日中対訳辞書の構築における漢字情報の利用

張 玉潔[*]　　馬 青[**]　　井佐原 均[*]
[*]通信総合研究所　[**]龍谷大学理工学部
[*]{yujie,isahara}@crl.go.jp　[**]qma@math.ryukoku.ac.jp

1　はじめに

対訳辞書の構築は機械翻訳、言語横断検索において重要な研究課題である。本稿は、日英・英中電子辞書を利用して日中電子辞書を構築する方法と結果について報告する。

対訳の自動獲得に関してさまざまな研究がある[1] [2] [3]。対訳獲得の手順は訳語候補の収集と訳語の選別からなる。訳語を選別するためのヒューリスティックな情報としては、元単語と訳語候補のそれぞれの品詞の対応関係が考えられる。また、元単語と訳語候補のそれぞれの英訳集合が共通する程度が考えられる。

我々はこれまでに、英語を介して日英・英中辞書から中国語訳語候補を得、その中から正しい訳語を選別する方法を提案した[4][5]。本稿は、訳語の選別に漢字情報を活用できることを示す。

2　訳語の獲得と選別

日中辞書を構築するには、EDR 日英辞書[6]と LDC 英中・中英単語対応表[7]を利用した。

2.1　日中対訳候補の獲得

EDR 日英辞書の各レコードに対して、日本語単語の英訳を英中単語対応表の英単語と照合し、対応する中国語訳語を日本語単語の中国語訳語の候補とする。中国語訳語の候補が得られたのが 144,002 個のレコードである。一部の例を表１に示す。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た。一方、不適当な訳語もたくさん含まれている。10 個以上の候補をもつレコードは 49.5%を占めている。候補数の一番多い場合では 256 個もある。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問題点になる。以下、訳語選別のためのスコアリング方法について述べる。

表 1　中国語訳語候補の例

例	日本語	中国語訳語
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	足輪	脚镯,脚镣,…
5	アジール	避难所,庇护,…

2.2　スコアリング方法

まず、スコアリングに用いられる情報を説明する。そして、スコアリング方法について述べる。

(1) 英訳の共通程度

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(1)により計算する[2]。

$$S_E(JW,CW_i)=\frac{2*|E(JW)\cap E(CW_i)|}{|E(JW)|+|E(CW_i)|} \qquad (1)$$

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。

(2) 品詞情報

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。EDR 日本語品詞体系には 37 個の品詞がある。 中国語訳語候補に対して、北京大学の形態素解析ツール[8]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。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[4]。たとえば、品詞が「普通名

詞」であるような日本語単語は「名詞」であるような中国語訳語を許容し、「助詞」であるような中国語訳語を許容しないとの拘束である。品詞の対応関係に「対応」、「準対応」、「不対応」と「未定」の四つの尺度を設け、それぞれに 1.0，0.8，0.2 と 0.0 を付与し、$S_{POS}(JW,CW_i)$ の値とする。

(3) スコアリング関数

上に述べた情報を用いて、中国語訳語候補をスコアリングする関数を次のように定義した。

$$Score(JW,CW_i) = W_E \times S_E(JW,CW_i) + W_{POS} \times S_{POS}(JW,CW_i). \quad (2)$$

W_E と W_{POS} は S_E と S_{POS} のそれぞれの重みであり、$W_E + W_{POS} = 1.0$ になる。よって、共通英訳が多いほど、*Score* が大きくなる。品詞が対応するほど、*Score* が大きくなる。

(4)スコアリング結果と評価

英訳の共通程度の情報と品詞情報の効果を分けて見るために、*Score* の計算に S_E のみと S_{POS} のみを用いた実験を行った。また、W_E と W_{POS} のいろいろな組み合わせで実験も行った。訳語候補の数が２０以上であるような日本語単語を無作為に 109 個選んでテストデータとした。各単語のそれぞれ中国語訳語候補に対して「正解」と「非正解」のラベルを人手で付与した。

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、*F-measure* を用いる。計算機実験の結果、S_E と S_{POS} をそれぞれ独立に用いた場合と両方の情報を用いた場合を比べると、両方の情報を用いた方がよかった。

3　漢字の対訳関係

日本語と中国語ともに漢字が使われている。訳語の選別には、漢字情報の利用が考えられる。

3.1 日本語漢字と中国語漢字

EDR 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 2 に示す。

表 2　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」を調べた[9]。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。

調査した結果、日本語漢字と中国語漢字の間に以下の二種類の対訳関係が存在することが分かった。(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 3 に示す。

表 3　字形が異なり、意味が同じ対訳の例

日本語漢字	稲	鋭	頭	郷	粧
中国語漢字	稻	锐	头	乡	妆

漢字の間の対訳関係は日本語単語と中国語訳語の間に反映される。たとえば、日本語単語「故郷」と中国語単語「故乡」では、「故」と「故」

は(1)の対訳関係で、「郷」と「乡」は表 3 に示されたように、(2)の対訳関係である。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訳関係を取り入れることが有効と考えられる。

3.2 漢字の対訳関係の獲得

EDR 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。第 2 節の方法を使って、日本語漢字と中国語漢字の間の対訳関係を以下のように自動的に獲得する。まず、見出しが単一の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る。一般に一個の日本語漢字には多数の意味があり、いくつかのレコードに格納されている。それぞれから集めた訳語をまとめて、その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 個の日本語漢字は中国語漢字の集合が得られた。次に、中国語漢字の候補をスコアリングする。そのために、スコアリング関数に字形の情報を加える。そして、スコアリング関数から品詞の情報を取り除く。これは、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。

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

$$IfUnicodeSame(JW, CW_i) = \begin{cases} 1, & \text{日本語漢字 } JW \text{ と中国語漢字 } CW_i \text{ は同じユニコードを持つ、つまり字形が同じ；} \\ 0, & \text{そうでなければ。} \end{cases} \quad (4)$$

英訳の共通程度の情報より、字形が同じであるかどうかの情報を重視し、W_E と W_{Orth} に 0.4 と 0.6 を設定した。各日本語漢字の中国語漢字候補をスコアリングして、5 位以内の候補を取った。一部の例を表 4 に示す。下線は正しい訳語を示し、[ ]は意味上で近いものを示す。この結果から次のことが分かる。(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があるし、「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。

表 4　得られた 5 位以内の漢字対訳関係の例

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

4　漢字情報の利用

得られた漢字の対訳関係をスコアリング関数に取り入れる。

4.1 単語の構造

日本語の語構成において、主述関係、修飾関係等の多様な関係が存在するが[10]、各関係における語構成要素の順序は中国語において同様となる場合が多い。

4.2 漢字情報の利用

上の観察結果により、日本語単語 JW と中国語訳語 CW_i が対訳になる可能性を、漢字の対訳関係により推定する。このように推定した可能性を $S_{Kanji}(JW, CW_i)$ で表す。まず、日本語単語 JW を中国語漢字での表現に直す。JW の漢字に対して、得られた漢字の対訳関係から対応する中国語漢字を取り、置換する。対訳関係に中国語漢字がない場合、ユニコードが同じである中国語漢字を取り、置換する。このようにできたものを JW^C で表す。次に、$S_{Kanji}(JW, CW_i)$ を次の式により計

算する。**EditDistance(** JW^C, CW_i **)**は JW^C と CW_i の間の編集距離である[11]。

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

4.3 スコアリング関数への漢字情報の追加

スコアリング関数に漢字情報を取り入れ、以下のように定義する。

$$Score(JW, CW_i) = W_E \times S_E(JW, CW_i) + W_{POS} \times S_{POS}(JW, CW_i) + W_{Kanji} \times S_{Kanji}(JW, CW_i). \quad (6)$$

ただし、$W_E + W_{POS} + W_{Kanji} = 1.0$ **.**

第２節のテストデータを用いて、訳語の選別実験を行った。各情報の効果を調べるために、それぞれの重みをいろいろな組み合わせで設定した。評価結果を表 5 に示す。これらの結果から、*F-measure* を見ると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。

(2) 二種類の情報を利用した場合、S_E と S_{POS}、また S_E と S_{Kanji} の組み合わせを利用した結果は S_E のみにより得られた結果よりよかった。このことから、品詞情報と漢字情報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、任意の重みで得られた結果は二種類の情報を利用した場合よりよかった。一番よい結果は W_E, W_{POS}, W_{Kanji} に 0.3,0.3,0.4 を設定して得られたものであった。漢字情報の導入により *F-measure* が約 8.9%上昇した。

表 5　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(case1)	80.73	66.67	**73.03**
0, 1, 0(case2)	75.23	46.84	65.10
0, 0, 1(case3)	91.74	48.09	63.10
0.9,0.1,0 (case4)	80.73	73.51	**76.95**
0,0.6,0.4 (case5)	92.66	58.92	72.04
0.4,0,0.6 (case6)	89.91	75.48	**82.07**
0.3,0.3,0.4(case7)	90.83	81.43	**85.87**

5 おわりに

正しい訳語を選別するためのスコアリングに、品詞情報と英訳情報のほかに、漢字情報も加えた。計算機実験の結果、漢字情報を利用したことにより、*F-measure* が 8.9%上昇した。今後は、スコアリングのさらなる改善を行うとともに、複合語の中国語訳語を求める方法を考案する。

参考文献

[1]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.

[2]Bond, F., Yamazaki,T., Sulong, R. B. and Okura K.(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.

[3]Shirai, S. and Yamamoto, K.(2001) Linking English Words in Two Bilingual Dictionaries to Generate Another Language Pair Dictionary. *ICCPOL2001*, pp.174-179.

[4]張玉潔,馬青,井佐原均(2003) 英語を介した日中対訳辞書の自動構築. 言語処理学会第 9 回年次大会発表論文集,pp.357-360.

[5]Zhang, Y., Ma, Q. and Isahara, H. (2003) Automatic Acquisition of a Japanese-Chinese Bilingual Lexicon Using English as an Intermediary. In *Proc. of International Conference on NLP and Knowledge Engineering'03*, pp. 471-476.

[6]日本電子化辞書研究所(1996)EDR 電子化辞書 1.5 版仕様説明書.

[7]LDC.http://www.ldc.upenn.edu/Projects/Chinese/

[8] 周強、段慧明(1994) 現代漢語語料庫加工中的切詞与詞性標注処理.中国計算機学報,Vol.８５.

[9]Yu, S. (1997). Grammatical Knowledge Base of Contemporary Chinese. Tsinghua Publishing Company.

[10]斉藤倫明、石井正彦(1997) 語構成.ひつじ書房.

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