実例を用いた類推による対応関係推定手法

タンテリ アンドリアマナカシナ† 荒木健治† 栃内香次†

本論文では原文と翻訳文間の対応関係の推定手法について述べる．実例に基づく機械翻訳では翻訳例に対応関係が含まれていれば翻訳パターンを容易に抽出することができる．従来の手法では統計に基づく手法であることから大量なコーパスが存在しないと良質な結果は得られない．したがって，大量なコーパスがまだ存在しない言語間では実例に基づく機械翻訳手法を適用するのが困難である．本手法では，対応関係が含まれている既存の翻訳例コーパスを用いて，新しい翻訳例中の対応関係を推定し，対応関係付きの翻訳例を増加させる．したがって，本手法を用いれば，翻訳例中の対応関係を利用することによって，一対多や多対多の対応関係をうまく推定することができる．本手法を用いた実験により仏日会話の文において，1,000翻訳例から80.0%以上の正対応関係が得られることが確認された．

Example-based Sub-sentential Alignment Method by Analogy

TANTELY ANDRIAMANANKASINA,† KENJI ARAKI† and KOJI TOCHINAI†

This paper describes a method for predicting word correspondences or links between a sentence and its translation sentence. In the Example-Based Machine Translation, translation patterns can be extracted easily if word correspondences between source sentence and translation sentence are determined. Statistical approaches are not able to produce reliable result unless a huge bilingual corpus is available. We propose a method for incrementing a link-included initial corpus automatically. It is appropriate for new languages whose huge corpus are still not available. The method was evaluated with French-Japanese spoken language texts. Links involving multiple tokens were predicted successfully. As the number of translation examples goes beyond 1,000, more than 80.0% of exact correspondence ratcs wcrc earned.

1. はじめに

実例に基づく機械翻訳では，バイリンガルコーパス中の対応関係を正確に見つけ出すことがきわめて重要である．翻訳例間の原文と訳文の対応関係が分かれば，翻訳パターンを容易に取り出すことができる．たとえば，フランス語文 **"je suis malade"**☆とその日本語の翻訳文 **"私 は 病気 です"** において，もし **"malade"** と **"病気"** の対応関係が決定されれば，翻訳パターン **"je suis X ： 私 は Y です"** を取り出すことができ，XとYをさまざまな単語対と置き換えることができる．

このような対応関係を見出すことは，辞書を利用して容易に行うことが可能に思えるが，登録されていない単語の出現，単語の活用形，辞書の見出し語と形態素解析結果との相違などの問題点がある．また，辞書の見出し語は1語単位であるのでさまざまな複合語には対応できない．さらに，同じ単語が同一の文に2回以上現れる場合それぞれの対応関係を正確に決定することは辞書では不可能である．そこで，辞書によらず，コーパスを用いて対応関係を求める手法が提案されている[1)~3)]．

コーパス中の対応関係の決定に関しては統計的な手法による研究が多く行われてきた[1)~3)]．統計的手法の問題点は，コーパスのサイズが限定されると，良質な結果が得られないことである．それゆえ，コーパスおよび辞書が整備されていない言語の場合，良好な結果が得られないことになる．さらに，バイリンガルコーパス中の一対多や多対多の対応関係をすべて正確に決定することはできないという問題がある．文献1)で

† 北海道大学大学院工学研究科
Graduate School of Engineering, Hokkaido University

☆ 文は形態素解析の結果で表示される．
フランス語の文や日本語の例文は太字で表示される．

は，一対多あるいは多対多の対応関係を決定できないことが失敗の原因であり，文献 2) では一対一の対応関係しか対象としていない．翻訳パターンの取り出しや機械翻訳という応用から見ると対応関係の誤りがシステム全体の性能に大きな悪影響を与える．一方，付属語の対応関係を決定しないでそれらを固定の言葉として扱う手法もある[3]が，その結果抽出される翻訳パターンは質が比較的良いが量は少ない．つまり，この場合も実用的な翻訳システムを作成するのに十分な翻訳パターンを抽出するためには大量のコーパスが必要になる．これらの点がこの手法の問題点である．

我々は対応関係を有する翻訳例を用いた仏日機械翻訳について研究を進めている[4),5)]．仏日機械翻訳は比較的新しい分野なので，大量なコーパスや辞書はまだ存在していない．上述のように，実例に基づく機械翻訳手法では翻訳例の数が大量にあればあるほど結果に対する信頼度が高まる．しかし，翻訳例の人手による作成には時間と労力がかかる．したがって，小規模な対応関係を有するコーパスを利用して自動的に大量の対応付けをする必要がある．

本論文では，原文と訳文の組の対応関係の推定において，すでに対応関係が決定済みの翻訳例から類似翻訳例を抽出し，その対応関係を基にその組の対応関係を推定する手法を提案する．ここで，既存の翻訳例コーパスの中にまったく同じ部分が現れなければ，品詞情報を利用して，対応関係を推定する．したがって，本手法では，少量のコーパスでは精度が低いという統計的手法の問題点を回避できる．

また，一対多や多対多の対応関係抽出問題も，あらかじめ翻訳例に対応関係が含まれていることで解決できる．たとえば，**"voulez - vous"** と **"いただけ ます か"** が現れたら，**"voulez"** だけが **"いただけ"** と対応するのか，**"voulez - vous"** に対応するのか，あるいは **"いただけ ます か"** と対応するのかいずれも考えられる．しかし，翻訳例の中にたとえば **"voulez - vous"** と **"いただけ ます か"** か，それに近い **"pouvez - vous"** と **"もらえ ます か"** があれば，それらの間の対応関係を参考にして容易に問題が解決できる．

次章で，本手法の概要を説明し，以下，実験方法と結果を述べ，結果の考察を行う．最後に今後の課題を述べる．

2. 対応関係の推定手法の概要

対応関係の推定の流れを，**図 1** に示す．原文とその翻訳文を入力し，翻訳例コーパスから類似翻訳例を抽出し，それらの対応関係を基に入力された原文と翻訳文間の対応関係を推定し，決定する．その結果に誤りが含まれていれば，人手で校正し，新しい翻訳例を翻訳実例コーパスに追加する．一方，対応関係の推定結果から何が正解か何が誤ったかによりフィードバック学習処理を行い，後の入力翻訳例の対応関係の推定を精度の向上させる．各文に，同手順を繰り返せば，大量なコーパスを少しずつ作ることができる．またフィードバック学習処理によって，同じ失敗を繰り返すのを避けたり，学習による成功例を応用することができる．本手法の過程は以下のとおり 3 つの段階に分けられる．

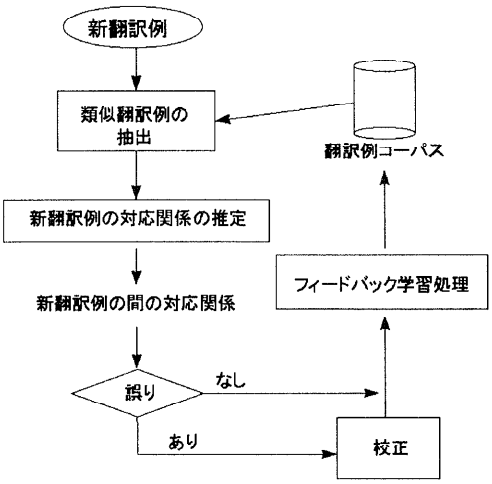

図 1 対応関係の推定手法の概要
Fig. 1 Overview of the method.

(1) 翻訳例コーパスからの類似翻訳例の抽出
(2) 新翻訳例の対応関係の推定
(3) フィードバック学習処理

翻訳例コーパスの中の実例の構造は，**表 1** に記述される．1 つの形態素は「単語/品詞」形式で表示される．フランス語の形態素解析では，INALF☆の EBTI プログラムを，日本語では CHASEN1.51[6)] を利用した．INALF が提供したフランス語の品詞の数は 40 である．その中の 12 が句読点や記号である．類似文の抽出にマイナスの影響を引き起こすことから，男性単語と女性単語，そして単数と複数は区別しないことにした．日本語では，品詞が木構造で分類されているが，我々は一番上のレベルに相当する 14 品詞を利用した．なお，本手法では構文解析，意味解析を行わず，形態素解析結果を用いている．構文解析や意味解析の結果を利用して，より正確な推定ができると思われるが，

☆ Institut National de la Langue Française.

表1 翻訳例の構造
Table 1 Structure of a translation example.

フランス語文	**"je/PRV suis/ECJ sans/PREP profession/SBC"**
日本語文	**"無職/6 です/4"**
対応関係	2/2 3,4/1

PRV: pronoun PREP: proposition SBC: common noun ECJ: verb "être"
6: 名詞 4: 判定詞

これらのツールの精度は依然として不十分である．一方，最近の形態素解析ツールは非常に精度が高い．そこで，品詞情報のみを利用することとした．

1組の対応関係は「WPf_1, WPf_2, ... / WPj_1, WPj_2, ...」形式で表示される．ここで，WPf_i はフランス語文の中の単語の位置で，WPj_j は日本語の中の単語の位置である．一対多や多対多は1組の対応関係に WPf_i あるいは WPj_j が複数あることで表示される．表1の例では，"3,4/1" は **"sans profession"** が **"無職"** と対応することを意味する．同様に **"suis"** は **"です"** と対応する．対応関係を人手で校正するとき，文の意味を理解したうえで，対応関係を決定した．したがって，さまざまな場合が存在している．日本語の **"は"** のような冠詞やゼロ代名詞のように対応がない形態素も考えられるし，**"s' il vous plaît"** と **"ください"** の関係のように複数の形態素と対応する場合も考えられる．また，**"voulez - vous"** と **"いただけます か"** の関係のように複数の形態素と対応する複数の形態素もあり，隣接しない部分は1つの形態素と対応する場合もある．たとえば，**" n' ai pas"** の **"n' pas"** と **"あり ません"** の **"ません"** を対応させる．

3. 類似翻訳例の抽出

3.1 抽出方法

類似文の抽出方法を以下に示す．

(1) 翻訳例コーパス中より入力翻訳例の原文と類似する原文を持つ翻訳例を検索する．
(2) 選択された翻訳例の訳文と入力文の訳文の類似を調査する．
(3) ここで，訳文同士が類似している場合に限りその翻訳例を類似翻訳例として抽出する．
(4) 翻訳例コーパス中より入力翻訳例の訳文と類似する訳文を持つ翻訳例を検索する．
(5) 選択された翻訳例の原文と入力文の原文の類似を調査する．
(6) ここで，原文同士が類似している場合に限りその翻訳例を類似翻訳例として抽出する．

このようにして抽出された翻訳例を合わせて対応関係の推定に利用する．

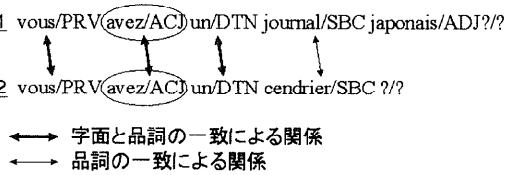

図2 原文の部分間のマッチング手法
Fig. 2 Partial matching method.

3.2 文の類似部分

ある言語で2つの文が類似していても他の言語ではそれらの翻訳文が類似するとは限らないことから，原文と翻訳文が類似している類似翻訳例を見つける可能性は低いと考えられる．そこで，類似文ではなく文の類似部分という考え方を導入する．文の一部だけを観察すれば，原文とも翻訳文とも類似している部分が見つかる可能性が高まる．最初に文の類似部分の検索を各言語で行う．入力原文の各形態素ごとにその形態素が入っている部分と類似する部分を n 個検索する．そのアルゴリズムは下記のとおりである．

> 入力文の各形態素に字面が同じ形態素を例文の中で検索する．見つかれば，その形態素の位置から，前方と後方の比較を行う．字面が一致しない位置から品詞を調べて比較を続ける．一致しない形態素が見い出されたら比較は終了する．

字面で一致する部分のうちその両側の品詞が一致するものの対応関係を決定する．**図2** に例を示す．式(1)の類似度を利用して最も類似する部分を選択する．

$$SM = \alpha \times NE + NP \qquad (1)$$

ここで，SM☆は類似度，NE☆☆は字面が一致する形態素の数，NP☆☆☆は品詞が一致する形態素の数である．

図2では形態素 **"avez/ACJ"** に関し，文の2番目の形態素が字面で一致する．そこから，前方で字面の一致 **"vous/PRV"** と **"vous/PRV"** を発見する．後方では字面の一致 **"un/DTN"** と **"un/DTN"** と品

☆ Similarity Metric.
☆☆ Number of Exact matches.
☆☆☆ Number of Part-of-speech tag matches.

詞一致 **"journal/SBC"** と **"cendrier/SBC"** を見つけ出す．ここで，隣接する部分しか検索しないので，**"japonais/ADJ"** という形態素を飛ばすことができない．したがって，"?" の対応関係は他の例文でカバーする．ここでは **"avez"** を例にあげているが上記のアルゴリズムで記述されたようにすべての形態素が対象である．合計では 3 つの字面一致と 1 つの品詞一致が存在する．したがって，類似度は式 (1) の NE に 3 を，NP に 1 を代入して，

$$SM = \alpha \times 3 + 1 \qquad (2)$$

となる．

各入力原文の各形態素ごとにその形態素が入っている部分と類似する n 個の部分を抽出する．たとえば，形態素個数が 6 の文の場合，各形態素に n の類似部分を検索し，その文に対し全部で $6n$ の類似部分を抽出する．ただし，同じ部分が取り出される可能性もあるし，字面一致がいつもあるとは限らないため類似部分が見つからない場合もある．たとえば，図 2 の文 1 を入力文とする．形態素 **"avez"** との類似によって，文 2 の **"vous avez un cendrier"** という部分が抽出される．また，この部分は形態素 **"vous"** の類似によっても抽出される．一方，翻訳例コーパスにまだ登録されていない形態素が出現すると，それに対し字面で一致する形態素がないので，類似部分は 1 つも抽出されない．つまり，実際に抽出された異なる類似部分の数は多くても n ということになる．SL☆は文の形態素の数とすれば，せいぜい $SL \times n$ の類似部分が抽出される．

1 文中では，同じ形態素が複数回現れる場合もある．この場合，各形態素は異なる部分に属するものと考えて扱う．たとえば，以下のフランス語の文を考える．**"la moitié des candidats sont des femmes."**☆☆ 形態素 **"des"** が 2 回出現している．したがって，他の文と比較するとき，それぞれの **"des"** が別々のものとして比較が行われる．

3.3 訳文中の部分間の類似

対応関係を取り出すために，原文か翻訳文の類似部分だけを探すのではなく，原文とも翻訳文ともに類似している部分の存在する文を抽出する．2 つの文とも類似している部分が存在していない場合には，その文は抽出しない．フランス語文の部分 `f1` と `f2` がそれぞれ日本語文の部分 `j1` と `j2` の翻訳文とする．もし，`f1` と `f2` および `j1` と `j2` が類似していたら，翻訳例 `(f1,j1)` と `(f2,j2)` が類似していると考える．ここで，3.2 節の文の対応関係検索アルゴリズムを再び利用して，入力翻訳文と類似文の翻訳文に類似部分があるか確かめる．

4. 新翻訳例に対する対応関係の推定

本章では，類似部分に存在している対応関係を利用して，入力原文と翻訳文の対応関係を推定する．推定方法の例を**図 3** に示す．`(f1,j1)` は新翻訳例で，`(f2,j2)` は取り出された類似例である．`f2` と `j2` の間の対応関係はあらかじめ翻訳例とともに与えられる．`f1` と `f2`，そして `j1` と `j2` の間の対応関係は類似翻訳例の抽出のときに決定される．ここでは，1 つの `f1` の形態素から 1 つの `j1` の形態素までの経路をすべて検索する．図 3 の例では，**"avez/ACJ"** から **"あり/2"** まで，そして **"journal/SBC"** から **"新聞/6"** までの 2 つの経路がある．それらが 2 つの文間の対応関係である．

いくつかの文の部分が取り出され，それらからすべての対応関係を推定しているので，結果的には同じ対応関係が何回も出現することがあったり，誤った対応関係も含まれたりすることがある．これらの誤りを取り除くために対応関係に優先順位を与える．そのために以下の 2 つのパラメータを利用する．

- 形態素と字面一致部分との距離，d
- 字面一致部分の長さ，l

たとえば，以下の 2 つのフランス語の文では

(1) **vous/PRV avez/ACJ un/DTN journal/SBC japonais/ADJ ?/?**

(2) **vous/PRV avez/ACJ un/DTN cendrier/SBC ?/?**

文 (1) の **"vous/PRV avez/ACJ un/DTN journal/SBC"** は字面と品詞列から **"vous/PRV avez/ACJ un/DTN cendrier/SBC"** という部分と一致する．字面一致部分は **"vous/PRV avez/ACJ un/DTN"** であって，その長さは 3 である．したがって，$l = 3$．たとえば **"journal/SBC"** という形態素を見ると，それと **"vous/PRV avez/**

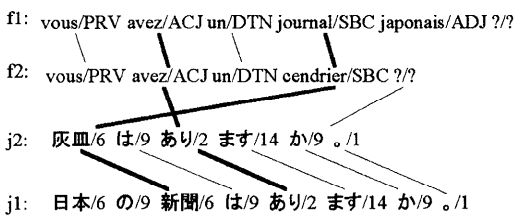

図 3 類推による対応関係の検索
Fig. 3 Search of word correspondences by analogy.

☆ Sentence Length.
☆☆ 日本語訳：受験者の半分は女性である．

ACJ un/DTN" という字面一致の部分との距離は 1 であるから $d = 1$ である．もしその形態素が字面一致部分に入っていたら $d = 0$ である．**"japonais/ADJ"** という形態素が一致している部分に入っていないことからそれに対する距離は計算されない．したがって，**"japonais/ADJ"** には別な類似文を利用する．

d と l で優先順位を定義する．(French, Japanese) が対応関係を表しているとすると，French と Japanese に対し，各距離 d の合計が小さいとき，その対応関係に高い優先順位を与える．他の対応関係と同じである場合には l が長い方を優先させる．たとえば，図 3 の例では，フランス語文では字面一致部分は **"vous/PRV avez/ACJ un/DTN"** で，日本語文では **"は/9 あり/2 ます/14 か/9 . /1"** である．それらの長さの合計は 8 形態素である．**"avez/ACJ"** と **"あり/2"** という可能な対応関係に関し，2 つとも字面一致部分に入っていることから距離 d の合計は 0 である．**"journal/SBC"** と **"新聞/6"** の場合では，各距離 d は 1 と 1 であって，合計は 2 である．

特に優先順位が必要なのは一対多や多対多の対応関係である．複数の対応関係の場合では，それらを構成している各対応関係が同じ部分から取り出されたら，それらをすべて正しい結果として見なす．すなわち，異なる部分から取り出された一対多や多対多の対応関係は優先順位の一番高いものを採用する．図 3 にこの例をあげる．一番高い優先順位で **"journal/SBC"** と **"新聞/6"** が対応しているとする．もし，**"japonais/ADJ"** と **"新聞/6"** の対応関係候補が出現したら，それを採用しない．**"新聞/6"** はすでに **"journal/SBC"** と一番高い優先順位で対応しているので，その対応関係が抽出された部分と同じ部分から抽出されない限り新たな対応関係は採用しない．

5. フィードバック学習処理

さまざまなユーザがシステムを使用するので，異なる分野の文や誤りのある文が入力される．また，対応関係の校正ときの間違いも存在する．システムがそれらの問題に耐えることができるようにここでフィードバック学習を行う．文献 7) で提案されたフィードバック学習処理の考えを本手法に導入した．各文にあるパラメータ FP[☆1]（初期値は -1）を利用する．校正結果と対応関係の推定結果を比較して，それらの対応関係の正誤を決定する．推定された対応関係が誤りであれば，使用された文の FP の値を 1 つ減らす．また，対応関係が正しければ，使用された文の FP を 1 つ増やす．FP は後の入力翻訳例に対しどの例を取り出せばよいかを決定するときに類似度に加えて用いる．ここでは，FP に関し 2 つの条件を導入する．

(1) 誤った対応関係が存在するとき，FP の減少はつねに行うが，正しいときには FP が初期値でない文のみ FP の増加を行う．これは，仮に正しいときに FP を増加させると特定の翻訳例のみが用いられるようになることを防ぐためである．

(2) FP はマイナスの値であるが，FP の絶対値を使用する．FP の絶対値が大きければ，その文が選択される可能性が低くなる．FP の値を類似文を抽出するときに類似度に加えて，以下の尺度を使用する．

$$NM = \frac{SM}{abs(FP)} \tag{3}$$

ここで，SM[☆2]は類似度で，NM[☆3]は類似文を選択するときの新しい尺度である．abs[☆4]は絶対値である．NM は文と文の類似度ではなく，入力翻訳例の対応関係の推定にどの文を使用すればよいかを決定するための尺度である．たとえば，ある翻訳例を 1 回使用して，その結果が誤りとなったとする．したがって，FP は 1 つ下がって -2 になり，新しい尺度は類似度の半分である．

表 2 利用したコーパス
Table 2 Data for the experiments.

翻訳例合計数	2,600
日本語文の長さの平均	7.74 形態素
フランス語文の長さの平均	7.84 形態素
1 文に対する対応関係の平均数	7.27

6. 実験と結果

実験に利用したコーパスを**表 2** に示す．このコーパスは仏日会話の本[8),9)]からとった．会話には短い文が多く存在することから文の平均長がやや短くなっている．翻訳例コーパスは最初空である．翻訳例を 1 つ 1 つシステムに入力し，対応関係を推定し，もし誤りがあれば人手により校正する．最初翻訳コーパスが空であるときを含めて，この実験では対応関係が推定できない場合も誤りとしている．最後に，入力された翻訳例を翻訳例コーパスに追加する．対応関係推定の適合率と再現率を以下の式で定義する．

☆1 Feedback Parameter.
☆2 Similarity Metric.
☆3 New Metric.
☆4 Absolute Value.

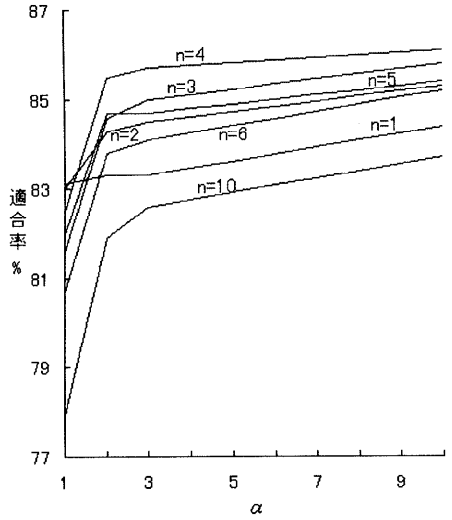

図4 パラメータの変化による適合率の変化
Fig. 4 Variation of the exact word correspondence rate by parameter variation.

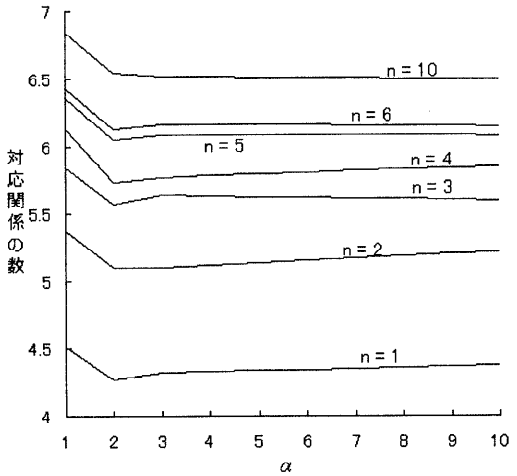

図5 パラメータの変化による対応関係の数の変化
Fig. 5 Variation of the number of extracted links by parameter variation.

$$\text{適合率 [\%]} = \frac{\text{正対応関係の数}}{\text{推定された対応関係の数}} \times 100 \quad (4)$$

$$\text{再現率 [\%]} = \frac{\text{正対応関係の数}}{\text{校正した後の対応関係の数}} \times 100 \quad (5)$$

式 (1) の α と 3.2 節の n の値は次に述べる予備実験により決定した．800 の対応関係付き翻訳例コーパスを使用し，新しい対応関係のない 200 翻訳例を入力した．α を 1, 2, 3, 10 まで，そして n は 1, 2, 3, 4, 5, 6, 10 に変化させて，結果の適合率と取り出された対応関係の数を観察した．適合率を**図 4** に示す．α が 1 になると高い適合率を得ることができない．α

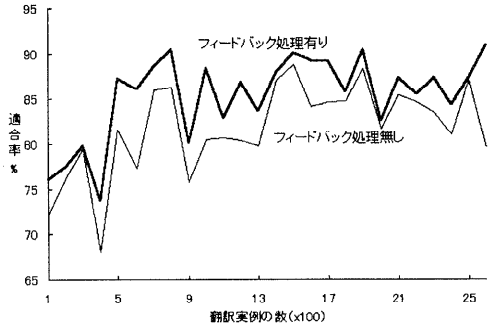

図6 適合率の変化
Fig. 6 Variation of the exact word correspondence rate.

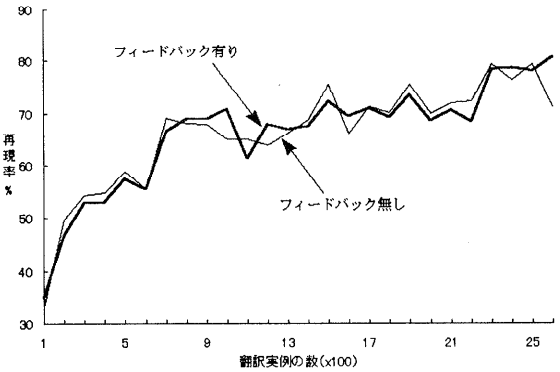

図7 再現率の変動
Fig. 7 Variation of the ratio of the number of extracted links.

が大きくなると適合率が高くなるが α が 3 のときからは大きな変化が見られない．n による変化を見ると，α が大きいとき n が 1, 2 または 10 の場合は，他の場合に比べ適合率が低い．取り出された対応関係の数を**図 5** に示す．n が一定のとき α が 1 のとき対応関係の数が一番多くなる．また，α が一定の場合には n が大きくなればなるほど対応関係の数が多くなる．高い適合率で多くの対応関係を抽出することを考えると α が 10，そして n が 5 となる．つまり，各形態素に最大 5 つの類似部分を抽出する．α は類似度に利用するパラメータで結局類似度は以下のようになる．

$$SM = 10 \times NE + NP \quad (6)$$

この式から α が大きければ類似度は字面一致の数の影響が大きくなる．字面一致の数がほぼ同じ場合に品詞一致の数の影響が出る．

次に，2,600 翻訳例を 1 つずつランダムにシステムに入力した．推定された対応関係が誤りの場合には，人手により校正してから翻訳例コーパスに追加する．結果を 100 例ずつにわけて適合率と再現率を観察した．

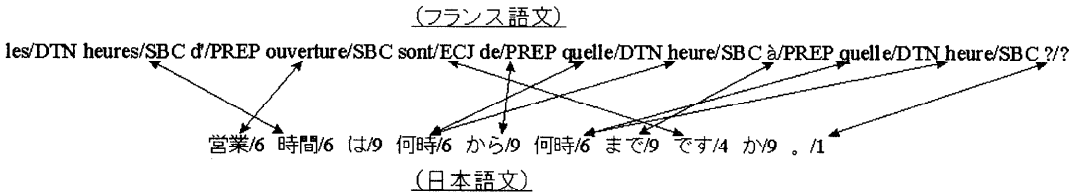

図 8 結果の例
Fig. 8 Example of result.

表 3 品詞情報の必要性
Table 3 Usefulness of the part-of-speech tags.

	すべての対応関係	正対応関係	適合率
品詞情報から	28.7%	21.2%	62.9%
字面から	71.3%	78.8%	94.5%
合計 / 平均	100.0%	100.0%	85.4%

フィードバック学習を行わない場合も実験し，フィードバックの効果も考察した．適合率を**図 6** に，再現率を**図 7** に示す．

次に品詞情報が本手法においてどう役立ったかを考察する．品詞情報によって推定された対応関係と字面一致で抽出された対応関係の数を数えた．その結果を**表 3** に示す．品詞情報によって推定された対応関係はすべての結果の 28.7%であるが，正対応関係中では 21.2%に下がった．品詞情報の利用の適合率は 62.9%である．

7. 考　　察

図 6 より，適合率は上下に変動するが全体的に翻訳例の増加にともなって少しずつ上昇して行く．1,000 の翻訳例からつねに 80%以上の適合率が得られた．また，2,600 までの翻訳例では適合率は最大 90.9%である．フィードバック処理がある場合の方がない場合よりも上回っていることより，フィードバック学習処理の有効性を示すことができた．一方，図 7 より，翻訳例が増えれば増えるほど再現率が増加していくことがはっきり確認できた．フィードバック処理の有無の影響は再現率についてはあまりなかったが，適合率への影響は大きかった．精度の高い翻訳パターンを少数でも取り出すことが翻訳システムにとってより重要であると考えられるので，フィードバックは本手法にとって有効である．再現率は 2,600 の翻訳例で最大の 80.7%が得られた．これは本手法が有効であることを示している．翻訳例の数の少なさ，そして非文が多い会話文ということを考えると比較的良い結果が得られたと考えられる．2 つのグラフを同時に見ると，再現率の上昇と高い値を示している適合率より，本手法の有効性を確認することができた．

実験結果の例を図 **8** に示す．**"quelle heure"** と **"何時"** の対応関係のように一対多や多対多の対応関係がうまく決定されている．論理的に **"何"** は **"quelle"** と，そして **"時"** は **"heure"** と一致する．しかし，**"何時"** が 1 つの形態素になったことで **"quelle"** と **"何時"** そして **"heure"** と **"何時"** という対応関係になった．類似文の部分がすでに翻訳例の中に存在することから，このような一対多や多対多の対応関係を推定することが可能である．

さらに，本手法は文の中に複数現れている同じ形態素のそれぞれの対応関係を決定することができた．図 8 では，**"quelle heure"** と **"何時"** はともに 2 度出現しているが，それぞれの対応関係を正しく決定できた．それが可能になったのは単語ではなく部分を見ることにより隣の単語などを考慮しているからである．単語単位で考えている統計的な手法ではこのような処理は行えない．

決定されたすべての対応関係の 28.7%は品詞情報によって決定された．これは品詞情報の必要性を示している．対応関係が推定できた未登録語のうちの正しいものは 62.9%であった．品詞情報を利用しなければ再現率が 70–80%から 50–55% に減少する．**図 9** に品詞情報の有効性を示す．図 9 で **"ouverture/SBC"** という形態素は品詞によって **"vol/SBC"** と一致した．同様に **"営業/6"** は **"飛行/6"** と一致した．したがって，**"vol/SBC"** と **"飛行/6"** の対応関係によって **"ouverture/SBC"** が **"営業/6"** と対応することを推定できた．選択された部分の中に出ているそれらの対応している部分は字面で一致しないにもかかわらず品詞情報を観察することによって正しい対応関係に決定できた．つまり，本手法では未登録の単語が新翻訳例に存在しても対応関係を決定することができる．適合率は字面で一致する対応関係の推定より低いが，品詞情報は未登録語の対応関係の推定のために必要な情報と考えられる．

適合率と再現率の図は大きく上下に変動しているように見えるが，すべての文を一度にランダムに入力したのではなく，会話本から取った一部分をランダムに

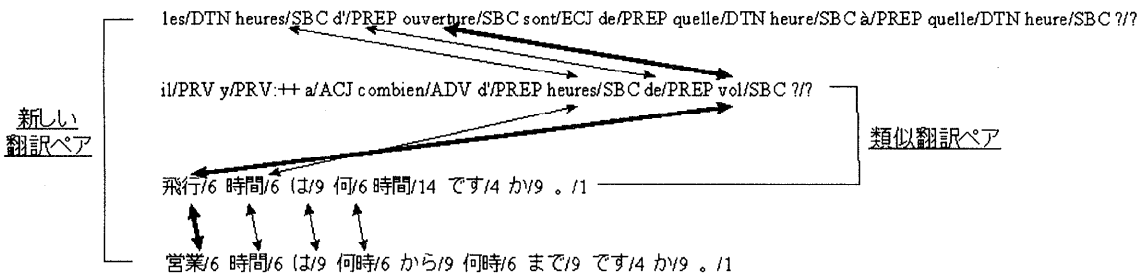

図9 品詞情報によって推定できた対応関係
Fig. 9 Link which was predicted from part-of-speech tags.

少しずつシステムに入力した．したがって，会話本の出現順の影響が少し存在すると考えられる．しかし，値が揺れてもほとんど10%以下である．文の多様性を考えるとその程度の変化はもともと存在すると考えられる．

類似部分の抽出結果は両言語の品詞の数に影響を受ける．品詞の数が少なければ長い類似部分を見つけることができる．しかし，少なすぎると高い回数で出現する品詞が誤抽出を引き起こす．一方，品詞の数が多いと品詞で一致する部分が短く，未登録語の対応関係を探すのが困難になる．それは推定できなかった対応関係の原因の1つであった．フランス語の40の品詞に対し日本語の14の品詞は少ないと考えられ，バランスを考慮する必要がある．本実験では表記の標準化による文の意味の変更の可能性を考えて，使用した形態素解析結果は文字区切りと品詞付与にとどめた．表記の標準化が結果にどんな影響を与えるかは今後の課題である．

もう1つの誤りの原因は類似部分の検索のときに対応関係のないものが出現することである．類似文の中に対応関係のない形態素があると再現率が低下する．一方，一対多や多対多の対応関係では，同じ翻訳例を利用して推定した場合，その対応関係を採用するが，異なる翻訳例を利用していた場合には採用しない．それらの問題は翻訳例の増加につれて少しずつ解決されていくが，手法に，類似文で対応関係がない形態素の対応関係の推定や異なる類似翻訳例を利用した一対多や多対多の対応関係の決定手法について研究する必要があると考えられる．

8. むすび

本論文では類推を用いた対応関係の推定手法について述べた．実例に基づく機械翻訳手法では大量の対応関係付きコーパスが必要である．ところが，十分に研究されていない言語間ではこの大量の対応関係付きコーパスが存在しない．一方，統計を用いた対応関係推定手法では，大量なコーパスがないと良質な結果は得られない．そこで，本論文では，対応関係が決定されている比較的少量の既存の翻訳例を用いて，新しい翻訳例の対応関係を推定し，対応関係付きの翻訳例を大量に増加させる手法を提案した．ここで，新しい翻訳例の対応関係を推定するときに類似文の対応関係を利用している．そうすることによって一対多や多対多の対応関係も1文に同じ単語が複数回現れる場合の対応関係を抽出できる．さらに，フィードバック学習処理の導入によって，失敗を繰り返すのを避けることもできる．

実験結果から抽出された対応関係の数が翻訳例の増加にともなって増加していくことが確認できた．さらに，適合率も徐々に増加していくことが認められた．2,600組の翻訳例において90%以上の適合率と80%以上の再現率が得られた．使用した会話文に非文の多いことを考えると結果は良好であると考えられる．一方，フィードバック学習処理を行った場合は行わない場合より良い結果が得られることからフィードバック学習処理の有効性を確認できた．さらに，一対多や多対多の対応関係や1文に数回出現する同じ単語の対応関係を抽出できることや品詞情報の必要性を確認できた．

今後の課題としては，両言語の品詞の数のバランスや類似文で対応関係がない形態素の対応関係の推定や異なる類似翻訳例を利用した一対多や多対多の対応関係の決定方法を考えられる．

謝辞 本研究の一部は文部省科学研究費補助金（第10680367号および第09878070号）によって行われている．フランス語の形態素解析を提供した“Institut National de la Langue Française”に，謹んで感謝の意を表する．

参考文献

1) Brown, P.F., Della Pietra, S.A., Della Pietra, V.J. and Mercer R.L.: The Mathematics of Statistical Machine Translation: Parameter Es-

timation, *Computational Linguistics*, Vol.19, No.2, pp.263–309 (1993).
2) Melamed, D.: A Word-to-Word Model of Translational Equivalence, *35th Conference of the Association for Computational Linguistics (ACL'97)*, Spain (1997).
3) 北村美穂子，松本裕治：対話コーパスを利用した対話表現の自動抽出，情報処理学会論文誌，Vol.38, No.4, pp.727–736 (1997).
4) Andriamanankasina, T., Araki, K., Miyanaga, Y. and Tochinai, K.: Method for Searching the Best-Matching Sentence in Example-Based Machine Translation, Technical Report of IEICE, Vol.NLC97-10, pp.15–20 (1997).
5) Andriamanankasina, T., Araki, K., Miyanaga, Y. and Tochinai, K.: Machine Translation Based on the Relations between Words, *Proc. NL Symposium Towards Useful Natural Language Procesing*, Japan (1997).
6) Yamashita, T.: ChaSen Technical Report, Nara Advanced Institute of Science and Technology (1996).
7) 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，電子情報通信学会論文誌，Vol.J79-D-II, No.4, pp.391–402 (1996).
8) Meguro, S.: *Manuel de Conversation Francaise*, Hakusuisha, Tokyo (1987).
9) Sato, F.: *Locutions de base*, Hakusuisha, Tokyo (1990).

(平成 10 年 10 月 12 日受付)
(平成 11 年 5 月 7 日採録)

タンテリ アンドリアマナカシナ（学生会員）

昭和 43 年生．平成 2 年マダガスカル情報工学大学卒業．平成 8 年小樽商科大学大学院商学研究科修士課程修了．現在，北海道大学大学院工学研究科博士後期課程在学中．自然言語処理，機械翻訳の研究に従事．

荒木 健治（正会員）

昭和 57 年北海道大学工学部電子工学科卒業．昭和 63 年同大大学院博士課程修了．同年，北海学園大学工学部電子情報工学助手．平成元年同講師．平成 3 年同助教授．現在，北海道大学工学部電子情報工学専攻助教授．学習を用いた自然言語処理の研究に従事．日本認知科学会，人工知能学会，言語処理学会，IEEE，ACL，AAAAI 各会員．

栃内 香次（正会員）

昭和 37 年北海道大学工学部電気工学科卒業．昭和 39 年同大大学院工学研究科電子工学専攻修士課程修了．現在，同大学工学部電子情報工学専攻教授．主として音声情報処理，自然言語処理の研究に従事．工学博士．電子通信学会，日本音響学会各会員．