2007－NL－177 (11)
2007／1／26

雑談を対象としたSeGA-ILSDの多言語に対する汎用性の評価

黒田 道友　荒木 健治
北海道大学大学院 情報科学研究科
E-mail: kuroda@media.eng.hokudai.ac.jp, araki@media.eng.hokudai.ac.jp

本稿では，対話の中から応答規則を自動獲得する「性淘汰遺伝的アルゴリズムを用いた帰納的学習による音声対話処理手法(Spoken Dialogue processing method using Inductive Learning based on Genetic Algorithm with Sexual selection : SeGA-ILSD)」の多言語における汎用性を確認する。SeGA-ILSDとは，人間との対話から応答規則を獲得，学習する仕組みを備えた雑談型対話システムである．本システムの学習方法は，言語に依存していない．SeGA-ILSDの多言語への汎用性を実証するために，本稿では日本語，英語，中国語，ドイツ語を用いて実験を行った結果について述べる。その結果，SeGA-ILSDは各国語で同様に学習を行うことができ，言語によることなく応答規則を獲得できることを確認した．

Evaluation of Generality of SeGA-ILSD for a Chat Using Different Languages

Michitomo KURODA and Kenji ARAKI
E-mail: kuroda@media.eng.hokudai.ac.jp, araki@media.eng.hokudai.ac.jp

This paper proposes SeGA-ILSD (Spoken Dialogue processing method using Inductive Learning based on Genetic Algorithm with Sexual selection) along with analysis of its generality for four languages : Japanese, English, Chinese and German. SeGA-ILSD achieves high generality using machine learning. SeGA-ILSD can show equivalent performance for different languages. To confirm it, we carried out experiments for SeGA-ILSD performance evaluation. Their results show that SeGA-ILSD has equivalent performance for four languages. Results verify that SeGA-ILSD has high generality for these four languages.

1 はじめに

計算機との対話処理は，計算機が開発された当初より研究されている．古典的なものとしてはELIZA[1]がある．ELIZAはキーワードマッチングを利用し，人手で作成した静的なルールを用いているので，どのような話題にも何らかの応答を返すことが可能である．しかし，静的なルールを利用しているため，同じような応答が多く，ユーザに適応した応答を返すことができないので使い続けるとユーザに不満が生じる．

最近の雑談型対話システムとしては，うつ病患者の治療を目的としたものがある[2]．これはELIZAよりも複雑な処理を行っており，うつ病患者の治療には有効であるが，全ての応答規則を人手で与えているため，その開発に膨大なコストを必要とする．

これらの問題点は，システムが人間の言語獲得の過程を模倣し，自動的にルールを獲得することができれば解決できるものと考えられる．対話例から学習するシステムとして，我々は『遺伝的アルゴリズムを用いた帰納的学習による音声対話処理手法(Spoken Dialogue processing method using Inductive Learning with Genetic Algorithm : GA-ILSD)[3]』という研究を行ってきた．この手法では，幼児から大人へと成長していく段階をシミュレートしている．

幼児は充分な言語的知識を獲得していないのでタスク指向型のような対話は行えず，雑談から学習をしていると考えられる．したがって，GA-ILSDでも雑談を対象としている．GA-ILSDでは，対話例からルールを獲得し，さらに遺伝的アルゴリズムを用いた帰納的学習(Inductive Learning with Genetic Algorithm:GA-IL)[4]によってルールを学習しシス

テム応答を生成するというものである．なお，帰納的学習については第3章で詳細を述べる．

GA-ILSD では，対話例から充分にルールが学習できていないときは，ELIZA 型のシステムで応答文を生成し，対話例を収集している．そして，学習が進みルールが獲得されると ELIZA 型応答は減少し，GA-IL による応答が増加するようになっている．このように動的にルールを獲得する仕組みによって，低いコストでユーザに適応した応答が可能となる．しかし，この手法では，適応度に精度を用いているため，そのルールが使用されないと淘汰が行われない．このため，GA によってランダムに組み合わせることで生成されたルールが毎世代ごとに効率よく淘汰されない．そのため，GA の特徴である進化が効率よく行われないという問題点がある．

この問題点を解決するために，「性淘汰理論[5]を用いた遺伝的アルゴリズム(GA with Sexual selection : SeGA)」に注目した．性淘汰理論とは，雌はより優秀な雄を選り好みすることで優秀な子孫を残そうとし，雄は雌に選ばれるために自らの形質を進化させるというものである．淘汰理論を一般的な GA に導入すると，適応度が高くなる方向だけでなく，雌の選り好みによる方向への探索も可能になるという結果が得られている[6]．性淘汰理論の考え方を利用した SeGA-ILSD[7]では，ルールに雌雄を設定し，雌ルールに帰納的学習を進める雄ルールを選り好みさせた．その結果，帰納的学習が進みやすい方向へと進化を導き，効率よく優秀な応答を生成することができた．

SeGA-ILSD は，人間の言語獲得の過程を模倣した学習型対話システムであるので，言語に依存することなく学習を行い，ルールを獲得することができると考えられる．本稿では，このことを実験によって検証する．実験に用いた言語は，日本語，英語，中国語，ドイツ語の4ヶ国語である．

以下，第2章では SeGA-ILSD の処理過程を，第3章では本稿で行った評価実験とその考察を，第4章ではまとめを述べる．

2 処理過程

2.1 概要

SeGA-ILSD の概要を図1に示す．

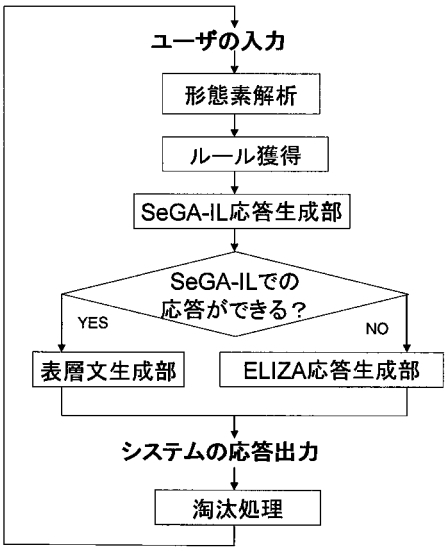

図1　　SeGA-ILSD の流れ

本システムは，まずユーザの音声発話による入力を受け，音声認識装置によって音声認識を行う．認識結果は，形態素解析ツールによって解析され，ルールが獲得される．そして，SeGA によってルールを増やし，帰納的学習により再帰的にルールを獲得する．なお，SeGA については **2.4.2**, **2.4.3** で，帰納的的学習については **2.4.4** で詳細に説明する．このようにして SeGA-IL により獲得されたルールから応答を返すことができる場合はその応答を返し，応答ができない場合は対話を続けるために ELIZA 型応答システムを用いて応答を行う．また，ルールには精度と使用回数で定義される適応度を設定しており，淘汰処理によって適応度の低いルールを削除している．

2.2 形態素解析

ユーザがシステムと雑談をしてもらう感覚で，自由に話しかけてもらうことで入力を行う．入力発話は，音声認識装置によってテキスト化される．なお，ここで用いた音声認識装置は，

・日本語･･･Microsoft Japanese Recognizer (Version 6.1)

・英語･･･Microsoft English Recognizer (Version 5.1)

・中国語･･･Microsoft Simplified Chinese Recognizer (Version 5.1)
・ドイツ語･･･ViaVoice (Release 10)
である．テキスト化された入力発話は，形態素解析ツールによって解析され，単語分割結果と品詞情報が得られる．ここで用いた形態素解析ツールは，
・日本語･･･JUMAN (Version 5.0)[8]
・英語･･･Apple Pie Paser (Version 5.9)[9]
・中国語･･･ICTCLAS [10]
・ドイツ語･･･TreeTagger (Version 3.1)[11]
である．

得られた品詞情報は，自立語と付属語の区別に利用される．

2.3 対話からのルールの獲得

SeGA-ILSD のルールには，応答文生成ルールと表層文生成ルールの二種類のルールが存在する．対話からのルールの獲得例を図 2 に示す．

応答文生成ルールは，図 2 のようにユーザとシステムの対話の形態素解析結果から，自立語のみを抽出し対とすることで獲得する．表層文生成ルールは，一発話の形態素解析結果から自立語だけを抽出したものと，付属語を含めた全単語列を抽出したものを対として獲得される．これ以降，ルール中の“:”で区切られる左側を左辺，右側を右辺と呼ぶ．

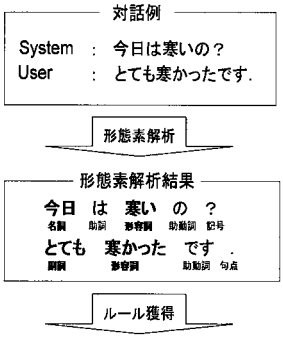

図 2　対話例からのルール獲得例

応答文生成ルールは，システム応答の自立語列を生成するのに用いる．具体的には，入力発話と左辺の字面的な一致率が高いルールの右辺を応答として選択する．そして，表層文生成ルールによって，自立語列に付属語を加えて自然な文を生成する．単独で意味を持つ自立語のみを応答文生成ルールに用いているのは，音声の誤認識への対策と微妙な表現の変化にも頑健な処理を行うためである．

2.4 SeGA を用いた帰納的学習

獲得されたルールに対して，SeGA を用いた帰納的学習を行うことで，既存のルールを組み合わせて新たなルールを生成する．

SeGA-IL によって得られたルールから応答文として選ばれるものは，ルール左辺の自立語列とユーザの入力文の自立語列との単語一致率が 65%以上で，最も適応度が高い応答文生成ルールである．ここでの単語一致率とは，入力された自立語列のうち，いくつがルール左辺中の自立語と一致したかの割合である．

2.4.1 性淘汰理論

性淘汰理論とは，ダーウィンによって提唱された理論であり，クジャクの尾羽やライオンのたてがみのような，自然淘汰で説明のつかない雌雄での形質の違いを説明する．この理論によると，雌は繁殖において雄よりもコストがかかるため，優秀な雄を選択する必要があるが，その際にある形質によって選択する[5]．雌に好まれた形質は，選択する側の雌には必要のないものなので，自然淘汰の力によって発現せず，雄のみに発現する．そのため，性差が生じる．

2.4.2 SeGA

上記のような性淘汰理論を一般的な GA に導入すると，適応度が高くなる方向だけでなく，雌の好みによる方向への進化が可能になる．そのため，本システムでは，雌の好みを「自身との共通部分が多い雄」と設定した．これによって，共通部分が多い方向へと進化が可能になる．詳細は **2.4.4** で述べるが，帰納的学習では，ルール同士を比較し共通部分と差異部分を見分けることで学習を行う．つまり，共通部分を持たないルール同士からでは学習を行

うことができない．そのため，共通部分を増やすことで帰納的学習が進みやすい方向への進化を促すことができる．

SeGA においては，各ルールを染色体，ルール中の単語を遺伝子に対応させている．本システムでは，帰納的学習によって得られる変数化された部分を持つ汎用性の高いルールを雄ルール，変数部分を含まないルールを雌ルールと設定した．これは，雄ルールは雌ルールに選ばれるように自らを装飾する必要があるため，変数部分を変化させることができるものの方が装飾に向いていると考えたからである．

2.4.3 SeGA の処理過程

SeGA は，以下の手順で実行される．

手順1．雄ルールが装飾を行う
手順2．雌ルールが装飾済み雄ルールから選択
手順3．交叉
手順4．突然変異
手順5．淘汰
手順6．手順 1〜5 の手順を繰り返し

まず，雄ルールは雌ルールに選択されやすいように自らを装飾する．具体的には，雄ルールは変数部分を持っているので，その変数部分に他のルールを代入することで自身を装飾する．このとき，より出現頻度の高い単語で構成されたルールで装飾することにより，一般的に雌ルールとの共通部分が多くなり，多数の雌ルールに選択してもらうことができるようになる．この際，文法的な問題が生じるのを防ぐために，対話履歴中の発話の品詞列と同じ品詞列になるような制限を設けて装飾を行っている．

具体的な装飾の例を図 3 に示す．この例では，@ で表される変数部分に“休日：テニス”というルールを代入することで，装飾を行っている．

次に，装飾後の雄ルールの中から，雌ルールは自身との字面上での共通部分が最も多い雄ルールを配偶者として選択する．その後，交叉を行うが，文の構造が失われるのを避けるため，両親の遺伝子に存在する共通品詞列に対して，ランダムにマスクをかけて一様交叉を行う．交叉の後，突然変異率 2% で突然変異をおこない，ルール中の単語を同品詞の異なる単語に置き換える．

最後に淘汰処理を行う．手順 1〜4 でできる子供ルールは，両親の適応度の平均値を受け継いでおり，適応度が 75 未満のルールをこの処理によって削除している．なお，適応度については，**2.7** で詳しく述べる．

このような操作を 5 世代繰り返すことで，雄ルール，雌ルールから子供ルールを作成している．5 世代と限定したのは，現実的な時間内に応答を返すためである．

2.4.4 帰納的学習

本システムにおける帰納的学習とは，二つのルールを比較し，共通部分と差異部分を見分ける能力を用いて，ルール中に存在する共通部分，差異部分をそれぞれ再帰的にルールとして獲得する学習方法である．ルールの獲得例を図 4 に示す．

図 4 で示す通り，共通部分とは字面上で一致している単語列を指し，差異部分とは字面上で異なる単語列を指す．帰納的学習は，ルールの両辺を比較して，両辺に共通部分を持ち，かつ差異部分を一つだけ持つルールのペアに対して行われる．帰納的学習が可能なルールのペアが見つかった場合，差異部分を変数“@”として置き換え，汎用化したルールを生成する．これを共通ルールと呼ぶ．また，抽出された差異部分は差異ルールと呼ぶ．

さらに，帰納的学習によって獲得された共通ルール同士も帰納的学習の対象となるので，再帰的に差異部分を抽出，変数化することによって，より汎用性の高いルールの生成が可能になる．

なお，現実的な時間での処理を可能にするため，最近獲得された 30 個のルールに対して帰納的学習を行い，再帰の回数も 10 回に制限した．

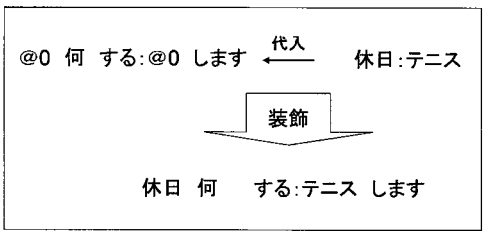

図 3　　装飾の例

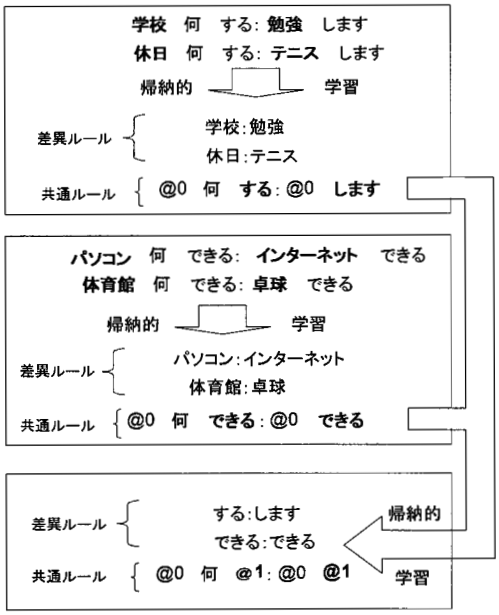

図 4　　帰納的学習でのルール獲得例

2.5 ELIZA 応答生成部

ELIZA 型応答は，SeGA-IL による応答を生成できなかったときに，対話を継続させルールを獲得するために行う．日本語では，GA-ILSD のときと同じ ELIZA 型対話システム[3]を用いた．英語では，英語版 ELIZA モジュール（ELIZA.pm Version 1.04）[12]を用いた．中国語，ドイツ語では，英語版 ELIZA モジュールをそれぞれネイティブの協力者に翻訳してもらって利用している．

2.6 表層文生成部

SeGA-IL による応答が可能だった場合，自立語列で表現されている応答文生成ルールを，この処理によって自然な文に復元している．表層文生成ルールにおいても，性淘汰 GA と帰納的学習を行い，より多くの表層文生成ルールを生成している．このようにして増加させた表層文生成ルールのうち，選択された自立語列に対応したものを用いて，自然な文を生成している．

2.7 淘汰処理

この処理では，獲得したルールを適応度に応じて淘汰している．ユーザは，直前のシステム発話を受け，発話を入力することになるが，ユーザの入力した発話に特定の誤りキーワードが含まれている場合，その直前のシステム発話に使用したルールの誤応答回数を 1 増加させる．誤りキーワードが含まれていない場合は，正応答回数を 1 増加させる．さらに，SeGA-IL によって作られたルールのうち，応答で使用されなかったルールは，連続未使用回数を 1 増加させている．これらの値を用いて，以下の式(1)でルールの適応度を定義する．式(1)に示すように，応答率に連続未使用回数を考慮に入れた α を乗算することで適応度を算出している．

$$\text{適応度}=\frac{\text{正応答回数}}{\text{正応答回数}+\text{誤応答回数}}\times\alpha\times 100 \quad \cdots(1)$$

$$\alpha=\begin{cases}1 & (\text{対話例から獲得したもの})\\ 1-0.08\times\text{連続未使用回数} & (\text{SeGA - ILから獲得したもの})\end{cases} \quad \cdots(2)$$

適応度が 75 未満のルールをこの処理で削除する．なお，式(2)中の「対話例から獲得したもの」とは，**2.3** で述べたように対話例もルールとして獲得しており，そのようにして対話から直接獲得したルールに対して行う処理である．同様に，「SeGA-IL から獲得したもの」とは，SeGA-IL によって獲得されたルールに対して行う処理である．したがって，SeGA-IL によって自動獲得されたルールは，正応答率が 100%でも連続未使用回数によっては削除されることがある．一方，対話から直接獲得したルールは，連続未使用回数が高くても，正応答率が低くない限り削除されることはない．これは，対話から直接獲得したルールは実際に対話に用いられたものであり，SeGA-IL によって自動獲得されたルールに比べて有用性が高いと考えられるためである．SeGA-IL によって自動獲得されたルールは，実際に対話に使用してみなければ有用性を判定できない．そのようなルールを保持しつづけることは，システムの処理時間を増加させることにつながる．従って，SeGA-IL によって獲得されたルールについては，正応答率以外の基準である連続未使用回数を取り入れて淘汰を行っている．

3 評価実験，考察

本システムを用いて．中国語，ドイツ語における汎用性の評価実験を行い，先に実験を行った日本語，英語の実験結果[7]との比較を行う．

3.1 実験方法

先に行った日本語，英語の実験[7]と同じように対話ターンは250ターンで，初期条件としてシステムがルールを全く獲得していない状態から対話を開始した．また，被験者は中国語を母国語とする理系の女子大学院留学生と，ドイツ語を母国語とする文系男子大学院留学生である．

3.2 応答の評価方法

システムの応答一つ一つに対して，被験者が評価を行った．応答の評価には以下のような基準を設けた．

・正応答・・・意味的に正しく，表現が自然
・準応答・・・意味的に正しいが，表現が不自然
・誤応答・・・意味的に誤っている

これらの基準に従って，被験者自身の判断でシステムの応答の評価を行った．したがって，応答の評価は主観的なものになるが，雑談には明確な正解が存在しないため，このような評価方法とした．

システムの応答には，ELIZA 型システムによるELIZA 型応答と，SeGA-IL によって生成されたルールで応答する SeGA-IL 型応答との二種類が存在する．したがって，システム応答の評価は，次の6種類に分類できる．

① ELIZA 型の正応答
② SeGA-IL 型の正応答
③ ELIZA 型の準応答
④ SeGA-IL 型の準応答
⑤ ELIZA 型の誤応答
⑥ SeGA-IL 型の誤応答

なお，正応答と準応答を合わせた①～④を「有効応答」と呼ぶ．

3.3 実験結果

各言語の音声認識の精度を，単語認識率で調査した結果を以下に示す．

・日本語・・・86.1%
・英語・・・62.0%
・中国語・・・62.4%
・ドイツ語・・・86.3%

また，各言語の SeGA-IL 型応答と ELIZA 型応答の応答率を表1に示す．なお，SeGA-IL 型応答の応答率，ELIZA 型応答の応答率は，それぞれ以下の式(3)，(4)で定義する．

$$\text{SeGA - IL型応答率} = \frac{\text{SeGA - IL型応答の数}}{\text{全ターン数}} \times 100 \quad \cdots(3)$$

$$\text{ELIZA型応答率} = \frac{\text{ELIZA型応答の数}}{\text{全ターン数}} \times 100 \quad \cdots(4)$$

各言語のシステム応答の精度を，表2に示す．なお，応答の精度は，以下の式(5)のように **3.2** で定義した①～⑥のシステム応答の評価の割合で定義する．

$$\text{評価の割合} = \frac{\text{①～⑥それぞれの評価数の数}}{\text{全ターン数}} \times 100 \quad \cdots(5)$$

3.4 考察

表1から，日本語は26.1%，英語は35.2%，中国語は28.0%，ドイツ語は50.8%の応答を SeGA-IL によって生成したルールで，応答していることがわかる．どの言語も日本語と同等以上にルールの学習を行い，SeGA-IL 型の応答を返すことが確認できた．また，文法構造が似ていて分かち書きされている英語(35.2%)やドイツ語(50.8%)に比べて，日本語(26.1%)と中国語(28.0%)の SeGA-IL 型応答率が低い．異なり自立語数を調べてみたところ，日本語は373個，英語は242個，中国語は565個，ドイツ語は344個となっており，異なり自立語数が多かった日本語，中国語については学習が進みにくかったと考えられる．

表1　応答率の比較

	ELIZA型応答	SeGA-IL型応答
日本語	73.9%	26.1%
英語	64.8%	35.2%
中国語	72.0%	28.0%
ドイツ語	49.2%	50.8%

表2　精度の比較

	①	②	③	④	⑤	⑥
日本語	38.2%	14.6%	13.5%	2.5%	22.3%	8.9%
英語	8.4%	4.0%	53.2%	16.0%	3.2%	15.2%
中国語	16.0%	25.6%	30.4%	13.6%	10.0%	4.4%
ドイツ語	19.2%	17.6%	20.0%	19.6%	10.0%	13.6%

応答の有効性については，表2から日本語の有効応答の割合は68.8%，英語では81.6%，中国語では85.6%，ドイツ語では76.4%であり，有効な応答であることがわかる．

以上より，SeGA-ILSDは多言語においても汎用的に学習を行い，有効な応答を返すことが確認できた．

4 まとめ

本稿では，対話から自動的にルールを獲得する対話システムSeGA-ILSDの多言語における汎用性の調査を行った．日本語，英語，中国語，ドイツ語の4言語について調査したところ，どの言語においても日本語以上にSeGA-IL型の応答を返すことができた．また，SeGA-ILSDの応答の有効性についても，各言語で平均78.1%の有効応答率を得ることができた．これらの結果から，多言語においても汎用的に学習を行い，有効な応答を返すことが可能であることを確認した．

今後の課題としては，スペイン語やイタリア語，ポーランド語などさらに他の言語におけるSeGA-ILSDの有効性を調査していく予定である．また，応答の有効性だけでなく，被験者の満足度を考慮に入れた対話システム全体としての評価を行うことを予定している．

謝辞

本研究の一部は，大川情報通信基金研究助成の援助を受けて行われた．

参考文献

[1] J.Weizenbaum, “ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine”, Communications of the Association for Computing Machinery, Vol.9, pp.36-45, 1966.

[2] K.M Colby, “Human-Computer conversation in a cognitive therapy program”, in Machine Conversations, e.d. Yorick Wilks, pp.9-19, Kluwer Academic Publishers, 1999.

[3] 木村泰知，荒木健治，桃内佳雄，栃内香次，“遺伝的アルゴリズムを用いた帰納的学習による音声対話処理手法”，電子情報通信学会論文誌，Vol.J84-D-2 No.9 pp.2079-2091, 2001.

[4] K.Araki, K.Tochinai, "Effectiveness of Natural Language Processing Method Using Inductive Learning", Proceedings of the IASTED International Conference ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING, pp.295-300, 2001.

[5] G.Miller, “*The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*”, New Yourk: Doubledaay, 2000.

[6] 大森清博，藤原義久，前川聡，澤井秀文，北村新三，“不均衡突然変異を導く性淘汰に基づく進化的計算法”，システム制御情報学会論文誌，第15巻 第8号 pp.422-429, 2004.

[7] K.Araki, M.Kuroda, “Generality of Spoken Dialogue System Using SeGA-IL for Different Languages”, Proceeding of the IASTED International Conference COMPUTER INTELLIGENCE, pp.70-75, 2006.

[8] 黒橋禎夫，長尾真，日本語形態素解析システムJUMAN, 1999.

[9] 関根聡，”英語構文解析システム「Apple Pie Parser」”，情報処理，vol.41 No.11 pp.1221-1226, 2000.

[10] Zhang, H.P., Yu, H.K., Xiong, D.Y., and Liu, “Q.HHMM-based Chinese lexical analyzer ICTCLAS”, In Proceedings of Second SIGHAN Workshop on Chinese Language Processing, pages 184–187, 2003.

[11] Schmid, H., Probabilistic part-of-speech tagging using decision trees. International Conference on New Methods in Language Processing, pp. 44–49, 1994.

[12] CPAN, Chabot::ELIZA, http://search.cpan.org/~jnolan/Chatbot-Eliza-1.04/Chatbot/Eliza.pm