英語を介した日中対訳辞書の自動構築

張 玉潔[*1]　馬 青[*1,*2]　井佐原 均[*1]

日中機械翻訳システム開発の一環として、日中翻訳辞書の自動構築に関する研究を行っている。今まで、機械翻訳の研究は英語を一方の言語とする研究が多く行われてきたため、英語と英語以外の言語の間の対訳辞書は豊富に蓄積されている。一方、英語以外の言語間の対訳辞書はあまり開発されていない。近年、英語以外の言語の電子データは増え続けており、英語以外の言語間の対訳辞書はますます必要になってきている。豊富に蓄積された英語と英語以外の言語の間の対訳辞書を利用して、英語以外の言語間の機械翻訳システムを効率的に開発することが研究の課題になっている。本稿では、日英・英中辞書を利用して日中辞書をゼロから構築する方法とその結果について報告する。英語を介して、新しい言語間の対訳辞書を構築するアプローチは新しいものではないが、大量の候補から正しい訳語を選択するという問題がまだよく解決されていない。本稿では、多数のヒューリスティックな情報を利用して訳語を選択する手法を提案した。提案手法においては、訳語の選別に品詞情報と漢字情報を活用して、中国語の訳語候補を評価するためのスコアリング関数に多数のヒューリスティックな情報を取りいれた。提案手法を用いて、EDR 日英辞書の約 14 万個のレコードに対し、レコードごとに中国語訳語候補の順位付けを行った。その順位付けについて、20 以上の中国語訳語候補を持つ日本語単語を対象とした評価実験を行った。その結果、一位に順位付けた訳語の正解率が 81.4%に達したことが分かった。

キーワード：対訳辞書、日本語、中国語、英語、品詞情報、漢字情報

Automatic Construction of Japanese-Chinese Translation Dictionary Using English as Intermediary

YUJIE ZHANG[*1], QING MA[*1,*2] and HITOSHI ISAHARA[*1]

Electronic translation dictionaries are indispensable language resources in machine translation, cross-language information retrieval, and e-learning. Many electronic dictionaries that translate between English and one other language have been thoroughly developed, while the ones between two languages other than English are still lacking. Because the construction of an electronic translation dictionary is usually expensive and time-consuming, a method must be developed for automatically constructing a translation dictionary for new language pairs by using the existing translation resources between English and other languages. This paper describes our research on

[*1]情報通信研究機構，National Institute of Information and Communications Technology
[*2]龍谷大学，Ryukoku University

constructing a Japanese-Chinese translation dictionary from two existing dictionaries, the EDR Japanese-English and LDC English-Chinese dictionaries, using English as intermediary. In automatic acquisition of bilingual lexicons, one major issue is how to select correct translations from among a large number of candidates. We have developed a method of ranking candidate translations by utilizing three sources of information, the number of English translations in common, the part of speech, and Japanese kanji information. We evaluated the method on 109 Japanese words, each of which has over 20 candidate Chinese translations. The proposed method achieved 81.4% precision.

KeyWords: Translation dictionary, Japanese, Chinese, English, Part of Speech, Kanji

1 はじめに

対訳辞書は機械翻訳、言語横断検索または第二言語の学習において不可欠な言語資源である。人手による対訳辞書の構築はたいへんコストがかかるので、その自動化が自然言語処理の重要な研究課題となる。本研究は、日中対訳辞書の自動構築に関する研究である。

対訳辞書の自動獲得に関してさまざまな研究があり、主に基本対訳辞書がある場合とない場合の二種類に分けられる。基本対訳辞書がある場合、新語と専門用語の訳語の獲得が問題となり、基本対訳辞書と両言語のそれぞれの単言語コーパスを利用する研究がある[Fung 1998]。一方、基本対訳辞書がない場合、対訳辞書をゼロから構築する必要があり、そのために対訳コーパスと非対訳コーパスを利用する研究[Brown 1997; Tanaka and Iwasaki 1996; Fung and Yee 1998]、そして英語を介した研究がある[田中, 梅村, 岩崎 1998; Bond, Yamazaki, Sulong and Okura 2001; Shirai and Yamamoto 2001]。

これまでの機械翻訳の研究は英語を一方の言語とするものがほとんどであった。そのため、英語と英語以外の言語(以降略して英外、または、外英と呼ぶ)の間では、対訳辞書、対訳コーパスなどの言語資源が豊富に蓄積されている。その中で特に対訳辞書は計算機を用いた自然言語処理のために加工されているので、次のような三つの特徴、すなわち、(1)機械翻訳システムの開発に直接利用できること;(2)原言語単語には品詞や意味などの情報が付与されていること;そして、(3)源言語単語をカバーするような規模であることを持っている。日本では、EDR 日英電子辞書[情報通信研究機構 2002]がその代表である。

近年、英語以外の言語の電子データが増え続けており、それらの間の電子対訳辞書がますます必要になってきている。しかし、英語以外の言語間の対訳辞書に関しては上記と同じようなものがまだ開発されていないのが現状である。例えば、日中辞書に関して言えば、上記外英辞書のような開発に用いられるものがなく、その電子化のコストが高い。Yahoo(日本語サイト)のトップページに辞書の項目があり、その中に日本語とインドネシア語、日本語と中国語、日本語とドイツ語の対訳辞書が備えられている。しかし、いずれも検索サービスのためのものであり、機械翻訳システムの開発には直接利用できない。また、紙ベースの辞書とウェッブ上の検索用辞書はいずれもその規模が、英外・外英の対訳辞書のそれよりはるかに小さい。この点については、後述の実験結果により示されている。したがって、いかにすでに豊富に蓄積されている英外・外英の電子辞書を利用して、英語以外の言語間の電子辞書を効率的に開発するかが重要な研究課題になっている[井佐原 2002]。

本稿では、日英・英中辞書を利用して日中辞書をゼロから構築する方法と結果について報告する。英語を介して、新しい言語間の対訳辞書を構築するアプローチは新しいものではない。[田中,梅村,岩崎 1998]では、和仏対訳辞書を、[Bond, Yamazaki, Sulong and Okura 2001]では、日本語とマレー語の対訳辞書を、そして、[Shirai and Yamamoto 2001]では、日本語と韓国語の対訳辞書を作成するにあたって、英語を介したアプローチが用いられた。しかし、このアプローチにおいて、大量の候補から正しい訳語を選択する問題が完全に解決されたわけではなく、その性能をさらに改善する余地が大きく残っている。本稿では、品詞情報と漢字情報を活用して、多数のヒューリスティックな情報を利用して訳語を選択する手法を提案する。それらのヒューリスティックな情報を中国語の訳語候補を評価するためのスコアリング関数に取りいれた。提案手法を用いて、EDR 日英辞書の約 14 万個のレコードに対し、レコードごとに中国語訳語候補の順位付けを行った。その順位付けについて、20 以上の中国語訳語候補を持つ日本語単語を対象とした評価実験を行った。その結果、一位に順位付けた訳語の正解率が 81.4%に達したことが分かった。

2 訳語候補の獲得

本研究では、英語を介して、日本語の中国語訳語候補を獲得するために、EDR 日英辞書[情報通信研究機構 2002]と LDC 英中・中英単語対応表[LDC 2002]を利用することとした。したがって、本稿での訳語候補の獲得とは、EDR 日英辞書の日本語単語に対して、中国語訳語を付与（獲得）することである。

2.1 EDR 日英辞書

EDR 日英辞書には 364,430 個のレコードがある。レコードにはレコード番号、見出し、品詞、概念コード、英語での概念の説明、日本語での概念説明、英訳などの情報が記述されている。同じ見出しでも意味によりいくつかのレコードがあり、また同じ意味でも見出しによりいくつかのレコードがある。英訳には訳語が複数個あり、それぞれの訳語は単語、句、説明文のいずれかの形になっている。そのレコードの一つの例を[2.1-1]に示す。

[2.1-1] JEB0387675　アーミン　JN1　3bf389　“an animal, called ermine”
オコジョという動物　“Mustela ermine <scientific name> |stout |ermine|”

2.2 LDC 英中・中英単語対応表

英中単語対応表には 110,834 個のレコードがある。英中単語対応表のレコードには英単語と中国語訳語のみが記述され、英単語と中国語訳語の品詞情報がない。中国語訳語は主に単語、句である。そのレコードの二つの例を[2.2-1]と[2.2-2]に示す。

[2.2-1] ermine /貂/貂的白毛皮/

[2.2-2] stout /强壮的/坚固的/坚强的/稳重的/勇敢的/激烈的/胖胖的/丰富的/烈啤酒/肥硕的/肥硕/

中英単語対応表には 128,366 個のレコードがある。中英単語対応表のレコードには中国語単語と英訳のみが記述され、中国語単語と英訳の品詞情報がない。英訳は主に単語、句である。そのレコードの一つの例を[2.2-3]に示す。

[2.2-3] 貂 /leopard/panther/pard/

2.3 日中対訳候補の獲得

EDR 日英辞書の各レコードに対して、その日本語単語の英訳を LDC 英中単語対応表の英単語と照合し、照合に成功した英単語の中国語訳語をその日本語単語の中国語訳語の候補とする。EDR 日英辞書の 364,430 個のレコードのうち、中国語訳語の候補が得られたのが約 40%、計 144,002 個のレコードである。残りの 60%は中国語訳語の候補が得られなかったが、そのうちの約 3%は、英訳が LDC 英中単語対応表にないためであり、ほかの約 57%は EDR 日英辞書に単語の形での英訳がないためであった。次にそれぞれのケースについて詳しく説明する。

表 1 得られた中国語訳語候補の例

例	日本語	中国語訳語候補
1	エニシダ	金雀花
2	選び直す	改选, 重选
3	受流す	避开, 使困惑, …
4	夷 1	外国人, 侨民, …
5	夷 2	乡下人, 农民, …
6	足輪	脚镯, 脚镣, …
7	アジ―ル	避难所, 庇护, …

(1) 中国語訳語の候補が得られた 144,002 個のレコードについて

その一部の例を表１に示す。下線は正しい訳語を表す。得られた中国語訳語候補を調べた結果、以下のことが分かった。

(1.1)ほとんどのレコードの中国語訳語候補には正しい訳語が含まれている。表１の例 1,2 のように訳語候補数が少ない場合においては、それらのほとんどが正しい訳語であった。

(1.2)日本語単語の異なる意味に対して、それぞれの正しい訳語が得られた。表１の例 4,5 に示されているように、“夷”の二つの意味、“外国人”と“情趣を解さない田舎者”に対して、前者の訳語候補に“外国人”が、後者に“乡下人”が含まれた。

(1.3) 表１の例 6,7 に示されているように、市販の日中辞典[倉石,折敷瀬 2001]にも載っていない対訳が得られた。

ただし、例 3 に示されているように、不適当な訳語もたくさん含まれている。多数の訳語候補から正しい訳語を選別することが問題点になる。

(2) 英訳が LDC 単語対応表にないため、中国語訳語候補が得られなかった 3%(計 11,381 個)のレコードについて

その英訳が次の形になったため、LDC に載っていない。

(2.1)複数形：例えば earthworks

(2.2)名詞化: 例えば pitifulness

(2.3)複合名詞：例えば icewall

(2.3)頭文字：例えば IGF

(2.5)地名：例えば Awa

(2.6)まれなもの：例えば argyle

(2.1) (2.2) (2.3) のケースは、英語の形態素解析を通して中国語訳語候補を求めることができる。ほかのケースは手作業で定義する。ただし、本稿では取り扱っていない。

(3) 残りの 57%(計 209,047 個)のレコードについて

EDR の英訳が句また説明文の形である。日本語単語を見ると、複合語のものが多い。これらの中国語訳語の獲得は、複合語の訳語を求めるという問題に帰着でき、そ

の扱いを今後の課題とする。

中国語訳語の候補が得られた 144,002 個のレコードにおいては、それぞれの候補数が異なっている候補数による分布を表 2 に示す。表 2 により、10 個以下の候補をもつレコードは 50.4%を占めている。つまり、約半分のレコードはその中国語訳語の候補が 10 個以下で、しかもその中に正しい訳語が含まれる。一方、10 個以上の候補をもつレコードは 49.6%を占めており、その中に 20 個以上の候補をもつレコードが約 25%を占めている。候補数が一番多い場合では 256 個もある。訳語候補が少ない場合では、不適切なものがあっても少ないので、そのまま使ってもそれほど問題がないだろうと思われる。また、さらに人手で訳語を選択しても、それほど手間がかからないと考えられる。しかし、訳語候補が多い場合では、訳語候補が多いことはそれらの中に不適切なものもたくさん含まれていることを意味するので、そのままでは使えない。したがって、多数の訳語候補から正しい訳語を選別することが重要な問題となり、その選別方法について述べる。勿論、この選別方法は少ない訳語候補を持つ日本語単語にも使える。

表 2　中国語訳語候補数の分布

中国語訳語候補数	レコード数	候補数≦N であるようなレコードの数(%)
N=1	15875	15875(11)
N=5	4786	46841(32.5)
N=10	6115	72511(50.4)
N=20	2314	108788(75.5)

3 正しい訳語の選別方法

訳語の選別は訳語候補が正しい訳語になる可能性をさまざまな方面から推定し、もっともらしいものを選ぶことになる。この可能性の推定には、英訳の共通程度に関する情報、品詞情報と漢字情報を用いることが考えられる。ここで、それぞれの情報により推定したスコアをS_E、S_{POS}、S_{Kanji}で表す。中国語訳語候補をスコアリングする関数にはこの三種類の情報により推定したスコアを用いる。具体的には、日本語単語JWと中国語訳語CW_iが

対訳になる可能性をスコアリングする関数を次のように定義する。

$$Score(JW, CW_i) = W_E \times S_E(JW, CW_i) + W_{POS} \times S_{POS}(JW, CW_i) + W_{Kanji} \times S_{Kanji}(JW, CW_i) \quad (1)$$

ただし、$W_E + W_{POS} + W_{Kanji} = 1.0$. W_E、W_{POS}、W_{Kanji}はS_E、S_{POS}、S_{Kanji}のそれぞれの重みである。以下では、この三種類の情報およびそれらを用いた推定方について述べる。

3.1 英訳の共通程度に関する情報

対訳候補を選別するには、もとの単語の英訳と訳語候補の英訳がどのくらい共通しているかを考慮に入れると有効であることが先行研究によって報告されている[田中, 梅村, 岩崎 1998]。共通する英訳が多ければ、訳語候補がもとの単語に意味的に近くなるので、その候補がより適切な訳語であると考えることができる。田中らは訳語候補の英訳ともとの単語の英訳が共通する単語の数を訳語候補の選別に用いた。Bond らは共通する単語の数に対して、ある正規化を行って、訳語候補の選別に用いた[Bond, Yamazaki, Sulong, and Okura 2001]。本研究では、Bond らと同じ方法を用いた。

日本語単語JWと中国語訳語CW_iが対訳になる可能性を、それぞれの英訳集合の共通する程度で推定する。このように推定した可能性を$S_E(JW, CW_i)$で表し、以下の 式(2)により計算する。

$$S_E(JW, CW_i) = \frac{2^* | E(JW) \cap E(CW_i) |}{| E(JW) | + | E(CW_i) |} \quad (2)$$

ここで、E(·)は単語の英訳集合を表し、|·|は集合の要素数を表す。中国語訳語候補の英訳はLDC中英単語対応表を検索することで得られる。

3.2 品詞情報

市販の日中辞典を調べると、日本語単語と中国語訳語は品詞において何らかの関係があることが分かる。訳語選択には、品詞の対応関係の情報が有効であると考える。Bond ら[Bond, Yamazaki, Sulong and Okura 2001]の日本語とマレー語の対訳辞書の研究においては、品詞について荒く 8 つのカテゴリーを設け、ある品詞の元単語に対して、同じ品詞の訳語しか選択しない。これは品詞情報を利用する一つの方法である。本研究では、EDR 日本語品詞体系の 37 個の品詞を全般的に考察して、中国語単語の品詞との対応関係を作成した。まず、日本語単語と中国語訳語候補の品詞の対応について調べた。中国語訳語候補の品詞情報を得るために、訳語候補に対し北京大学の形態素解析ツール[Zhou and Yu 1994]を用い

て単語分割及び品詞付与を行った。中国語の品詞体系においては 39 個の品詞が定義されている。表 3 は日本語の品詞と中国語のそれ(一部)の対応付けを示す。中国語の品詞には日本語の「形容動詞」に対応するものがないため、「なし」と書いてある。

表 3 日本語と中国語の品詞の対応付け

日本語	普通名詞	動詞	形容詞	形容動詞	接続詞	副詞	数詞	感動詞
中国語	名詞	動詞	形容詞	なし	連詞	副詞	数詞	嘆詞

各レコードの日本語単語とそのレコードの中国語訳語候補の一つ一つを単語ペアとし、それらの品詞を取り出し、品詞ペアとする。訳語候補が複数の単語からなる場合、その最後の単語の品詞を取り出す。取り出した品詞ペアをパターンごとにカウントした。異なる品詞パターンは全部で 222 個あり、表 4 には頻度のもっとも高い 8 つのパターンを示している。

表 4　品詞ペアのパターン分布

順位	品詞ペア(日本語:中国語)	カウント
1	普通名詞 ： 名詞	446,941
2	普通名词 ： 动词	338,682
3	动词 ： 动词	244,264
4	普通名词;动词 ： 动词	128,563
5	形容動詞 ： 助詞	97,003
6	普通名詞 ： 助詞	77,964
7	普通名词 ： 形容词	77,853
8	普通名詞 ： 名詞語素	67,074

表 4 から分かるように、1 番目、3 番目、4 番目のパターンは、訳語の品詞と日本語単語の品詞が対応しているものである。5 番目のパターンは日本語の「形容動詞」と中国

語訳語の最後の文字が主に“的”になった「助詞」とのペアである。このような訳語は日本語の「形容動詞」の連体形の用法と似ているため、「形容動詞」に対応していると言える。また、8 番目のパターンは、その中国語訳語の最後の文字の品詞が「名詞語素」であるため、日本語の「普通名詞」に準対応していると言える。これらの結果から、得られた訳語候補の品詞とそのもとの日本語単語の品詞が互いに対応しているものが多いことが分かった。

一方、2 番目、6 番目と 7 番目の品詞パターンのように、中国語訳語の品詞が日本語単語の品詞と対応していないものもある。このような場合には中国語訳語候補がその日本語単語の訳語として適当でないものが多い。例えば、「普通名詞」“アーミン”の訳語候補の中に、「名詞」の訳語の“貂”、“貂的白毛皮”が正しいが、「形容詞」の訳語の“肥硕”、“强大的”が正しくない。実際「形容詞」の候補は英訳“stout”の形容詞の意味から得られたものである。

以上の調査結果により、不適切な訳語が生じた原因の一つは英語を介した際にもとの日本語単語の品詞に対応しなくなったためだと考えられる。そこで、訳語候補を絞るために、日本語単語の品詞から中国語訳語の品詞への拘束規則を定義した。拘束規則とは計 222 個の異なる品詞パターンに対し、必要に応じ判断条件を加えて、対応関係を定義したものである。対応関係は「対応」、「準対応」、「不対応」と「未定」の四つの尺度のいずれかである。判断条件は中国語訳語の最後の文字が特定の文字「的」あるいは「地」であるなどである。表 5 は表 4 の品詞パターンに対して定義した拘束規則の一部である。たとえば、表 5 の 5 番目の拘束規則は日本語単語の品詞が形容動詞であり、中国語訳語の品詞が助詞である場合では、中国語訳語の最後の文字が「的」であるなら、このような訳語は日本語の「形容動詞」に対応しているとする。拘束規則の例をもう一つ挙げると、日本語単語の品詞が普通副詞であり、中国語訳語の品詞が助詞である場合では、中国語訳語の最後の文字が「地」であるなら、このような訳語は日本語の「普通副詞」に対応しているとする。これは、中国語では文字「地」で終わる文節がよく副詞の役割を担うからである。「準対応」は訳語の最後の単語が日本語単語の品詞に対応している品詞として働く語素(例えば、表 5 の 8 番目の拘束規則)、あるいは訳語が熟語のようなものである。「不対応」は品詞が対応していないものである。「未定」は現在判定できないものである。

表 5　品詞拘束規則

順位	品詞ペア(日本語:中国語)	判断条件	対応関係
1	普通名詞:名詞		対応
2	普通名词:动词		不対応
3	动词:动词		対応
4	普通名词;动词:动词		対応
5	形容動詞:助詞	最後の文字が「的」であるなら	対応
6	普通名詞:助詞		不対応
7	普通名词:形容词		不対応
8	普通名詞:名詞語素		準対応
24	普通副詞:助詞	最後の文字が「地」であるなら	対応
40	普通名詞:数量詞		未定

そして、日本語単語 JW と中国語訳語 CW_i が対訳になる可能性を、品詞の対応関係から推定する。このように推定した可能性を $S_{POS}(JW, CW_i)$ で表す。品詞の対応関係の四つの尺度「対応」、「準対応」、「不対応」、「未定」に対して、それぞれに 1.0、0.8、0.0、0.2 を付与し、$S_{POS}(JW, CW_i)$ の値とする。

3.3　漢字情報

日本語と中国語にはともに漢字が使われているから、訳語の選別には、漢字情報も有効利用できると考えられる。

3.3.1　日本語漢字と中国語漢字

EDR 日英辞書において、日本語漢字を調査した結果から、以下のようなことが分かった。

(1) 半分以上のレコードはその見出しが漢字を含み、28%のレコードはその見出しが漢字のみからなる。

(2) 漢字を含む単語は 196,412 個あるが、異なる漢字は 4,893 個しかない。漢字ごとに、

それを含む見出しをカウントした。順位 5 番以内の漢字を表 6 に示す。

表 6　レコードの見出しに含まれた順位 5 番以内の漢字

漢字	見出しにその漢字を含むレコードの数
合	4397
出	3966
上	3799
手	3772
切	3754

また、中国語の「現代漢語語法信息詞典」[俞，朱，王，张　1997]を調べた。その中に 61,135 個のレコードがあり、異なる漢字は 6,483 個ある。

訳語候補の中の中国語漢字が日本語単語の中の日本語漢字と同じ意味を持つなら、その候補が正しい訳語になる可能性が高いと思われる。したがって、このような漢字間の関係は訳語の選択に使うことが可能である。同じ意味を持つことに着目すれば、日本語漢字と中国語漢字の間に以下の二種類の対応関係が得られる。

(1) 字形が同じかつ意味が同じである。たとえば、故、国、家、山、兄、子、雄、器。

(2) 字形が異なるが、意味が同じである。このような例を表 7 に示す。

表 7　日本語漢字と中国語漢字において、その字形が異なり、意味が同じ例

日本語漢字	稲	鋭	頭	郷	粧
中国語漢字	稻	锐	头	乡	妆

日本語漢字や中国語漢字は一個以上の意味を持つ場合が少なくない。例えば、漢字「故」は、広辞苑[新村 1998]によれば、日本語では 6 個の意味をもち、中日辞典[小学館　1999]によれば、中国語では 7 個の意味を持っている。上の(1)と(2)において「意味が同じである」ということは共通する意味が少なくとも 1 つあることを意味する。そして、日本語単語の中のある漢字と中国語訳語の中のある漢字が「意味が同じである」ならば、この二つ

の単語が同じ意味を持つ可能性が高いと考える。

実際に調べたところ、日本語漢字と中国語漢字は共通する意味を多数持つものが少なくない。例えば、上の例「故」に関して、日本語と中国語の意味はそれぞれ以下になり、５番までの意味はそれぞれ同じである。{　} の中にはその意味を取っている単語の例である。

日本語漢字「故」の意味：

（JS１）古いものごと。{故事、典故}

（JS２）古い知り合い。なじみ。{故郷、縁故}

（JS３）死ぬこと。{故人}

（JS４）わざとすること。{故意}

（JS５）いわれのあることがら。{事故、故障}

（JS６）使い古い。{故紙、反故}

中国語漢字「故」の意味：

（CS１）もとの。昔の、以前の、古い。{故事，故知，故址}

（CS２）古なじみ、友たち、友人、友情。{故乡，亲故，沾亲带故}

（CS３）死ぬ，死亡した（人)。{故人，病故，故去}

（CS４）わざと、ことさらに、故意に。{故意，明知故犯}

（CS５）事故、事件。{事故、变故}

（CS６）わけ、理由、原因。{不知何故}

（CS７）ゆえに、だから、したがって。{无私故能无畏}

日本語単語を構成する日本語漢字と中国語単語を構成する中国語漢字の間のこのような対応関係は両単語間の意味の近さ（つまり、両単語が訳語関係にあるか否か）を推測することに利用できる。たとえば、日本語単語「故郷」と中国語単語「故乡」を構成する漢字において、日本語の漢字「故」と中国語の漢字「故」は(1)の対応関係で、「郷」と「乡」は表７に示されたように、(2)の対応関係である。したがって、これらの単語は対訳関係にある。このように、中国語訳語を選別するには日本語漢字と中国語漢字の対応関係を取り入れることが有効である。

3.3.2 漢字の対応関係の獲得

EDR 日英辞書の 4,893 個の漢字のうち、2,847 個の漢字は単独でレコードの見出しとして定義されている。日本語漢字と中国語漢字の間の対応関係を以下のように自動的に獲得する。

(1) まず、見出しが単一の日本語漢字であるレコードに対して、中国語訳語が一文字の候補を集め、その日本語漢字に対応する中国語漢字の候補の集合とする。その結果、2,586 個の日本語漢字については中国語の対応漢字の候補の集合が得られた。

(2) 次に、中国語の対応漢字の候補をスコアリングする。スコアリング関数は字形の情報と英訳の共通程度の情報[1]を用いて構成する。この関数には、品詞の情報を用いない。それは、中国語漢字が大体多数の品詞を持ち、品詞上の拘束効果があまりないからである。したがって、スコアリング関数は次のようになる。

$$Score(JW, CW_i) = W_E \times S_E(JW, CW_i) + W_{Orth} \times IfUnicodeSame(JW, CW_i). \quad (3)$$

$IfUnicodeSame(JW, CW_i) =$

1, 日本語漢字 JW と中国語漢字 CW_i は同じユニコードを持つ、つまり字形が同じ；

0, そうでなければ。 (4)

英訳の共通程度の情報より、字形が同じかどうかの情報を重視し、W_E と W_{Orth} の値を固定してそれぞれ 0.4 と 0.6 に設定した。各日本語漢字に対しその中国語の対応漢字の候補をスコアリングして、5 位以内の候補を取り出した。その結果の一部を表 8 に示す。下線は正しい訳語を示し、[ ]は意味的に近いものを示している。その結果から次のことが分かった。

(1)大部分の日本語漢字については、その 1 位の中国語漢字が正しい訳語になっている。その中に、「波」のように字形が同じものもあれば、「後」と「后」のように字形が異なるものもある。

(2)一部の日本語漢字については、その 1 位の中国語漢字は正しい訳ではないが、「搏」(組み打ちをする)のように「戦」と意味的に近いものになっている。

[1] ここでいう英訳の共通程度の情報の利用は個々の漢字を用いてLDC中英単語対応表を検索して得られた情報を利用するものであり、3.1 節に述べた単語の英訳の共通程度の情報の利用とは異なる。

表 8　得られた 5 位以内の漢字対応関係の例

日本語漢字	中国語漢字(5 位以内)
波(wave)	波，浪，［漪］，圜, 涞
辞(word)	辞，话，言，［语］，［词］
後(back)	后，底，础，［终］，［末］
悪(evil)	恶，仇，［歹］，［坏］，害
塊(mass)	［群］，块，派，［团］，坨
戦(fight)	［搏］，［仗］，军，赛，战

自動的に得られた漢字対応関係に対し、人手により一位の結果をチェックし、必要に応じて訂正した。たとえば、表 8 においては、日本語漢字「戦」について、対応している中国語漢字の 5 番目の「战」を一位に移した。

前にも述べたが、訳語の選別においては、訳語候補の中の中国語漢字が日本語単語の中の漢字と同じ意味を持つなら、その候補が正しい訳語になる可能性が高いと考えられる。したがって、このような漢字間の関係は訳語の選択に使うことが可能である。一方、個々の日本語漢字と個々の中国語漢字が (漢字レベルで) 字形が同じでも異なっていても、意味が異なっていれば、これらの漢字の組み合わせでできた日本語と中国語の単語が同じ意味を持ったとしても、個々の漢字からその単語の対応関係（つまり同じ意味を持つ単語対であるか否か）を推定することが不可能なため、このような情報は訳語選択に利用できない。したがって、ここでは、同じ意味を持つ漢字にのみ着目し、漢字間の意味上の対応関係を求めた。その中には、字形が同じのものもあれば、字形が異なるものもある。字形が同じで、意味が異なる場合は非常に少ないので、字形が同じという情報を意味上の対応関係を求めるときに利用した。字形が同じで、意味が異なるというごくまれな場合については英訳の共通程度の情報の利用により低い順位に配置されることになった。

3.3.3 漢字の対応関係の利用

得られた漢字の対応関係をスコアリング関数に取り入れる前に、日本語と中国語のそれぞれの語構造において、漢字構成要素はどのようになっているかを見てみる。日本語漢字と中国語漢字はそれぞれ語構成要素であり、日本語漢字と中国語漢字はいずれも意味を

表す文字である。日本語の語構成において、漢字構成要素の間には次のような多様な関係が存在する[斉藤、石井 1997]。

(ア)　主述関係：地震、国営、事変

(イ)　補足関係：文選

(ウ)　修飾関係：英語、家族、作品

(エ)　補助関係：椅子、国内

(オ)　客体関係：愛国、結婚

上の各関係における漢字構成要素の順序は中国語においても同様となる場合が多い。したがって、日本語単語と中国語単語の意味的な近さを測るには編集距離を使うことができる。具体的には、日本語単語 JW と中国語訳語候補 CW_i が対訳になる可能性 $S_{Kanji}(JW,CW_i)$ を表 8 に示しているような漢字の対応関係を用いて次のように推定する。まず、日本語単語 JW を中国語漢字による表現に直す。すなわち、JW の各漢字に対して、式(3)により得られた漢字の対応関係(表 8 を参照)から対応する中国語漢字を取り、日本語漢字を置換する。ここでは、対応関係の中に 1 位の漢字のみを使用した。また、それら(JW の漢字)が本研究で得られた対応関係に存在しない場合、ユニコードが同じである中国語漢字を用いて置換する。このようにできたものを JW^C で表す。例えば、日本語単語「故郷」の場合、「故」と「郷」のそれぞれに対して、中国語漢字「故」と「乡」で置換する。その結果、JW^C ＝「故乡」が得られる。次に、$S_{Kanji}(JW,CW_i)$ を式(5)により計算する。

$$S_{Kanji}(JW,CW_i)=1-\frac{\mathrm{EditDistance}(\mathrm{JW}^C,\mathrm{CW}_i)}{\max(|JW^C|,|CW_i|)} \qquad (5)$$

ただし、EditDistance(JW^C,CW_i)は JW^C と CW_i の間の編集距離である[Levenshtein 1965]。

4　スコアリング結果と評価

3 章で提案した訳語候補の選別手法に対し、評価実験を行った。

評価に用いたテストデータは訳語候補の数が 20 以上の日本語単語(計 37,528 個)の中から 0.3%の単語を無作為に取り出して得られたものであった。その結果、109 個の単語が得られた。このうち、単語「重大だ」の訳語候補の数がもっとも多くて、計 145 である。各単語のそれぞれの中国語訳語候補の「正解」と「非正解」のラベルは人手で付与した。「正解」の判断はその日本語単語のレコードに記載されている日本語での概念説明また英

語での概念の説明により行った。

スコアリングの結果を評価するには、次の三つの評価基準を用いた。*OneRecall* は *n* 位以内の結果の中に少なくとも一つの正解を含むテストデータの割合である。EDR において、一つのレコードは一つの概念しか持っていないので、対応の中国語訳語を一つ取れればよいと考えられる。*Precision* は *n* 位以内の結果の中の正解の割合である。そして、総合的に評価するために、*OneRecall* と *Precision* についての *F-measure* を用いる。本稿では、最も厳しい評価として $n = 1$ とした。

スコアリング関数に用いる三種類の情報、またはそれらの組み合わせの効果を調べるために、それらの情報の重みを 0 から 1 まで 0.1 の刻みで変化させ、$W_E + W_{POS} + W_{Kanji} = 1$ になるような $(W_E, W_{POS}, W_{Kanji})$ のすべての組み合わせを用いて、訳語選別の評価実験を行った。得られた実験結果に対し評価を行い、その中から、一つの情報のみ利用した場合（case 1, case 2, case 3）の結果と、二つの情報を利用した場合(case 4, case 5, case 6)と三つの情報を利用した場合(case 7)においてそれぞれもっともよい結果のみを選び、表 9 に示す。

表 9　三種類の情報を合わせて利用した結果

W_E, W_{POS}, W_{Kanji}	*OneRecall* (%)	*Precision* (%)	*F*-measure (%)
1, 0, 0　(case 1)	80.73	66.67	**73.03**
0, 1, 0　(case 2)	96.33	55.72	70.60
0, 0, 1　(case 3)	92.66	46.55	61.97
0.9, 0.1, 0　(case 4)	80.73	73.51	76.95
0, 0.6, 0.4　(case 5)	92.66	57.87	71.24
0.4, 0, 0.6　(case 6)	89.91	75.48	**82.07**
0.3, 0.3, 0.4　(case 7)	90.83	81.43	**85.87**

これらの結果を *F-measure* で評価した場合、次の結論が得られた。

(1) 一種類の情報のみを利用した場合

S_E を用いて得られた結果が一番よかったことから、英訳の共通程度に関する情報がも

っとも有効であることが分かった。

(2) 二種類の情報を利用した場合

S_E と S_{POS} の組み合わせを利用した結果の中に、(W_E, W_{POS}) を(0.9, 0.1)に設定した場合に得られたものが一番よかった。S_E と S_{Kanji} の組み合わせを利用した結果の中に、(W_E, W_{Kanji}) を(0.4, 0.6)に設定した場合に得られたものが一番よかった。S_{POS} と S_{Kanji} の組み合わせを利用した結果の中に、(W_{POS}, W_{Kanji}) を(0.6, 0.4)に設定した場合に得られたものが一番よかった。一種類の情報 S_E を用いた場合と、S_E と S_{POS}、S_E と S_{Kanji} の両方の情報を用いた場合とを比べると、両方の情報を用いた方がよかったことが分かった。以上のことから、英訳の共通程度に関する情報を利用した上で、品詞情報と漢字情報をそれぞれ加えることは訳語選別に有効であることが分かった。

(3) 三種類の情報を利用した場合

この場合は、任意の重みで得られた結果でも二種類の情報を利用した場合のそれよりよかった。そして、その中で結果がもっともよかったのは $(W_E, W_{POS}, W_{Kanji})$ を(0.3, 0.3, 0.4)に設定した場合であった。以上のことから、多数のヒューリスティックな情報を統合的に利用する方法が有効であることが分かった。

単語「重大だ」の計 145 個の訳語候補へのスコアリング結果(10 位まで)を表 10 に示す。但し、これらは表 9 の case1、case6 そして case7 で指定した三種類の情報の重みでの組み合わせで得られたものである。また、ここでは英訳の共通程度の情報のみを利用して得られた結果を結果 1 と呼び、英訳の共通程度の情報と漢字情報を利用して得られた結果を結果 2 と呼び、三種類の情報すべてを利用して得られた結果を結果 3 と呼ぶ。結果 1 においては、3 位までの訳語は正しかったが、4 位から 10 位までの訳語が正しくなかった。結果 2 においては、3 位までの訳語が正しかったことに加え、8 位と 10 位の訳語も正しかった。8 位の訳語「重要地」と 10 位の訳語「重要的」と「严重的」はそれぞれ「重大だ」と同じ漢字「重」を持っているため、漢字情報を加えることにより、スコアが上げられ、順位がそれぞれ 8 位と 10 位に上昇した。この例からも、漢字情報の有効性が分かる。結果 3 においては、3 位までの訳語が正しかった以外に、7 位、8 位の二つの訳語、9 位の訳語も正しかった。これは、品詞情報を加えることにより、結果 2 の正しくない訳語、5 位の「重視」と 6 位の「重」が動詞であるため、それぞれのスコアが下げられ、それぞれの順位が結果 3 の 16 位と 19 位(表に示していない)までに降下したからである。また、結果 1

の 9 位の「钥匙」のような名詞が結果 3 の 29 位(表に示していない)の低い順位になった。これらの例から、品詞情報の有効性が分かる。

表 10「重大だ」の中国語訳語候補のスコアリング結果

順位	中国語訳語		
	英訳の共通程度の情報のみを用いた場合(結果 1)	英訳の共通程度の情報と漢字情報を用いた場合(結果 2)	三種類の情報すべてを用いた場合(結果 3)
1	严重	重大	重大
2	重大	重大的	重大的
3	重要	重要	重要
4	主用	大	大
5	沉重	伟大, 重视	伟大
6	认真	重	自大
7	要	自大	重要地
8	大	重要地	重要的, 严重的, 庄重的, 巨大的, 较大的, 自大的
9	钥匙, 主科, 钥, 键, 批评性地, 用钻研眼光地	重要性	严重
10	本金, 精密地, 钜, 冢, 主修, 岸然, 必需, 弘, 有活力, 要紧, 郑重	重要的, 严重的, 巨大的, 自大的, 较大的, 庄重的	沉重

$(W_E, W_{POS}, W_{Kanji})$を(0.3, 0.3, 0.4)に設定した評価実験において、評価実験に用いた109個の日本語単語のうち、1 位の中国語訳語候補リストに一つ以上の正解が含まれた日本語単語が99個あり、1位の中国語訳語候補リストに一つの正解も含まれなかった日本語単語が10個あった。ここで、109個の日本語単語のうち、その中国語訳語の文字構成（つまり、個々の文字の字形）と明らかに異なるようなもの（例えば、中国語訳語「歩伐」と文字構成の異なる「足並み」）を取り上げ（つまり、日本語単語とその中国語訳語の文字構成がほぼ同じようなもの、例えば「調査」、「混合」と「新鮮だ」のようなものは取り上げない)、それらの中国語訳語候補のスコアリング結果の一部(3位まで)を表11に候補数の順に示す。ただし、1 位の中国語訳語候補リストに一つの正解もなかった 10 個の日本語単語は表 11 の最後の10行に置いた。また、正解の訳語は下線で示している。

表11を詳細に見ると、例えば、表にある最初の単語「羈絆」については79個の中国語訳語候補が得られた。それらをスコアリングした結果、1 位の中国語訳語候補リストには「羁绊」、2位の中国語訳語候補リストには「绊」、3位の中国語訳語候補リストには「加枷锁,桎梏」の訳語候補が入っている。それらはいずれも正解であった。1位の中国語訳語候補リストに一つの正解も含まれなかった10個の日本語単語については、そのうちの「清淑だ」、「やさしい」、「点」、「立ちゆく」、「明き」の五つの単語は2位の中国語訳語候補リストにその正しい訳語があり、「艶色」、「晴れの」、「小突く」、「掠り」の四つの単語は3位の中国語訳語候補リストにその正しい訳語があった。残りの単語「とはいうものの」は 3 位までの中国語訳語候補リストに正しい訳語がなかった。

表11についてさらに詳しく調べると、最初の三つの単語、「羈絆」、「産み出す」、「郭大する」に関しては、それらを構成する日本語漢字「羈」、「絆」、「産」、「大」とそれらの中国語訳語を構成する漢字「羁」、「绊」、「产」、「大」はそれぞれ漢字対応関係にあることが分かる。したがって、漢字情報は正しい訳語選別に大きく寄与したと思われる。また、次の三つの単語、「枉惑だ」、「嘸や」、「勘定」に関しては、それらを構成する日本語漢字と中国語の訳語候補を構成する漢字は対応関係にないため、英語の共通程度の情報が大きく寄与したと思われる。

Yahoo（日本語サイト）にある日中対訳辞書[三省堂]を利用して、表11のすべての日本語単語について調べた。その結果、計64個の単語のうち、22個の単語がその辞書に存在し、残りの42個の単語が存在していなかったことが分かった(存在していなかった単語

については、四角で囲んで示している)。このように、提案手法を用いることにより既存の電子日中辞書に載っていない対訳が得られた。

前も述べたように、「正解」の判断はその日本語単語のレコードに記載されている日本語での概念説明また英語での概念の説明により行われたものであり、単純に日本語と中国語の単語そのものを見て判断しているわけではない。例えば、表 11 の中の「変化する」は、日本語での概念説明と英語での概念の説明はそれぞれ「単語の語尾が変化する」と″of the form of a word, to change″である。したがって、その中国語訳語候補「格变化」と「词尾变化」 は正しいと判断した。

以上のように、多数のヒューリスティックな情報を統合的に利用する方法の有効性は評価実験により検証された。訳語候補は、EDR 日英辞書の各レコードに対して、その日本語単語の英訳を LDC 英中単語対応表の英単語と照合し、照合に成功した英単語の中国語訳語を取り出して得られたものである。この段階においては、拘束規則がまだ使われていない。拘束規則はその次の段階、つまり、複数の訳語候補から正しい訳語を選別するための候補の順位付けに使われている。また、三種類のヒューリスティックな情報を統合的に利用しているので、そのうちの一部分が順序付けに寄与できなくても、ほかの情報との併用で訳語候補に対して順位を付けることができる。例えば、ある日本語単語に対し二つの中国語訳語候補が得られたとする。それらの訳語候補の品詞と日本語単語の品詞との対応関係が全部「不対応」の場合、品詞情報により推定したスコアはともに$S_{POS}=0$でスコアが同じであるため、順位付けができない。しかし、ほかの情報により推定したスコアS_EとS_{Kanji}が異なっているならば、総スコアも異なり、二つの中国語訳語候補に順位を付けることができる。もし、三種類の情報により推定したスコアがいずれも同じ、あるいは、計算した総スコアが同じであれば、二つの訳語候補は同じ順位になる。この場合、正しい訳語は選別できない。ただし、これは候補の選別ができないことを意味するものであり、候補が挙られないことを意味するものではない。また、選別できないというのは、二つの訳語候補がもともといずれも正しい場合もあれば、いずれも正しくない場合もある。

表 11　20 個以上の中国語訳語候補を有する日本語単語から無作為に取り出した 109 個の日本語単語を用いた評価実験において得られた順位付きの、元の日本語単語と文字構成上異なる中国語訳語候補の例。（四角で囲んでいる単語は日中対訳辞書[三省堂]に載っていないものである）

日本語単語	中国語訳語			
(64 個)	候補数	1 位	2 位	3 位
羈絆	79	羁绊	绊	加枷锁, 桎梏
産み出す	59	产	产生	想出
郭大する	55	扩大	夸大, 放大	可放大
枉惑だ	54	尖, 高明	聪	妙, 乖
嘸や	53	一定	保管, 定当	定然
勘定	49	代价	价钱, 价	估定成本
押さえる	49	压制, 憋	限于	按捺, 镇压
みる	48	盯	睐	注意看
有難い	47	有利	祯	有望的, 有利的
研き上げる	46	好转, 炼	栽培	磨光
くぼ地	45	盆地	洼地	完全地
居る	43	停留, 处于	延缓	逗留, 呆
育み育てる	42	养育	护	扶植, 振兴
引き取る	42	取	得到, 受到	非. . . 不可
取りすてる	39	略去, 撤除	摈除, 摈弃	丢开, 搬
入魂だ	38	蔼, 睦	相近	友善
補完する	37	补充	把. . . 补足	补遗, 补角
受け入れる	37	受到	受理, 受像	接受
思う	36	揣	猜度, 揣测	臆测, 猜想
変化する	35	格变化	词尾变化	音调变化
ひるむ	34	踌	收缩	犁田

表 11　20 個以上の中国語訳語候補を有する日本語単語から無作為に取り出した 109 個の日本語単語を用いた評価実験において得られた順位付きの、元の日本語単語と文字構成上異なる中国語訳語候補の例。（四角で囲んでいる単語は日中対訳辞書[三省堂]に載っていないものである）（続き）

日本語単語	中国語訳語			
	候補数	1 位	2 位	3 位
心持	34	心情	心	情绪, 情感
流麗だ	33	雅	流动的	婉, 秀, 优美
年若だ	33	年青	年轻	青年的, 年青的
下等だ	33	低等	劣	卑下, 卑
うっとうしげだ	33	不愉快	暗淡	阴郁
縛める	32	缚	束缚	捆，绑住
らん惰だ	32	惰	懒惰	懒惰的
代え	32	代替者	代人	代
取りまわす	31	对待, 治疗	筹	持, 处理
似合う	30	合	适合	相配
だらし無い	30	不端庄的,	马虎的	懒散的
嫋々たる	30	纤巧, 纤弱, 纤小	精美	纤
捻回す	30	拧, 扭	搓	绞
書きあらわす	28	撰写, 编著	写, 表达	描述
売れゆき	28	发行额	买卖契约	销路, 销售额
笑味する	28	美味	鉴赏，品味	升值
陥る	27	凋谢	殂	薨
疑わしげだ	27	可疑	疑心的	真假可疑的
すべからく	27	一定要, 定当	必得	必须
一路	26	一心一意	一心一意地	心无旁物地
碌すっぽう	26	满意地	足够地	很好

表 11　20 個以上の中国語訳語候補を有する日本語単語から無作為に取り出した 109 個の日本語単語を用いた評価実験において得られた順位付きの、元の日本語単語と文字構成上異なる中国語訳語候補の例。（四角で囲んでいる単語は日中対訳辞書[三省堂]に載っていないものである）（続き）

日本語単語	中国語訳語			
	候補数	1 位	2 位	3 位
足並	25	步伐, 步子	步调, 足迹	步骤, 措施, 脚步
外使	25	大使	外交使节	特使
惻惻たる	25	凄	悲切, 萧, 恻	哀怨, 哀愁
囲い	24	篱笆, 围墙	栏位, 笆	樊
色差	23	色	着色	色带, 色彩
会釈する	23	点头振动, 鞠躬	弓	点头
推考する	22	推论	推断, 推测	猜谜儿, 揣测
叩き合う	22	吵嚷	争议	争执
俄だ	22	及时	容易地	连忙
復す	21	重回	归, 回来	回归, 退还
賞翫する	21	欣赏	鉴赏	享有
截り口	21	断口	刀口	捷径, 节略,
清淑だ	63	谧	娟, 幽, 袅 婉	晏, 谦虚, 徐
やさしい	51	轻	安逸, 嫩, 简易	柔
艶色	48	光亮	使有光泽, 晖	锦
とはいうものの	43	可是	但	还是
明き	36	宇宙	空间	穷
点	34	目	件	一块, 条
晴れの	29	冢	钧	大的, 华丽的,
立ちゆく	25	去世	渡过	经过
小突く	23	拨开	搠	戳
掠り	23	猱	葡萄疮,	抓伤

5 おわりに

本稿では、英語を介して日中辞書を自動的に構築する手法について述べた。多数の訳語候補から正しい訳語を選別するために、それぞれの英訳がどの程度共通しているかに関する情報、日本語単語と中国語訳語候補の品詞対応関係、そして日本語と中国語の漢字対応関係などの情報を用いてスコアリングする方法を提案した。提案手法を用いて、EDR 日英辞書の 14 万個のレコードに対し、それぞれ順位のつけられた中国語訳語候補を自動的に獲得した。20 個以上の中国語訳語候補を持つ日本語単語に対して、その順位付けの評価実験を行った。その結果、一位に順位付けられた訳語候補の正解率が 81.4%に達したことが分かり、提案手法の有効性が検証された。

提案手法は以下に述べるように、一定の汎用性を持っており、他言語への応用は可能と考える。中国語の訳語候補をスコアリングする関数には英訳の共通程度に関する情報、品詞情報、そして漢字情報を用いている。英語以外の諸言語は英語との対訳辞書があるので、源言語単語の英訳と目的言語の訳語候補の英訳との共通程度を求めることができる。英訳の共通程度に関する情報の有効性は本研究だけでなくすでに和仏対訳辞書を構築する研究によっても検証されている。品詞情報に関しては、英語以外の諸言語にも、品詞体系と形態素解析ツールがある。したがって、二つの品詞体系を対応させるには、提案手法のように、任意の品詞対の対応関係に「対応」、「準対応」、「未定」、「不対応」を設定すればよいであろう。漢字情報に関しては、漢字が両方に使われていない、または、いずれにも使われていない言語対において、直接利用できない。しかし、その考え方はほかの言語対にも適用できる。漢字は日本語と中国語において、意味概念を担う最小の言語単位である。ほかの言語にも、意味概念を担う最小の言語単位が存在する。例えば、英語には、“mono”と“phone”などの最小の意味単位の単語があり、韓国語には、漢字の一部分の意味を表す言語単位がある。二つの言語においてそれらの最小の意味単位の間の対応関係を獲得できれば、それらを訳語候補のスコアの推定に用いることができる。

EDR 日英辞書の残りの 57%(計 209,047 個)のレコードについて、訳語候補が一つも得られなかった。日本語単語を見ると、複合語のものが多い。これらの中国語訳語の獲得は、複合語の訳語を求めるという問題に帰着する。今後は、複合語について、その訳語の求め方を考案していく予定である。

参考文献

Brown, Ralf. D.(1997). Automated Dictionary Extraction for “Knowledge-Free” Example-Based Translation. *In Proceedings of 7th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp.111-118.

Bond, Francis, Yamazaki, Takefumi, Sulong, Ruhaida Binti, and Okura, Kentaro (2001). Design and Construction of a machine-tractable Japanese-Malay Lexicon.言語処理学会第 7 回年次大会発表論文集,pp.62-65.

Fung, Pascale (1998). A statistical view on bilingual lexical extraction: from parallel corpora to non-parallel corpora. *Lecture Notes Computer Science*, Vol.1529, pp.1-17.

Fung, Pascale and Yee, Lo Yuen (1998). An IR Approach for Translating New Words fro m Nonparallel, Comparable Texts. *In Proceedings of* Coling-ACL98, pp.414-420.

井佐原均(2002). 第三言語翻訳システム. 言語処理学会第 8 回年次大会発表論文集, pp. 37-40.

情報通信研究機構(2002). EDR 電子化辞書 2.0 版仕様説明書.

Tanaka, Kumiko and Iwasaki, Hideya (1996). Extraction of Lexical Translation from Non-Al igned Corpora, *In Proceedings of* Coling96, pp.580-585.

倉石武四郎, 折敷瀬興(2001). 日中辞典第二版. 岩波書店.

LDC 英中・中英単語対応表(2002). http://www.ldc.upenn.edu/Projects/Chinese

斉藤倫明, 石井正彦(1997). 語構成. ひつじ書房.

Levenshtein, V.I. (1965). Binary Codes Capable of Correcting, Deletions, Insertions and Reversals. *Doklady Akademii Nauk SSSR* 163(4), pp.845-848.

三省堂. 日中辞書. http://www.excite.co.jp/dictionary/japanese_chinese/

小学館(1999). 中日辞典.

新村出(1998). 広辞苑第五版. 岩波書店

Shirai, Satoshi and Yamamoto, Kazuhide(2001). Linking English Words in Two Bilingual Dictionaries to Generate Another Language Pair Dictionary. *In Proceedings of 19th International Conference on Computer Processing of Oriental Language*, pp.174-179.

田中久美子, 梅村恭司, 岩崎英哉(1998). 第三言語を介した対訳辞書の作成. 情報処理学会論文誌, Vol. 39, No. 6, pp. 1915-1924.

俞士汶, 朱学峰, 王惠, 张芸芸(1997). 現代漢語信息辞典. 清華大学出版社.

Zhou, Qiang and Yu, Shiwen (1994). Blending Segmentation with Tagging in Chinese Language Corpus Processing. In Proc. of COLING-94, 1274-1278.

略歴

張　玉潔：1983 年北方交通大学計算機科学科卒業。1986 年中国科学院計算技術研究所修士課程修了。同年中国科学院計算技術研究所に勤務。機械翻訳の研究開発に従事。国家科学技術進歩一等賞受賞。1999 年電気通信大学博士課程情報工学専攻修了。工学博士。1999 年電気通信大学助手。2000 年ＡＴＲ音声言語コミュニケーション研究所研究員。2002 年情報通信研究機構けいはんな情報通信融合研究センター自然言語グループ専攻研究員。自然言語処理、機械翻訳の研究に従事。言語処理学会会員。

馬 青： 1983 年北京航空航天大学自動制御学科卒業。1987 年筑波大学大学院修士課程理工学研究科修了。1990 年同大学院博士課程工学研究科修了。工学博士。1990～93 年株式会社小野測器勤務。1993 年郵政省通信総合研究所入所。1994 年同所主任研究官。2003 年龍谷大学理工学部教授。2004 年情報通信研究機構専攻研究員（短期）兼任。自然言語処理の研究に従事。言語処理学会、情報処理学会、電子情報通信学会、日本神経回路学会、各会員。

井佐原 均:1978 年京都大学工学部卒業。1980 年同大学院修士課程修了。工学博士。同年通商産業省電子技術総合研究所入所。1995 年郵政省通信総合研究所。現在、独立行政法人情報通信研究機構けいはんな情報通信融合研究センター自然言語グループリーダーおよびタイ自然言語ラボラトリー長。自然言語処理、語彙意味論の研究に従事。言語処理学会、情報処理学会、人工知能学会、日本認知科学会、ACL、各会員.