ＮＩＣＴ多言語コーパスにおける日中対訳データの構築

張 玉潔[*1]　　馬 青[*1,*2]　　内元　清貴[*1]　　井佐原 均[*1]
[*1]情報通信研究機構
[*2]龍谷大学

{yujie,qma,uchimoto,isahara}@nict.go.jp

1　はじめに

機械翻訳の研究またはシステムの開発の一環として、情報通信研究機構（ＮＩＣＴ）では、ＮＩＣＴ多言語コーパスの構築を行っている。本稿は、その中の日中対訳データの構築について述べる。

対訳コーパスは、異なる言語の対を収集したもので、文章、段落あるいは文を単位として構成される。対訳コーパスは源言語と目的言語との間の異なるレベルでの対応関係を網羅することができるので、例えば機械翻訳の研究や機械翻訳システムの開発において翻訳知識を抽出する際などに必須となる言語資源である。近年、市販の翻訳ソフトが入手できるようになったが、翻訳の質に関する多数の問題が未解決のまま残っている。したがって、基礎的な研究面においては、翻訳にかかわる言語現象を綿密に分析した上で、よりよい翻訳方法を提案することが期待される。一方、システムの開発などの実用的な研究面においては、用例ベースの方法と統計的手法が盛んに研究されている。このように、対訳コーパスは、基礎と実用の両面から必要とされる、重要な言語資源である。ヨーロッパ言語の間の対訳コーパスについてはすでに完備されており、ＬＤＣ[1]から入手可能である。ところが、英語とアジア言語の間の対訳コーパスはまだ少ない。特に、アジア言語の間の対訳コーパスがあまり存在しないというのが現状である。

このような状況を鑑み、ＮＩＣＴでは、アジア言語を含む多言語コーパスを構築するプロジェクトを数年前にスタートした[2]。現在は、日本語と英語、日本語と中国語の対訳コーパスを統一した仕様で構築中である。日本語データは京大コーパス（Version3.0）[3]と同一のものとＬＤＣの英語データ（Wall Street Journal）の一部を翻訳したものから構成される。京大コーパスは、１９９５年の毎日新聞の記事から抜粋された約４万文からなる。採用した英語データは上記京大コーパスを英訳したものとＬＤＣの英語データの一部である。中国語データは上記日本語データを中国語に翻訳したものである。翻訳はいずれもプロの翻訳者によって行われた。日本語データに対しては、日本語話し言葉コーパスの仕様に基づいて、単語分割、品詞タグ付け作業、そして構文構造タグ付け作業を行う。英語データに対しては、Penn Treebank style II の仕様に基づいて品詞タグ付け作業と構文構造タグ付け作業を行う。中国語データに対しては、本稿で述べる仕様に基づいて、単語分割と品詞タグ付け作業を行う。以上の作業を経て、日中対訳データが構築される。

本稿は、中国語データの準備及び、形態素解析を中心に、解析・タグ付けの詳細と問題点、得られたデータに対する分析などについて述べる。

2　日中翻訳の作業

中国側のデータは毎日新聞記事中の日本語文約４万文を中国語に翻訳したものである。実際の翻訳作業はプロの翻訳者によって行われた。

2.1　翻訳の基準

翻訳者は以下の基準に従って、翻訳する。

(1)一文の日本語を単位として訳す。

(2) 構造が原文に近い訳文を優先する。ただし、これは多数の訳し方がある場合の制限条件である。

(3)意味が通じるように、必要があれば前の文から情報を補う。特に、日本語には、主語がよく省略されるが、中国語には主語が必要なので、訳文には主語を補うことが多い。

(4)流暢性、読みやすさを考慮し、必要に応じて語順の入れ替え、カンマ「,」の挿入などを行う。これは、基準(2)に対する拘束である。つまり、流暢性を優先する。日本語には、文が長くても構造を示す格助詞があるが、中国語にはこのような表層的な情報がないため、長い文を理解するには支障が出る。解決の一手法として、適当なところに区切りをし、カンマ「,」を入れる。

2.2 訳文の質の確保

訳文の質を確保するために、一次翻訳のあと、以下の精密化を行った。(1)翻訳者による精密化：原文と訳文の両方を見て、原文の意味が忠実に訳されたか否かをチェックし、必要があれば、修正する。(2)中国語のネイティヴによる精密化：訳文のみを見て、理解できるかをチェックする。そして、流暢性もチェックする。問題のある個所に記しをつけ、翻訳者に渡し修正してもらう。

2.3 固有名詞の翻訳

地名と人名の中に中国語と同じ字形の文字があれば、それを採用して訳す。同じ字形の文字がなければ、字形が似ている文字を採用して訳す。例えば「埼玉」の「埼」、人名の「鼻」に対しては、「琦」、「鼻」を採用する。日本文化を表す特有な固有名詞、例えば「大相撲」や「春闘」などは個別に処理する。「大相撲」は中国語の簡単字に直したものとして「大相扑」がすでに存在しているのでそれを使う。一方、「春闘」を「春斗」と訳しても意味が通じないので、この場合は、「春斗」と訳しその後ろに説明として「春季劳资纠纷」を付け加える。

現在、一次翻訳作業はすでに終了している。その作業には約１年間かかった。結果として、計38,840の日本語と中国語の文対が得られた。しかし、精密化にはすでに2年間かかっているが、まだ完成していない。

3 中国語文の形態素解析と人手修正作業

中国語訳文に対して単語分割及び品詞タグ付け作業を行った。まず、解析ツールを用いて単語の分割と品詞タグ付けを行う。そして、得られた結果に対し人手により修正を加える。

3.1 解析ツールを用いた単語分割と品詞タグ付け

自動解析に用いたツールは北京大学で開発されたものである。北京大学は10年以上かけて、中国語の単語の定義、単語の分割基準を研究してきた。 その研究成果として、「現代漢語文法情報辞書」[4](約7万単語)や「人民日報のタグ付きコーパス」が作成され、自動単語分割と品詞タグ付けツールが開発された[5]。これらの研究成果は中国語処理のデファクトスタンダードとして広く用いられている。また、上記コーパスは規模が大きく、自然言語処理の研究に与える影響が大きい。以上のことから、中国語の単語分割と品詞タグ付けには、北京大学が制定した基準を採用し、北京大学の解析ツールを用いた。

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。その一部を表1に示す。

表 1. 中国語品詞タグの一部分

タグ	説明
Ag	形容詞語素
a	形容詞
ad	副詞として使われる形容詞
an	名詞の役割を持つ形容詞
b	区別詞
c	接続語
Dg	副詞語素
d	副詞
e	感嘆語
f	方向と位置を表す
g	語素
h	接頭辞

3.2 人手による修正

人手による修正作業を補助するツールは日本語用としてすでに開発されており、今回、これを中国語に適用した。このツールには、文字列を任意のところで分割したり、任意の二つの単語を一つに結合するための編集機能がある。単語分割を直した単語に、上記の中国語品詞体系から任意の品詞を選び、それを付与する操作もできる。そのほかに、以下の便利な機能を持っている。

(1) 「現代漢語文法情報辞書」への単語検索ができる。この機能は、単語の確認や品詞を調べる際に便利である。

(2) チェック済みコーパスに対し単語検索ができ、検索結果をその単語の左右の文脈によってソートすることができる。この機能はチェック済みのデータを参照するとき便利である。

(3) 直前に変更された単語と同じ属性を持つすべての単語に対し、同じ変更を一括で行うことができる。この機能は単語の分割と品詞タグ付けにおいて一貫性を確保するときに便利である。

3.3 チェック済みデータ

現在までに、中国語データに対し約２万文の単語と品詞タグのチェックが終了した。タグ付けの例文を表２に示す。一文目の“[大阪/ns　大学/n　医学部/n]nt”が示しているように、個々の単語に品詞(上記の例では、ns,n など)を付与した上で、必要があれば複数の単語を一括りにしてさらに品詞(上記の例では nt)を付与する。

人手で修正したのは主に日本の地名、人名、組織名、機関名といった、日本に関する固有名詞であった。1 万文のチェック済みデータを調べた結果、3.1 節に言及した北京大学の辞書に載っていない、新しい単語が約 7,000 個得られた。その一部を挙げる。

[例] 丰田,小松,宫城,大野,三浦,羽田,海部,坂田,西武队,法务局,警视厅,检察厅,总务厅,NTT,NHK,三泽,上京区,中央区,涩谷区,镰仓,加茂,难波,秩父宫,橄榄球场,超市,动画,空洞化,低迷,连休,DNA,全日空,民主化,大藏,皇居,本愿寺,信息化,佛教,市立,私营化,全球化,协和,问候信,倡导者,争抢,意欲,撞上,喝醉,内装,点点头

新しく得られた単語の品詞について、頻度が高いものから順に８位以内のものを表 3 に示す。

表 3.　新しく得られた単語の品詞分布

品詞	数
人名	2384
組織名、団体名	1450
地名	1261
名詞	1259
ほかの固有名詞	273
動詞	225
略語	70
名詞用動詞	65

4　おわりに

本稿では、NICT 多言語コーパスにおける、日中対訳データとその構築方法を紹介した。我々が定めた基準に従って、訳文の質の確保を工夫しながら、毎日新聞記事の約４万文の日本語文を中国語に翻訳した。そして、自動と人手修正の二段階の処理で単語分割と品詞タグ付けを行った。現時点までに、約 4 万文対の日中対訳文が得られている(中国語訳文は約 1,410,892 文字, 926,838 単語からなる)。そのうち、チェック済みの中国語訳文のタグ付きデータは 2 万文以上である。今後、引き続き残りのデータの人手でチェックを行うとともに、単語や句レベルでの対応関係の付与作業を行っていく予定である。

参考文献

[1]LDC. Linguistic data consortium. http://www.ldc.upenn.edu

[2]内元清貴,须藤清,村田真樹,関根聡,井佐原均(2004) 文脈を考慮したタグ付き対訳コーパス. 言語処理学会第 10 回年次大会発表論文

集, pp. 592-595.
[3] 黒橋禎夫, 長尾眞(1997) 京都大学テキストコーパス・プロジェクト. 言語処理学会第 3 回年次大会発表論文集, pp. 115-118.
[4] 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. (中国語)
[5] Qiang Zhou, Shiwen Yu (1994). Blending Segmentation with Tagging in Chinese Language Corpus Processing. In Proc. of COLING-94, 1274-1278.

表 2. 日本語の原文と対応している単語と品詞タグ付き中国語訳文

日本語原文	単語分割と品詞タグ付きの中国語訳文
大阪大学医学部は今年度中に、同学部の全研究室にコンピューターを設置、医療情報や研究論文のデータベース化作業を始める。	[大阪/ns 大学/n 医学部/n]nt 在/p 今年/t 内/f 将/d 在/d 该/r 学部/n 的/u 所有/b 研究室/n 设置/v 电脑/n ,/w 开始/v 建立/v 医疗/n 信息/n 及/c 研究/vn 论文/n 的/u 数据库/n 。/w
各コンピューターを世界最大のコンピューターネットワーク、「インターネット」と接続し、データベースも公開する。	各/r 电脑/n 都/d 与/p 世界/n 最/d 大/a 的/u 电脑/n 网络/n “/w 因特网/n ”/w 连接/v ,/w 数据库/n 也/d 将/d 公开/v 。/w
マルチメディア時代に合わせて積極的に情報提供し、阪大の存在を世界にアピールするのがねらい。	其/r 目的/n 是/v 根据/p 多媒体/n 信息/n 时代/n 的/u 需要/n 积极/ad 提供/v 信息/n ,/w 向/p 世界/n 宣传/v [大阪/ns 大学/n]nt 的/u 存在/vn 。/w
同様の試みは東京の大学などで一部始まっているが、医学部全体としては例がないという。	据说/v 与/p 此/r 相同/a 的/u 尝试/vn 虽然/c 在/p 东京/ns 的/u 大学/n 等/u 部分/m 地区/n 开始/v 实施/v ,/w 但/c 作为/p 医学部/n 整体/n 来说/u 尚/d 无/v 先例/n 。/w
コンピューター八十五台を購入。	购/Vg 入/v 八十五/m 台/ns 电脑/n 。/w
各研究室では、研究論文を英語と日本語で入力、診断に用いる画像情報などもデータベース化する。	在/p 各/r 研究室/n ,/w 研究/vn 论文/n 以/p 英文/n 和/c 日文/n 输入/v ,/w 还/d 将/d 建立/v 用于/v 诊断/v 的/u 图像/n 信息/n 库/n 等/u 。/w
費用は一億円規模になる見込みで、関西の政財界などで作る大阪大学医学部整備拡充委員会からの寄付でまかなうという。	据说/v 预计/v 所/u 需/v 费用/n 为/v 一亿/m 日元/n 左右/m ,/w 将/d 由/p 关西/ns 财政界/n 等/u 组成/v 的/u [大阪/ns 大学/n 医学部/n]nt 整备/vn 扩充/vn 委员会/n 筹集/v 捐款/n 。/w
また、構築したコンピューターネットワークは、インターネットを使って米国の大学と結び、研究者や学生同士の討論などに利用する。	此外/c ,/w 建立/v 起来/v 的/u 计算机/n 网络/n 将/d 通过/p 因特网/n 与/p 美国/ns 的/u 大学/n 实行/v 联网/v ,/w 以/c 供/v 研究/vn 人员/n 及/c 学生/n 间/f 开展/v 讨论/v 用/v 。/w
インターネットは一九六九年に米国防総省が軍の通信用に開発。	因特网/n 是/v [美国/ns 国防部/n]nt 为了/p 军队/n 通信/v 于/p 一九六九年/t 开发/v 的/u 。/w
その後、学術研究や商業用などに普及し、現在百五十カ国以上で推定約四千万人が利用。	之后/f ,/w 因特网/n 普及/v 到/v 学术/n 研究/vn 及/c 商务/n 等/u 方面/n ,/w 据/p 估计/v 现在/t 在/p 一百五十/m 多/m 个/q 国家/n 约/d 有/v 四千万/m 人/n 利用/v 因特网/n 。/w
ゴア・米国副大統領が推進役を務める情報ハイウエー構想の中心とされる。	一般/ad 认为/v 这/r 是/v 由/p 美国/ns 副/b 总统/n 戈尔/nr 主持/v 推进/v 的/u 信息/n 高速公路/n 构想/vn 的/u 中心/n 内容/n 。/w
米国の大学ではこのネットワークで積極的に学術情報を公開している。	美国/ns 的/u 大学/n 利用/v 这/r 一/m 网络/n 积极/ad 公开/v 学术/n 信息/n 。/w
井上通敏・阪大医学部教授は「インターネットを使って日本の研究者も海外から多くの情報を得ているが、情報発信は少なく一方通行。情報発信していかないと取り残されてしまう」と話す。	[大阪/ns 大学/n 医学部/n]nt 教授/n 井上/nr 通敏/nr 指出/v :/w “/w 日本/ns 的/u 研究/vn 人员/n 也/d 利用/v 因特网/n 从/p 海外/s 获得/v 大量/m 信息/n ,/w 但是/c 发布/v 的/u 信息/n 则/d 很/d 少/a ,/w 只/d 进/v 不/d 出/v 。/w 如/c 不/d 发布/v 信息/n 将/d 会/v 落后/a 。/w ”/w