ＥＤＲ日英辞書の中国語への拡張

張 玉潔[*1]　　馬 青[*1,*2]　　井佐原 均[*1]
[*1]情報通信研究機構
[*2]龍谷大学
{yujie,qma,isahara}@nict.go.jp

1. はじめに

対訳辞書は機械翻訳や言語横断検索において欠かせない言語資源である。ＮＩＣＴでは、平成１８年度から、科学技術文献の日中・中日機械翻訳システムの開発というプロジェクトをスタートした。機械翻訳の仕組みとして用例ベース手法を採用しているため、大規模な日中対訳コーパスに対して、単語・句レベルでのアライメントを行い、翻訳知識としての対訳関係を抽出することが必要である。基本対訳辞書は単語・句レベルでのアライメントにおいても欠かせないものである。本研究は付与情報が豊富でＮＩＣＴが著作権を所有するＥＤＲ日英辞書[1]を中国語へ拡張し、日中の基本対訳辞書の構築を行う。それにより、日英中三言語の対訳辞書が得られる。

2. 作業内容

作業はＥＤＲ日英辞書の各レコードに対して、中国語訳語と関連情報を付与することである。

2.1　ＥＤＲ日英辞書

ＥＤＲ日英辞書は、日本語単語をよく網羅しており、単語の多義性に対し一つの概念を一個のレコードとして格納している。レコードには、その概念についての説明と英訳などの豊富な情報が付与されている。レコードは全部で 364,430 個ある。レコードには次のような情報が記述されている。

＜レコード番号＞
＜見出し情報＞：＜漢字見出し＞＜かな見出し＞
＜文法情報＞：＜品詞＞
＜意味情報＞：＜概念識別子＞＜日本語概念見出し＞＜英語概念見出し＞＜日本語概念説明＞＜英語概念説明＞
＜英訳情報＞：＜訳語種別＞＜訳語表記＞＜訳語品詞＞
＜管理情報＞

この中で、＜概念識別子＞は単語の多義性を識別するための数字であり、概念の同一性を保持するためのものである。＜日本語（英語）概念見出し＞には、その概念を代表するにふさわしい単語見出しが記述される。＜日本語（英語）概念説明＞は概念を説明する文章であり、その概念と他の概念との識別を人間が行い易くするためのものである。＜日本語（英語）概念見出し＞と＜日本語（英語）概念説明＞は単語、句、文章のいずれかの形である。例１に、レコードの例を示す。

[例 1]

＜レコード番号＞　JEB0368581
＜漢字見出し＞ さえずる
＜かな見出し＞　サエズ・ル
＜品詞＞　動詞
＜概念識別子＞　3bbd74
＜日本語概念見出し＞ さえずる[サエズ・ル]
＜英語概念見出し＞
＜日本語概念説明＞　舞楽において、朗詠する
＜英語概念説明＞ in Japanese court dance and music, to recite
＜訳語種別＞ 0
＜訳語表記＞　recite
＜訳語品詞＞　動詞及び動詞句
＜管理情報＞　 DATE=”95/2/15”

同じ＜見出し情報＞でも＜概念識別子＞によりいくつかのレコードがあり、また同じ＜概念識別子＞でも通常いくつかのレコードがある。

2.2　中国語への拡張

ＥＤＲ日英辞書の各レコードに対し、以下のような情報を付与する。

（１）レコードの＜意味情報＞に基づいた中国語訳語

（２）中国語訳語ごとへの品詞・構文情報

（３）中国語訳語ごとへのレジスター情報

（４）＜日本語概念見出し＞への中国語訳

（５）＜日本語概念説明＞への中国語訳

3. 情報付与基準

3.1 中国語訳語の付与基準

レコードの＜日本語概念見出し＞と＜日本語概念説明＞に基づき中国語訳語を付与する。

その概念が中国語に存在する場合、それを表す中国語単語を付与する。ただし、その概念に対していくつかの中国語訳語があるとき、その概念との一致性の高い順にすべて付与する。ただし、例えば日本語単語「叔父」に対し、中国語では場合によって「叔叔」(「父の弟」) と「舅舅」(「母の弟」) を訳し分けする必要がある。したがって、そのような場合は一つの概念に対し、異なる意味の中国語訳語に対して訳し分け情報を付与する。

日本語単語の概念が中国語に存在しない場合、以下の方式で訳語を付与する。

- 言い換え：その表現を使用することにより、元の表現と同様の効果が得られる（類似の概念が想起される)。
- 逐語訳：字面どおりに直訳する。
- ローマ字：ローマ字の発音を中国語の表音文字で表す。
- 説明文：中国語文章で説明する。
- 地名、人名、組織など固有名詞の場合
  - 漢字の場合

    その漢字が中国語簡体字にあれば、その文字を使う；その漢字が中国語簡体字になく中国語繁体字にあれば、対応する簡体字を使う。日本語漢字が中国語文字にない場合、必要であればその文字を導入し、日本語の発音をPinYin で表記し付け加える。
  - カタカナの場合

    それが英語やドイツ語などからの単語であれば、元の発音を中国語の表音文字で表す。

3.2 日本語見出しと日本語概念説明の翻訳基準

＜日本語概念見出し＞は、日本語単語見出しと異なる場合が多いが同じ場合も少なくない。したがって、訳した中国語概念見出しも中国語訳語と同じ場合が多い。＜日本語概念説明＞は、より長い文章になるが、修飾節を含む名詞句レベルのものもあれば一つの文になる場合もある。中国語に訳すとき、中国語も同じ構文構造になるように訳す。この作業により、同時に日中対訳文・句を得ることが期待できる。

3.3 レジスター情報の付与基準

中国語訳語に以下のような付加的な情報も付与する。

- 中国語訳語の種別：「逐語訳」、「説明文」、「音訳」など
- スタイル：「古語」、「口語体」、「文語体」、「俗語」
- フォーマル：「尊敬」、「謙譲」、「軽蔑」

3.4 品詞・構文情報の付与基準

日中機械翻訳において、文法的に正しい中国語文を生成するには、品詞・構文情報が必要である。そのために、中国語訳語に品詞・構文情報を付与する。中国語訳語に対し、まずその訳語が単語か句かを判別する。次に、単語あるいは句の分類に基づき品詞・構文情報を付与する。単語の品詞情報は１８種類とし、句の構文情報は１１種類とした[2]。具体的な品詞・構文情報は以下のとおりである。

単語の品詞：名詞、時間詞、場所詞、方向詞、代詞、動詞、形容詞、状態詞、副詞、介詞、連詞、助詞、語気詞、区別詞、数詞、量詞、感嘆詞、似声詞

句の構文情報：名詞句、動詞句、形容詞句、介詞句、数量句、時間句、場所句、区別詞句、副詞句、独立句（熟語、慣用語）、単文（主語と述語がある）。

この作業は、＜日本語概念説明＞を訳したあとに行うため、その訳である＜中国語概念説明＞を見ることができる。さらに、日本語単語と英訳の品詞情報を参照することもできる。したがって、この作業は、日本語が分からない中国語ネィーティブによっても行うことができ、中国語言語資源の構築に従事する専門家を活用することができる。

4. 作業補助ツール

上に述べたように、中国語訳語を付与し、＜日本語概念見出し＞と＜日本語概念説明＞を中国語に訳す翻訳作業と、中国語訳語に品詞・構文情報を付与する単言語における作業がある。作業の特徴を考慮し、それぞれの補助ツールを開発した。

4.1 翻訳作業の補助ツール

作業は、千個のレコードを一個の作業ファイルとする。翻訳作業の補助ツールの操作画面を図１に示す。ツールは以下のような機能がある。

１）作業対象とするファイルを選択すると、左側にレコードの見出しリストが表示される。特定の見出しを選択すると、そのレコードの＜見出し情報＞、＜品詞＞、＜概念識別子＞、＜日本語概念見出し＞、＜英語概念見出し＞ 、＜日本語概念説明＞、＜英語概念説明＞が表示される。

２）中国語訳語を編集し、その訳語のレジスター情報を付与することができる。複数の訳語がある場合にも対処できる。

３）＜日本語概念見出し＞と＜日本語概念説明＞の中国語訳を編集することができる。

４）二つの検索機能を持つ。一つは、同じ＜概念識別子＞を持つレコードを検索する。これは、同じ＜概念識別子＞のレコードでは＜日本語概念見出し＞と＜日本語概念説明＞の内容が同じである場合が多いため、訳した内容を参照できるようにするためである。もう一つは、日本語単語を入力し、その単語と同じ見出しのレコードを検索する。

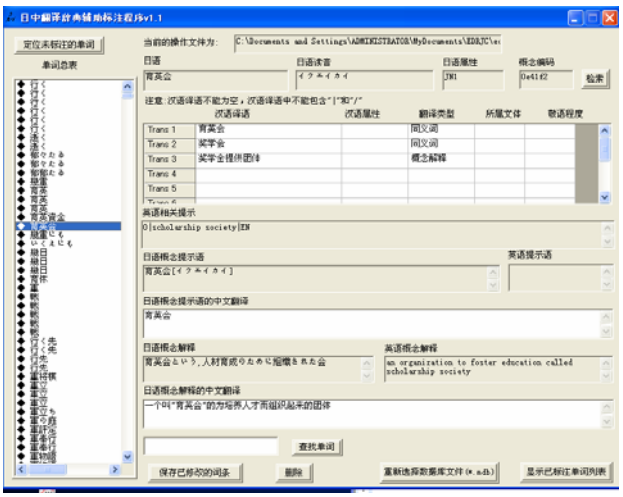

図１．翻訳作業補助ツールの操作画面

4.2 品詞・構文情報の付与作業の補助ツール

品詞・構文情報付与作業の補助ツールの操作画面を図２に示す。ツールは主に、以下のような情報表示、情報付与及びレコード検索の機能を持つ。

１）中国語訳語の品詞付与時に参照できるよう、日本語単語の＜品詞＞、＜中国語概念説明＞、英訳の＜訳語表記＞と＜訳語品詞＞を表示する。

２）中国語訳語が単語か、句かを判別してから、それぞれ対応している品詞・構文情報の一覧表から選択できる。

３）品詞・構文情報が付与されていないレコードと品詞・構文情報付与済みのレコードから、それぞれ同じ中国語訳語を含むレコードを検索することができる。この検索機能により、多数の同じ中国語訳語に対して、品詞・構文情報の付与や付与された品詞・構文情報の修正をまとめて行うことができる。

４）人手作業の負担を減らすために、ツールを使う前に中国語訳語に対する品詞の自動付与も実

施している[3]。このような場合に、品詞の付与結果は表示できるが編集を禁止する機能も備えている。その自動付与の方法は以下のとおりである。中国語訳語がただ一個の単語であり、かつその単語が「現代漢語文法情報辞書」[2](約7万単語)において一つの品詞しか持っていない場合、その訳語にその品詞を付与する。

図 2. 品詞情報付与作業補助ツールの操作画面

5. 作業の実施と現状

この作業は、平成１７年度から始めた。低コストかつ高品質の中国語訳と関連情報を得るために、中国国内の有力な専門翻訳者と豊富な中国語処理資源を活用するようにしている。作業のはじめに、概念識別子と日・英概念見出しと日・英概念説明の内容により、364430 個のレコードを整理した。その結果、概念識別子と日・英概念見出しと日・英概念説明の異なるレコードを、計 265304 個得て、それらに対して翻訳作業を行っている。ほかのデータは見出しのみが違うため、その中国語訳の情報は上記のレコードからコピーして使う。また、高品質の中国語訳を保証するために、翻訳作業のあとに訂正作業もあり、それはレベルのより高い翻訳者により行う。

平成１７年度に、翻訳・訂正作業において、40374 個のレコードが完成された。得られた中国語訳語、中国語概念見出し、中国語概念説明はそれぞれ、255013、171760、565588 文字である。平成１８年度に、頻度の高い日本語単語（JUMAN の日本語単語リストと重なる部分）をはじめ、約 11 万個のレコードを完成する予定である。また、中国語訳語に品詞・構文情報を付与する作業も始めた。以下には、“日本語単語[中国語訳語] 日本語概念見出し[中国語概念見出し] 日本語概念説明[中国語概念説明]”の形で例を挙げる。

育英会 [育英会;奖学会;奖学金提供团体]

育英会 [育英会]

育英会という,人材育成のために組織された会 [一个叫"育英会"的为培养人才而组织起来的团体]

異口同音 [异口同声]

異口同音 [异口同声]

おおぜいが同じことを言うこと [许多人同时说出相同的话]

居溢れる [拥挤不堪;人多坐不开]

居溢れる[拥挤不堪;人多坐不开]

(ある場所に)人が大勢集まり入りきれない[(某场所)聚集了许多人,已经挤不下了]

エクステンションコース [大学公开讲座;公开课;广播讲座]

エクステンションコース[大学公开讲座;广播讲座]

一般の人々に公開されている大学の講座[大学里向一般人群公开的讲座]

6. おわりに

本稿では、ＥＤＲ日英辞書の中国語への拡張作業において、情報付与の基準、作業補助ツール及び現時点の作業結果を報告した。平成１９年度までに全部完成して公開する予定である。

参考文献

[1]情報通信研究機構(2002). EDR 電子化辞書 2.0 版仕様説明書.
[2]俞士汶, 朱学峰, 王惠, 张芸芸(1997). 现代漢語信息辞典. 清華大学出版社. (中国語)
[3] 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.