Mem. Fac. Sci. Eng. Shimane Univ.
Series A
34, pp. 77-86 (2000)

簡易型の古文機械翻訳システムの試作

矢田　佳子*・山口　政道**・松尾　美一***・佐藤　匡正****

*数理・情報システム学専攻　総合理工学研究科　島根大学大学院
**情報科学専攻　理学研究科　島根大学大学院
***情報科学専攻　理学研究科　島根大学大学院
****数理・情報システム学科　総合理工学部　島根大学

A Simplified Machine Translation System for Classic Nippon-Go Documents

Yoshiko YATA
Department of Mathematics and Computer Science, Graduate School of Science and Engineering Shimane University
Masamichi YAMAGUCHI
Department of Computer Science, Graduate School of Science Shimane University
Yoshikazu MATSUO
Department of Computer Science, Graduate School of Science Shimane University
Tadamasa SATOU
Department of Mathematics and Computer Science, Interdisciplinary Faculty of Science and Engineering Shimane University

概　　要

古文を身近なものにする手法として機械翻訳が考えられる．古文の機械翻訳システムの開発には，形態素の分解や，翻訳に必要な大規模な辞書が必要である．こうした辞書の作成には多大な費用と時間を要する．

そこで，このような大規模な辞書をもたない簡易型の古文機械翻訳システムを考案し，試作と評価を行った．本提案システムは実現の容易さを狙った簡便な翻訳方式であり，方式の妥当性を確かめるには方式に対する評価が重要となる．そこで，翻訳能力について単純単語置換翻訳方式によるものと比較して評価した．その結果，本提案の翻訳方式の能力が単純単語置換翻訳方式より著しく向上し，またそれは第三者による客観的な結果であることから，提案システムが有効であり，かつ汎用性があるという見通しが得られた．

1. 序　　論

現代において古文が縁遠くなっているが，文化の伝承という点で，古文を身近なものにする手法が望まれる．この方法として，機械翻訳が考えられる．しかし，分かち書きされていない日本語文は，語の構成や文の解析などに必要となる処理が，英文などと比べると複雑で，機械処理技法の実現に困難を伴う．現在，一般的に行われているALTJAWSを始めとした代表的な解析方法[1)]は，単語辞書により形態素解析を行う手法であり，解析の確度は用

いる辞書に依存している．ここで用いる辞書は正確な解析を求める程，収録すべき単語数が膨大となり，その実現には多大な費用と時間が必要である．

そこで，このような大規模な辞書が不要である簡易型の古文機械翻訳システムを考案し，試作と評価行う．本提案システムは実現の容易さを狙った簡便な翻訳方式であり，翻訳能力の妥当性を確かめるための評価が重要である．翻訳能力の評価には様々な方法が考えられるが，本方式においては，提案する簡便な方式の有効性を確かめ，将来性について把握するのが第一である．そこで，本方式の評価には次の方策を考える．

(1) 単純単語置換方式との比較

(2) 官能検査の適用

第(1)項については，適用の限定された最も原始的な単純単語置換方式の翻訳能力を基本とした時，提案システムが，その能力として適用の広さをもつかという点から評価するものである．また，第(2)項は，その評価についての客観性を与える方策である．

2. 翻 訳 方 式

本システムで実現した翻訳方式の特徴について述べる．

2.1 構成の考え方

古文を翻訳する場合，コンパイラ処理と同様に，文及び語を認識し，単語単位で現代語に変換する．そのためには，正確な単語分けをして，確定した単語について，品詞や活用形などを解釈しなければならない．しかし，文章を直接単語に分ける場合，文字列上での単語分けの起点の確定や起点での活用形が決定できないという問題が生じ，実現が難しい．

これを避ける方法として，文字列を鍵語によって字句に分ける方法を考える．この方法では，字句分けによって文が認識でき，更に字句から単語に分けることが出来る．つまり，字句を単語分けの起点とする．また文末や字句末の語句の性質から，起点の活用形が決定される．そして，単語間の接続法則に従って単語分けを行うことで，適用する単語辞書の簡素化を図る．

2.2 字句分解法

翻訳するに当たって，翻訳の対象となる文章を字句に分けてから単語に分ける．字句分けは鍵語に着目した字句分解方法を用いて行う．

1) 考え方

本格的な単語辞書を用いずに，文章中に存在する格助詞や句読点などの鍵語（keyword）に着目して字句分けを行う方式である．まず，文章中に含まれる特定の文字列からなる語句を鍵語として決定しておく．そして単語に付属語，または他の単語が付属したものを鍵語で機械的に分解する．この分解された語句を字句とする．

文章に書かれた日本語文は，格助詞，副詞，句読点などの特定の語句によって区切ると，切れ目ごとにまとまりをもち，意味が損なわれることがない．日本語文を字句に分解すると

き，これらの語句を鍵語として区切ると，単語辞書を用いずに簡便なシステムで実現できる．

2)　鍵語の選択

鍵語に着目した字句分解方法ではその正確さは，選択した鍵語によって変わるので，この状況を把握するための調査を行う．

①　調査方法

字句解析率を目安として，字句分解法の正確さを調べる．ここで字句解析率は

$$(\text{字句解析率}) = (\text{正当な字句数}) / (\text{全字句数}) \times 100$$

ここでは，選定する鍵語による字句の正当性を確かめる．鍵語として活用のない品詞14種を選定した場合と，37種類を選定した場合である．前者の内訳は格助詞12種と句読点であり，後者は前者に助詞5種，代名詞8種，連語2種，接続詞3種を追加したものである．字句分けの正当性について資料1[注1)]と資料2[注2)]について分析を行った．

②　調査の結果

鍵語を格助詞，句読点のみで構成した場合，字句解析率は，資料1で92.3%，資料2で90.0%である．鍵語に格助詞だけを用いて字句分解を行うと9割以上の精度がある．鍵語に副詞，代名詞等を追加した場合，字句解析率は，資料1で96.9%，資料2で97.9%である．鍵語に副詞，代名詞等を含めると，95%以上の精度が得られる．

3)　翻訳方式への適用

古文に対してこの字句分解法式を適用した．鍵語としては格助詞7種，係助詞1種，記号5種，句読点の15種を選定した．この場合，9割以上の確度があり，翻訳方式への適用が可能である．

2.3　活用形の決定法

文末または字句末単語の活用形の決定方法である．文章を字句に分解することで，字句末が明らかになる．また，句点までの語句のまとまりを文とすることから，文末も明らかになる．それらを単語分けの起点とする．そして，文末または字句末の単語の活用形は鍵語によって決定する．

例えば，格助詞，句読点を鍵語とする場合を考える．これらの語句は，次の例文中では実際このように使われる．

例文：物 を 思 ひ 、 願 を 立 つ 。
（格助）　（読点）　（格助）　（句点）

格助詞には体言，もしくは体言に準ずる語（活用語の連体形）につくという接続法則がある．読点は文の中で語句の切れ目に打つ点であり，句点は文の終わりに打つ記号である．各語句のこのような性質をもとに，字句末の単語の活用形を決定する．

注1）日経バイト：11月号（1998年）

注2）診療録管理：日本診療録管理学会刊　Vol. 9（1997年）

鍵語が格助詞の場合，鍵語の前は体言かもしくは活用語の連体形とする．読点の場合，活用形の連用形が文を一旦中止し，後に続ける用法をもつことから，読点の前の活用形は連用形とする．特に，文末，つまり鍵語が句点の場合は，活用形の終止形が文を言い切る用法をもつことから，活用形を終止形とする．実際に文末単語の活用形を調査すると，8割以上の活用形が終止形であった．この調査結果からも翻訳方式への適用が可能だと言える．これを表1に示す．

2.4　活用語の単語分け

辞書を用いない方法で活用語を単語分けし，基本形変換する方式である．単語分けの起点と活用形が決まっているので，そこから単語分けを行う．単語間の接続法則から次に来る単語の品詞や活用形が決まる．それに従って字句を順に単語分けする．また，活用のある古語単語は一度基本形に変換してから現代語に翻訳する．現代語変換する時に必要とされる単語は基本形を基準とする．

例えば，「捕へさせず。」という字句を考える．句点の前の単語は終止形とすることから，「ず」は助動詞「ず」の終止形と単語分けされる．助動詞「ず」の前の単語は未然形であるという接続法則から「させ」は助動詞「さす」の未然形と単語分けされる．これらの単語分けには助動詞辞書を使用する．助動詞「さす」の接続法則から前述単語の活用形は未然形となるが，「捕へ」は辞書に存在しないことから動詞と判別される．これは活用形であるので基本形に変換され，動詞「捕ふ」の未然形と単語分けされる．

2.5　辞書の構成

本システムでは，単語分けで使用する単語辞書と翻訳で使用する翻訳辞書をもつ．各辞書

表1　文末調査結果

活用形	割合（%）	
	資料3[注3)]	資料4[注4)]
末然形	0	0
連用形	0	0
終止形	86.7%	82.7%
連体形	13.3%	2.7%
已然形	0	4.0%
命令形	0	10.7%

注3)「竹取物語」(第一章) 小学館
注4)「竹取物語」(第二章) 小学館

の関係を図１に示す．この場合，類似した項目をもつ辞書が複数あると，辞書内容の変更や追加しにくいという問題が生じる．しかし，実現性の面からは，辞書が単純な構造になり，容易に実現することができる．また，単語数を抑えることで問題を回避する．

1)　単語分け辞書

単語分けは連語辞書，助詞辞書，助動詞辞書，形容詞語幹辞書の４種類の単語辞書を用いて行う．辞書は見出し語の他，その見出し語の品詞や単語間の接続関係を示す番号をもつ．単語分けを行う場合，単語辞書を検索して行き，文が辞書中の単語と一致したら品詞と活用形を与えて単語に区切る．辞書と一致しない語は辞書を用いない方法で単語分けを行う．特に動詞の場合，文章中に出現する単語の種類が膨大で，後の単語の追加などが困難である．そのため単語辞書を用いない方法で単語分けを行い，古語の活用語尾変換表を用いて終止形に変換する．その他の活用語は動詞に比べて遥かに単語の種類が少なく，助動詞については全単語数が決まっている．これによって少ない辞書でシステムを実現することができる．

2)　翻訳辞書

翻訳は連語辞書，助詞辞書，助動詞辞書，形容詞辞書，名詞辞書，動詞辞書の７種類の翻訳辞書を用いて行う．名詞など活用のないものは辞書に見出し語と，それに対応する現代語をもち，動詞のような活用語はそれに加えて活用の種類と，活用行をもつ．単語分けをした時に，単語には品詞が与えられているので，使用する翻訳辞書が決まる．辞書中の単語と一致したら対応する現代語に変換する．活用語は終止形に変換されているので，単語に与えられている活用形に従って現代語を活用させる．活用形の変換には現代語の活用語尾変換表を用いる．この場合，活用形の翻訳辞書は見出し語として終止形だけをもてばよく，辞書の単語数を軽減することができる．

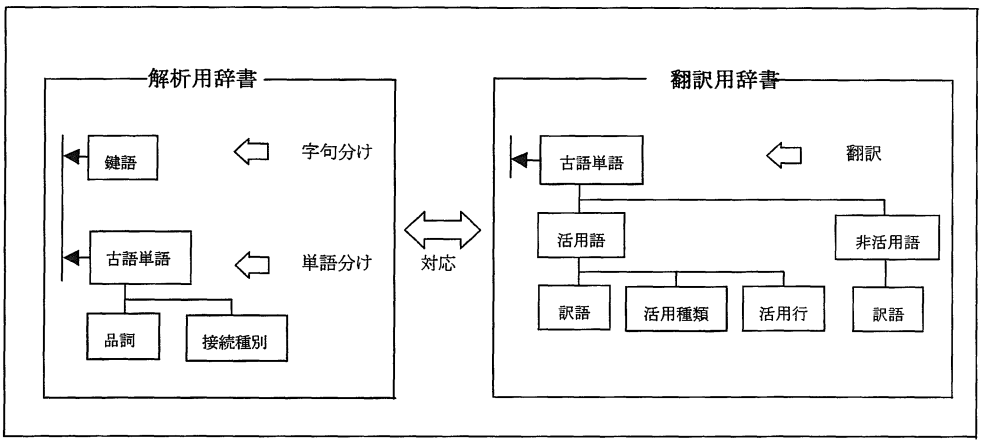

図１　辞書相関図

2.6 例　　示

「物を思ひけり。」という文を例とする．この文は，字句分解法によって「物/を（鍵語)/思ひけり/。(鍵語)/」と字句分けされる．字句単位に単語分けを行うので，まず字句「物を」を単語分けする．字句「物を」は鍵語が格助詞「を」であるから前に来る単語は体言もしくは活用語の連体形となる．前にある単語が漢字１文字の「物」であるからこれは名詞と単語分けされる．

次に，字句「思ひけり。」を単語分けする．句点の前の単語は終止形とすることから，「けり」は助動詞「けり」の終止形と単語分けされる．助動詞「けり」の前の単語は連用形であるという接続文法則から動詞の連用形「思ひ」が単語分けされる．また，「思ひ」を現代語に変換するには，まず動詞「思ひ」を古語の活用語尾変換表を用いて「思ふ」という基本形に変換する．

よって，「物/を/思ひけり/。/」は「物（名詞)/を（鍵語)/思ふ（動詞・連用)/けり（助動詞・終止)/。(鍵語)」に単語分けされた．単語の意味を翻訳辞書から引き，現代語に翻訳する．「物」と「を」は古語を現代語の意味が変わらないので，そのまま「物を」とする．一方，「思ふ」は，意味を翻訳辞書から引き，現代語「思う」に翻訳する．この時古語が連用形であるので，現代語の活用語尾変換表を用いて活用形「思い/思っ」に変換する．特に，この古語「思ひ」を現代語「思い/思っ」に変換する方式を図２に示す．また助動詞「けり」は「た」と現代語に変換され，「物を思った。」という現代語訳が得られる．

3.　システムの評価

本提案システムの評価では，単語辞書によって現代語に単純置換する翻訳方式と比較する．単純単語置換翻訳方式は，文章を字句分けした状態で辞書だけを使用して現代語に置き換えるだけの，最も簡素な方式である．これは本提案システムに対する比較基準として試作した．

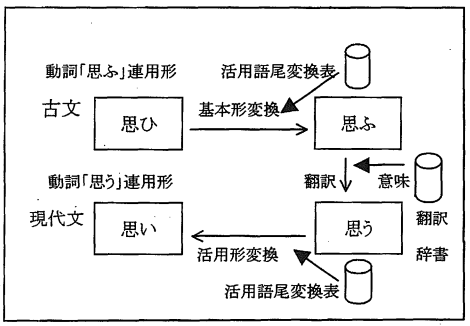

図２　基本形変換の方式

3.1 評価方法

翻訳方式に対する評価は，人の感覚器官を計測器として行う官能検査の考え方に基づいて行う．人は総合的な判断を迅速に行うことが可能である．判断基準として，次の4水準の評価尺度を設ける．自然な場合，ややぎこちない場合，ぎこちない場合，誤っている場合の4つである．評価の対象は訳語のみで，翻訳の基礎となる字句について実施する．各字句に対して4つの評価尺度のいずれかを与え，その個数を数える．正当性は，全字句数に対する各評価尺度の割合であり，次で与える．

$$(\text{各評価尺度の数})/(\text{字句数})\times 100$$

ここでは，単純単語置換翻訳方式と，本提案での翻訳方式に対する評価を行う．なお，資料は小学館版の『竹取物語』第一章，第二章とした．単純単語置換翻訳方式の結果を表2に，本提案での翻訳方式の結果を表3に示す．

3.2 評　　価

単純単語置換翻訳方式では，第一章での「誤り」でないものの正当性は70.4%に対し，第二章では54.8%と低下している．これは単語辞書に制限を与えたことによって生じたものである．一方，本提案方式では，「誤り」でないものの正当性は第一章で86.2%，第二章で82.0%と偏りがない．

表2　単純置換方式の評価結果

評価対象	評価単位	字句数	評価尺度			
			自	や	ぎ	誤
第一章	一字句	142	87	7	6	42
	正当性（%）		61.3%	4.9%	4.2%	29.6%
第二章	一字句	228	88	4	33	103
	正当性（%）		38.6%	1.7%	14.5%	45.2%

表3　本提案での翻訳方式の評価結果

評価対象	評価単位	字句数	評価尺度			
			自	や	ぎ	誤
第一章	一字句	137	99	10	9	19
	正当性（%）		72.3%	7.3%	6.6%	13.8%
第二章	一字句	339	247	18	13	61
	正当性（%）		72.9%	5.3%	3.8%	18.0%

4. 異なる評価者による評価

官能検査の考え方に基づいた評価では，人の感覚器官を計測器として行うため，評価する者による影響が無視できない．人は総合的な判断を迅速に行うことが可能であるが，同時に，各人の個人的な主観が入る危険性も生じる．評価に客観性を与えるためには，評価者による偏りの程度を確かめておく必要がある．

そこで，本提案システムの評価を，異なる者が行った場合の評価結果を示す．評価は先に述べた評価方法と同様にして行う．この結果を図3，図4に示す．

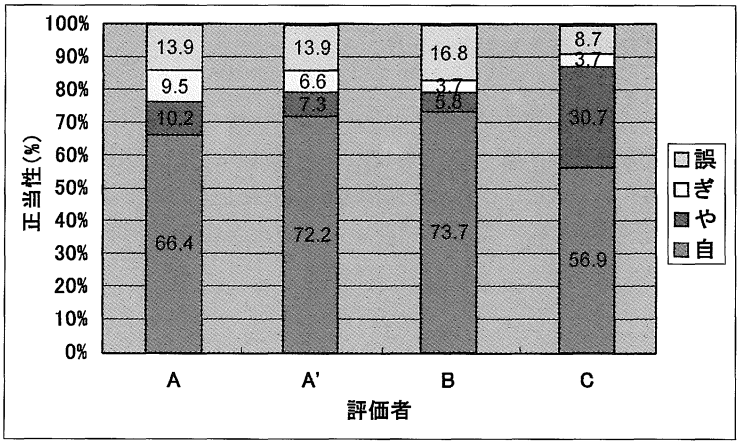

図3　異なる評価者による評価結果（第一章）

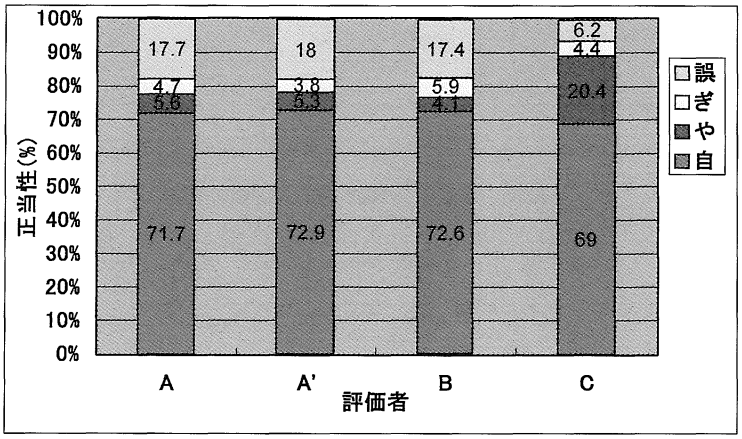

図4　異なる評価者による評価結果（第二章）

4.1 評 価 者

評価を行うのは次の3人である.

- 評価者A：システムの開発者
- 評価者B：今年新たにシステム開発に加わった者
- 評価者C：システム開発とは無関係な者

開発者であるA, Bは共に古文に関心があり知識もあるが，それでは開発者側の視点による評価になりがちである．そこで一般的な立場であるCを評価者として加えた．時間経過による評価基準の変化を見るために評価者Aが昨年行った評価結果をA′で示す.

4.2 分 析

1) 個人的な傾向

① AとA′

「誤り」でないものの割合が変化している．これは同一人物であっても時間の経過で，評価基準が厳しくなったため生じた．しかし，差は僅かであり，評価結果に信憑性があると言える.

② A(A′), BとC

A, A′, Bの分布は似通っているのに対し，Cの分布には差が見られる．前者は少なくとも開発者であるという共通の意識や先入観をもつため，それが評価基準に反映されていると想像される．しかしCは，その意識や先入観をもたないことから評価基準に反映されない.

③ C

CはA, A′, Bより「自」の割合が減り「や」の割合が増えている．これは古文への知識，関心度の違いが原因と想像される.

2) 全体の傾向

評価者の「誤り」の正当性は，16.8%, 16.6%, 17.2%, 7.1%と何れも2割以下の幅に収まっている.

評価者が変わることにより結果に個人差が生じている．しかし今回の調査からは，共通した意識をもつ者では分布が類似する傾向にある．また，「誤り」であるか否かという点から見れば，4つの評価結果はある程度の幅にある.

5. 結 論

本提案の翻訳方式の能力が単純単語置換翻訳方式より著しく向上し，資料を変えた翻訳結果に差が見られない．また評価者を変えて評価を行っても，個人差はあるが，ある程度の範囲に収まっている．これは開発者の主観的な評価結果であるだけでなく，第三者の客観的な結果であることから，評価方法として妥当であると言える．これらのことから，このシステムは有効であり，かつ汎用性がある.

参 考 文 献

1) 池原 悟：「機械翻訳研究の現状と課題」池原研究室・年度報告（平成10年度）Natural Language Processing Vol. 3 pp240–252
2) 山口，佐藤：「鍵語に着目した字句分解法の試み」情報処理学会第59回全国大会 講演論文集（1999）
3) 山口，佐藤：「鍵語による字句分け法の適用法」電気・情報関連学会中国支部第50回連合大会 講演論文集（1999）
4) 山口，佐藤：「医療関係論文における文記述特性の分析」日本診療録管理学会（1999）
5) 松尾，佐藤：「伝統的かな遣い変換システムの試作」情報処理学会第59回全国大会 講演論文集（1999）
6) 矢田，佐藤：「簡易型の古文機械翻訳システム」情報処理学会第61回全国大会 講演論文集（2000）
7) 矢田，佐藤：「古文機械翻訳システムにおける翻訳評価の個人差」電気・情報関連学会中国支部第51回連合大会 講演論文集（2000）