自 然 言 語 処 理　146－8
（２００１．１１．２０）

帰納的学習を用いた機械翻訳手法における数字表現の利用方法について

松原 雅文　荒木 健治　栃内 香次
北海道大学大学院工学研究科電子情報工学専攻
E-mail:{matuhara, araki, tochinai}@media.eng.hokudai.ac.jp

本手法における翻訳は，原言語テキストを抽象度の高い数字表現に変換してから行われる．そのため，ここでの翻訳結果は目的言語に対応した数字表現となる．この数字表現をさらに目的言語テキストに変換することにより，最終的な翻訳結果となる．あるテキストに対応する数字表現は，複数のテキストに対応しており，抽象度が高くなっている．しかしながら，このような抽象度の高い数字表現に帰納的学習を適用することにより，より多くの翻訳ルールを獲得することができる．本手法においては，この数字の割り当て方法が重要となる．割り当て方法を検討した結果，帰納的学習の利点を損わないように，言語間の共起情報に基づき数字を割り当てることとした．その結果，翻訳精度の向上が示され，数字表現を利用する本手法の有効性が確認された．

Usage of Number Representation for Machine Translation Method by Inductive Learning

Masafumi Matsuhara, Kenji Araki and Koji Tochinai
Graduate School of Engineering, Hokkaido University

A source language is translated into a target language via number representation on our proposed method. A text in the source language is translated into a number representation text. The number representation text for the source language is translated into a number representation text for the target language. The number representation text for the target language is translated into a text in the target language. A number representation is more abstract than a language because the number representation text corresponds to several texts in the language. The system based on our proposed method is able to acquire more translation rules on number representation than that on language because of its own abstraction. It is important how to assign number representation. The correct translation rate increases by use of number representation based on cooccurrence.

1　はじめに

世界規模のネットワークの発達により，さまざまな言語を母国語とする世界中の人たちとの距離が近くなった現在，コミュニケーションツールとして機械翻訳システムの需要が高まってきている．現在，最も一般的な商用の機械翻訳システムは，解析的な知識に基づくものである．このような機械翻訳システムにおいて翻訳精度を向上させるためには，大規模な単語辞書，対訳辞書等の知識をあらかじめシステムに与えてやればよい．しかしながら，これらの知識は翻訳対象となる言語や分野に大きく依存し，すべての言語現象を適切な形式でシステムに与えることは困難であると考えられる．このような問題を解決するために，統計的手法に基づく翻訳 [2]，用例に基づく翻訳 [3][4] が提案されている．これらの手法において精度の良い翻訳を行うためには，一般に大量のコーパスが必要となる．しかしながら，

コーパスが整備されていない言語も数多く存在することなどから，このような手法でさまざまな言語に適用可能な翻訳システムを作成することは困難であると考えられる．

これらの問題を根本的に解決するためには，システムが学習により自分自身で翻訳ルールを獲得していく方法が考えられる．これを実現するために我々は，表層的な情報のみから学習を行うことが可能な帰納的学習による翻訳手法 (IL-MT) を提案している [5][6]．この手法においては，字面情報から共通部分，差異部分を手掛かりに翻訳ルールを獲得していく．このように表層的な情報のみを利用しているので，この手法は種々の言語へ適応可能である．その反面，表層的な情報のみを利用しているので，字面上で 1 文字でも異なる文字列は共通部分とはなり得ないため，ルールとして獲得することができない．これを補うためにはより多くのコーパスを収集することが考えられるが，統計的な手法と同様の問題を抱え込むこととなる．そこで少量のコーパスからでも数多くの翻訳ルールを獲得できる手法として，遺伝的アルゴリズムを適用した帰納的学習による機械翻訳手法 (GA-ILMT) を提案している [7]．しかしながらこの手法においては，多大に生成された誤った翻訳ルールをフィードバック処理によっても完全に淘汰できないという問題点があった．

我々はすでに，情報が縮退した数字表現からその失われた情報を復元し漢字かな混じり文に変換する，帰納的学習を用いた数字漢字変換手法 (IL-NKT) を提案している [8][9]．この手法において 1 つの数字表現は複数の漢字かな混じり文に対応しており曖昧である．しかしながら，本手法の持つ高い適応能力によりこの曖昧さを排除し，80[%] 程度の精度で変換可能であることが確認されている．ここで，数字表現が曖昧であるということは，漢字かな混じり文に比べて抽象度が増大していることを意味する．

そこで，我々はこの数字表現の持つ抽象度の高さを利用した機械翻訳手法を提案している [10]．帰納的学習を用いた翻訳ルールの獲得は，字面情報より，共通部分，差異部分を手掛かりとして行われる．よって，抽象度の高い数字表現に帰納的学習を適用することにより，字面情報で一致する文字数が増加するので，獲得可能なルール数を増加させることができると考えられる．これを実現するために，本手法において，原言語テキストは，まず対応する数字表現に変換される．数字表現はその抽象度の高さから，複数の原言語テキストに対応している．原言語テキストに対応する数字表現は，目的言語に対応した数字表現に変換され，さらにこれを目的言語に変換することにより，最終的な翻訳結果となる．この際，数字表現の持つ曖昧さについては，帰納的学習の持つ高い適応能力を用いて解消することにより，正しい翻訳結果を得る．このようにして，本手法においては，帰納的学習を最大限に活用した機械翻訳手法の実現を目指している．

本手法においては，変換された数字表現をもとにして翻訳，学習が行われるので，数字表現への変換方法は翻訳精度に直接的に大きな影響を与えるものと考えられる．そこで本稿では，本手法における数字の割り当て方法を説明し，さらに本手法に基づくシステムを作成し実験を行った結果から，数字表現の利用により翻訳精度の向上が可能となることを述べる．

2 基本的な考え方

本手法で用いている帰納的学習においては，字面情報から共通部分，差異部分を手掛かりに翻訳ルールを獲得する．以下の例においては下線部分が共通部分となる．

He is Taro.：彼は 太郎 です.
He is Takuya.：彼は 拓哉 です.

よって，この 2 組の翻訳例から以下の 3 組の翻訳ルールが獲得される．

He is @1.：彼は @2 です.
Taro：太郎
Takuya：拓哉

ここで，@x は変数を表しており，翻訳の際に他のルールを代入することが可能である．

このように表層的な情報のみを利用することにより，この手法は種々の言語へ適応可能である．その反面，表層的な情報のみを利用しているので，字面上で 1 文字でも異なる文字列は共通部分とはなり得ない．よって，次のような例では，英語文に共通部分が含まれないため，翻訳ルールを獲得することができない．

He is Taro.：彼は太郎です.
I am Takuya.：私は拓哉です.

しかしながら，これらの翻訳例中にも，字面が異なってはいるが，有効な翻訳ルールが含まれているものと考えられる．そこで，このような翻訳ルールを獲得することを本研究の目的とする．

この目的を実現するために文字列の抽象度を増加させることを考える．以下のように，文字列が記号に割り当てられているものとする．

α = He, I, ...　　Θ = 彼は, 私は, ...
β = is, am, ...　　Λ = 太郎, ...
γ = Taro, ...　　Ψ = 拓哉, ...
δ = Takuya, ...　　Ω = です, ...

この対応関係を用いて，前述の翻訳例は以下のように表すことができる．

$\underline{\alpha\beta\gamma.}$：$\underline{\Theta}\Lambda\underline{\Omega.}$
$\underline{\alpha\beta\delta.}$：$\underline{\Theta}\Psi\underline{\Omega.}$

下線部分が共通部分となるので，以下のような翻訳ルールを獲得することができる．

$\alpha\beta$ @1.：Θ @2 Ω.
γ：Λ
δ：Ψ

ここで獲得されたルールは複数の文字列に対応しており，曖昧である．しかしながらこの曖昧さは，本手法の持つ高い適応能力により解消可能である [8][9][10].

このように本手法においては，抽象度の高い数字表現を介することにより，帰納的学習を最大限に活用した翻訳を行うことが可能となっている．

3 数字表現からの帰納的学習を用いた機械翻訳手法

本手法は基本的に種々の言語に適応可能であるが，本稿では英日翻訳を対象としている．

本手法の全体の処理過程を図 1 に示す．使用者により入力された英語文は，英語数字変換処理により数字表現に変換される．変換された英語数字表現に対して，翻訳処理が行われる．翻訳処理の処理過程を図 2 に示す．英語数字表現は，数字翻訳処理により目的言語である日本語に対応した日本語数字表現に翻訳される．ここでの翻訳は，学習処理により獲得された数字翻訳辞書を用いて行われる．数字翻訳辞書には翻訳ルールとして，英語と日本語のそれぞれに対応する数字表現の組が登録されている．日本語数字表現は，数字漢字変換処理により日本語文に変換され，最終的な翻訳結果となる．ここでの変換は，数字漢字変換辞書を用いて行われる．数字漢字変換辞書には変換ルールとして，日本語のセグメントとそれに対応する数字表現の組が登録されている．最終的な翻訳結果に誤りが含まれている場合には，人手により校正が行われる．次に，漢字数字変換処理により，校正済み翻訳結果は校正済み翻訳結果数字表現に変換される．これは，数字表現を用いて学習処理を行うための処理である．もちろん，この数字表現も複数の日本語文に対応している．このように抽象度が高い数字表現を用いて学習処理を行うことにより，学習効率を高めることができる．学習処理では，校正済み翻訳結果

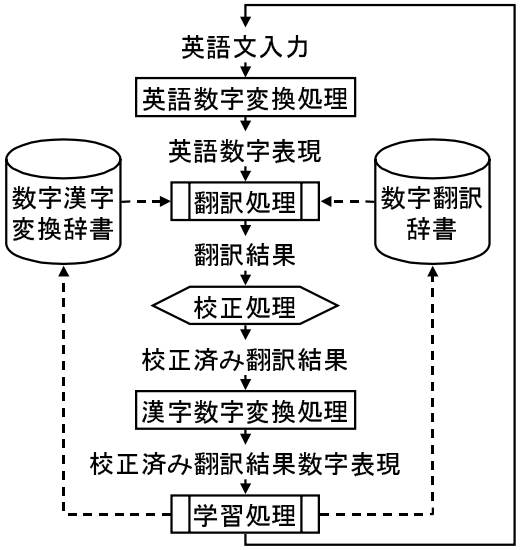

図 1: 処理過程

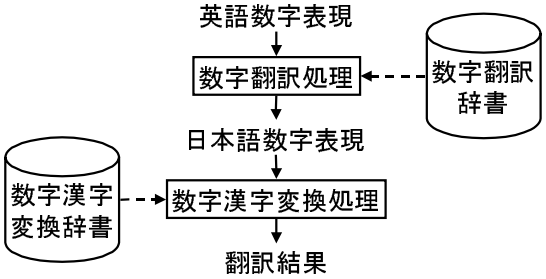

図 2: 翻訳処理過程

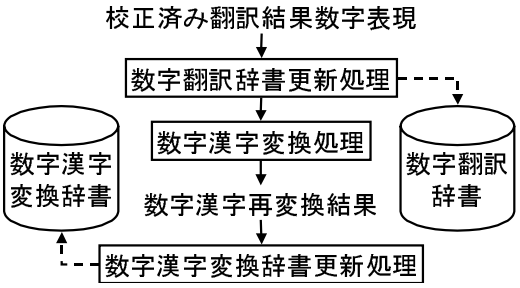

図 3: 学習処理過程

表 1: 数字とアルファベットの対応関係

1:XRUFq	2:zKLjJ	3:DTPBA
4:EHNOx	5:MSWYC	6:bvIkw
7:fmpyd	8:gchru	9:ilnst
0:その他		

とそれに対応した校正済み翻訳結果数字表現を用いて，数字翻訳辞書，数字漢字変換辞書の更新が行われる．学習処理の処理過程を図 3 に示す．ここでは，帰納的学習によるルールの獲得と，すでに辞書に登録されているルールの尤度の更新が行われる．この処理により，翻訳の際に誤りとなったルールの尤度は低下する．このようにして更新された辞書を用いて，次回からの翻訳が行われる．よって，これらの処理を繰り返すことにより，次第に翻訳精度が向上するシステムとなっている．

4 数字の割り当て方法

本手法で用いている帰納的学習においては，原言語文とそれに対応した目的言語文の組から，言語間の対応関係を保持したまま翻訳ルールを獲得する必要がある．よって，言語間の共起情報に基づき数字表現への割り当てを行うこととした．

英語の単語である “bank” を考える．“bank” に対応する日本語の単語としては，“堤防”，“銀行” などがあり，これらが翻訳例中に同時に出現する確率は他の単語のそれに比べて高いものと考えられる．よって，この “堤防”，“銀行” のような共起確率の高い単語に同一の数字表現を割り当てる．これにより，これらが共通部分の候補となり，言語間の対応関係を保持したまま，翻訳ルールを獲得することができると考えられる．

表 2: 数字とかなの対応関係

1:ユヅペパィヤゲ	2:ボヘプポメギャフ
3:グビズベュケヒソ	4:ゴミジブムザホラ
5:ョルゾバヲツモヨ	6:レネダガチゼタオ
7:ワテサドナッエキ	8:ロアセハトシクノ
9:ニリマコデーカウ	0:スンイ
*:その他	

しかしながら，実際のデータにおいては，このような共起する単語が出現する確率は非常に低いものと考えられる．信頼性の高い値を獲得するために大量のコーパスを利用することも考えられるが，この場合，コーパスに基づく手法と同様の問題を抱え込むことになる．よって，本手法においては，少量のコーパスで頑健さを保持するために，数字への割り当てを文字単位で行うこととした．なお，同様の理由により，日本語については，よみがなを対象とする．“ていぼう”，“ぎんこう” に対して，以下のように数字の割り当てが行われているものとする．

1：て，ぎ，... 2：い，ん，...
3：ぼ，こ，... 4：う，...

この場合，“てんこう” なども同様の数字表現 “1234” によって表現されるため，これらが共通部分となり得る．

このように，本手法においては抽象度を高めた数字表現に帰納的学習を適用することにより，学習効率の向上を図っている．

5 評価実験

本手法の有効性を確認するために，前述の処理過程に基づいたシステムを作成し，評価実験を行った．なお，共起情報に基づく数字の割り当てを決定するために予備実験を行った．

5.1 予備実験

数字の割り当て方法を決定するために予備実験を行った．実験データとしては，後述する旅行者用英会話文の「機内」のデータを用いた．このデータから前述の数字の割り当て方法に従い，対応関係を決定した．決定された数字とアルファベットの対応関係を表 1 に，数字とかなの対応関係を表 2 に示す．それぞ

表 3: 実験データ

	入力文字数	校正済み翻訳結果文字数
機内	8,463	4,203
空港	16,809	7,395
チェックイン	12,667	5,763
合計	37,939	17,361

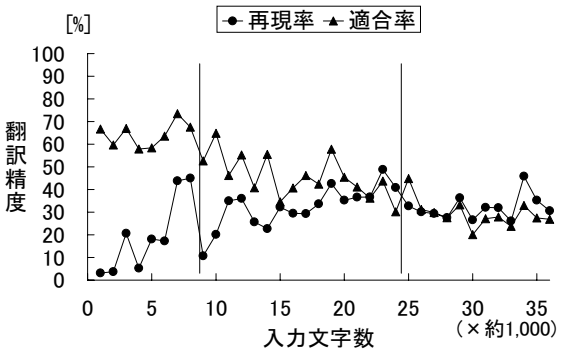

図 4: 翻訳結果における翻訳精度

れ，10 種類，11 種類の数字に割り当てられており，数字表現に変換したデータのエントロピーは，ともに 3.0[bit] となった．

5.2 実験データ及び実験手順

実験データとして，旅行者用英会話文の「機内」「空港」「チェックイン」の 3 つの場面を用いた [11]．それぞれ 316 文，587 文，427 文があり，合計 1,330 文を実験データとした．実験データを表 3 に示す．

実験は 1 文単位で行った．すなわち，まず英語文 1 文を入力し，図 1 に示した処理過程に従い翻訳を行う．そして，学習処理によりそれぞれの辞書を更新する．更新された辞書を用いて，次の 1 文の翻訳を行う．このように入力，翻訳を 1 文ごとに繰り返して実験を進めていき，入力文字数約 1,000 文字ごとの入力データに対し，以下に示す再現率，適合率により評価を行った．

$$\text{再現率} = \frac{\text{正翻訳文字数}}{\text{校正済み翻訳結果文字数}}$$

$$\text{適合率} = \frac{\text{正翻訳文字数}}{\text{翻訳結果文字数}}$$

なお，英語数字表現からの数字翻訳結果である日本語数字表現に対しても，同様に再現率，適合率で評価している．

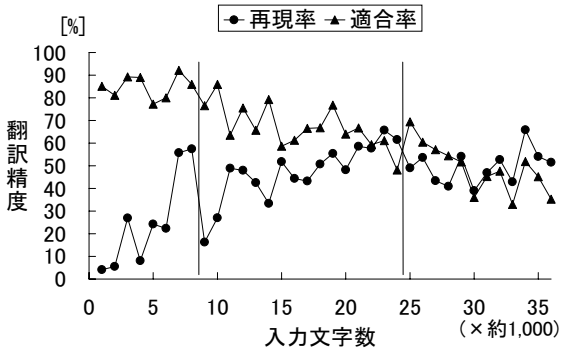

図 5: 日本語数字表現における翻訳精度

評価方法としては，理解容易度，忠実度 [12]，合文法性 [13] などがあるが，少なからず評価者の主観が入ってしまうものと考えられる．また，字面上から獲得される本手法の翻訳ルールの有効性を確認するため，今回は正解となる日本語文は使用者が意図する 1 文のみであるものとして，字面上で一致した文字数で評価している．意味的に正解であっても字面が異なるとその文字は誤変換となるため，非常に厳しい評価基準であるが，これにより客観的な評価が可能である．

なお，本手法はあらゆる対象に動的に適応可能であるので，その適応能力を確認する必要がある．よって，種々の対象において初期状態を一定に保つために，辞書は空の状態から実験を行った．

5.3 実験結果

本手法の翻訳結果における翻訳精度の推移を図 4 に示す．全体の再現率は 29.0[%] であった．数字表現を利用しない場合のシステムでも同様の実験を行っており，その際の再現率は 24.4[%] であったので，数字表現を利用することにより 4.6 ポイントの向上が確認された．また，本手法において数字漢字変換を行う前の段階である，日本語数字表現における翻訳精度を図 5 に示す．この場合の全体の再現率は 37.7[%] であった．

5.4 考察

図 4 から分かるように，実験の初期の段階では各辞書が空なので再現率は低い値となっている．しかしながら，入力データ数の増加

に伴い，再現率は次第に上昇していく．場面が変化するときに再現率は一時的に低下するが，その後，現在の対象に適応した翻訳ルールを学習することにより，再現率は再び上昇していく．最終的に 45[%] 程度までの上昇が確認された．しかしながら，後半部分の再現率の上昇度合は低く，適合率は低下傾向にある．また，図 5 から分かるように，数字翻訳結果である日本語数字表現の再現率は上昇傾向にあるが，適合率は低下傾向にあり，後半部分では再現率を下回っている．これは，数字表現の持つ曖昧さから過度に翻訳ルールが適用された結果である．このように，日本語数字表現に多くの誤りが含まれると，その後の数字漢字変換処理を正しく行うのが困難となる．そのため，数字漢字変換処理における誤変換の割合が増大し，結果として翻訳結果における再現率の低下を招いている．

6 おわりに

本稿では，数字表現の持つ抽象度の高さに着目して学習，翻訳効率の向上を目指す「数字表現からの帰納的学習を用いた機械翻訳手法」を提案した．本手法においては，抽象度の高い数字表現を利用して学習，翻訳を行うので，その数字表現の割り当て方法が重要となる．帰納的学習では言語間の対応関係を保ったまま翻訳ルールを抽出する必要がある．そこで，言語間の共起情報に基づいた数字の割り当て方法により，数字表現へ変換するものとした．その結果，数字表現を利用しない場合に比べて，再現率で 5 ポイント程度の上昇が確認され，本手法の有効性が確認された．しかしながら，日本語数字表現を目的言語に復元する際の変換精度が低下していることも確認された．

よって，今後はこの復元精度の低下を抑える手段を検討し，さらなる翻訳精度の向上に向けてシステムの改善を進める予定である．

参考文献

[1] 田中穂積：自然言語処理–基礎と応用–, 社団法人電子情報通信学会編 (1999).

[2] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer and P. S. Roossin：“A Statistical Approach to Machine Translation,” Computational Linguistics, Vol.16, No.2, pp.79–85(1990).

[3] 古瀬蔵, 隅田英一郎, 飯田仁：“経験的知識を活用する変換主導型機械翻訳,” 情報処理学会論文誌, Vol.35 No.3, pp.414–425(1994).

[4] 佐藤理史：“実例に基づく翻訳,” 情報処理学会誌, Vol.33, No.6, pp.673–681(1992).

[5] 荒木健治, 栃内香次：“多段階共通パターン抽出法を用いた翻訳例からの帰納的学習による翻訳,” 情報処理北海道シンポジウム’91, pp.47–49(1991).

[6] 内山智正, 荒木健治, 宮永喜一, 栃内香次：“帰納的学習による機械翻訳手法の評価実験,” 情報処理学会研究報告, NL93-4, pp.23–30(1993).

[7] 越前谷博, 荒木健治, 桃内佳雄, 栃内香次：“実例に基づく帰納的学習による機械翻訳手法における遺伝的アルゴリズムの適用とその有効性,” 情報処理学会論文誌, Vol.37, No.8, pp.1565–1579(1996).

[8] 松原雅文, 荒木健治, 桃内佳雄, 栃内香次：“文字情報縮退方式を用いた帰納的学習によるべた書き文の数字漢字変換手法の有効性について,” 信学論 (D-II), Vol.J83-D-II, No.2, pp.690–702(2000).

[9] M. Matsuhara, K. Araki, Y. Momouchi and K. Tochinai：“Evaluation of Number-Kanji Translation Method of Non-Segmented Japanese Sentences Using Inductive Learning with Degenerated Input,” Proceedings of 12th Australian Joint Conference on Artificial Intelligence(AI’99), pp.474–475(1999).

[10] 松原雅文, 荒木健治, 栃内香次：“数字表現からの帰納的学習を用いた機械翻訳手法の有効性について,” 信学技報, TL2000-46, pp.49–56(2001).

[11] 地球の歩き方編集室：旅の会話集 2 米語/英語，ダイヤモンド社 (1993).

[12] 長尾真：“機械翻訳文の質の評価と言語の表現,” 情報処理学会誌, Vol.26, No.10, pp.1197–1202(1985).

[13] 吉見毅彦, Jiri Jelinek, 西田収, 田村直之, 村上温夫：“日英機械翻訳システム TWINTRAN の言語知識と翻訳品質の評価,” 自然言語処理, Vol.7, No.4, pp.143–162(2000).