E_030

日英翻訳における訳語統一支援システム

A Support System for Unifying Terms used in Japanese-English Translation

今井一裕† 小川泰弘† 外山勝彦†
Kazuhiro Imai Yasuhiro Ogawa Katsuhiko Toyama

1 はじめに

近年、様々な分野において日本語の文書が外国語に翻訳されている。その際、同一分野内の翻訳であれば、特に専門用語については、文書間で訳語が統一されていることが望ましい。しかし、現実には翻訳者が異なると、同じ語に対しても異なる訳語が用いられることが多い。例えば、これまで日本の法令は、関係府省や民間により個別に英訳されてきたため、同一分野にある関連する法令であっても、同じ法令用語が異なる訳語に英訳されていた。そのため、内容を正確に理解できない場合や、混乱を生じる場合があった。そうしたことから、国際取引の円滑化や、対日投資の促進などのため、信頼できる英訳への要望が国内外から出されている。この問題を解決するため、政府において、法令英訳の基準となる、法令用語用の標準対訳辞書の作成が進められている [1]。今後は、この標準対訳辞書に準拠して法令を英訳することになるが、既に翻訳されたものに対しては、最初から翻訳し直すよりも、標準対訳辞書の訳語と異なる箇所を修正する方が、翻訳にかかるコストは小さくなる。

そこで本稿では，日英翻訳における訳語の統一を支援するシステムを提案する。

2 訳語統一支援システム

提案するシステムは、日本語文書と英訳文書が1文ずつ対応付けられた対訳コーパスを入力すると、1文ごとに訳語が標準対訳辞書に準拠しているかどうかを検査し、その結果を出力する。検査結果の出力を図1に示す。

2.1 訳文の検査

訳文の検査は以下の手順により行う。まず、日本語文中から左最長一致法によって、標準対訳辞書の見出し語を獲得する。その際、辞書構造に TRIE[2] を用いることにより、辞書引きを高速化した。

次に、得られた見出し語に対して、訳文中に標準対訳辞書の訳語 (以下、標準訳と呼ぶ) があるかどうか照合する。標準訳が訳文中にある場合には、日本語文中の見出し語と、訳文中の訳語の両方の箇所を太字で表示する (図 1(a))。標準訳が訳文中にない見出し語に対しては、訳文中における訳語を推定し、標準訳に置換する (図 1(b))。推定できなかった場合には、日本語文中の見出し語の箇所に、背景色を付ける (図 1(c))。なお、訳文中には、名詞の複数形や、動詞の三人称単数現在形といった、原形と異なる形で出現することもあるため、それらの変化形とも照合する。

また、いずれの場合においても、日本語文中の見出し語や、訳文中の訳語の箇所にカーソルを合わせると、ポップアップにより辞書の内容を表示する (図 1(d))。

2.2 不適訳リスト

訳文中の訳語が標準訳と異なる場合、訳文中のどの箇所が訳語であるかがわからなければ標準訳に置換できない。そこで、本システムでは、訳文中で用いられている訳語を自動的に推定する。その際、推定した訳語のうち、標準訳とは異なる訳語を不適訳としてあらかじめ収集する。そのようにして構築した訳語のリストを本システムでは不適訳リストと呼ぶ。

不適訳の推定には、標準対訳辞書の見出し語 x と、訳文中の単語列 y の類似度を、次に示す Dice 係数を用いて計算した。

$$Dice(x,y) = \frac{2f_{xy}}{f_x + f_y}$$

f_{xy} : 見出し語 x と単語列 y が同時に出現する回数
f_x : 文書中における見出し語 x の出現回数
f_y : 文書中における単語列 y の出現回数

†名古屋大学大学院情報科学研究科

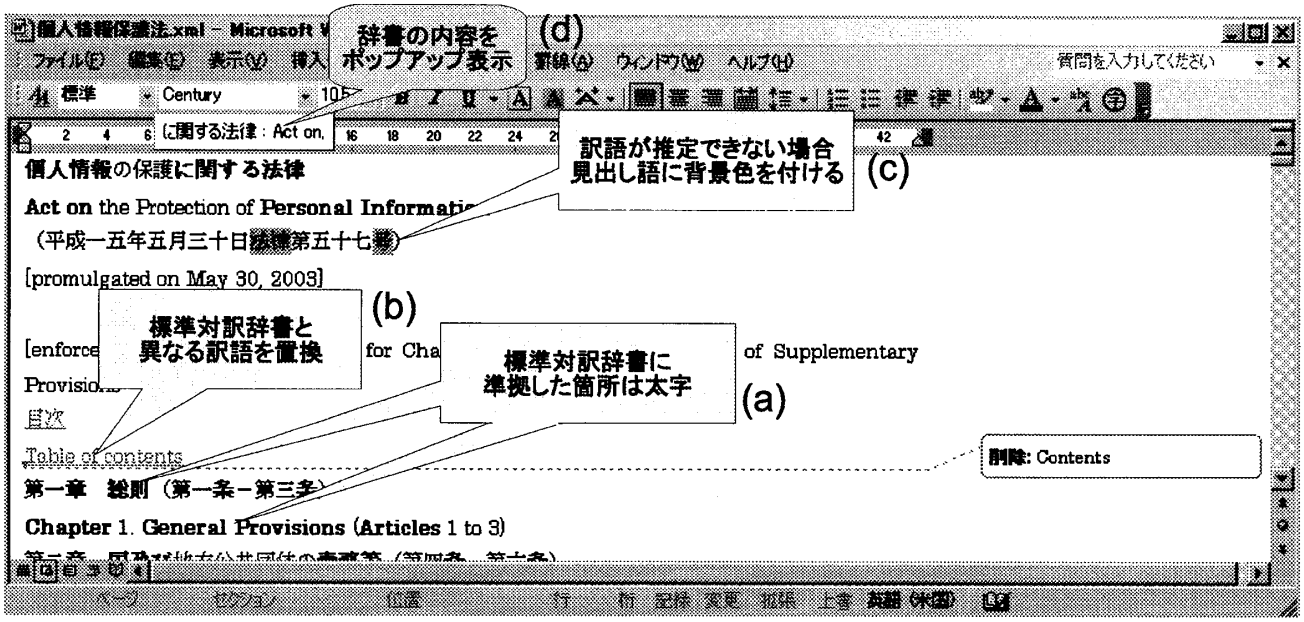

図 1: 訳語統一支援システムによる検査結果の出力

なお、日本語文書中における見出し語の出現回数と、Dice 係数の値にそれぞれ閾値を設け、その閾値を越えた単語列を不適訳リストに登録した。

2.3 wordML を用いた検査結果の出力

検査結果の出力には、MS-Word を利用した。これはユーザが使い慣れていることと、インタフェース部作成のコスト削減のためである。MS-Word 2003 では、XML によって Word 文書を記述できるようになっている。また、wordML は、その記述のためのスキーマ言語である [3]。

図 2 に wordML による文書の記述例を示す。まず、段落は要素 w:p によって記述する。w:p は子要素 w:r を持ち、w:r は文字のプロパティを表す子要素 w:rPr と、文字列を表す子要素 w:t を持つ。太字や背景色といった文字修飾は、w:rPr の子要素として記述する。例えば MS-Word 上で太字表示をするには、4 行目のように、要素 w:b の属性 val の値を “on” にする。背景色を付けるには、16 行目のように、要素 w:shd の属性 w:fill の値に色を指定する。

また、MS-Word には、脚注内容をポップアップによって表示する機能があり、表示される内容は 6 行目から 8 行目のように、要素 w:footnote によって記述される。

文字列の置換には削除・挿入の変更履歴機能を利用した。変更履歴機能は 22 行目からの要素 aml:annotation によって記述される。その属性 w:type には、削除する場合は ‘Word.Deletion’ を、挿入する場合は ‘Word.Insertion’ を、それぞれ記述する。

3 評価実験

本システムの性能評価のために実験を行った。実験では、不適訳リストを構築した後、「労働基準法」443 文とその英訳を入力として用いた。また、標準対訳辞書には、内閣官房司法制度改革推進室から提供された辞書 (見出し語数 3,324 語) を用いた。

3.1 不適訳リストの構築

まず、以下の法令 10 本とその英訳 (計 4,594 行) を用いて不適訳リストを構築した。

> 改正独占禁止法、刑法、行政手続法、種苗法、
> 消費者契約法、情報公開法、製造物責任法、
> 著作権法、労働基準法、個人情報保護法

今回の実験では、予備実験の結果から、不適訳リストに登録する閾値は、日本語文書中における見出し語の出現回数 18 回以上、Dice 係数 0.8 以上とした。

その結果、日本語文書中に出現した見出し語 1,579 語のうち、86 語に対して、訳語数 152 語の不適訳リストを得た。また、そのうち不適訳として適当な訳語は 37 語であった。この結果は不十分ではあるが、不適訳の充実は今後の課題とし、今回は人手により修正を加

```
1<w:p>
2 <w:r>                                          ← 太字
3  <w:rPr>
4   <w:b w:val="on"/>
5  </w:rPr>
6  <w:footnote w:suppressRef="on">               ┐
7   <w:p><w:r><w:t>個人情報</w:t></w:r></w:p>    ┘ ポップアップ
8  </w:footnote>
9  <w:t>個人情報</w:t>
10 </w:r>
11 <w:r>
12  <w:t>の保護</w:t>
13 </w:r>                                        ← 背景色
14 <w:r>
15  <w:rPr>
16   <w:shd w:val="clear" w:color="auto" w:fill="99CCFF"/>
17  </w:rPr>
18   <w:t>に関する法律</w:t>
19 </w:r>
20</w:p>
21<w:p>
22 <aml:annotation w:type='Word.Deletion' aml:author='' aml:id=''>   ┐
23  <aml:content>                                                    │
24   <w:r>                                                           │ 削除
25    <w:delText>Contents</w:delText>                                │
26   </w:r>                                                          │
27  </aml:content>                                                   │
28 </aml:annotation>                                                ┘
29 <aml:annotation w:type='Word.Insertion' aml:author='' aml:id=''>  ┐
30  <aml:content>                                                    │
31   <w:r>                                                           │ 挿入
32    <w:t>Table of contentses</w:t>                                 │
33   </w:r>                                                          │
34  </aml:content>                                                   │
35 </aml:annotation>                                                ┘
36</w:p>
```

図 2: wordML による検査結果の記述

え、見出し語 79 語に対して、訳語数 116 語の不適訳リストを得た。次節からの実験では、この修正を加えた不適訳リストを用いた。

3.2 評価

本実験では、

1. 日本語文書からの見出し語の獲得
2. 訳語の対応付け

の 2 項目について評価した。

3.2.1 日本語文書からの見出し語の獲得

本システムは、日本語文書中から獲得した見出し語が標準対訳辞書に準拠しているかどうかを検査するため、獲得できなかった見出し語は検査できない。したがって、見出し語を漏れなく獲得することが要求される。また、本来見出し語でないものを見出し語として獲得することは望ましくない。そこで、上述の評価項目 1 に対しては、次に定義する精度と再現率を用いる。

$$\text{精度} = \frac{\text{システムが見出し語として獲得したうち適切なものの総数}}{\text{システムが見出し語として獲得した総数}}$$

$$\text{再現率} = \frac{\text{システムが見出し語として獲得したうち適切なものの総数}}{\text{文書中の見出し語の総数}}$$

表 1: 見出し語の獲得における精度と再現率

評価項目	精度 (%)	再現率 (%)
見出し語の獲得	97.0	100.0

なお、本実験では、字面のみが一致する場合、例えば、「法律、法令」の意味で標準対訳辞書に登録されている見出し語「法」を、「…で定める方 法 により…」のような、「方法」の一部であるにもかかわらず獲得してしまったような場合を誤りとする。

3.2.2 見出し語に対する訳語の対応付け

本システムが獲得した見出し語に対して、その訳語を訳文中の適切な箇所と対応付けたか評価する。

まず、日本語文から獲得した見出し語に対して、その訳語を訳文と対応付けた結果は、次の A、B、C のいずれかとなる。

A：訳文中の正しい箇所と対応付けた
B：訳文中の誤った箇所と対応付けた
C：訳文中に対応する箇所が見つからなかった

これより、評価項目 2 に対して、正解率を次のように定義する。

$$\text{正解率} = \frac{A\text{の総数}}{A\text{の総数} + B\text{の総数} + C\text{の総数}}$$

なお、出現回数、対応付けの正誤は人手により判定した。

3.3 結果と考察

3.1 節で述べた不適訳リストと、「労働基準法」443 文を用いて実験したところ、システムは、日本語文書中から、見出し語として 4,432 語を獲得した。なお、「労働基準法」中にある見出し語の総数は 4,298 語であり、システムはその 4,298 語全てを獲得した。これより、表 1 に示すように、精度は 97.0%、再現率は 100%となった。

この結果について考察する。再現率 100%を得て、なおかつ 97%の精度を確保できたことは、本システムが有用であることを示している。また、見出し語の獲得における誤りとしては、前述した「方法」中の「法」の他に、「事項」中の「項」、「疾病にかかつた」等の中に出現する「かつ」といったものがあった。

次に、訳語の対応付けに対しては、表 2 に示す結果となり、正解率は、70.5%となった。対応付けの誤りは、見出し語「当該」に対する誤りが最も多く、対応

表2: 訳語の対応付け

	標準訳	不適訳	合計	割合 (%)
A: 正しい箇所と対応付けた	2,926	106	3,032	70.5
B: 誤った箇所と対応付けた	132	2	134	3.1
C: 対応付けがない			1,132	26.3
合計			4,298	100.0

付けを誤った箇所全体の34%にあたる45箇所あった。これは、「当該」の標準訳として“the”、“that”といった英文中に出現しやすい語が登録されており、正しい箇所と対応付けることが容易ではないためである。

また、1,132箇所(26.3%)の見出し語に対して、対訳となる箇所を訳文中に見つけられなかった。元の訳文中において、標準訳に訳されていない箇所は、この1,132箇所に、不適訳を正しく対応付けた106箇所、誤って対応付けた2箇所を加えた、計1,240箇所であった。つまり、元の訳文中における不適訳を、標準訳に正しく置換できたのは、8.5%(= 106/1240)にすぎない。これは、辞書に登録されている見出し語の総数3,324語に対して、79語の見出し語にしか不適訳が登録されていないことが原因である。しかし、対訳となる箇所が推定できなかった場合でも、2.1節で述べたように、日本語文書中の見出し語の箇所には背景色を付けているため、ユーザは訳語が標準訳でないことを知ることができる。このことを踏まえると、対応付けを誤った箇所は、3.1%のみであり、本システムは、訳語の統一を支援するシステムとして、有用であると考えられる。なお、不適訳リストの登録語数を増やすことにより、この結果は改善できる。その手段としては、不適訳を人手により容易に追加できるインタフェースを開発し、ユーザが本システムを用いて翻訳を修正する作業と並行して、発見した不適訳を不適訳リストに追加することが考えられる。

4 関連研究

訳語統一という視点の研究として、翻訳支援システムTER-GET[4]がある。TER-GETは、新たに翻訳しようとする日本語文に対して、専門用語の対訳辞書を引き、日本語文の見出し語箇所に訳語を付与して表示する。翻訳者はこれを見て翻訳することにより、訳語が統一された翻訳を作ることができる。それに対して、本稿で提案するシステムは、既に作成された翻訳文書に対して、辞書と異なる箇所を指摘することによって訳語統一を支援する。

また、翻訳文を検査・添削するシステムとしては、教育支援という視点から提案された英作文問題の添削システム[5]がある。このシステムは、問題ごとに正解データベースを持ち、ユーザの回答文と正解データベースとを比較し、一致するものがあれば正解とする。一致しない場合は、ユーザの回答と最も近似した正解文を用いて、回答の余分な箇所を削除したり、足りない箇所を挿入したりして添削する。また、単語の誤用に対しては、誤用である単語とその用途からなる誤答データを持ち、回答に誤答データの単語が含まれている場合には、ユーザに誤用した単語の語意を提示する。誤答データを用いる点では、提案システムで用いた不適訳リストと類似しているが、このシステムは訳語が統一された翻訳を作るのではなく、既にある正解と比較することで、ユーザの学習を促すという点において、提案システムとは異なる。

5 おわりに

本稿では、訳語の統一を支援する手法を提案し、訳語統一支援システムを開発した。また、「労働基準法」443文に対して実験を行い、提案手法の有効性を示した。

なお、現在、政府が進めている法令英訳における訳語統一に対して、本システムが使用されている。

今後の課題としては、不適訳リストの充実が挙げられる。そのためには、不適訳の推定方法を改善することや、人手により不適訳を獲得する方法が考えられる。また、日英以外の言語への応用も今後の課題である。

参考文献

[1] 法令外国語訳実施推進検討会議：最終報告, http://www.cas.go.jp/jp/seisaku/hourei/houkoku.pdf. (2006).

[2] 青江順一：キー検索技法-IV トライとその応用, 情報処理, Vol.34, No.2, pp.244-251 (1993).

[3] Word 2003 Object Model, http://msdn.microsoft.com/library/default.asp?url=/library/en-us/odc_2003_ta/html/odc_ancword.asp

[4] 新井立夫：翻訳支援システム TER-GET の概要, 情報処理学会研究報告, 自然言語処理,62-9, pp.61-68 (1987).

[5] 西村則久, 安村通晃：外国語作文における自動添削手法について, 情報処理学会研究報告, 人文科学とコンピュータ,41-1,pp.1-6 (1999).