2007－ICS－148 (3)
2007／7／4

ＷＥＢ情報を基づく英日訳語選択

呉 浩東

獨協大学
国際教養学部

あらまし 本稿では、WWW ホームページのテキストを言語資源として利用する新しい訳語選択方法を提案する。まず、統語関係を持つ訳語候補を対訳辞書から取り出し、その組み合わせをウェブ上から各種の頻度情報を抽出して連想度の推定する式を定義してスコアを計算する。さらに、訳語選択のアルゴリズムを用いてもっとも適切的な訳語を決定する。一意的に訳語選択できない場合、シソーラスの同義語を用いて選択の信頼性を高める。ここで提案したモデルは、文内文脈情報の利用に有効であり、横断的言語検索にも利用することも期待できる。
キーワード 訳語選択、 統語関係、ウェブ情報、シソーラス

English-Japanese Translation Word Selection Model Using Web Information

Kotoh GO

School of International Liberal Arts, Dokkyo University

Abstract We in this paper present a novel model for using text information from the web for translation word selection. Using this approach, we collect word candidates in some syntactic relation using a parallel dictionary and compute the scores that show various possible compounds of word pairs in the target language. We design an algorithm and use it to determine the most likely word candidate as the correct one for translation. If the translation can not be determined as the word does not occur frequently, we will replace it by its synonym in thesauri. This approach is efficient to use contextual information in sentence and is expected to be useful in cross-language information retrieval.

Keyword Translation Word Selection, Syntactic Relation, Web Information, Thesaurus

1. はじめに

機械翻訳において、原言語の意味を正しく目的言語に再現するためには、原言語表現の意味に適した訳語の選択が必要である。例えば、英語の名詞の sentence を日本語に訳す場合、「文」に訳すか「判決」に訳すか二つの可能性があり、sentence のまわりに現す単語群（すなわち、文脈）を手掛かりとして訳語を決める。

近年、インターネットでの情報交換の普及に伴い、ホームページの翻訳結果が簡単に得られる

WWW用の翻訳ソフトウェアが大量に登場によって、情報の受送信の支援道具として機械翻訳に対するニーズと期待感が急速に高まっている一方、現在の機械翻訳システムにおいて訳質の点で改良の余地が大きいも明らかである。この訳質の飛躍的に向上を実現するため、訳語選択の精度の改善に大きく関係している。

本稿では、英語・日本語機械翻訳の際、ウェブから巨大なデータからより信頼性の高い共起情報を得て翻訳規則の抽出することによって文脈を考慮した訳語選択の方策を論究する。

2. 機械翻訳からより自然な訳文を得るために

現在の機械翻訳ソフトの大半は形態素解析と構文解析技術を利用する変換型機械翻訳方式を採用している。文脈情報や意味情報を無視しているため、英日・日英機械翻訳システムなどの訳質はまた十分に満足できない、改善余地がまた非常に大きい。たとえば、英文"I visited a patient at the hospital."を複数な機械翻訳ソフトに入力してみると、「私は病院で患者を訪問した。」と「私は病院で患者を訪ねました。」の日本語訳しか得られない。ここで、visit の訳語をもっとも使われている「訪問する」か「訪ねる」を選択する。動詞 visit のまわりの統語的な関係を持つ単語（patient）や句（at the hospital）を無視されている。ここで、われわれは動詞共起パターンを定義し、統計的な推定により別の訳語を選択する可能性がある。たとえば、"I visited a patient at the hospital."を「私は病院で患者を見舞った」という自然な訳文、すなわち、人間の判断に一致する、意味を通す訳文の生成を目指す。

3. コーパスとしての WEB 資源

訳語を決めるのは、文脈情報の活用ために頻度の高い統計データの確保が重要である。そこで、有効な手掛かりを獲得するために膨大なコーパスを必要である。翻訳の精度を向上するために、コーパスのサイズが大きければ大きいほどよいである。ここで、三つの問題に直面している。

1) 現在の段階において、日英二言語コーパス（コンパラブルコーパス）のサイズがまた十分に大きいと言えない。出現頻度の少ないタームについては、訳語候補を列挙するにとどまり、訳語候補の有効な順位付けが難しいという問題が残る。
2) 大規模コンパラブルコーパスの開発にはコストが非常に高く、その入手も極めて困難である。
3) コーパスにある言葉は十分に新しいとは限らない。常に最新の用語を含んだ数百万文以上のものも収録されたコーパスを更新し配布し続けることは非現実的である。

一方、ウェブページに、統計処理を行うのに十分な大きなテキストがあり、常に最新な状態に保たれていて、どこからいつでも利用できる。インターネット全体を見れば一つの巨大なコーパスを構成していると言える。これらのページの多くが頻繁に更新されており、毎日新しいページも次々と登場する。最新の用語やその使用法が常に反映されている。文章の質にはばらつきがみられるが、誤用やミスなどのケースも珍しくないとは言え、統計上の処理に工夫すれば十分に対処できる。

ウェブ資源を用いて最も簡単な言語処理はスペルチェックと表記ゆれチェックである。たとえば、*thesaurus* と *thesaulus* のどちらが正しいか、Google を用いて調べ、前者は 7840 万回に対して、後者がわずか 81 回であった。頻度情報から正解が得られる。

ウェブを自然言語処理に利用するも一つ重要な理由は、世界中の大多数の言語は実用に使えるサイズを持つコーパスは存在しておれず、二言語コーパスはさらに稀である。

以上の視点から、WWW は翻訳知識を獲得するための有効な言語資源の一つと言える。

4．訳語選択モデル

本稿では、翻訳を実施する際、文から適切的な訳語を確定するために、統語関係を持つ単語を一緒に抽出し、それぞれの訳語候補の組み合わせの中から適切なものを選択する。

特に役立つ単語間統語関係は、「形容詞＋名詞」や複合名詞や「述語＋名詞」などがあげられる。以下は動詞と名詞、形容詞と名詞の統語関係を動詞共起パターンとして抽出し、訳語選択の仕組みを述べる。

4.1 共起パターン

4.1.1 動詞と名詞の共起パターン

動詞は文の意味を決定するのはもっとも重要な品詞である。動詞は複数の語義を持つ場合が多く、一つの語義に一つの動詞共起パターンを対応させる。ここで、動詞共起パターンは原言語の動詞のまわりに統語関係を持つ名詞や前置詞句から、目標言語にそれぞれの語義の組み合わせである。さらに、ウェブ上から得た統計情報によって、適切な動詞共起パターンを抽出し、多義性をもつ動詞の訳語を決定する。

ここでは、動詞 visit を例として訳語選択の仕組みを説明する。

英和辞書によると、【動】visit は以下の日本語動詞に対応する。

. 訪問する
. 訪れる
. 見舞う
. 見学する
. アクセスする
. 滞在する

そして、visit の文内の統語関係もつ単語を一緒にとる。これから下記の例を用いてわれわれの方法を説明する。

1) visit a dentist
2) visit New York
3) visit the homepage
4) visit a company
5) visit a museum
6) visit a patient

dentist(歯医者)、New York(ニューヨーク)、homepage(ホームページ)、 company（会社）、museum(博物館)は述語 visit の目的語である。動詞の訳語候補と名詞との共起パターンは以下の例の通りである。

1)【名詞】を訪問する
2)【名詞（人間、場所)】を（に）訪れる
3)【名詞（人間)】を見舞う
4)【名詞】を見学する
5)【名詞】にアクセスする
6)【名詞（場所)】に滞在する

4.1.2 形容詞と名詞の共起パターン

形容詞と名詞の共起パターンは通常以下の二種類が挙げられる。

（1） 形容詞＋名詞
（2） 名詞＋to be＋形容詞

4.2 訳語選択モデルの設計

本論文で提案する訳語選択モデルの構成を説明する。以下では、源言語Sの用語を、文脈に一致する目標言語Tに翻訳する場合を考える。

共起頻度が非常に低い場合、誤用や入力のミスや非ネイティブ者による乱用現象が考えられる。排除しないと安全性を確保できない。こいうケースは以下のように対処する。

$$S(t_1,t_2) = \log_2 P\frac{f(t_1,t_2)^2}{f(t_1)f(t_2)} \qquad (1)$$

最大共起頻度に比べると非常に低い場合、以下の三条件のどちらを満たす動詞共起パターンを除外し、対応する訳語共起頻度を 0 にする。

条件1： 最大共起頻度は 30～500、かつ、共起頻度 < 最大共起頻度/100
条件2： 最大共起頻度は 500～10000、かつ、共起頻度 < 最大共起頻度/500
条件3： 最大共起頻度>10000、かつ、共起頻度 < 最大共起頻度/3000

日本語ページ数が現時点 Google（http://www.google.com/）で調べたものである。訳語を推定するために、各共起パターンを下記の式で点数化する。式の中に、P は日本語ページ数である。T は term の意味であり、単語や句を指す。 f は t をウェブに現す頻度である。訳語の候補を選択するために、下記の式で各候補の尤度を計算する。

動詞 visit を例とすると、式 1 を用いて共起パターンのスコアを示す。-8 は visit の目的語に対して、動詞の n 番目の訳語選択がありえないことを意味する。たとえば、“ホームページを見舞う”、“博物館をアクセスする”は通常に使えないものを示唆する。表の一行において、スコアが高い方は選択の優先度が高いことになっている。例えば、“visit homepage”の優先順位は「ホームページをアクセスする」、「ホームページを訪ねる」、「ホームページを訪問する」になる。訳語を選択するとき、スコアが 0 以下あるいは選択される可能性は最高のものより 1 /10000 以下になっている訳語対を選択候補から除外することを提案する。具体的に、ある訳語対のスコアは順位一番の訳語対のスコアより 13.29 低い場合に適用する。そうすると、「患者をアクセスする」、「患者を（に）

訪ねる」を除外する。

4.3 訳語選択アルゴリズム

訳語選択の手順は以下のアルゴリズムにする：

STEP 1: 原言語の単語を対訳辞書から訳語候補を抽出する。
STEP 2: 式１を用いて統語関係を持つ訳語対のスコアを計算する。
STEP 3: スコア高い順で候補リストに入れる。
STEP 4: 最高なスコアより値は 13.29（10000 倍の差を意味する）低い場合、候補リストから削除する。
STEP 5: 候補リストに訳語候補が一つの場合、訳語として選択する後、STEP 6 へ；さもなければ、STEP 7 へ。
STEP 6: 共起パターンは対訳辞書に登録されていない場合。共起パターンの二言語表現を対訳辞書に登録してから終了する。
STEP 7: 二つ以上の訳語候補からもっともふさわしい訳語を特定する。他の訳語候補のスコアと最高スコアの差>6.65 の場合、最高のスコアも持つ訳語候補を訳語に決める。さもなければ、主語や副詞句や前置詞句と動詞の共起関係の連想度をウェブ上に調べ、式１と式２を用いてスコアを計算し、二つのスコアを総合比較する上訳語を決める。

4.4 機械翻訳のための言い換え

構文構造の複雑さと語彙の曖昧さに対処することは従来の機械翻訳システムの難点である。ここでは、機械翻訳の前編集と後編集において、換言処理によって、曖昧な文型表現を簡単なものに分解することと、訳語選択の複雑さを軽減させるために単語を入れ替えるなどの対処を行う。まず、語彙レベルの同義語の言い換えは、訳語の候補を減らすことに有効である。例えば、

(1) 妻 ⇔ 家内 ⇔ ワイフ
(2) 入手する ⇔ 手に入れる

語彙的な言い換えは、シソーラスから簡単に得られる。
さらに、構文レベルの言い換えも訳語の選択に役立つ。

(3) s. Bill sold Tom a car.
t. Bill sold a car To Tom.
(4) s. 花子が太郎に愛されている。
t. 太郎は花子を愛している。
(5) s．いい　です　か？
t. よろしい　です　か？
t. いい　でしょう　か？
t. よろしい　でしょう　か？

上記のような文型変形によって機械翻訳の出力はより単純化になり、訳質の改善につながることが可能になる。

5．あとがき

本稿では、ウェブをコーパスとして用いる訳語選択手法を提案した。この発想は現在の時点で大規模な対訳コーパスの入手が極めて困難である一方、有名なＷＥＢサイトから複数言語の対応テキストがあり、訳語の対応付けの手掛かりになることが多いからである。

この手法は機械翻訳システムの開発に有用性が持つ。また、機械翻訳支援や語学教育支援システムの開発にある程度役立つ。応用例のとして、作文支援システムに言葉の選択にある種の自由度を持たせ、適当な単語の組み合わせを提示することができるという点で、作文支援だけでなく、入れ替え型の練習問題の自動生成にも活用することも期待できる。

参考文献

1) 池野篤司、村田稔樹、下畑きより、山本秀樹：“インターネット自然言語資源を利用した機械翻訳”，沖電気研究開発，第182号，Vol.67，pp.49-52, 2000

2) Dongli Han, Haodong Wu (Kotoh Go), Teiji Furugori: “Resolving Overlapping Ambiguities and SelectingCorrect Word Sequence in Chinese Using Internet Corpus.” Journal of Natural Language Processing, Vol.8, No.3, 2001

3) Han Li, Cong Li: “Word Translation Disambiguation Using Bilingual Bootstrapping.”Computational Linguistics, Vol.30, No.1, pp.1-21, 2004

4) 宇津呂武仁、日野浩平、堀内貴志：“日英関連記事を用いた訳語対応推定”，自然言語処理, Vol.12, No.5, pp.43-69, 2005

5) Philip Resnik and Noah A. Smith: “The Web as a Parallel Corpus.” Computational Linguistics, Vol.29, No.3, pp.349-380, 2003

6) Andy Way and Nano Gough: “*w*EBMT: Developing and Validating an Example-Based Machine Translation System using the World Wide Web.” Computational Linguistics, Vol.29, No.3, pp.421-458, 2003

7) 山田節夫、山本和英、飯田仁：“「協調融合機械翻訳」における訳語選択”，言語処理学会第４次全国大会，pp.508-511，1998

8) Wessel Kraaij, Jian-Yun Nie, and Michel Simard: “Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval.” Computational Linguistics, Vol.29, No.3, pp.381-419, 2003

9) 富士秀、大倉清司、長瀬友樹：“翻訳支援における訳語信頼度”，言語処理学会第 13 回年次大会論文集，pp.388-391, 2007

10) Nano Gough, Andy Way, and Mary Hearne: “Example-Based Machine Translation via the Web.” Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002.

11) 外池昌嗣、宇津呂武仁、影浦峡、佐藤理史、阿辺川武：“ウェブを用いた専門用語翻訳支援における多様な情報源からの信頼度情報の提示”，言語処理学会第 13 回年次大会論文集，pp.400-403，2007