文のより広い範囲の情報を用いた係り受け関係の改善方法を提案する.

本稿ではまず, 先に提案した前処理手法の概要を説明する. その後, 新たに提案する係り受け関係の改善方法を述べ, 最後に実際にシステムを作成して行った評価実験の結果から提案手法の有効性とその汎用性を確認する.

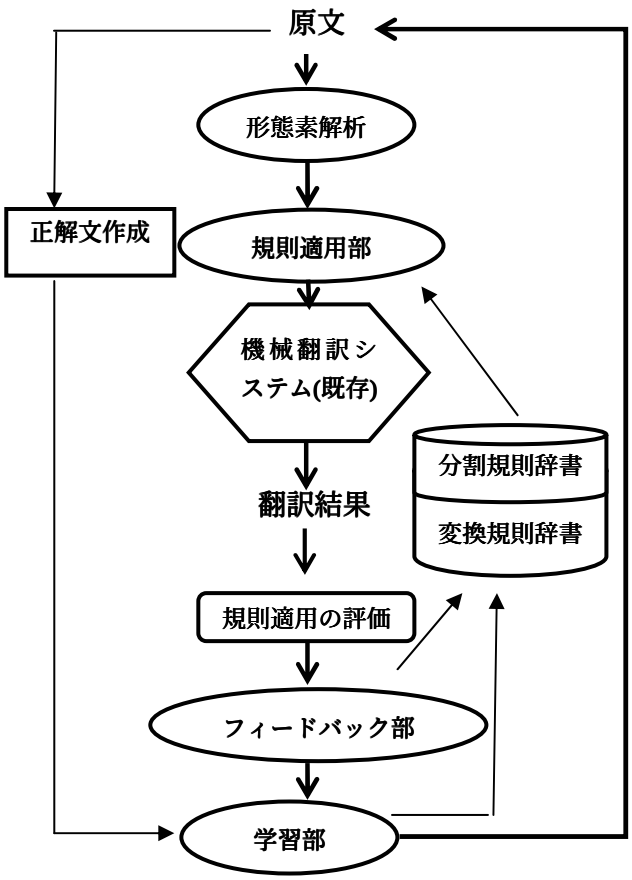

図 1　処理の概要

2　システムの概要

本手法の全体的な処理過程を図 1 に示す. 各部での処理の詳細は文献[11]に述べられているので, 本稿では本手法の概要と今回行った変更点について述べる. ここで, 本手法では特定の言語の情報を用いていないため基本的に種々の言語に適用可能であると考えられるが, 本稿では英日翻訳を対象とする. また本研究では, 意味が理解できる範囲においては, 翻訳文の自然性に関しては前処理の対象としない. これは, 翻訳文の自然性という問題の改善は, 同じ補助システムで考えた場合, 前処理よりも後処理で行った方が格段に効果的なことが明白であるからである. これについては, 我々は別のテーマとして研究を行っている[12].

本システムに原文が入力されると, まずこの原文に対して形態素解析を行う. ここで, 本手法で用いる情報は形態素解析結果のみとしている. これは, そもそも機械翻訳における誤りというもの

[1] 機械翻訳の前処理の帰納的学習 Inductive Learning for Preprocessing in Machine Translation

入力文１：In/IN a/DT text/NN ,/, there/EX exists/VBZ a/DT variety/NN of/IN surface/NN features/NNS // which/WDT can/MD be/VB considered/VBN …
形態素列：IN DT NN , EX VBZ DT NN IN NN NNS // WDT MD VB VBN…

入力文２：Recently/RB ,/, there/EX have/VBP been/VBN some/DT approaches/NNS to/TO use/VB information/NN of/IN lexical/JJ cohesion/NN ,/, // which/WDT are/VBP computed/VBN …
形態素列：RB , EX VBP VBN DT NNS TO VB NN IN JJ NN , // WDT VBP VBN…
下線部が共通部分,
獲得規則：@1 , EX @2 DT @3 IN @4 NN @5 // which/WDT
“@数字”は変数
RB：副詞，MD：助動詞，VBN：過去分詞，NN：一般名詞，VBZ：一般動詞，IN：前置詞，
DT：冠詞，CC：接続詞，WDT：関係代名詞

図 2　分割規則獲得例

は, 構文解析や意味解析の解析誤りによるものが大きいと考えられるので, ここでそれらの処理を用いることは矛盾している. 続いて形態素解析結果に対して変換規則辞書から適用可能な規則を適用する．本稿では IL-PP で用いている規則を変換規則, 今回提案する係り受け関係改善のための規則を分割規則と呼ぶ. 変換規則適用の際, 係り受け関係に関する変換規則が適用されていない場合には分割規則辞書から適用可能な規則を適用する. 次に, 変換結果と変換を行っていない入力文を機械翻訳システムにより翻訳し, 翻訳結果同士を比較し判定を行う．その判定結果から, 使用された変換規則, 分割規則に対してフィードバックが行われる．最後に, 人手により作成された正変換結果と入力文の形態素解析結果から変換規則, 分割規則が獲得される. 以下で変換規則, 分割規則の獲得処理, フィードバック処理について説明する.

2.1　変換規則

学習部において, 入力英語文の形態素解析結果と, 人手により作成された正解文を比較し, 差異部分を抽出することにより変換規則を獲得する. ここで, 共通部分とは 3 単語以上連続して同一である部分, 差異部分とは共通部分にはさまれている部分と定義する.

変換規則は, 差異部分, 差異部分前後の単語と品詞，正変換度数，誤変換度数から成る．なお, 獲得された時点では，正変換度数 1，誤変換度数 0 ではなく，正変換度数，誤変換度数共に 1 とする．これは，本稿で提案するような既存のシステムに対して付加的に処理を行う補助システムの場合は, いかに多くの正解を作るかということよりも, いかに誤りを抑えながら正解を作るかということに重点を置くべきであると考えられるからである．このような観点から，2.3 節で述べる規則適用正解率が初期状態において中間の適用条件である 80%未満 50%以上となるようにそれぞれの数値を設定した.

2.2　分割規則

本研究では文の分割，及び“ , ”の付加を係り受け関係改善のための基本的な処理としている. それに加えてこれらの処理による主語の喪失などを防ぐため単語の付加, 削除を行うこともある. 分割規則は文の分割または“ , ”の付加が行われている文同士から, 分割点以前の形態素列の共通部分を抽出することにより獲得される. 図 2 に分割規則獲得例を示す．各規則は正変換度数，誤変換度数を持つと同時に，規則獲得元となった 2 文の分割以降の単語の付加, 削除の処理内容も保持している. 分割規則においても変換規則と同様な理由から，獲得された段階では正変換度数 1, 誤変換度数 1 とする. さらに次節で述べる規則適用正解率が 50%未満となった規則は，不適切な規則として分割規則辞書から削除される.

2.3　規則適用条件と規則適用正解率

規則適用正解率は, 以下の式によって計算される.

$$\text{規則適用正解率} = \frac{CF}{(CF + EF)} \times 100 \ [\%]$$

ここで，CF は正変換度数，EF は誤変換度数である．この規則適用正解率によって, 以下のように規則の適用条件が決定される.

変換規則
- 規則適用正解率 80%以上
  差異部分の一致
- 規則適用正解率 50%以上 80%未満
  差異部分および前後 2 単語ずつの品詞の一致

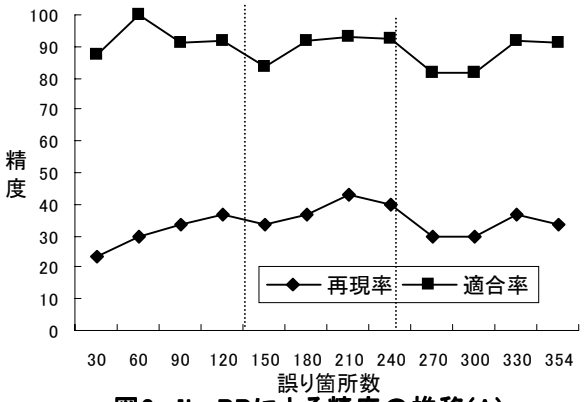

図3 IL-PPによる精度の推移(A)

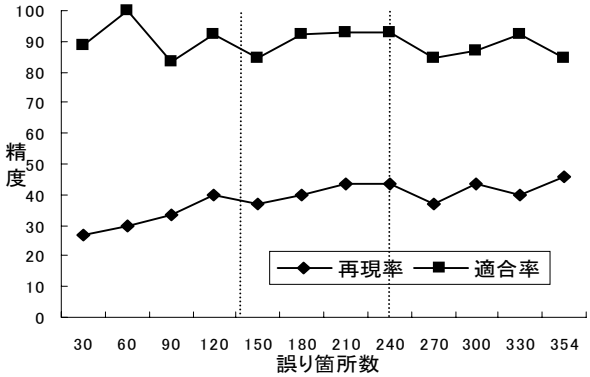

図 4 改善手法による精度の推移(A)

・ 規則適用正解率 50%未満

差異部分および前後 2 単語ずつの単語の一致

分割規則

分割規則の適用条件は一定で, 形態素列の一致, 分割点直後の単語の一致, さらに規則獲得元となった2文のどちらかの分割以降の単語の付加, 削除の処理が適用できることが条件となる.

2.4 フィードバック処理

各変換規則, 分割規則は正変換度数, 誤変換度数から計算される規則適用正解率に基づいた規則適用条件を持っている. フィードバック処理では適用した規則に対して翻訳結果が改善された場合にはその規則の正変換度数を＋1, 逆に翻訳結果が誤りへ変化した場合には誤変換度数を＋1とする. また, 規則の適用によって翻訳文の正誤に変化が生じなかった場合には, 正変換度数, 誤変換度数は変化しない.

3 実験

実際にシステムを作成し 2 つの実験を行った.

3.1 実験 1（性能評価実験）

実験 1 は今回新たに提案した係り受け関係改善方法の性能を評価する目的で行う. 実験には IL-PP を用いたシステムと, そのシステムに今回提案する手法を組み合わせたものの 2 つを用いて比較検証を行った. テキストには自然言語処理に関する 3 編の論文 A, B, C[13][14][15]計 514 文を用いた. 1 文あたりの平均単語数は 19.2 単語であった. システムの入力のための形態素解析には「Brill tagger」[16]を, 機械翻訳システムには 2 種類の商用のシステム A, B を用いた. 初期条件を一定とするため, 初期状態の変換規則辞書, 分割規則辞書は空の状態から行った.

表 1 処理前後のテキストの状態(A)

		処理前	処理後
誤り箇所総数		354 箇所	246 箇所(235)
誤りの内訳	訳語選択	167 箇所	101 箇所(101)
	係り受け関係	138 箇所	107 箇所(94)
	ユーザーの入力	49 箇所	24 箇所(24)
改悪箇所			14 箇所(16)

表 2 処理前後のテキストの状態(B)

		処理前	処理後
誤り箇所総数		449 箇所	335 箇所(319)
誤りの内訳	訳語選択	221 箇所	148 箇所(148)
	係り受け関係	179 箇所	145 箇所(126)
	ユーザーの入力	49 箇所	23 箇所(23)
改悪箇所			19 箇所(22)

3.1.1 実験結果

図 3 は機械翻訳システム A において IL-PP のみを用いた場合の, 図 4 はそれに加えて改善手法を組み合わせた場合の誤り箇所 30 箇所ごとに算出した再現率, 適合率の推移である. 同様に図 5 と図 6 は機械翻訳システム B における再現率, 適合率の推移である.

表 1 は機械翻訳システム A に IL-PP を適用した場合の実験の処理前と処理後のテキストの状態の変化である. それに対して, 括弧内の数字はさらに改善手法を適用した場合の処理後の値である. 同様に表 2 は機械翻訳システム B における実験結果である.

3.1.2 考察

図 3 と図 4, 図 5 と図 6 を比べてみると, 図 4, 図 6 の方が全体的に再現率が高い結果となっており, 特に後半部分では約 5%程度の向上が見ら

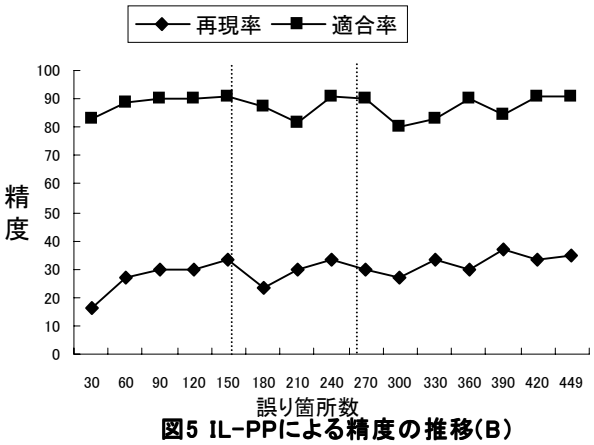

図5 IL-PPによる精度の推移(B)

れる．一方，適合率もわずかではあるが上昇している．これらのことから，改善手法は誤りをほとんど生成することのない悪影響が少ないシステムであるといえる．次に誤りの種類ごとの処理前と処理後の変化をみてみると，機械翻訳システム A では IL-PP だけでは訳語選択の誤りとユーザーの入力誤りは 40～50%改善されているのに対し，係り受け関係の誤りは 22.5%しか改善されていない．しかし，改善手法を適用することによって 31.9%まで改善されている．同様に機械翻訳システム B においても係り受け関係の改善が 19.0%から 29.6%へ 10.6 ポイント程度向上している．これらの結果から改善手法の複数の機械翻訳システムへの適応能力と有効性が確認できた．

改悪箇所についてみてみると改善手法を用いることによる改悪箇所は，機械翻訳システム A では 2 箇所，B では 3 箇所存在した．それに対して IL-PP の係り受け関係改善のための規則による誤りは，機械翻訳システム A では 8 箇所，B では 10 箇所存在した．今回作成したシステムでは，規則の過適用を考慮し先に変換規則が適用された場合には分割規則の適用は行っていない．しかしこの実験結果から規則の適用順序の変更，あるいは規則ごとに優先順位を設けるなどの対策が必要であると考えられる．

表 3　処理前後のテキストの状態

		処理前	処理後
誤り箇所総数		643 箇所	418 箇所
誤りの内訳	訳語選択	328 箇所	200 箇所
	係り受け関係	247 箇所	157 箇所
	ユーザーの入力	68 箇所	35 箇所
改悪箇所			26 箇所

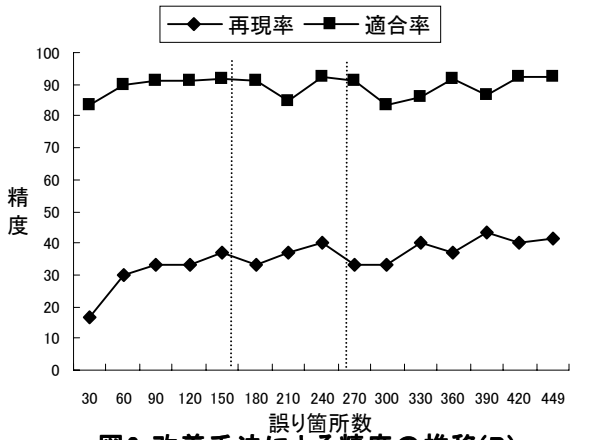

図6 改善手法による精度の推移(B)

3.2　実験 2（適応能力評価実験）

実験 2 はシステム全体の適応能力を評価する目的で行う．我々は既に文献[11]において異なる機械翻訳システムによる適応能力の評価は行っているが，テキストの種類を変えた場合の適応能力は検証していない．そこで本節では 3.1 節の実験の後，新たに計算機マニュアルを対象テキストとして実験を行った．実験文数は実験 1 で用いたものと同じ論文 A，B，C の 514 文と計算機マニュアル 392 文の計 906 文，機械翻訳システムにも実験 1 と同じ商用のシステム A を用いた．

3.2.1　実験結果

図 7 に実験結果を示す．図 7 は誤り箇所 40 箇所ごとに算出した再現率，適合率の推移である．表 3 は実験の処理前と処理後のテキストの状態の変化である．

3.2.2　考察

誤り箇所数 360 以降で一度精度が低下するが，その後回復し，若干ではあるが低下する前よりも精度が向上していることがわかる．これはテキストの種類の変化に対応して，学習能力によりシステムが新しいテキストに適した規則を獲得できていることを意味している．

表 4 はテキストの種類が変わる 354 箇所目から 400 箇所目までの誤りの種類ごとにみた改善の様子である．訳語選択の誤り，ユーザーの入力の誤りについては表 3 で算出した全体の値に比

表 4　テキストの変化直後の各改善率

	処理前	処理後	改善率
訳語選択	43 箇所	31(12)箇所	27.9%
係り受け関係	36 箇所	24(12)箇所	33.3%
ユーザーの入力	7 箇所	5(2)箇所	28.6%

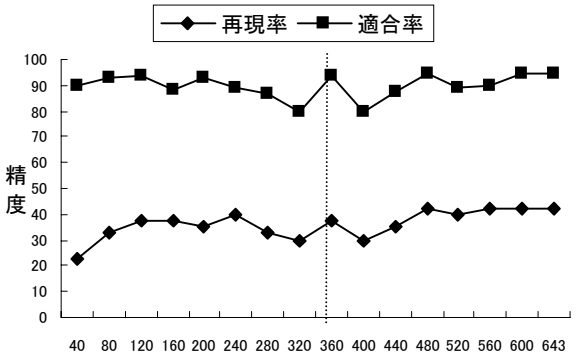

図7 再現率適合率の推移

べ約 10～20%ほど低くなっている．それに対して係り受け関係の誤りは 4%程度しか低下していない. これは規則の適用方法の違いによるものと思われる．2.1 節で述べているように，変換規則は差異部分とその前後の限られた情報しか用いておらず, さらに規則の適用には差異部分の完全な一致が条件となっており, テキストの種類が変わった場合, そもそも適用できる規則が少なくなってしまう. 一方分割規則は文の広い範囲の情報を用いて, その適用に際しても形態素列と分割点直後の一単語の一致を条件としている．また，変数への入力も条件を設定していないため汎用的に適用可能な規則が獲得されている. 以上の点から改善率に差が現れたと思われる.

4 おわりに

我々はこれまでに機械翻訳における翻訳誤りを改善する汎用的な前処理手法を提案してきた. 本稿では, これまでの手法で改善効果が弱かった係り受け関係の誤りをさらに改善する手法を提案し, その効果と汎用性を評価するための実験を行った. 実験の結果, 従来の我々の手法に比べ係り受け関係の誤りを 10%程度改善し，再現率，適合率においても数%の改善が見られた. さらに複数の種類のテキストと機械翻訳システムにおいてその適応能力を確認した.

今後は分割処理と変換処理の適用順序の問題を検討し, 高い適合率を維持しながら再現率の向上を図る予定である.

参考文献

[1]Yamauchi Satoshi：A Method of Evaluation of the Quality of Translated Text, MT Summit Ⅶ, pp564-567, Sept.1999

[2]成田一：翻訳ソフトの性能評価，情報処理学会研究報告，自然言語処理　研究報告 No.125-14,pp123-130,1998

[3]佐藤理史：論文表題を言い換える，情報処理学会論文誌，Vol.40，No.7，pp2937-2945，1999

[4] 張玉潔, 尾関和彦:分類木を用いた日本語長文の自動分割, 言語処理学会第４回年次大会発表論文集，pp390-393，1998

[5]Dras，M.：Reluctant Paraphrase: Textual Restructuring under an Optimization Model, Proc. Pacling97, pp.48-55, 1997

[6]田中穂積　監修：“自然言語処理―基礎と応用―”電子情報通信学会編，1999

[7]吉見毅彦，佐田いち子，福持陽仕：頑健な英日機械翻訳システム実現のための原文自動前編集, 自然言語処理, Vol.7, No.4，pp99-117，2000

[8]木村真理子，野村浩一，平川秀樹：日英機械翻訳における日本語分割処理について，情報処理学会自然言語処理研究会報告，96-8，pp57-64，Oct.1993

[9]白井諭，池原悟，河岡司，中村行宏：日英機械翻訳における原文自動書き換え型翻訳方式とその効果，情報処理学会論文誌，Vol.36，No.1，pp12-21，Jan.1995

[10]金淵培，江原輝将：日英機械翻訳のための日本語ニュース自動短文分割と主語の補完，情報処理学会自然言語処理報告，93-3，pp15-22，1993

[11]Yasuto Nagashima, Kenji Araki, Koji Tochinai：Evaluation of Generality of Inductive Learning for Preprocessing in Machine Translation，Proc. IEEE SMC2001，Natural Language Processing and Knowledge Engineering, pp921-926, Tucson, Arizona，USA，Oct.2001

[12]尾崎正行，荒木健治，栃内香次：帰納的学習を用いた自然な日本語文生成手法の評価，情報処理学会研究報告，自然言語処理　研究報告 No.141-10,pp57-62,2001

[13] Y. Masatomi，K. Araki，K. Tochinai：Acquisition System of Deep Case Using Queries and Replies，Proc. Applied Informatics 2000，Innsbruck，Austria，pp194-197，Feb.2000

[14]H. Haga，Y. Miyake：Grammar-Based Document Structure Analysis and Its Application to Document Conversion, Proc. Applied Informatics 2000, Innsbruck, Austria，pp194-197，Feb.2000

[15]Kenji Araki，Hiroshi Echizen-ya，Koji Tochinai：Performance Evaluation in Travel English for GA-ILMT, Proc. IASTED International Conference ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING，pp117-120, Banff，Canada，July　1997

[16]Eric Brill：A Corpus-Based Approach to Language Learning, University of Pennsylvania, 1993