英語を介した日中対訳辞書の自動構築

張 玉潔　　馬 青　　井佐原 均
通信総合研究所
{yujie,qma,isahara}@crl.go.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を獲得するには、基本対訳辞書と両言語のコーパスを利用する研究がある[1]。基本対訳辞書がない場合、基本対訳辞書をゼロから構築するために、対訳コーパスを利用する研究[2]と、英語を介した研究がある[3][4][5]。対訳獲得の手順としては、まず訳語候補を作り出し、次に統計情報を利用して適当な訳語を選別する。

機械翻訳の研究は英語を一方の言語とする翻訳研究が多く行われてきた。英語と他の言語の対訳辞書、言語コーパスなどの言語資源は豊富に蓄積された。それらをいかに利用して、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かが研究の課題になっている[6]。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。

2 訳語候補の獲得

英語を中継言語として日中辞書を構築するには、EDR 日英辞書[7]と LDC 英中・中英単語対応表[8]を利用した。当面の課題は EDR 日英辞書の日本語単語に対して、中国語訳語を付与することとする。

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、意味コード、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。英訳には訳語がいくつかあり、訳語は単語、句、説明文のいずれかの形である。

2.2 LDC 英中・中英単語対応表

英中単語対応表には 110,834 個のレコードがあり、中英単語対応表には 128,366 個のレコードがある。レコードには原単語と訳語のみが記述され、原単語の品詞情報がない。訳語は主に単語、句である。

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、日本語単語の英訳を英中単語対応表の英単語とマッチして、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。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%の 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 単語対応表にないためであり、ほかの約 57%は単語の形での英訳がないためであった。次にそれぞれについて詳しく説明する。

(1)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、そのそれぞれの候補数が異なっていて、一番多い場合では 256 個もある。この 144,002 個のレコードにおいて、候補数による分布を表1に示す。表1に示すように、10 個以下の候補をもつレコードは 50.4%を占める。得られた中国語訳語候補の例を表2に示す。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。

表1 中国語訳語候補数の分布

中国語訳語候補数 n	レコード数 (候補数=n)	レコード数 (候補数≦n(%))
1	15,875	15,875(11)
5	4,786	46,841(32.5)
10	6,115	72,511(50.4)
20	2,314	108,788(75.5)

表 2　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,…

(ア)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た。例 1,2 のような候補数の少ない結果には、ほぼ正しい訳語であった。

(イ)日本語単語の異なる意味に対して、それぞれの中国語訳語が得られた。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。

(ウ)例 6,7 に示されているように、EDR と LDC の電子データから市販の日中辞典にも載っていない語の訳が得られた。

(エ)意味が同じで見出しが異なる単語は同じ中国語訳語が得られた。例えば、日本語単語“受け手”、“受手”と“請け手”の訳語は同じく、“接收者”になっている。

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適当な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)のレコードは、その英訳が次の形になったため、LDC にはない。(a)複数形：例えば earthworks (b)名詞化：例えば pitifulness (c)複合名詞：例えば icewall (d)頭文字：例えば IGF (e)地名：例えば Awa (f)まれなもの：例えば argyle

(a) (b) (c) のケースは、中国語訳語候補を英語の形態素解析を通して求めることができる。ほかのケースは手作業で定義する。

(3)残りの 57%(計 209,047 個)のレコードは、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すると考えられる。

以下では、得られた中国語訳語候補に対して、その中から適当な訳語を選別する方法を述べる。

3　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して、次にスコアリング方法について述べる。

3.1 品詞

日中辞書を引くと、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。ここで、日本語単語と得られた中国語訳語候補の品詞の対応について調べた。EDR 日本語品詞体系は 37 個の品詞がある。中国語訳語候補の品詞情報を得るために、訳語候補に対して単語分割及び品詞付与を行った[9]。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。日本語の品詞の一部分に中国語の品詞を対応付けて見た(表 3)。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、“＊”で示している。

表 3　日中品詞の対応情況（一部分）

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	＊	連詞	副詞	数詞	嘆詞

各レコードの日本語単語と一つ一つの訳語候補とをペアリンクして、日本語単語とそれぞれの訳語候補との品詞ペアを取り出した。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した。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 には四つのパターンの例を示す。表 4 から分かるように、もっとも多いパターンは、普通名詞と名詞であり、日本語単語の品詞と中国

表4　品詞ペアのパターン分布（一部分）

順位	品詞ペア(日:中)	カウント
1	普通名詞：名詞	446,941
5	形容動詞：助詞	97,003
6	普通名詞：助詞	77,964
8	普通名詞：名詞語素	67,074

語訳語の品詞が対応しているものである。5 番目に多い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」に対応すると言える。また、8 位になるパターンは、その中国語訳語の最後文字の品詞が「名詞語素」であるため、日本語の「普通名詞」に相当すると言える。これらの結果から、得られた訳語候補の品詞とその日本語単語の品詞が互いに対応しているものがかなり多いことが分かった。

一方、6 位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”の訳語候補の中では、「名詞」の訳語の“貂”や“貂的白毛皮”が正しく、「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、不適当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、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」のいずれかである。「対応」の判定には、最後の文字が“的”である訳語は日本語の「形容詞」と「形容動詞」に対応しているとし、最後の文字が“地”である訳語は「副詞」に対応しているとした。「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、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。

3.2 共通する英訳の数

対訳候補を選別するには、もとの単語の英訳と得られた訳語の英訳がどのくらい共通しているかも考慮に入れた[3]。LDC 中英単語対応表を検索することで、中国語訳語候補の英訳を得た。中国語訳語候補の英訳と日本語単語の英訳との共通する単語の数を中国語訳語候補のスコアリングに用いた。共通する英訳が多ければ、訳語候補が日本語単語に意味的に近くなるので、訳語候補がより適当な訳語であると考えることができる。

3.3 スコアリング方法

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。中国語訳語候補 C を日本語単語 J の訳語とするペナルティ値を次の式により計算する。

$$Pen(J,C) = F1(J,C) + F2(C) - F3(J,C)$$

F1 の値は 3.1 で定義した品詞拘束規則により表 5 のように定める。よって、品詞が対応するほど、F1 が小さくなり Pen は小さくなる。F2 の値は候補 C の長さと正比例にするように定める。訳語としては、単語、句の方が説明文より好ましいと考えられる。よって単語数が少ないほど、F2 が小さくなり Pen は小さくなる。F3 の値は J と C のそれ

表5　品詞拘束規則により F1 の値を定める規準

J と C の品詞	F1 の値
「対応」	一番小さい値
「準対応」	二番目に小さい値
「不対応」	一番大きい値
「未定」	二番目に大きい値

ぞれの英訳で共通する単語の数と正比例にするように定める。よって、共通単語が多いほど、F3が大きくなりPenは小さくなる。

4 スコアリング結果と評価

品詞情報と共通の英訳情報の効果を分けて見るために、Penの計算にF1とF2だけを入れる実験Ⅰと、F3を加える実験Ⅱを行った。実験Ⅰと実験Ⅱでは、それぞれの計算式により、訳語候補をスコアリングして順位を並べた。スコアリングの結果を評価するには、正しい訳語を得るための候補数がどのくらい減ったかについて調べた。そのため、訳語候補の数が２０以上で、正しい訳語が含まれる日本語単語を無作為的に172個選んだ。そのうち、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。

訳語候補の中に正しい訳語を含ませるために訳語候補は次のように決める。指定順位以内に正しい訳語が一個あれば、その順位以内の候補を取り出し；そうでなければ、すべての候補を取り出す。実験Ⅰの結果について指定順位を１とし、実験Ⅱの結果について順位を２とした。スコアリングなしで得られた候補は同じく１位である。各場合の平均候補数を表6に示す。

単語“重大だ”の145個の訳語候補へのスコアリング結果の一部分を表7に示す。実験Ⅰにおいては、正しい訳語“重大”が１位に入り、“归结”のような動詞、“钥匙”のような名詞が低い順位に入った。実験Ⅱにおいては、１位の候補の数は2に減り、いずれの候補も正しい訳である。

表6　スコアリングによる平均候補数の変化

日本語単語数	中国語訳語候補数の総和	平均候補数		
		スコアリングなし	スコアリング	
			Ⅰ	Ⅱ
172	6109	35.5	19.6	8.6

表7　“重大だ”の訳語候補のスコアリング結果

実験	順位	候補数	訳語
Ⅰ	1	27	要紧,重要,重大,紧张,...
	7	16	归结,铭刻,主修,终结,...
	8	33	钥匙,大小,竟争者,...
Ⅱ	1	2	重大,严重
	2	3	重要,沉重,大

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方法について述べた。多数の訳語候補から正しい訳語を選別するために、日本語単語と中国語訳語候補の品詞対応関係、それぞれの英訳がどの程度共通するか、そして訳語候補の長さなどの情報を用いてスコアリングする方法を提案した。提案手法を用いることで、EDRの14万個のレコードについて順位のつけられた中国語訳語候補が得られた。実験を行った結果、正しい訳語を確保できる候補数は平均35.5から8.6に減り、提案手法の有効性が検証された。今後の課題は、スコアリング方法の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ことである。

参考文献

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

[2] R. D. Brown(1997)Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation. 7th TMI, pp.111-118.

[3]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.

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

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

[6]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.

[7]通信綜合研究所(2002)EDR 電子化辞書 2.0 版仕様説明書.

[8]LDC英中・中英単語対応表.http://www.ldc.upenn.edu/Projects/Chinese/

[9] 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.