自然言語処理 154－9
（2003．3．6）

# 日本語－ウイグル語対訳辞書拡充のための日本語言い換え処理

釜谷 聡史† 小川 泰弘† ムフタル・マフスット‡ 稲垣 康善†

†名古屋大学大学院工学研究科 ‡名古屋大学大学院国際開発研究科
kamatani@inagaki.nuie.nagoya-u.ac.jp

**概要**

本稿では、対訳辞書における未登録語を登録単語に言い換えることにより、辞書を拡充する手法を提案する。本手法では、言い換え処理を、収集段階と選択段階の二段階に分割する。収集段階では、国語辞書の語義文を用いて、言い換え候補を収集する。選択段階では、語構成の見地から、言い換え前後の語句間の類似度を計算し、意味的に適切な候補を選択する。また、本手法により、実際に日本語–ウイグル語対訳辞書を拡充する実験を行った結果、未登録語300単語中68.3%が、対訳として用いることの可能な言換えを獲得できた。

# Paraphrasing Japanese Words to Expand a Japanese–Uighur Bilingual Dictionary

Satoshi KAMATANI†, Yasuhiro OGAWA†, Muhtar Mahsut‡ and Yasuyoshi INAGAKI†

†Graduate School of Engineering, Nagoya University
‡Graduate School of International Development, Nagoya University
kamatani@inagaki.nuie.nagoya-u.ac.jp

**ABSTRACT**

In machine translations, the number of words in bilingual dictionary has an important influence on the translation. However, a cost of a dictionary development is very expensive. In this paper, we resolve this problem by paraphrasing a non-entry word into the entry words. We divide the paraphrasing process into two steps: collecting and screening. In the collecting step ,we make paraphrasing expressions of an original word by using its lexical descriptions in a Japanese monolingual dictionary. In the following screening step , we calculate the similarity between the original word and each of its paraphrasing expressions, and choose the best one. We applied this method to our Japanese-Uighur bilingual dictionary. As a result , for 68.3% of non-entry words, the appropriate Uighur words were given.

## 1 はじめに

グローバル化の進展や、インターネットの普及などにより、多言語間の機械翻訳に対する要求が高まっている。これに伴い、日本語・英語などの代表的な言語間で盛んに機械翻訳に関する研究が進められており、ある程度実用的なシステムが構築されている。反面、研究の対象となっていない言語や、研究が始まっていても、萌芽段階にある言語が多くある。このような言語では、翻訳知識を得るための言語データが十分蓄積されておらず、また、翻訳の要である対訳辞書の整備が進んでいないことが多い。特に、対訳辞書の規模は、システムが扱える文の数に大きな影響を与えるにもかかわらず、コストの高さから、容易に拡充できないという問題がある。例えば、我々が開発を進めている日本語–ウイグル語機械翻訳システム[1][2]においても、元となった日本語辞書の登録単語数が35万語であるのに対して、日本語–ウイグル語対訳辞書は3.6万語しかない。

ここで、人手による翻訳を考える。人間が翻訳作業をする場合、自身が知らない単語に出会うと対訳辞書を引く。このとき、その単語についての記載が対訳辞書になかった場合、人間は、これを解決するために、次の二つの手段をとると考えられる。一つは、他の人に助言を求める方法であり、一つは、翻訳できない単語を、意味の同じ表現に言い換え、改めて辞書をひく方法である。本研究では、人間の後者の行動を真似ることで、すなわち、未登録語を登録語に言い換えることにより、その訳語を獲得する。これにより、二言語間の言語知識が必要な問題を、一言語内で扱える問題にすることができる。

そこで、本稿では、日本語–ウイグル語対訳辞書の拡充を目的として、対訳辞書における未登録語を言い換える手法を提案する。このとき、翻訳に適用することを考え、言い換えの前後で、できるだけ意味の過不足のない表現を獲得する。

## 2 言い換え獲得の枠組み

言い換え処理は、次の二段階に分割できる [3]。それは、可能な限り言い換えの候補を集める『収集段階』と、その候補の内、言い換えとして不十分、あるいは、不適切なものを削除する『選択段階』である (図 1)。

このように分割することで、『収集段階』では、言い換え表現の多様性を重視した再現率の高い収集をし、『選択段階』では、言い換えとして不適格なものを削除する精度の高い篩分けをすることが可能となり、多様性と質のバランスの取れた言い換えが獲得できる。さらに、各段階を独立させて扱うことができるため、そこに適用する手法も柔軟に変更が可能である。

本手法では、収集段階において、見出し語と意味的に等価な関係にある国語辞書の語義文を利用して、言い換えの候補を収集する。さらに、選択段階において、単語間の類似度を用いて、意味情報を必要十分に保存している言い換えを選択する。

## 3 言い換え先候補の収集

辞書の語義文は、見出し語の語義を説明したものであると同時に、意味を十分保存した言い換えであると見ることができる。ただし、その記述の目的は、あくまでも語義の説明であるために、冗長な表現となっている場合も多い。そのため、語義文における、見出し語の同等句のみを取り出して、言い換えとするのが望ましいが、その同等句を直接決定することは難しい。そこで、言い換えの収集段階では、語義文を構成している文節を単位として、その全ての組み合わせを考え、言い換え先候補とする。

具体的には、以下の手順で言い換え先候補を収集する。

1. 語義文を、日本語形態素解析ツール JUMAN[4] によって形態素解析し、それを、日本語構文解析ツール KNP[5] によって係り受け解析する。これによって、文節情報と係り受け情報を得る。
2. 言い換え元が用言の場合は、獲得すべき言い換え先（候補）もまた、用言で終わるべきであることから、語義文末に「こと」「さま」がある場合、これを削除する。
3. 語義文の末尾の文節（ただし、「こと」「さま」は、これを除く）を言い換えの中心とし、言い換え先候補を作成するときに、必ず含める単語とする。これは、語義文末尾の文節が、その見出し語を説明する上で最も重要な役割をしているという傾向に基づいている。
4. 語義文を構成する文節の全ての組み合わせの内、元の語義文中の係り受け関係を崩さないものを言い換え先候補とする。

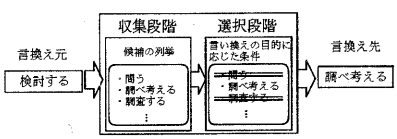


図 1: 言い換え処理モデル

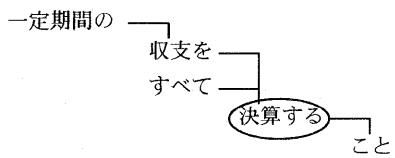


図 2: 「総決算する」の語義の解析結果

この言い換え先候補収集の手順を、言い換え元「総決算する／語義：一定期間の収支をすべて決算すること」を例として説明する。

まず、手順 1. において、KNP で係り受け解析がなされ、図 2 のように、文節情報、係り受け情報が得られる。さらに、言い換え元「総決算する」は用言であることから、手順 2. において、文末の “こと” が削除される。結果、言い換え先候補の収集に利用される文節は、“一定期間の”、“収支を”、“すべて”、“決算する”、となる。手順 3. で、言い換えの中心となる単語を “決算する” とし、手順 4. で、“決算する” を必ず含む、全ての文節の組み合わせが作られる。ただし、元の係り受け関係が崩れてしまうような組み合わせである「一定期間の決算する」などは、候補から外す。以上により得られる言い換え先候補を、以下に示す。

1. 決算する
2. すべて決算する
3. 収支を決算する
4. 収支をすべて決算する
5. 一定期間の収支を決算する
6. 一定期間の収支をすべて決算する

# 4 言い換えの選択

## 4.1 語の構成に基づく意味推定

3節で収集した言い換え先候補から、言い換えとして意味的に過不足のない、言い換え元の同等句を選び出す。このためには、どの候補が最も適切な言い換えであるか、その指標を定めなければならない。

例として、言い換え元「点火する」の言い換え先候補として、次の3つが挙がっている場合を考える。

1. 灯す
2. 火を灯す
3. 物に火を灯す

これらの中で言い換え元の同等句である、と判断できるのは2.の「火を灯す」である。

日本語の単語の意味は、そこに含まれる部品の意味から構成されている場合が多い。「点火する」の例では、漢字“点”の意味「点ける」と、漢字“火”の意味「火」から「火を灯す」という語義が構成されているといえる。このように、その単語の意味を構成している意味の部品とも言うべきものを、本研究では**意味因子**と呼ぶことにする。

この考え方を導入すれば、本研究における適切な言い換えとは、次の二点を満たすものである。

- 言い換えの効率性 (efficiency)

  意味因子をなるべく少ない単語数で表現し、冗長な、すなわち類似度の低い単語を含まないほど、効率の良い言い換えである。

- 言い換えの充足性 (sufficiency)

  言い換え元の単語が持っている、全ての意味因子について、対応する語が言い換え先に含まれていれば、元の意味を充足する言い換えである。

以下では、まず、意味因子について述べ、次に、効率性と充足性の評価の際に用いる、単語間類似度の計算手法について述べる。その後、推定した意味因子と、類似度計算手法を用いて、言い換えの効率性と充足性、そして、これら二つの指標をもとにした、言い換えの妥当性の評価手法を提案する。

## 4.2 意味因子

意味因子を利用して、言い換えの適切さの指標を定めるには、言い換え元の表記から、それが持っている可能性のある意味因子を推定する必要がある。本研究では、以下の三つを意味因子候補とする。

- **[漢字]** 言い換え元を構成している漢字一字ごとの意味

  漢字は、表音文字であると同時に、表意文字であり、その一字一字に意味がある。また、この漢字によって構成された単語は、その漢字と関係のある語義を内包していると考えられる。前節で述べたように、「点火する」は“点”の意味と“火”の意味の合成から成っていると考えられ、これを最小の意味因子としてみることができる。また、漢字の意味は、5.1節で作成する漢字意味辞書に基づき、そこに記載された語義を意味因子として捉える。

- **[部分]** 言い換え元の意味のある部分 (単語)

  言い換え元に単語として見ることができる部分が含まれている場合、これを意味因子として考えることができる。例えば、「一括払い」という単語について考えてみると、「一括」という単語と「払い」という単語の複合語となっていることが分かる。このように、単語として認識できる語が含まれていれば、その言い換え元は、その単語の意味を内包していると考えるべきであり、これは意味因子であるといえる。また、意味因子として用いるか否かは、その部分がEDR日本語単語辞書 [7] に登録されているかで判定する。

- **[全体]** 言い換え元そのもの

  言い換え元自身が持っている意味も、また、意味の部品であるといえる。例えば、「右往左往する」という単語は、単純な漢字の語義の組み合わせで語義が構成されているのではなく、組み合わせたことによって、新たに「混乱する」といった意味が生まれたと考えられる。

表 1: 「総決算する」の意味因子候補

| 観点 | 意味因子 |
|---|---|
| [漢字] | “総”の語義，“決”の語義，“算”の語義 |
| [部分] | 総，決，決算，決算する，算，算する |
| [全体] | 総決算する |

実際に意味因子となるのは、これら三つの内の一つ、あるいは、その組み合わせである。例えば、言い換え元「総決算する」について、意味因子候補を列挙した例を、表1に示す。言い換え「総決算する → 収支をすべて決算する」においては、実際に意味因子となるのは、[部分]“決算する”と、[漢字]“総の語義”である。その求め方については、次節以降で説明する。

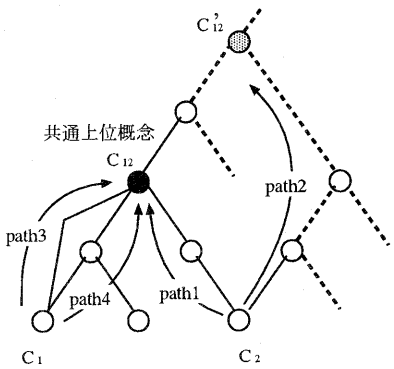


図 3: EDR 概念体系

## 4.3 類似度計算手法

### 4.3.1 単語間類似度

実際に使われている意味因子候補を選ぶために、言い換え先との類似度を計算する。本研究では、類似度計算手法として、長尾 [6] によって紹介された、シソーラス上の概念間距離から類似度を計算する、式 (1) を用いて、概念間の類似度 $SIM_c(c_1, c_2)$ を計算した。

$$SIM_c(c_1, c_2) = \frac{2 \times \mathrm{depth}(c_{12})}{\mathrm{depth}(c_1) + \mathrm{depth}(c_2)} \quad (1)$$

$c_1, c_2$ : 概念
$c_{12}$ : $c_1, c_2$ の共通上位概念
$\mathrm{depth}(c)$ : 概念 $c$ の根からの深さ

また、本研究では、シソーラスとして EDR 概念体系辞書を用いる。しかし、この辞書では、上位概念の多重継承を許し、一つの概念が二つ以上の上位概念を持つ場合もある。ゆえに、一般のシソーラスなどの概念分類とは異なり、概念全体を完全な木構造の形に体系化していない。このため、図 3 のように、概念の深さが経路によって複数考えられ、類似度も複数計算できる。例えば、図 3 では、共通上位概念として $C_{12}, C'_{12}$ の二つが考えられ、また、$C_1$ の深さを測る経路も二つある。そこで、計算される類似度が最も高くなるように、すなわち、図 3 の例では、共通上位概念を $C_{12}$ に定め、$C_1$ の深さは path3 の経路で測る。

また、EDR 日本語辞書において、多義語には複数の概念が付与されている。このため、一つの単語の組について、複数の類似度 $SIM_c(c_1, c_2)$ が計算される場合がある。本研究では、類似度 $SIM_c(c_1, c_2)$ の最も高い概念のペアを選び、単語間の概念の類似度 $SIM_w(w_1, w_2)$ を式 (2) のように計算する。ここで、$0 \leq SIM_w \leq 1$ である。

$$SIM_w(w_1, w_2) = \max_{i,j} SIM_c(c_{w_1 i}, c_{w_2 j}) \quad (2)$$

$w_1, w_2$ : 単語
$c_{w_1 i}, c_{w_2 j}$ : $w_1, w_2$ の概念 (の一つ)

### 4.3.2 意味因子候補と言い換え先候補の類似度

意味因子の各候補と言い換え先候補間の類似度 $GM$ を、次式で計算する。

$$GM = \sqrt[n]{\prod_i^n \max_j SIM_w(cm_i, pm_j)} \quad (3)$$

$cm_i$ : 意味因子を構成する $i$ 番目の自立語
$pm_j$ : 言い換え先候補中の $j$ 番目の自立語
$n$ : 意味因子を構成する自立語数

例えば、「総決算する → 収支をすべて決算する」について計算したものを表 2 に示す。

表 2 において、意味因子「総決算する [全体]」に着目すると、言い換え先の中の自立語について、その類似度類 $SIM_w$ は、「収支」との間で 0、「すべて」との間で 0、「決算する」との間で 0.2 である、と計算され、式 (3) によって、$GM = 0.2$ と計算されたことを示している。また、意味因子「糸を束ねる [漢字 "総"]」に着目すると、この意味因子は、二つは自立語から構成されている。この場合は、まず、その意味因子中の各自立語について、言い換え先中の自立語との間で、先程と同様に $SIM_w$ を計算する。次いで、式 (3) により、「糸」と「収支」、「束ねる」と「決算する」の間の $SIM_w$ が相乗平均され、$GM = \sqrt{0.29 \times 0.43} = 0.35$ が計算されている。

以上のようにして、全ての意味因子候補と言い換え先の間で類似度が計算する。そして、言い換え先の各自立語において、類似度 $SIM_w$ が 0 よりも大きく、かつ、$GM$ の値が最も大きなものを、その自立語と最も関連がある意味因子とみなす。表 2 の例では、「収支」は「決算 [部分]」が、「すべて」は「総 [漢字]」が、「決算する」は「決算する [部分]」がそれぞれ関連する意味因子としてみなされる。

## 4.4 言い換えの効率性

言い換え先が、意味因子と関連性の高い、すなわち、類似した語のみで構成されているほど、良い言

い換えであるといえる。この指標とするために、前節で定めた関連する意味因子と言い換え先の自立語間の類似度を、各自立語の**有用度** $e$ として与える。表2の例において、有用度は、表3のように計算できる。各自立語に対して計算された有用度 $e$ の相乗平均をとり、言い換えの効率 $(Eff)$ とする。

$$Eff = \sqrt[n]{\prod_{i}^{n} e_i} \quad (4)$$

$e_i$ : 言い換え先候補中 $i$ 番目の自立語の有用度

$n$ : 言い換え先候補を構成する自立語数

$Eff$ は、言い換え先が有用度の高い語のみで構成されていれば1に、有用度の低い語を含むほど0に近い値をとる。表3の例の効率は、$Eff = \sqrt[3]{0.88 \times 1.00 \times 1.00} = 0.96$ と計算される。

### 4.5 言い換えの充足性

言い換え先で、言い換え元が持っていた意味因子が、全て反映されていることを確認する。4.1節で述べた通り、単語の意味は、漢字を基本として構成されているといえる。よって、言い換え元を構成している漢字を、全て被覆するような意味因子の組み合わせがあればよい。そこで、漢字一字ごとに、その意味が言い換え先においてどれだけ反映されているかを評価する。言い換え先での意味因子の反映の度合い、すなわち、反映度 $s$ は、4.3節で計算した類似度によって表すことができる。つまり、意味因子と類似度の高い語が言い換え先に含まれていれば、その類似度分だけ、反映されていると捉える。

以上により、言い換え元中の各漢字に対して反映度を与え、次式により言い換えの充足率 $(Suf)$ を計算する。

$$Suf = \sqrt[n]{\prod_{k}^{n} s_k} \quad (5)$$

$s_k$ : 言い換え元の $k$ 番目の漢字の反映度

$n$ : 言い換え元を構成する漢字数

これを「総決算する → 収支を全て決算する」について計算したものを表4に示す。

表 2: 「総決算する」における類似度計算

| 意味因子 | 言い換え先の自立語 | | | GM |
|---|---|---|---|---|
| | 収支 | すべて | 決算する | |
| 総決算する [全体] | 0 | 0 | 0.2 | 0.2 |
| 決算する [部分] | 0 | 0 | 1.00 | 1.00 |
| 決算 [部分] | 0.88 | 0 | 0 | 0.88 |
| 算する [部分] | 0 | 0 | 0.66 | 0.66 |
| ⋮ | | | | |
| すべて [漢字 "総"] | 0 | 1.00 | 0 | 1.00 |
| 糸 (を) | 0.29 | 0.18 | 0 | |
| 束ねる [漢字 "総"] | 0 | 0 | 0.43 | 0.35 |

表 3: 「収支をすべて決算する」の効率性

| 自立語 | 意味因子 | 有用度 $e_i$ |
|---|---|---|
| 収支 | 決算 [部分] | 0.88 |
| すべて | 総 [漢字] | 1.00 |
| 決算 | 決算する [部分] | 1.00 |
| 効率 $Eff$ | | 0.96 |

表 4: 「収支をすべて決算する」の充足性

| 漢字 | 意味因子 | 反映度 $s_i$ |
|---|---|---|
| 総 | 総 [漢字] | 1.00 |
| 決 | 決算する [部分] | 1.00 |
| 算 | 決算する [部分] | 1.00 |
| 充足率 $Suf$ | | 1.00 |

### 4.6 言い換えの妥当性

本手法における、妥当な言い換えとは、言い換えの効率性 (efficiency) と言い換えの充足性 (sufficiency) の両方が満たされているものである。よって、効率 $Eff$ と充足率 $Suf$ の二つの値の相乗平均を、言い換えの妥当性 (validity) $V$ とする。

$$V = \sqrt{Eff \times Suf} \quad (6)$$

「総決算する → 収支を全て決算する」の例では、表3,4より、$V = \sqrt{0.96 \times 1.00} = 0.98$ と計算でき、妥当な言い換えがなされていると判定できる。

## 5 実験

提案した言い換え獲得手法に基づいて、日本語–ウイグル語対訳辞書における未登録を言い換え、これ

を翻訳することで、対訳語を獲得する実験を行った。

### 5.1 漢字意味辞書の作成

本手法では、4.2節で述べたように、漢字の語義を利用するため、漢字の語義を記述した漢字意味辞書が必要になる。本実験では、広辞苑第四版 (CD-ROM版)[8]から漢字一文字の意味を収集、整理し、これを漢字意味辞書とした。辞書構築の手順を以下に示す。

1. 漢字ごとに、その読みを区別しないで、広辞苑中の語義文を取り出す。
2. 意味番号、出展、用例など、辞書特有の付記情報を削除する。
3. 句点ごとに語義文を分解し、それぞれを一つの語義とする。
   例:“思慮。おもわく。”→ “思慮”+“おもわく”
4. 辞書特有の表現である、文末の “さま”,“こと” を削除する
5. “もの”,“人” 及び、それに係る語を削除する。これらの語は、一般的過ぎる語であり、また、説明のために補助的に用いられていること多いことから、これを削除した。
6. 人手によって、不自然な部分を修正する。
   ここで、不自然な部分とは、削除によって生じた、文としての不自然さのことであり、語義に関わる部分には、人手を加えなかった。
7. 語義一つ一つについて、日本語形態素解析ツール JUMAN[4] によって形態素解析し、語義を構成している自立語と、その品詞を得る。

以上の作業によって、漢字意味辞書を作成した。この辞書の登録漢字数は605字であり、総意味数は10,502である。また、漢字一字あたりの最多意味数は114であり、同最少意味数は1である。

### 5.2 言い換え先候補の収集とその選択

実験対象となる言い換え元を、EDR日本語単語辞書から以下の条件に従って取り出した。

1. 日本語–ウイグル語対訳辞書における未登録単語
2. 5.1節で作成した漢字意味辞書内の漢字のみで構成されている単語
3. EDR日本語コーパス[7]中の出現頻度が、各品詞の上位1,000単語に入っている単語

結果、名詞473単語、動詞514単語、サ変名詞429単語が実験対象として取り出された。ただし、EDR日本語単語辞書では、表記が同じ単語でも、概念が異なれば別エントリとして登録されている。よって、取り出される実験対象も、概念の異なり数に基づいて区別し、単語数を集計している。

次に、各言い換え元のEDR日本語単語辞書に記載されている概念説明を語義文とみなし、3節で述べた手法により、言い換え先候補を収集した。その結果、表5のように言い換え先候補が収集された。

収集した言い換え先候補の内、4.6節に示した手法で計算した、言い換えの妥当性 $V$ が、最も高いものを言い換え先として選択した。ただし、同じ言い換え元において、妥当性 $V$ の値が等しいものは、言い換え先として同等に適切であると判定する。このときには、一つの言い換え元に対して、複数の言い換え先を与える。

### 5.3 日本語言い換えとしての評価

得られた言い換えのうち、各品詞300個をランダムに抜き出し、日本語言い換えとして直感的に適切であるか、人手で評価した。その際、以下の観点で、言い換えの成否を判定した。

まず、言い換え成功としたものは、次の2つの項目に含まれるものである。

- **妥当**
  以下に示す問題が見当たらず、適切に言い換えられていると判断できるもの。
- **文脈**
  言い換えとしての成否が、文脈に大きく依存すると考えられる場合、言い換え自体は成功していると判定する。
  例えば、「参画する → 計画に加わる」という言い換えについて見ると、「経営に参画し てい

表5: 言い換え先候補

| 品詞 | 言い換え元 | 候補総数 | 候補数 | |
|---|---|---|---|---|
| | | | 最多 | 最少 |
| 名詞 | 473 | 2,897 | 129 | 1 |
| 動詞 | 514 | 2,541 | 82 | 1 |
| サ変 | 429 | 2,087 | 55 | 1 |

る」という文脈においては、適用することができない。しかし、文脈次第では言い換えが可能となるため、言い換えは成功していると見る。

言い換え失敗としたものは、次の項目に属するものである。

- **意味過分**

  意味的に過分のある言い換え先が得られたり、言い換え元の意味因子と対応するとは考え難い語が言い換え先に含まれている場合。

  例えば、「本土 → その国の中心をなす国土」という言い換えなどが、これに分類される。

- **意味欠損**

  言い換え元の意味情報が言い換え先において欠落した場合。

  例えば、「苦戦する → 戦いをする」という言い換えでは、言い換え元における "苦しい" という意味が落ちていると考えられる。

- **国語辞書の不十分な記述**

  本実験は、一般の国語辞書の語義文ではなく、EDR 日本語単語辞書の概念説明を利用している。概念説明は、ある概念と他の概念との識別を、人間が行い易くするために付加されたものであり、完全な語義説明文となっているわけではない。そのため、本手法によって、言い換えが得られない場合がある。

  例えば、「会話する → 会話する」という言い換えは、語義が見出し語自身を用いて説明されていたために得られた。このように、言い換え先と言い換え元が同じなど、辞書固有の問題として扱えるものは、これに分類する。

- **その他**

  上記以外の理由で、正しく言い換えられなかったと判定される場合。

  「落ちつく → なる」という言い換えは、言い換え元の語義「心が安定した状態になる」から得たものである。本手法は、言い換え元の意味因子と関連する自立語を、係り受けを保ったまま取り出す。よって、「状態」という語と類似する意味因子がなかったために、「安定した」という文節が抜き出せなかった。

以上の判断基準で、評価した結果を、表 6 に示す。

## 5.4 辞書拡充

5.3 節で取り出した、各品詞 300 個の言い換えについて、その言い換え先を日本語-ウイグル語翻訳システム [1] によって翻訳した。翻訳の成否に関する結果を、表 8 に示す。ここで、翻訳成功としたものは、言い換え先が全てウイグル語に翻訳されたものである。解析失敗としたものは、日本語-ウイグル語翻訳システム [1] が入力文の解析に失敗し、出力を得られなかったものである。

さらに、翻訳した結果が言い換え元の訳語として利用可能か、ウイグル語ネイティブによって、人手によって評価した。評価対象は、それぞれの品詞で翻訳が成功したものうち、各 100 個をランダムに取り出したものである。結果を表 7 に示す。表中で、"条件付可能" としたものは、文脈によっては利用可能であると判定されたものである。

## 5.5 考察

今回提案した言い換え手法は用言向きの手法であるといえる。体言の国語辞書の語義文は、その見出し語がどのようなものであるか、言葉を尽くして説明してある傾向が強い。例えば、「売り場」に対する語義は、「物を売る一定の場所」となっており、同等句を取り出しにくい。このように、体言に向かない傾向は、表 6 によく現れている。逆に、用言の言い換えでは、「一体化する」に対して「まとめる」のように、比較的易しい言葉で言い換えられることから、翻訳の目的に照らしても、有効な言い換えが得られている。

また、表 7 から、言い換えた結果が、おおむね翻訳可能であることが分かる。このことから、未登録語を言い換えることの有効性が示せた。

表 6: 日本語の言い換えとしての評価

| 言い換え評価 | | 名詞 | 動詞 | サ変 |
|---|---|---|---|---|
| 言い換え成功 | 妥当 | 80 | 144 | 146 |
| | 文脈 | 43 | 38 | 47 |
| 言い換え失敗 | 意味過分 | 23 | 21 | 22 |
| | 意味欠損 | 37 | 31 | 18 |
| | その他 | 85 | 29 | 10 |
| | 国語辞書 | 32 | 37 | 57 |
| 計 | | 300 | 300 | 300 |

表 7: 翻訳結果

| 品詞 | 翻訳 | | 解析失敗 | 計 |
|---|---|---|---|---|
| | 成功 | 失敗 | | |
| 名詞 | 245 | 48 | 7 | 300 |
| 動詞 | 273 | 16 | 11 | 300 |
| サ変名詞 | 221 | 68 | 11 | 300 |

表 8: 対訳語としての適切さ

| 品詞 | 可能 | 条件付可能 | 不可能 | 計 |
|---|---|---|---|---|
| 名詞 | 37 | 26 | 37 | 100 |
| 動詞 | 41 | 26 | 33 | 100 |
| サ変名詞 | 41 | 34 | 25 | 100 |

次に、本手法を、辞書拡充の面から評価する。表8より、前後の文脈などの条件付きで可能としたものを含めれば、68.3%が対訳として用いることができる、と判定された。このことから、本手法の利用可能性が確認できた。

不可能であると判定されたものについて、日本語の言い換えとしての評価が、どのようになされていたかを表9にまとめる。これによれば、日本語の言い換えとしては適切であるにもかかわらず、翻訳した結果が対訳語として認められなかったものが多くあることが分かる。この原因としては、次の二点が挙げられる。一つは、日本語–ウイグル語機械翻訳システムが解析を誤ったために、正しい訳語が付与できなかったものである。例えば、「言い渡す → 命じる」は、日本語の言い換えとしては妥当であると判定した。しかし、翻訳システムが「命」を名詞として解析したために、正しく翻訳することができなかった。二つ目は、ウイグル語における単語の概念が異なるために、正しく翻訳できなかったものである。例えば、「出国する → 国を出る」は、は、日本語の言い換えとしては妥当であると判定した。しかし、「出国する」は、ウイグル語において「国**から**出る」と表現すべき単語であったために、対訳語として用いることができないと判定された。

表 9: 対訳語として利用不可能な言い換え

| 日本語における評価 | | 名詞 | 動詞 | サ変 |
|---|---|---|---|---|
| 言い換え成功 | 妥当 | 6 | 9 | 8 |
| | 文脈 | 3 | 4 | 6 |
| 言い換え失敗 | 意味過分 | 2 | 8 | 5 |
| | 意味欠損 | 6 | 3 | 0 |
| | その他 | 17 | 8 | 0 |
| | 国語辞書 | 3 | 1 | 6 |
| 計 | | 37 | 33 | 25 |

## 6 おわりに

本稿では、国語辞書の語義文を基に言い換え先の候補を収集しておき、語の構成に基づく意味因子を考えることにより、類似性と意味の等価性を持った言い換え先を選択する手法について述べた。また、この言い換え手法を用いて、日本語–ウイグル語対訳辞書を拡充し、その利用可能性を確認した。

## 参考文献


[1] 小川泰弘, ムフタル・マフスット, 杉野花津江, 外山勝彦, 稲垣康善, “日本語–ウイグル語機械翻訳における派生文法に基づくウイグル語動詞句の生成”, 自然言語処理,Vol.7,No.3,pp. 57-77,(2000).

[2] ムフタル・マフスット, 小川 泰弘, 稲垣 康善, “日本語－ウイグル語機械翻訳のための格助詞の変換処理”, 自然言語処理,vol.8,No.3,pp123–142,(2001).

[3] 釜谷聡史, 小川泰弘, 稲垣康善, “語句間の類似度と語義を組み合わせた言い換え獲得手法”, 電気系学会東海支部連合大会講演論文集,p.244,(2002).

[4] 黒橋 禎夫:日本語形態素解析システム JUMAN version3.61, 京都大学大学院情報学研究科,(1999).

[5] 黒橋 禎夫:日本語構文解析システム KNP version2.0b6, 京都大学大学院情報学研究科,(1998).

[6] 長尾真 ほか編,“自然言語処理”, 岩波講座ソフトウェア科学 15, 岩波書店,(1996).

[7] 日本電子化辞書研究所, ＥＤＲ電子化辞書仕様説明書, (1996).

[8] 新村出編, 広辞苑第四版 EPWING CD-ROM 版, 岩波書店,(1996).