

既存目的言語文書からの訳語の自動学習方式

2C-6

野上 宏康 熊野 明 田中 克己 天野 真家
(株) 東芝 総合研究所

1. はじめに

機械翻訳システムにおいて高品質翻訳を達成するためには、「ユーザの好みに関わる訳語の曖昧さに対しても、システムが正しく解消できる」ことが不可欠である。例えば「display」の訳語を、「表示装置」、「表示」、「ディスプレイ」等の中からユーザに応じて正しく選択できる必要がある。この機能を実現するためには、「システムに対して、正しい訳語を指示する作業(以下「学習作業」と呼ぶ)」が不可欠である。

従来この学習作業は、ユーザ自身が行っていたが、本稿では、この作業をシステムが既存の目的言語文書のみを利用することにより自動的に行う新しい訳語学習方式について述べる。2章で、現状の訳語学習方式の概要とその問題点を指摘し、3章で、原文書の内容に関連のある既存の目的言語文書のみを利用して訳語を学習する方式について述べる。4章で、この学習方式の有効性を確認するために行った実験について述べ、その結果に基づいて本方式の有効性を考察する。

2. 従来の訳語学習方式

従来一般的に行われている訳語の学習方式は、「システムが原言語の単語に対する訳語候補を提示し、ユーザに現在の文書に適切な訳語を選択させ、その訳語をそれ以降の翻訳で優先して出力する」というものである。

この従来の訳語学習方式では、ユーザが訳語の学習作業を体系的に行えるようなユーザ独自の辞書を持たない場合には、次のような問題が生じる。

- ・訳語の学習作業を文書に依存して出現する単語に対して逐語的に行わねばならないこと
- ・少量多品種の翻訳を行う場合、その少量の文書のためにユーザが学習作業を行う事は期待できないこと

3. 目的言語文書からの訳語学習方式

3-1. 目的言語文書からの訳語学習方式

本方式は、「最初に原文書の内容に関連のある既存の目的言語の文書を用意しておき、複数の訳語候補が存在する場合に、各訳語候補が用意した目的言語文書に存在するかを検索し、存在する訳語候補を優先する」というものである。

この方式を英日翻訳の場合で具体的な例を用いて説明する。英語の原文書として図3-1に示すコンピュータのマニュアルを、この文書の内容に関連のある既存の目的言語文書として図3-2に示す日本語文書を用いて翻訳する場合を考える。例えば図3-1の文書中の「window」には、「ウインド」、「ウイン

ド」、「ウインドウ」等の訳語候補が存在する。これらの訳語はユーザの文書に依存して決まるものである。これらの各候補に対して各々が図3-2の文書に出現するかどうかを検索する。この場合は「ウインドウ」という語が図3-2の文書に出現しているので、「window」の訳語としては「ウインドウ」を優先する。また図3-1の文書中の下線を引いた単語「command」、「terminal」、「default」についても、その訳語候補が図3-2の文書中に出現している。本方式の有効性については、4章で実験結果に基づいて考察する。

3-2. 本方式の利点

本方式には、ユーザが訳語の学習作業を体系的に行えるようなユーザ独自の辞書を持たない場合でも、次のような利点がある。

- ・ユーザが、翻訳する原文書の内容に関連のある既存の目的言語文書さえ持っていれば、従来の逐語的な学習作業を全く行わなくても目的言語文書に依存した学習効果が期待できること
- ・用意する目的言語文書は翻訳された文である必要はないので用意しやすいこと
- ・ユーザによる学習作業があまり期待できなかった少量多品種の翻訳の場合にも、目的言語文書に依存した学習効果が期待できること
- ・従来の学習作業に有効であった、対訳文を用意する必要がないこと

...

When output is being sent to the terminal, you can type any characters. This command stops sending output, and displays the usual prompt. The option sets the size of the window instead of the default.

...

図3-1 原文書の例

...

XXXコマンドの代わりに、YYYコマンドを使用すると、端末のディスプレイにテキストの内容が表示されます。通常は、メッセージも表示されます。XXXコマンドは、端末の特性を決めるためと、デフォルトウインドウサイズを決めるため、別のファイルを参照しさらに、必要なフラグをあらかじめセットするために環境変数も参照します。

...

図3-2 目的言語文書の例

・実際の後編集では、直接ワープロ機能を用いて修正する機会が多いが、この結果を直接学習に使用できること

4. 実験および考察

4-1. 実験概要

3章で述べた学習方式の有効性を確認するため、原文書として「コンピュータの操作コマンドの英語マニュアル(約1,300単語)」(以下「文書1」), 正しい訳語を得るために「文書1の日本語翻訳文書」(以下「文書2」), および目的言語文書として「上記コマンドと同レベルのコマンドの日本語マニュアル(32,000文字)」(以下「文書3」)の3文書を用意した。そして「文書1」に出現する名詞の専門用語に対する訳語(各用語に対する、「文書2」での訳語)が「文書3」にどの程度出現するかを調べる実験を行った。

4-2. 実験結果

実験結果を図4-1に示す。「文書3」に正しい訳語(「文書2」での訳語)が存在した「文書1」の専門用語の中で、(1)はその専門用語に対する他の訳語候補(最大5個まで認めている。以下同様)が「文書3」に存在しなかった割合(「文書1」に現れた専門用語の総数に対する。以下同様)であり、(2)はその専門用語に対する他の訳語候補も「文書3」に存在した割合である。「文書3」に正しい訳語が存在しなかった「文書1」の専門用語の中で、(3)はその専門用語に対する他の訳語候補も存在しなかった割合、(4)はその専門用語に対する他の訳語候補が「文書3」に存在した割合である。なお、各割合は「同一単語は1回だけカウントした場合(「文書1」に現れた専門用語の総数も同一単語は1回だけカウント)」と「同一単語も別々にカウントした場合(「文書1」に現れた専門用語の総数も同一単語を別々にカウント)」の2通りで算出した。前者の割合は、実際の訳語学習作業を行う回数の観点から、後者は、訳語学習作業を行わなかった場合の翻訳品質の観点から、本方式の有効性を考察する場合に有効である。

また(1)(2)(3)(4)の具体例を図4-2に示す。図の左側に英単語の例、右側にその訳語候補の例を挙げている。太字は、「文書2」に現れた訳語、下線を引いた訳語は「文書3」に現れた訳語である。また「文書3」に正しい訳語が存在した「文書1」の専門用語の割合と「文書3」の文書量(文字数)との関係を図4-3に示す。実線は「同一単語は1回だけカウントした場合」、破線は「同一単語も別々にカウントした場合」である。

4-3. 考察

今回行った実験から以下のことが考察できる。

- ・「文書1」に現れる専門用語(名詞)の50%((1)の場合)に対して、本方式による学習効果が期待できる。しかし「文書1」に現れる専門用語(名詞)の8%((2)の場合)と11%((4)の場合)に対しては、本方式により逆に間違った学習を行ってしまう可能性がある。このうち(2)の8%に対しては、この場合には学習を行わないようにすることで、少なくとも間違った学習を行ってしまうことは回避できる。
- ・(1)の場合に、「同一単語も別々にカウントした場合」の割合の方が高くなったことより、頻繁に出現する単語ほど、本方式により学習される確率が高くなることが予想できる。これは、頻繁に出現する単語ほど目的言語文書に出現する可能性が高くなると期待できるので一般的な現象であると考えられ

る。このことは、本学習方式の学習効率が良いことを意味している。

- ・本方式に対し有効な目的言語文書の量は、今回の実験では約25,000文字程度であろう。「文書3」よりさらに関連の高い目的言語文書を用いれば、この量はさらに少なくなると考えられる。特に、原文書の途中まで翻訳が終了している場合に、残りの部分に対しそれまでの翻訳結果を、またバージョンが上がったマニュアルに対し、前バージョンの翻訳結果を用いると非常に高い訳語の学習効果が期待できる。

「文書3」に訳語が存在した「文書1」の専門用語	(1)他の訳語候補非存在	50% (59%)
	(2)他の訳語候補存在	8% (4%)
「文書3」に訳語が存在しない「文書1」の専門用語	(3)他の訳語候補非存在	31% (26%)
	(4)他の訳語候補存在	11% (11%)

図4-1 実験結果 * () はのべ数の場合

(1) terminal	: 端末, 端末装置, ターミナル
(2) signal	: 信号, 符合, シグナル
(3) screen	: 表示画面, 表示面, 画面, スクリーン
(4) character	: 文字, 字, キャラクタ

図4-2 実験で現れた英単語とその訳語の例

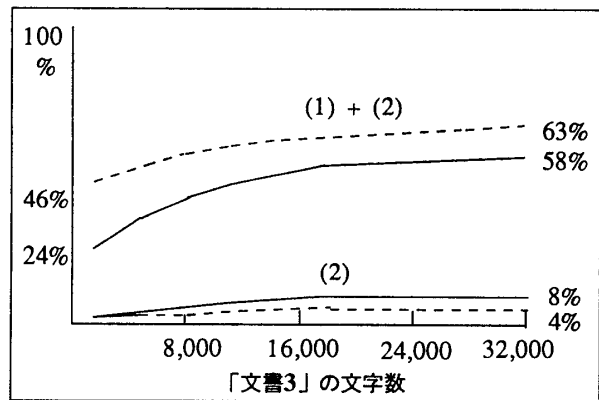


図4-3 (1)(2)の値と「文書3」の文書量との関係

5. おわりに

本稿では、既存の目的言語文書の情報を利用して、訳語を自動的に学習する新しい訳語学習方式を提案し、その有効性を実験結果に基づいて考察した。

今後さらに、複数の訳語候補が目的言語文書に存在する場合の処理方法、名詞以外の品詞に対する本学習方式の有効性、有効な目的言語文書の量とその選択方法、目的言語文書が大規模になった場合の高速な検索方法等を検討していく予定である。

参考文献

[1] 中島,他:対訳テキストを利用した訳語選択のための共起関係の自動抽出,情報処理学会第39回全国大会論文集, pp.706-707, 1989