テレコム99特集

UDC　681.39.014：519.768：681.324(100)

インターネット自然言語資源を利用した機械翻訳

Machine Translation Using the Internet Natural Language Resources

池野篤司 Atsushi Ikeno　村田稔樹 Toshiki Murata　下畑さより Sayori Shimohata　山本秀樹 Hideki Yamamoto

要　旨

本稿では，WWWホームページを自然言語処理の資源として利用する「インターネット時代の機械翻訳」について述べ，以下の２つの技術を具体例として取り上げる。(１) 複数の訳語候補から最も妥当な語を選択する際に，他の語との「相性」の良さを考慮して決定する技術。(２) 常に決まった形に翻訳することができる「常套句」を，ある言語の文書と，それと同一の内容を別の言語で書いた文書(対訳文書)から発見し獲得する技術。

1．ま　え　が　き

インターネットでの情報交換の普及に伴い，自国語での受発信に対するニーズはますます高まっている。1999年11月現在で，２億２千万のインターネット人口のうち，英語以外でのアクセスは１億人を超えるという[1)]。

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[2)]の登場によって，情報の受発信の支援技術として機械翻訳は広く知られるようになったが，同時に訳質の点で改良の余地を残していることも明らかとなった。

この訳質の飛躍的向上を実現する新方式として近年注目されているのが「用例を利用した翻訳」である。従来の方式では，文法や辞書といった翻訳規則は開発者の試行錯誤によって作り上げられるので，修正は簡単ではない。それに対して新方式は，文章を統計的に処理して翻訳規則を抽出するので，統計の元となる文章を取捨選択したり追加したりすることで容易に修正することができる。この手法の実用に際しての唯一の問題は，有効な規則を獲得するのに膨大な量の文章(コーパス)を必要とする点である。

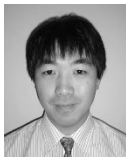
一方，インターネットではWWWで公開される文書情報を代表とする豊富な自然言語(人間が日常使用する言語)資源が日々更新・追加されている。これらの情報は，刻々と変化する語彙・使用法の現在の姿を確実に反映しているという点において，従来の翻訳辞書ではカバーできない知識を内包していると言える。

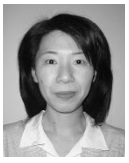
本稿では，「訳語選択技術」と「翻訳対獲得技術」の２つの技術を具体例として，この「生きた」データをリアルタイムに利用する「インターネット時代の機械翻訳」について述べる。

2．用例を利用した翻訳 - 機械翻訳の新しい流れ

機械翻訳はある言語（原言語）を解析し，別の言語（目的言語）でそれとほぼ同じ内容の文章を生成するソフトウェアである。コンピュータ科学および人工知能の分野では興味深い課題の一つであったが，50年の研究開

池野篤司
研究開発本部
マルチメディアネットワークラボラトリ


村田稔樹
研究開発本部
情報サービスシステムラボラトリ


下畑さより
研究開発本部
情報サービスシステムラボラトリ

山本秀樹
研究開発本部
マルチメディアネットワークラボラトリ

発の歴史を経て，ようやく一般に広く知られるようになってきた。

機械翻訳は技術的な文書を対象にした場合は使用に耐えるようになってはきたが，様々な種類の文書を翻訳するにはまだ改善の余地がある。ところが，従来の方式の機械翻訳システムは，開発者の試行錯誤によって微妙にバランスされた規則と辞書で構成されているので，それらを乱さずに修正することは容易ではない。

そのような「品質と保守の問題」を解決する，従来と全く異なる手法として「用例を利用した翻訳」が注目を集めている。コーパス中の文（用例）から規則（知識）を獲得して翻訳する方式には様々なものがあり，統計を使うか否か，用例あるいはコーパスをどのように指定するかなどの点でそれぞれ異なるが，本稿ではそれらをまとめて「用例を利用した翻訳」と呼ぶものとする。中でも統計をベースにする手法は翻訳品質向上の打開策となると考えられている。

「用例を利用した」機械翻訳では，訳文は現実に存在する用例そのもの，あるいは用例に類似したものから構成されるため，生成文の品質は保証される。従来の文法・辞書の修正は，本方式ではコーパスの修正や選択に相当するため，翻訳品質を比較的容易に維持することができる。

3．コーパスとしてのインターネット

新方式の唯一の問題点は，有効な知識を獲得するために膨大なコーパスを必要とすることである。この方式は，用例とそれを処理して得られる統計を翻訳に利用するので，コーパスのサイズが大きければ大きいほど翻訳品質は向上する。

普通の電子辞書にも用例は存在するが，使われている言葉が十分に新しいとは限らない。常に最新の用例を含んだ何千万文も収録されたコーパスを配布しつづけるのも非現実的である。

それでは，統計処理を行うのに十分なほど大きくて，常に最新の状態に保たれていて，ローカルに保持しなくてよい，そんなコーパスは存在するのだろうか。

WWWのホームページは，画像やデータのみのページを除いては自然言語で記述される。それぞれのページに含まれる文は多くないが，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は頻繁に更新されており，新しいページも次々と現れている。つまりこれは，最新の用語やその使用法が常に反映されていることを意味する。文章の質にはばらつきがあるが，統計的に処理することにより無視することができる。

以上の特徴から，WWWホームページは翻訳知識を獲得するための理想的な言語資源であると言える。

次節では，インターネットをコーパスとして利用することで，より効果的に働く手法をいくつか説明する。

4．インターネットに適した手法

4.1　訳語選択技術

一般に，ある単語が翻訳されるとき訳語の候補は複数あるので，複数単語の翻訳に際しては，それらの組み合わせの中から適切なものを選択する必要がある。

「つらい仕事」という日本語を英語に翻訳する場合を考えよう。「つらい」は “tough”, “difficult”, “hard”, “painful” などと訳すことができ，一方，「仕事」は “work”, “task”, “job”, “business” などの訳が候補に挙げられる（図１）。翻訳結果はそれぞれの候補にある単語を組み合わせたものとなるが，これらを適当に組み合わせると，たとえば “hard job” といった収まりの悪い句ができてしまう。

従来の方法では，たとえば “difficult task” を出力させたいときには，「つらい仕事」という句を丸ごと辞書の見出しとして登録するより他に手段はなかった。しかしこの手段をとると，どの訳語候補の組み合わせが最適か，原文のどの句を見出しとして登録すべきか，などの判断がすべてシステムの開発者あるいはユーザに委ねられる。そのため現実的には，辞書に見出しのない句に対してはデフォルトの組み合わせを出力せざるを得なかった。

最適な組み合わせを決定するための手法の一つに，組み合わせを構成する各要素間の相性／調和に着目する手法がある。要素間の相性は，最も単純な場合，同一文（章）中に共に出現する（共起する）頻度によって求め

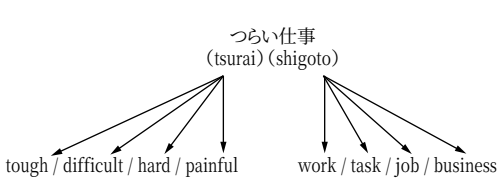


図１　「つらい」と「仕事」の訳語候補
Fig. 1　Translation candidates of "tsurai" and "shigoto"

ることができる。

本手法では，WWWホームページをローカルに保持したコーパスと見なして単語間の相性を計算する。インターネット検索エンジンの中には，検索要求した語句の出現頻度（あるいはページ数）を結果として返すことができるものがあるので，それらを用いてデータを収集する。

本手法の手順を以下に示す。

（1）訳語候補から可能なすべての組み合わせを生成する。

（2）検索エンジンに対して各々の組み合わせの共起頻度を問い合わせる。

（3）最も頻度の高い組み合わせを最適な訳語とする。

表１に前述の句を対象とした実験結果を示す。AltaVista[*1)][http://www.altavista.com/]で検索した結果，“hard work”という組み合わせは303,434ページに出現し，“difficult task”という組み合わせは約45,000ページにしか出現しなかったことがわかる。この結果から，本手法では最適な訳語として“hard work”を選択する。

本手法で用いる検索エンジンに要求される機能を以下に示す。

－ 検索するページの使用言語を指定できる。

－ 検索するページの内容（トピックス）を指定できる。

検索するページのトピックスを指定することによって最適な訳語は変わる可能性がある。このような指定を行うことは分野依存の知識/結果を得たい場合に有効である。Yahoo![*2)][http://www.yahoo.com]のような「ディレクトリ型」の検索エンジンは，ページの内容を分類してデータを格納するので，このような目的に利用するのに適している。

これに加えて以下の手順を実行してもよい。

（4）得られた頻度情報のすべて，あるいは最適訳語を決定するのに使われたデータのみを個人用の辞書データ（学習辞書）として保持しておく。

学習辞書を利用することによって，検索エンジンに同じ検索要求を繰り返す必要がなくなるため，処理が高速化される。

表１　検索結果：訳語候補を含むページ数
Table 1　Number of pages containing translation candidates

	work	task	job	business
tough	2,015	2,762	8,916	2,268
difficult	3,891	44,875	7,174	2,083
hard	303,434	3,476	3,966	768
painful	288	579	127	134

＊1）AltaVista は米国 AltaVista 社の登録商標。　＊2）Yahoo! は米国 Yahoo! 社の商標。

4.2　翻訳対獲得技術

ある決まった形式で頻繁に用いられ（かつ常に同じような形式で翻訳され）る定型表現を本稿では「常套句」と呼ぶ。このような表現を解析や生成に用いることにより，機械翻訳の精度を大きく向上させることができる。それらがもし名詞句や前置詞句といった普通の(文法で定義されうる)句ではなくて，文全体にわたって様々な単語と関連を持つ「文に枠をはめる」ような表現であれば，その効果はさらに大きい。

しかし，そのような長い変則的な句を見つけるのは簡単ではない。文中のどの単語が「常套句」の要素となるかを判断するには，数多くの文を見書きした経験が必要となる。

また仮に原言語でそのような表現が発見できたとしても，機械翻訳で使うためには，目的言語からそれに対応した表現を拾ってこなければならない。

巨大な２カ国語のコーパス（対訳コーパス）の統計処理によってこれを自動化する手法が提案されているが，一般にはそのような対訳コーパスの入手は困難である。

ところが，有名なWWWホームページは複数の言語で記述されていることが多いので，ここでもまたホームページをコーパスと見なすことで，これらの手法を有効に利用することができる。

以下に，

（1）コーパスから「常套句」を抽出する手法[3)]

（2）対訳コーパスから翻訳対を抽出する手法[4)]

の２つの手法について述べる。

（1）「常套句」の抽出

本手法では，まず「常套句」の構成要素となり得る連続する単語列を「ユニット」として抽出し，それらを組み合わせることで「常套句」を得る。

ユニットの範囲確定には，各単語をキーワードとみなしたとき，その直前および直後に配置される語のばらつきをエントロピー値を使って評価する方法を用いる。エントロピーは，様々な語が配置されているとき（ばらついているとき）に高い値を示し，そうでないときは低くなる。

キーワードの直後の語がばらついているとき，ユニッ

トの最後の要素はそのキーワードとなる。前方に関しても同様の処理を行い，ユニットの範囲を確定する。

通信分野のマニュアルから例を取り上げる。“a local area” という処理中のユニットがあったとき，その直後の語は “network”であると推測できる（“a local area telephone” や “a local area is ...” という単語列はおそらく存在しないため，ばらつきが小さい）。この場合，ユニットは後方に伸展されることになる。

一方で，“a local area network” の直後はかなりばらついていると考えられる（“is” “has” “had”など様々な語が配置されているはずなので，ばらつきが大きい）。この場合，ユニットの終端が決定され，“network” が最後の要素となる。

ユニットが抽出されたら，それらを文での出現順に組み合わせて，より長い共起表現を構築する。

表現を構成するユニットの間に穴（変数要素）を持つ離散的共起表現はさらに大きく翻訳品質向上に寄与するが，本手法はそのような表現も得られる可能性がある。これについては実験中である。

（2）翻訳対の抽出

本手法は，それぞれの単言語コーパスでの出現頻度と，対訳コーパスでの対応文中での共起頻度（対訳共起）を用いて，頻出する単語列とその翻訳対を発見する手法である。

まず，それぞれの単言語コーパスでの出現頻度から処理候補となる単語列を，それぞれの言語でリストアップし，一方の言語の候補単語列と，他方の言語の候補単語列とで，可能なすべてのペアを作成する。

次に，単言語コーパスでの出現頻度の類似性と，対訳共起の頻度から，すべてのペアに対して「類似度」を計算する。

そして「言語Aのある単語列」と，それと組み合わせたときに最も高い類似度を持つ「言語Bの単語列」を取り出す。その「言語Bの単語列」と組み合わせたときに最も高い類似度を持つ「言語Aの単語列」が，最初の「言語Aのある単語列」と同一であれば，それらの単語列は翻訳対であると認定される。

獲得された翻訳対は獲得済みリストに登録され，それぞれの言語の候補リストから削除される。

翻訳対を獲得するステップは，閾値を徐々に下げながら再帰的に実行される。これにより，より確からしい翻訳対から順に獲得されることになる。

本手法では，対訳コーパス中の各文があらかじめ対応付けられていることを想定している。HTMLタグを用いて記述されたWWWホームページは，タグ情報が文の対応付けに有効に働くので，本手法に適用する対訳コーパスとしては最適である。

5. あ と が き

本稿では，インターネットをコーパスとして用いる機械翻訳の手法について述べた。「用例を利用した翻訳」のアプローチをとるこれらの手法に必要な十分に大きいコーパスを，インターネット自然言語資源に求めることにより，翻訳の品質が向上することがわかった。WWWホームページは特に，常に最新の用語が供給されるという点でも理想的である。

現状では英語の文章の入手が最も容易なので，実験は英語と他の言語で行ったが，本稿の手法はいかなる言語対にも応用可能である。

これらの手法の機械翻訳以外への応用例として，作文支援システムが挙げられる。従来の作文支援システムには穴埋めをして文章を完成させるためのテンプレートを提示するものがあるが，ある程度の自由度を持って作文をする場合には有効に働かない。本稿で提案した訳語選択技術は，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，作文支援システムに有用なコンポーネントになりうる。

6. 参 考 文 献

（1）Global Reach：Global Internet Statistics（byLanguage），http://www.euromktg.com/globstats/，1999

（2）村田稔樹，他：WWWのリアルタイム翻訳を可能とする機械翻訳システム，電子情報通信学会論文誌，J79-B-1，No. 5，pp. 364〜372，1996

（3）下畑さより，山本秀樹：隣接文字のエントロピに基づく定型表現の自動抽出，沖電気研究開発，Vol. 65，No. 1，pp. 19〜22，1998

（4）北村美穂子，松本裕治：対訳コーパスを利用した対訳表現の自動抽出，情報処理学会論文誌，Vol. 38，No. 4，pp. 727〜736，1997

本論文は Forum Proceedings of World TELECOM 99+Interactive 99に掲載された論文を日本語に翻訳したものである。本論文の内容はすべて著者の見解であり，国際電気通信連合（ITU）の見解とはまったく関係を持たない。