日米特許対訳コーパスの自動構築

独立行政法人情報通信研究機構　自然言語グループ
内山　将夫

PROFILE 筑波大学大学院工学研究科修了。博士（工学）。現在、情報通信研究機構自然言語グループ主任研究員

mutiyama@nict.go.jp

1 はじめに

対訳コーパスとは、日本語と英語などの複数言語からなる対訳文を格納したテキストデータベースのことである。たとえば、ここで紹介する日米特許対訳コーパスには、「The present invention relates to a gas purification capability measuring method for a gas purification apparatus of a getter system used in the manufacture of semiconductor devices and the like, and the gas purification apparatus.」と「本発明は、ゲッタ方式のガス精製装置に関するもので、特に半導体装置等の製造に用いられるガス精製能力計測手段付きのガス精製装置に係るものである。」のような対訳文が含まれている。

対訳コーパスは、様々な用途に利用できる非常に重要な言語資源である。たとえば、人間の翻訳者がある文を訳したいときに、そこに出てくる単語の訳に自信がないときがある。そのような場合には、対訳コーパスにおける当該の単語の訳され方を調べることにより、コーパス中の対訳文に即した訳をすることが可能である。また、対訳コーパスから半自動的に対訳辞書を作成することもできる。さらに、最近では、コーパスに基づいた機械翻訳の研究が盛んであり、そこでは、対訳コーパスが与えられれば、自動的に、機械翻訳システムを作ることを目標とした研究が行われている。

このように、対訳コーパスは有用な言語資源であるが、これまでのところ、日英についての大規模な対訳コーパスは存在していなかった。そのため、日英に関して、対訳コーパスに基づく大規模な研究をすることは、不可能だった。

このような背景の中で、情報通信研究機構自然言語グループでは、日米の特許から大規模な日英対訳コーパスを作成した。これにより、日英について、対訳コーパスに基づく大規模な研究を行うことが可能になった。

2 対訳コーパスの作成に利用したデータ

特許の対訳コーパスを作成する元データとしては、国立情報学研究所主催のNTCIR-6の特許検索タスクに利用された次のデータを用いた。

（1）日本公開特許公報全文 1993-2002年発行分
（約350万件）

（2）米国特許全文 1993-2000年発行分
（約100万件）

これらの特許データから、日本と米国に同時出願された特許を、米国特許に記述されている優先権番号により

同定した結果として、84677件の特許の対応を得た。これらを観察した結果として、特許の「実施例」と「背景」の記述が、日米について、比較的直訳されていることが分かったので、これらの部分を利用して対訳コーパスを作成することにした。

そのために、簡単なパターンマッチングにより、これらの対応を抽出したところ、「実施例」については77014件、「背景」については　72589件の対応を得た。以下では、これら149603件から文対応データを作成したときの方法について述べる。なお、以下では，「実施例」と「背景」を共に「文書」と呼ぶ。

3 対訳文の対応スコア

対訳文対応は、対訳文書中の日英の文をプログラムにより対応付けることによりなされる。対応付けにおいては、日本語文と英語文との類似度を、EDR日英辞書を利用して求め、そのような類似度の総和が最大となるような文対応を求めている。最終的な文対応のスコアは、日本語文と英語文の類似度に加えて、それら対訳文を含む対訳文書の類似度も顧慮して付与する。

4 文対応の抽出

このような文対応のスコア付けを、149603件の全文書に対して適用した結果、約700万の文対応を得た。これらの文対応から、1対1の文対応のみを抜き出すと、約420万である。これら1対1の文対応から、ノイズを低減するために、日本語文が句点で終っていないものを除去した。また、重複する文対応は一つを除き除去した（日英文が共に同じ場合に重複とした）。この結果として、約390万の1対1文対応が得られた。

これらの文対応をスコアの降順にソートした。次に、これらの上位を利用して対訳コーパスを作成したのだが、このときに、どの程度の順位までの文をコーパスに採用するかを決めるために、まず、1999981～2000000位の20文における文対応の品質を調査した。その結果は、17文が日英がほぼ直訳の対応をしており、2文が50%以上の内容の重複があった。次に、2499981～2500000位の20文を調べたところ、13文が互いに直訳の関係にあり、6文が50%以上の重複があった。この結果より、我々は、上位の200万文を対訳コーパスに採用した。更に、これらから、日本語文か英語文のどちらかが100単語を越えているものと、日本語文と英語文の長さが大きく違うものを除去した。その結果として、1988732文が得られた。

これらの約200万の文から訓練、開発、開発テストおよびテストデータを作成した。なお、訓練データとは、機械翻訳システムの基本的な部分を作成するために利用する大規模なデータであり、開発、開発テストデータとは、機械翻訳システムの微調整に利用するデータであり、テストデータとは、機械翻訳システムの性能をテストするときに利用するデータである。

これらの作成方法は以下の通りである。まず、特許文には重複が多いことから、同一特許に含まれている文が、訓練とテストに出現することは避ける必要がある。そのために、この約200万の文対応を、同一の特許に属するものは一つのグループにした。そして、このグループを単位として、全体を、91%を訓練、3%を開発、3%を開発テスト、3%をテストに無作為に分割した。

その結果、約180万文の訓練データが得られた。また、開発、開発テスト、テストデータについては、これらからそれぞれ無作為に抽出した1000文について、後述する対訳評価がA判定のものを選び、それぞれをDEV、DEVTEST、TEST とした。それぞれの文数は、DEVが916、DEVTESTが927、TESTが899である。

これらのDEV、DEVTEST、TESTを選択するために、翻訳業者に、2段階からなる判定作業を依頼した。第1段階では、マッチングの評価として

（A）　文全体としてマッチ

（B）　半分くらいマッチ

（C）　文は、ほとんどもしくは全く対応していない

を評価した。次に、対訳評価として、英語文が日本語文の内容を正しく反映しているかを次のように評価した。

（A）　（ほぼ）完璧

（B）　8割程度は正しい

（C）　精度はB未満

（X）　A，B，C以外

これらの評価結果を表1に示す。表の「マッチング」の列より、97～98%の文対応が、文全体として一致していることがわかる。また、「対訳」の列より、90～93%の文対応が英語文が日本語文の内容をほぼ完璧に反映していることがわかる。これらより、抽出した文対応データは、機械翻訳のデータとして、十分使える品質であると言える。

5 日米特許対訳コーパスの利用

このようにして作成された日米特許対訳コーパスは、NTCIR-7の特許翻訳タスクにおいて、コーパスベースの機械翻訳システムの開発のために提供される予定である。

また、平成19年8月29～30日には、この対訳コーパスの一部を利用して、「統計的機械翻訳による特許文翻訳に関する講習会」が開催された。この講習会は、筑波大学山本幹雄准教授、藤井敦准教授、宇津呂武仁准教授、および筆者が主催したものであり、国内における統計的機械翻訳の発展を目的として、学生や研究者を対象として講義や実習を行ったものである。実習では、英文

表1　途上国への情報化協力

	マッチング			対訳			
	A	B	C	A	B	C	X
DEV	974	23	3	916	57	24	3
DEVTEST	983	16	1	927	56	16	1
TEST	973	24	3	899	72	26	3

特許を日本語に機械翻訳するという課題が行われた。

我々は、今後、この日米特許対訳コーパスが、特許の機械翻訳の研究だけでなく、対訳コーパスを利用した一般の研究に有効に活用されることを期待している。