4-2　対訳データの効率的な構築方法

4-2 Efficient Technologies for Creating Parallel Corpora

内山将夫
UTIYAMA Masao

要旨

大規模な対訳コーパスは、コーパスベースの機械翻訳にとって、必須の言語資源である。しかしながら、それらを作成することは非常に困難である。本稿では、大規模な対訳コーパスを作成可能な2つの方法を述べる。1つめの方法は、既存の日本語と英語の翻訳テキストから、対訳文を抽出する手法である。2つめの方法は、既存の対訳テキストを収集するのではなくて、ボランティア翻訳者の翻訳を支援することにより、新規の対訳を直接獲得する方法である。これらの手法を併用することで、包括的な対訳コーパスを作成可能である。

A large scale parallel corpus is essential language resource for corpus-based machine translation. However, it is very difficult to construct them. This paper discusses two methods for creating parallel corpora. The first is extracting parallel sentences from existing Japanese-English parallel documents. The second is supporting volunteer translators for creating new translations in order to obtain new parallel sentences, instead of gathering existing translations. These methods enable us to develop comprehensive parallel corpora.

[キーワード]
機械翻訳，言語資源　対訳コーパス，ボランティア翻訳者，みんなの翻訳
Machine translation, Language resources, Parallel corpora, Volunteer translator, Minnano Hon'yaku

1　はじめに

大規模な対訳コーパスは、コーパスベースの機械翻訳にとって、必須の言語資源である。しかしながら、それらを作成することは非常に困難である。本稿では、大規模な対訳コーパスを作成可能な2つの方法を述べる。これらの方法は、言語を問わずに適用可能であるが、本稿では、日英対訳コーパスに適応した場合について述べる。

1つめの方法は、既存の日本語と英語の翻訳テキストから、対訳文を抽出する手法である。2つめの方法は、既存の対訳テキストを収集するのではなくて、ボランティア翻訳者の翻訳を支援することにより、新規の対訳を直接獲得する方法である。

2　既存の日本語と英語の翻訳テキストからの対訳文の抽出

日英対訳テキストの例としては、たとえば、日本語で書かれた新聞記事と、それを英語に翻訳した新聞記事がある。しかし、一般に、対訳テキストは直訳関係にあることは少ない。たとえば、この場合には、英語の新聞記事は、英語のスタイルで書かれているので、たとえ日本語新聞記事の翻訳であっても、文字通りの翻訳ではなく、なんらかの変更が含まれているのが普通である。

その他のノイズの多い対訳テキストの例には、日本と米国に同時に出願された特許がある。これをパテントファミリーという。パテントファミリーは大量に存在するため、ここから大規模対訳コーパスが獲得できる。また、このようにして作成された対訳コーパスは、特許翻訳用の自動翻訳

エンジンを作成するためにも利用可能である。実際、NTCIR-7、8、9、10の特許翻訳タスクにおいては、本稿の方法で作成された対訳コーパスが利用されている[1][2]。

2.1 対訳コーパス作成に利用したパテントファミリー

本稿では、NTCIR-6の特許検索タスク[3]で利用された特許データから対訳コーパスを作成する。そのデータは

- 1993–2002年の日本特許公開公報約350万文書
- 1993–2000年の米国特許約100万文書

からなる。

これらから、米国特許に記述されている優先権番号に基づいて、84,677のパテントファミリーを得た。これらのパテントファミリーを調べたところ、「発明の詳細な記述」と「発明の背景」の部分が、直訳されていることが多かった。そのため、これらの部分から対訳コーパスを作成することにした。

抽出されたパテントファミリーに、単純なパターンマッチを適用して、合計で149,603の「発明の詳細な記述」と「発明の背景」の部分を得た。これらを以下では文書と呼ぶことにする。

2.2 文アライメント

本稿では、内山・井佐原[4]の方法を利用して対訳文アライメントを実行した。その手順は以下の通りである。まず、標準的な動的計画法[5]を利用して各対訳文書における文アライメントを実行した。このときに、各文対J(i)とE(i)の類似度として以下を利用した[4]。

$$\mathrm{SIM}(\mathrm{J(i)},\ \mathrm{E(i)}) = \frac{2\times\sum_{j\in J(i)}\sum_{e\in E(i)}\frac{\delta(j,e)}{\deg(j)\ \deg(e)}}{|J(i)|+|E(i)|}$$

ただし、jとeは文中の日本語単語と英語単語であり、

|J(i)|はi番目のアライメントにおける日本語単語の数

|E(i)|はi番目のアライメントにおける英語単語の数

$\delta(j,\ e)$はjとeが対訳単語であるときに1、そうでないときに0

deg(j) = jと対訳関係にある英単語の数

deg(e) = eと対訳関係にある日本語単語の数

である。この類似度を利用して最適なスコアの文アライメントを動的計画法により獲得したあとに、日本語文書Jと英語文書Eの類似度を以下のように計算する[4]。

$$\mathrm{AVSIM(J,E)} = \frac{\sum_{i=1}^{m} SIM(J(i),E(i))}{m}$$

ただし、(J(1), E(1))、(J(2), E(2))、…(J(m), E(m))は、動的計画法により得られた文アライメントである。AVSIM(J, E)が高い値となるのは、JとEに含まれている文アライメントの類似度が高い場合である。この場合には、文書JとEの類似度も高いと考えられる。

また、文書EとJにおける文数の比もスコアとして利用した。すなわち、

$$\mathrm{R(J,E)} = \min\left(\frac{|J|}{|E|},\ \frac{|E|}{|J|}\right)$$

ただし、|J|と|E|は、それぞれ、文書JとEの文数である。この値は、2つの文書の文数が近いときに高いスコアとなる。

これらを総合して、文J(i)とE(i)のスコアを以下のように定義した。

$$\begin{aligned}\mathrm{Score(J(i),E(i))} = {} & \mathrm{SIM(J(i),E(i))} \\ & \times\ \mathrm{AVSIM(J,E)} \\ & \times\ \mathrm{R(J,E)}\end{aligned} \qquad (1)$$

このスコアは、文の類似度および文書の類似度が高いときに高くなる。

2.3 対応精度の高い文アライメントの抽出

2.1で抽出された149,603の対訳文書について、上述の手法を適用した結果として、約700万文の文アライメントが抽出された。ここから、1文と1文が対応している文アライメントのみを選ぶと約420万文であった。また、これから日本語文が句点で終わっているなど、適切な文である可能性の高い文アライメントを抽出すると390万文であった。

更に、上位200万文付近の文を20文調べたところ17文がほぼ文字通りの翻訳であり、上位250万文付近を調べたところ13文がほぼ文字通りの翻訳であった。これらに基づいて、上位200万文の文アライメントコーパスを対訳コーパスとして抽出した。そこから更に100単語以上の文と日英の単語数が大きく異なる文を除くと合計で

1,988,732 文が抽出された。

このようにして抽出された文の対訳としての精度を調査するために、無作為に 1,000 文を抽出した。そして、次の 2 段階により精度を調査した。まず、最初のステップでは、各文アライメントについて、（A）日英文が全体的に一致している、（B）日英文の 50％以上が一致している、（C）それ以外の 3 段階で人手評価した。その結果、（A）が 973、（B）が 24、（C）が 3 であった。

次のステップでは、各文アライメントについて、（A）日英文がほぼ完全に意味的に一致している、（B）日英文が 80％程度一致している、（C）日英文が 80％以下一致している、（X）それ以外、と人手評価した。その結果、（A）が 899、（B）が 72、（C）が 26、（X）が 3 であった。これより、抽出された対訳コーパスのアライメント精度は高いといえる。

次に、これら 1,000 文について、式（1）のスコアと上記の人手評価との関係を図 1 に示す。図 1 は、スコアの降順にならべられた文アライメントの順位と B、C、X の累積数の関係を示している。実線が示すとおり、ノイズとなる文アライメントは下位のランクである。なお、もし、ノイズとなる文アライメントがスコアと関係なく分布している場合には、点線のように分布する。これより、式（1）のスコアは、適切な文アライメントに高いスコアを与えているといえる。

2.4 機械翻訳実験

次に、上記の 390 万文の対訳文から更にノイズと考えられる対訳文を除去した結果の 350 万文を利用して、訓練に利用した対訳文の総量と翻訳精度の関係を調査した。翻訳実験には、Pharaoh デコーダ [6] を利用した。また、翻訳精度は％ BLEU [7] で評価した。

本実験は、訓練文数のみを変えて％ BLEU の評価をするという観点から、以下の 3 点については、同一条件とした。すなわち、（1）対訳文データにおける単語アライメントの情報、（2）言語モデル、（3）翻訳モデルや言語モデルの重み、（4）テストデータ。そして、（1）の対訳文データの量を変更させての実験をした。

これらの共通の設定は以下のようにして得た。（1）単語アライメントは GIZA＋＋ [8] を利用して計算した。（2）3-グラム言語モデルを 350 万文から構築した。（3）あらかじめ、200 万文の対訳データで訓練したモデルを開発データでチューンして得た重みを、全ての訓練データ量について共有した。（4）テストデータには全ての訓練データで共通の 2,000 文を利用した。

このような条件下で、訓練データ量を、50 万、100 万、150 万、200 万、250 万、300 万、350 万文と変化させて、そのときの％ BLEU を求めた。その結果が図 2 である。

図 2 より、英日方向の機械翻訳では、250 万文付近で％ BLEU 値の伸びが鈍化している。また、日英方向では、300 万文までは％ BLEU が伸びているが、350 万文では、％ BLEU が低下していることがわかる。

これより、日英・英日双方向ともに、ある量ま

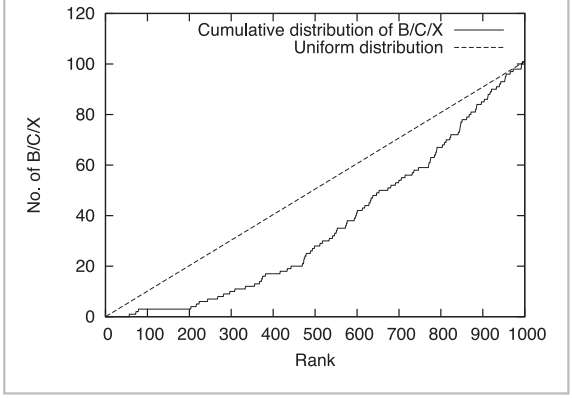

図 1 文のスコアの順位と文アライメント B、C、X の数

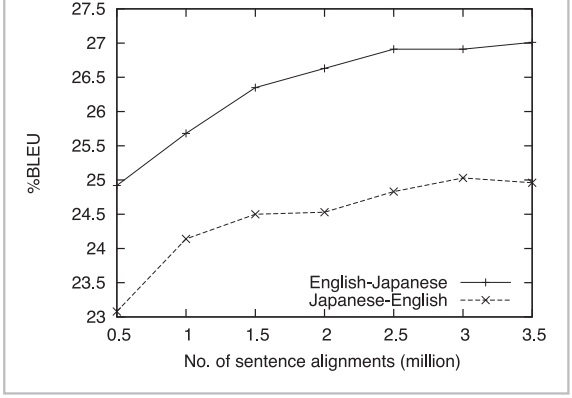

図 2 文アライメントの訓練文数と％ BLEU の関係

では、対訳文を増やすと翻訳精度が向上するが、一定量を超えると翻訳精度の伸びが鈍化したり低下するといえる。これより、式（1）のスコアは、適切な文アライメントに高いスコアを与えているといえる。

3　翻訳支援による対訳文獲得

2 で述べた方法は、既存の対訳テキストから対訳コーパスを作成する方法である。一方、本節では、ボランティア翻訳者を支援することにより、対訳コーパスを獲得する方法を述べる。

ボランティアの翻訳者を支援するツールとしては、既に、翻訳支援用のエディタである QRedit [9] など、様々な翻訳支援用のツールがある。しかし、これらのツールの支援対象は個々の翻訳者であるので、世の中の多数の翻訳者を一度に支援することはできない。

そこで、NICT MASTAR プロジェクト多言語翻訳研究室は、東京大学図書館情報学研究室と共同で、世の中の多数の翻訳者を一度に支援するためのシステムを構築することを考えた。具体的には、我々のプロジェクトにおいては、ボランティアの翻訳者をホスティングすることを目標とし、そのための Web サイト（http://trans-aid.jp）を構築した。図 3 は、我々のプロジェクトで構築した「みんなの翻訳」サイトのスクリーンショットである。

ホスティングにより、世の中の多数の翻訳者を支援できると考えた理由は以下のものである。

図 3　「みんなの翻訳」サイト（http://trans-aid.jp）

（1）まず、他分野における成功例として、オープンソースの世界においては、sourceforge.net のように、オープンソースプロジェクトをホスティングすることにより、オープンソースの開発や普及を促進している例がある。そのため、ボランティアの翻訳者をホスティングすることにより、同様な成功が望めるのではないかと考えた。

（2）次に、文献 [9] からわかるように、オンラインのボランティアの翻訳者は、翻訳支援ツールをあまり使っていない。一方、みんなの翻訳サイトにおいては、QRedit 等を組込むことにより、みんなの翻訳サイトの利用者が、自然に、翻訳支援ツールを使える環境を提供する。これにより、利用者が、特に意識せずとも、翻訳支援を受けることができる。

（3）また、みんなの翻訳サイトにおいては、ボランティアの翻訳者が翻訳したテキストは、原文と共に保存されているので、みんなの翻訳の利用者は、自分の翻訳結果だけでなく、他の翻訳者の翻訳結果を共有できる。そのため、他人の翻訳を自分の翻訳に利用できる。

（4）最後に、みんなの翻訳サイトにおいては、ボランティアの翻訳者が翻訳したテキストは、サイト上に保存され、みんなに公開されるので、自分の翻訳を公開したい人にとっては、公開する場所を提供することになる。

以上の理由により、もし、多数のボランティアの翻訳者がみんなの翻訳を利用すれば、翻訳支援ツールの提供や、翻訳の共有と再利用という利点により、多数のボランティアの翻訳者をサポートできると考えた。

みんなの翻訳の特徴は、（1）高機能な翻訳支援エディタ QRedit を誰もが利用できることと、（2）みんなの翻訳で公開されている翻訳には、「2 次的著作物を作成し、それを公開しても良い」というライセンスが付与されているため、適切な使用であれば、翻訳を利用できるということと、（3）三省堂の協力により「グランドコンサイス英和辞典（36 万項目収録）」が翻訳支援に利用できることである。

3.1　高機能な翻訳支援エディタ QRedit

翻訳支援エディタ QRedit の基本設計理念は、以下の 4 点に集約される。(1) 新たな情報・機能を提供するのではなく、翻訳者が現に行っている作業の手間を省く、(2) システムが決めるのではなく翻訳者が決めるのに必要な情報を提供する、(3) 翻訳者の発想を豊かにする情報を表示する、(4) できるだけシンプルにする。これらの方針は、翻訳者へのインタビューおよび現状の翻訳支援技術の水準に基づいて決められた。

QRedit では、入力された原文に対し、複数の辞書や翻訳者が登録した用語を対象に辞書引きを行い、翻訳者は単語をクリックすることで簡単にその訳語を把握することができる。また、高度なイディオム検出機能を備えていて、語間に挿入を許した異形イディオムを検出できる。

これらのイディオムに対して、QRedit では、図 4 のように、下線を引いて示すなどして、翻訳者が見落さないようにしている。イディオムは、熟練翻訳者でも誤訳する可能性があるので、このように、エディタ側から何らかの警告を与えることは、有用である。

その他にも、QRedit では、Web 検索ができるとか、用語を登録できるとかの機能がある。特に、用語を登録すると、その用語を QRedit 内から検索できる。この検索は、自分が登録した用語だけでなく、他の人が登録した用語も同様にできるので、みんなの翻訳で公開されている用語が増えれば、辞書に載っていない用語であっても、辞書引きができるようになる。

3.2　翻訳の共有

翻訳結果を共有するためには、原文と翻訳文の使用許諾について考慮する必要がある。たとえば、当然であるが、原文の著者が翻訳文の公開を許可していない場合には、翻訳文は公開できないので、翻訳結果を共有することはできない。

そのため、みんなの翻訳の利用者には、原文と翻訳文の使用許諾について確認を求めている。また、みんなの翻訳の利用者には、各自が翻訳した文は、2 次的利用ができるように許可することを求めている。そのために、システムは、みんなの翻訳の利用者が翻訳文を保存するときに、以下のようにして、使用許諾などを確認している（図 5）。

(1) まず、システムは、「あなたが翻訳の対象とした文書（原文）は、原文著者が明示的に許可している場合を除いて、私的な利用その他など、著作権法で認められている範囲でしか利用できません。原文著者は、その翻訳を公開しても良いと（あなたや他の人に）許可をしていますか？」と確認する。(2) それが「はい」の場合には、システムは、クリエイティブ・コモンズ等の「あなたの文書から 2 次的著作物を作成し、それを公開しても良い」という条件に矛盾しない使用許諾条件を設定してもらうようにしている。

このようにして、みんなの翻訳では、原著者や翻訳者の著作権を尊重しつつ、翻訳を共有できる仕組を準備している。

3.3　今後の展開

みんなの翻訳は 2009 年 4 月 8 日に一般公開した。2012 年 6 月時点のユーザは 2,100 人以上、

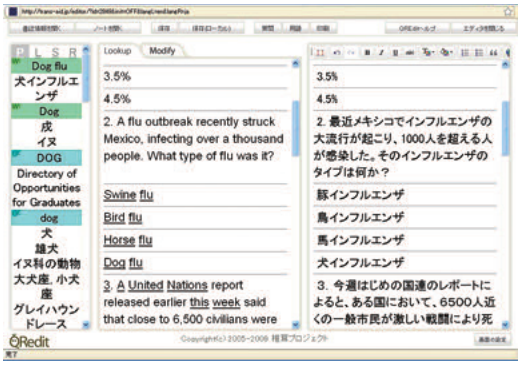

図 4　翻訳支援エディタ QRedit

図 5　文書の使用許諾権の設定

翻訳された文書数は約 10,000 文書である。また、日英中 3 言語間、および、英カタラン、日独、日韓の翻訳を支援している。今後は、みんなの翻訳の教育利用を重点的に研究することにより、翻訳者に早い段階からみんなの翻訳に親しんでもらって、より多くの人に使ってもらうようにしたい。

4 おわりに

本稿では、対訳データを作成する技術として、既存の翻訳テキストから対訳テキストを抽出する手法と、翻訳支援により、新規の対訳テキストを獲得する手法を述べた。これらの手法を併用することで、包括的な対訳コーパスを作成可能である。

参考文献

1 Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro, "Overview of the Patent Translation Task at the NTCIR-7 Workshop," Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp. 389–400, Dec. 2008.

2 Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya, and Sayori Shimohata, "Overview of the Patent Translation Task at the NTCIR-8 Workshop," NTCIR-8, pp. 371–376, 2010.

3 Atsushi Fujii, Makoto Iwayama, and Noriko Kando, "Overview of the Patent Retrieval Task at the NTCIR-6 Workshop," Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, pp. 359–365, May 2007.

4 Masao Utiyama and Hitoshi Isahara, "Reliable measures for aligning Japanese-English news articles and sentences," Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics, 2003.

5 William A. Gale and Kenneth W. Church, "A program for aligning sentences in bilingual corpora," Computational Linguistics, 19(1): 75–102, 1993.

6 Philipp Koehn, "Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models," Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA), pp. 115–124, 2004.

7 Kishore Papineni and Salim Roukos and Todd Ward and Wei-Jing Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), 311–318, 2002.

8 Franz Josef Och and Hermann Ney, "A Systematic Comparison of Various Statistical Alignment Models," Computational Linguistics, Vol. 29, No. 1, pp. 19–51, 2003.

9 Abekawa Takeshi and Kageura Kyo, "QRedit: An integrated editor system to support online volunteer translators," Digital humanities, pp. 3–5, 2007.

（平成 24 年 6 月 14 日 採録）

内山将夫（うちやまま さお）
ユニバーサルコミュニケーション研究所
多言語翻訳研究室主任研究員
博士（工学）
自然言語処理、機械翻訳
mutiyama@nict.go.jp