法令翻訳における対訳表現の統一性評価指標

今井 一裕，小川 泰弘，外山 勝彦
名古屋大学大学院 情報科学研究科

1 はじめに

近年，国際取引の円滑化や，対日投資の促進，海外での法整備支援などのために，日本法令を外国語に翻訳することが必要とされている．しかし，従来の翻訳作業は，法令を所管する府省や，法令出版社など民間による個別の活動として行われてきた．そのため，同じ原文に対して様々な訳し方があり，その中には利用者に誤解を生じさせるものもあった [1]．例えば,「弁護士」という語には，“attorney”, “barrister”, “lawyer” 等，複数の対訳表現が見られた．英語では，これらの語の意味は少しずつ異なるため，利用者は，いずれの語も「弁護士」という同じ意味で用いられていることが分からない場合があった．したがって，法令翻訳においては，日本語の原文の内容を利用者に誤りなく伝えるために，対訳表現を適切なものに統一することが要求される．

そこで，日本政府は法令用語の対訳表現を統一するために，法令用語や法令文に頻出する典型的な言い回しに関して,「標準対訳辞書」[2] を構築した．現在は，この標準対訳辞書に従って，政府の主導による法令の翻訳作業が進められており，作成された翻訳法令は，Web にて公開されている [3].

標準対訳辞書の利用により，単語レベルでの対訳表現の統一に関しては一応の解決を見ることとなった．しかし，標準対訳辞書の見出し語には，複数の対訳表現を持つものがある．例えば,「免除する」に対しては，標準対訳辞書には “release”，“exculpate”,“waive” などが登録されており，文脈に応じてこれらを使い分けるとされている．しかし，あらゆる文脈に対して，どの対訳表現を用いるかを標準対訳辞書にすべて登録することは容易ではない．したがって，対訳表現が統一されているかどうかを判断するには，標準対訳辞書だけでは不十分であり，文脈を考慮する必要がある．

そこで，本稿では，法令翻訳文に対して，対訳表現の統一性を評価する指標を提案する．一般に，翻訳において評価される観点として，妥当性と流暢性がある．妥当性とは，訳文中において，適切な対訳表現が用いられているかどうか，流暢性とは，その言語，さらにはその分野における文として，自然な文かどうかを評価する観点である．妥当性，流暢性を兼ね備えた翻訳評価指標として，機械翻訳システムの評価指標 BLEU[4] が知られている．本稿では，BLEU を利用し，統一性の観点から妥当性，流暢性を捉え直すことにより，同一法令に対する複数の翻訳文書に対して，優劣を判別できる指標を提案する．

2 BLEU

BLEU[4] では，機械翻訳システムが出力した訳文と，人手により翻訳した参照訳を比較することにより，システムを評価する．BLEU スコアでは，基本的には，訳文中の n グラムが参照訳中に出現する割合 p_n を用いる:

$$p_n = \frac{\sum_{\text{訳文 } S\in \text{ 訳文集合}} \sum_{n\text{-}gram\in S} Count_{clip}(n\text{-}gram)}{\sum_{\text{訳文 } S\in \text{ 訳文集合}} \sum_{n\text{-}gram\in S} Count(n\text{-}gram)}$$

$Count(n\text{-}gram)$ = 訳文 S 中における n グラム $n\text{-}gram$ の出現回数

$Count_{clip}(n\text{-}gram)$ = min(max(各参照訳における$n\text{-}gram$の出現回数), $Count(n\text{-}gram)$)

訳文が参照訳より短い場合には，上式の分母が小さくなり，p_n の値が大きくなってしまうため，ペナルティとして次式の BP を与える．ここで，c は訳文の長さ，r は参照訳の長さである．

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \le r \end{cases}$$

さらに n の値に応じた重み w_n を導入し，BLEU の最終的なスコアは次式で定義される:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (1)$$

一般に，n の上限として $N = 4$，重みとして $w_n = 1/N$ が用いられる．

式 (1) に示すように，BLEU では，n を 1 から N まで変化させている．低次の n においては単語に対する適切さ，すなわち妥当性が評価される．一方，高次の n においては語と語の繋がりに対しても考慮されるため，流暢性が評価される．

BLEU を利用した評価指標として，自動要約の評価指標 ROUGE-N[5] がある．BLEU は訳文に出現した n グラムが参照訳に出現したかどうかに基づく精度指向の指標であるのに対して，ROUGE-N は，人手による要約 (BLEU の参照訳に相当) に出現した n グラムが，システムの出力した要約文に出現したかどうかによってスコアを計算する再現性指向の指標である．

3 統一性を重視した法令翻訳の評価指標

機械翻訳システムの評価指標である BLEU には，翻訳の正解となる参照訳が必要である．しかし，今回は人手により作成した法令翻訳文を評価するものであり，原文に対する参照訳は用意されていない．

ここで，一般に，法令文には，定型的な表現によって記述されるという特徴がある．例えば，施行期日を規定する文であれば「この法律は、公布の日から起算して…月を経過した日から施行する。」「この政令は、…法の施行の日から施行する。」など，記述のための表現が決められている [6]．対訳表現を統一する観点からは，同じ表現が用いられている原文であれば，訳文においても同じ対訳表現に統一されることが望ましい．

そこで，法令文対訳コーパスを利用する．対訳コーパスに含まれる法令文から，原文に類似した文を獲得し，その訳文を参照訳の代わりに利用する．以下，これを「擬似参照訳」という．この対訳コーパスには，1 章で述べた政府による翻訳法令 [3] を用いる．これは，統一的で，信頼できる翻訳法令を目標に作成されているため，翻訳法令として適切な対訳表現が使われた自然な訳文であると仮定できる．したがって，この訳文と同様の訳し方が行われている翻訳法令文は，統一性という点において妥当性，流暢性が高いと考えられる．

3.1 擬似参照訳の獲得

擬似参照訳を獲得するために，法令対訳コーパスに含まれる日本語法令文集合に対してクラスタリングを行った．これにより，原文との距離が小さいクラスタは原文に類似した表現が用いられている文の集合であると考えられる．よって，そのクラスタに属する法令文の訳文を擬似参照訳とすることができる．クラスタリングは以下の方法により行った．

まず，文末の形態素により，法令文集合を分割した．これは，文末に現れる述語が法令文の内容を表す手がかりとなるためである．次に，定型的でない部分をあらかじめ減らすために，文末の文節から 2 段階までに係る文節のみを残し，それ以外を削除した．例えば，図 1 に示すように，「この法律は、会社法の施行の日から施行する。」という文に対しては，文末の文節「施行する」と，それに係る「法律は」と「日から」，さらに「法律は」に係る「この」，「日から」に係る「施行の」を残し，それ以外の「会社法の」という文節は削除する．この結果得られる「この法律は、施行の日から施行する。」を用いてクラスタリングを行う．

このように分割した法令文集合に対して，階層型クラスタリングを行った.その際，文間の距離関数には，正規化した形態素単位での編集距離を用いた．クラスタリングの結果，1 文だけからなるクラスタが生成されたが，これは擬似参照訳として用いないものとする．

また，法令文には，多くの法令に出現する定型文が存在する．例えば，「この法律は、公布の日から施行する。」という文は，法令の公布日と施行日を同じ日に定めているどの法令にも出現する．このように，法令文集合に同一の文が複数出現した場合には，前述の擬似参照訳ではなく，その定型文に対する訳文の集合のみを参照訳として用いる．

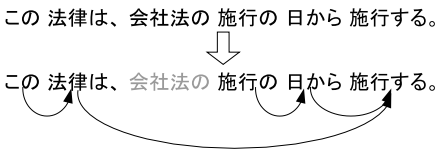

図 1: 係り受けによる文節の削除

このようにして擬似参照訳を獲得したが，擬似参照訳は，本来，原文とは異なる文の翻訳である．このことから，BLEU をそのまま用いたのでは，不十分な点がある．3.2 節では BLEU の拡張により，3.3 節では，標準対訳辞書を用いて妥当性評価を補強することにより，この問題を解決する．

3.2 BLEU の拡張

3.2.1 擬似参照訳中の出現頻度による重み付け

擬似参照訳は，一般に，原文を翻訳したものではないため，原文の訳文中に出現する n グラムが擬似参照訳中にも出現するとは限らない．したがって，このような n グラムにより，BLEU のスコアを不必要に下げてしまうという問題が生じる．

しかし，擬似参照訳に多く出現する n グラムであれば，原文の翻訳にも用いられる可能性は高いと考えられる．そこで，次の式 (2) に示す重み付き BLEU **(BLEU-W)** を考案し，擬似参照訳において n グラム *n-gram* が出現する文数に基づく重み $w(\textit{n-gram})$ を p_n に付与した．

$$w(\textit{n-gram}) = \frac{\text{クラスタ中で } n \text{ グラム } \textit{n-gram} \text{ が出現する文数}}{\text{クラスタ中の文数}}$$

$$p_n = \frac{\sum_{\textit{n-gram} \in S} Count_{clip}(\textit{n-gram}) * w(\textit{n-gram})}{\sum_{\textit{n-gram} \in S} Count(\textit{n-gram}) * w(\textit{n-gram})} \quad (2)$$

ところが，この重み付けでは，擬似参照訳に出現しない n グラムの重みは 0 となり，スコアに寄与しなくなる．そのため，スコア計算に用いられる n グラムは，訳文に出現し，かつ擬似参照訳に出現する n グラムに限られる．この範囲の n グラムは擬似参照訳に少なくとも 1 度は出現するので，$Count_{clip}(\textit{n-gram})$ の値が 0 になることはない．したがって，一つの n グラムが高々1 度しか出現しない訳文に対しては，$p_n = 1$ になってしまい，複数の文書間において，スコアの比較ができない文が増加するという問題が発生した．

3.2.2 再現性を考慮した n グラム範囲の拡張

前節の問題を解決するためには，スコア計算に用いる n グラムの範囲を拡張する必要がある．ここで，擬似参照訳は原文と類似した文を翻訳したものであるため，擬似参照訳の多くの文に出現している n グラムは，原文を翻訳した際にも出現しているべきで

表 1: 原文–クラスタ間の距離別によるスコア

原文–クラスタ間	擬似参照訳+BLEU			擬似参照訳+BLEU-W			擬似参照訳+BLEU-RW		
距離 d	政府訳	法令出版社訳	Google訳	政府訳	法令出版社訳	Google訳	政府訳	法令出版社訳	Google 訳
$0.0 < d \leq 0.1$	0.352	0.322	0.126	0.852	0.869	0.559	0.551	0.451	0.157
$0.1 < d \leq 0.2$	0.307	0.320	0.109	0.821	0.784	0.425	0.435	0.423	0.104
$0.2 < d \leq 0.3$	0.298	0.299	0.125	0.738	0.675	0.427	0.351	0.351	0.131
$0.3 < d \leq 0.4$	0.153	0.143	0.076	0.588	0.579	0.273	0.262	0.231	0.079
$0.4 < d \leq 0.5$	0.138	0.116	0.069	0.518	0.483	0.233	0.185	0.140	0.081
$0.5 < d \leq 0.6$	0.089	0.087	0.063	0.344	0.285	0.215	0.117	0.109	0.059
$0.6 < d \leq 0.7$	0.074	0.069	0.059	0.331	0.286	0.190	0.110	0.101	0.066
$0.7 < d \leq 0.8$	0.076	0.072	0.061	0.300	0.282	0.268	0.112	0.094	0.079
$0.8 < d \leq 0.9$	0.054	0.052	0.071	0.128	0.128	0.188	0.048	0.046	0.043
$0.9 < d \leq 1.0$	0.059	0.058	0.053	0.093	0.093	0.093	0.053	0.055	0.051
平均	0.133	0.127	0.081	0.456	0.428	0.259	0.190	0.168	0.077

あると考えられる．

そこで，擬似参照訳 Ref に出現する n グラムのうち，出現する文数の割合が $\alpha(0 \leq \alpha \leq 1)$より大きい$n$グラムからなる集合を $TopRef(n,\alpha)$とし，$TopRef(n,\alpha)$ に含まれる n グラムも，式 (3) に示すようにスコア計算に用いる．これ以降，式 (3) の p_n を用いた BLEU を，**BLEU-RW** と呼ぶ:

$$TopRef(n,\alpha) = \{n\text{-}gram \in Ref | w(n\text{-}gram) > \alpha\}$$
$$p_n = \frac{\sum_{n\text{-}gram \in S \cup TopRef(n,\alpha)} Count_{clip}(n\text{-}gram) * w(n\text{-}gram)}{\sum_{n\text{-}gram \in S \cup TopRef(n,\alpha)} \max(Count(n\text{-}gram), 1) * w(n\text{-}gram)} \quad (3)$$

$TopRef(n,\alpha)$ の導入は，ROUGE-N[5] と同様に，原文の翻訳に用いられるべき対訳表現が，訳文において用いられているかどうかに関する評価である．すなわち，精度指標である BLEU に，再現性の視点からの評価を加えたと言える．

3.3 標準対訳辞書への準拠率による妥当性評価

擬似参照訳は，一般に，原文とは異なる法令文の翻訳である．したがって，原文中に現れるすべての語について，適切な対訳表現が用いられているかどうかは評価できておらず，妥当性の評価としては不十分である．

しかし，妥当性，すなわち単語に対して適切な対訳表現を用いているかどうかという点においては，先に述べたように標準対訳辞書の対訳表現に統一することが求められる．

そこで，妥当性の評価においては，BLEU を利用した指標の代わりに，次式に示す標準対訳辞書への準拠率により評価する:

$$\text{準拠率} = \frac{\sum_{\text{原文}} \text{訳文中に標準対訳辞書の訳語が出現する見出し語数}}{\sum_{\text{原文}} \text{文中に出現した標準対訳辞書の見出し語数}}$$

4 評価実験

労働基準法に対する複数の翻訳文書に対して，提案指標によるスコアを計算し，評価した．

4.1 実験方法

評価対象の翻訳法令として，労働基準法 242 文に対して，政府が作成した翻訳 [3]，法令出版社による翻訳 [7]，Google の翻訳ツール [8] を用いた翻訳の 3 種類を用いた．各翻訳に対して，BLEU，BLEU-W，BLEU-RW の 3 つの指標によるスコアと，標準対訳辞書への準拠率をそれぞれ計算した．

政府訳は，標準対訳辞書を用いて作成したものである．法令出版社訳は，専門家が翻訳しているが，標準対訳辞書は用いられていない．Google 訳は機械翻訳である．したがって，翻訳法令としての適切さは，政府訳，法令出版社訳，Google 訳の順に高いものと考えられる．したがって，指標によるスコアもこの順に高くなることが期待できる．擬似参照訳の作成には，政府による翻訳法令から，労働基準法 242 文を除いた日本語文 17,793 文 (異なり数) と，その対訳である英文 20,154 文 (異なり数) を用いた[1]．なお，$TopRef$ のパラメータである α の値は 0.5 とした．

4.2 結果・考察

実験においては，22 文に対して，文末の形態素が他の法令に出現しないか，1 文のみからなるクラスタしか生成できなかったため，擬似参照訳を得ることができなかった．そのため，残りの 220 文に対してスコアを計算した．原文–クラスタ間距離とスコアの関係を表 1 に示す．また，標準対訳辞書への準拠率を表 2 に示す．

まず，表 1 に示すスコアの平均値により，BLEU では，Google 訳と他の翻訳に対する差は見られたものの，政府訳と法令出版社訳の間に有意な差は見られなかった．しかし，BLEU-W および BLEU-RW では，政府訳と法令出版社訳のスコアにおいても有意な差が見られた．また，BLEU-W では，220 文のうち 53 文において政府訳と法令出版社訳のスコアが等しく

[1]日本語文より英文の文数が多いのは，同じ日本語文に対して，複数の訳文が存在する場合があるためである．

表 2: 標準対訳辞書への準拠率

(労働基準法中に出現する見出し語の総数:2,620 語)

翻訳文書	訳文に辞書訳が出現した見出し語数	準拠率
政府訳	2,042	0.779
法令出版社訳	1,765	0.674
Google 訳	1,533	0.585

原文: 前項の委員会は、次の各号に適合するものでなければならない。

政府訳: The committee set forth in the preceding paragraph must conform to the following items:

法令出版社訳: The committee mentioned in the preceding paragraph shall conform to each of the following items:

擬似参照訳 1: The statement of the detailed explanation of the invention as provided in item 3 of the preceding Paragraph shall comply with each of thefollowing items:
(前項第三号の発明の詳細な説明の記載は、次の各号に適合するものでなければならない。)

擬似参照訳 2: The statement of the scope of claims as provided in paragraph 2 shall comply with each of thefollowing items:
(第二項の実用新案登録請求の範囲の記載は、次の各号に適合するものでなければならない。)

擬似参照訳 3: The statement of the scope of claims as provided in paragraph 2 shall comply with each of thefollowing items:
(第二項の特許請求の範囲の記載は、次の各号に適合するものでなければならない。)

図 2: 政府訳より法令出版社訳のスコアが高い例文

なったが，BLEU-RW では両者のスコアが等しい文は 15 文に減少した[2]．さらに，BLEU-RW では，原文–クラスタ間距離が小さいほどスコアの差が大きいという傾向が見られた．

表 2 に示す標準対訳辞書への準拠率に対しても，政府訳，法令出版社訳，Google 訳の順となった．

BLEU-RW によるスコアの平均値では，政府訳が法令出版社訳を上回ったが，個々の文においては，法令出版社訳が政府訳を上回った訳文が存在した．この原因は，原文–クラスタ間の距離により，異なる 2 つに分けられる．

まず，原文–クラスタ間の距離が小さい場合には，法令出版社訳における対訳表現が，政府訳よりも適切であることが原因であった．すなわち，より適切な訳文である法令出版社訳に高いスコアを与えたのであり，指標に問題はない．例えば，「前項の委員会は、次の各号に適合するものでなければならない。」という文の類似文には，「… の … は、次の各号に適合するものでなければならない」という文が含まれている (図 2)．政府訳では，「なければならない」の対訳表現として “must” が用いられている．しかし，擬似参照訳を見ると，「なければならない」の対訳表現としては，“shall” を用いるべきであることが分かる．法令出版社訳では，擬似参照訳と同様に “shall” が用いられていたため，政府訳に比べてスコアが高くなった．これと同様に，「次の各号に」に対しても，擬似参照訳から，“with each of the following items” と訳すべきであることが分かるが，政府訳では，“to the following items” となっていたため，スコアを下げる原因となった．なお，この例文に対して，BLEU-W では，政府訳，法令出版社訳ともに，$p_n = 1(n = 1,2,3,4)$ となった．すなわち，適切な対訳表現が用いられていない政府訳と，適切な対訳表現が用いられている法令出版社訳のスコアが等しくなり，優劣を判別できなかった．

次に，原文–クラスタ間の距離が大きい文では，原文とクラスタ内の文が類似していないことが原因として挙げられる．そのために，原文の翻訳に用いられるべき n グラムが擬似参照訳に出現せず，また，$TopRef(n,\alpha)$ に含まれる n グラムが訳文に出現しなくなり，スコアが小さくなった．表 1 からも，提案指標では，距離が大きくなるにつれてスコア自体が小さくなり，政府訳と法令出版社訳の差も小さくなることがわかる．このことから，クラスタとの距離が大きい原文に対しては，本指標の信頼性は低いと考えられる．この問題を解決するために，原文との距離が小さいクラスタをいかにして獲得するかが今後の課題である．

5 おわりに

本稿では、法令翻訳において，適切な対訳表現に統一するための指標として，擬似参照訳の利用，BLEU の拡張，標準対訳辞書への準拠率を提案した．また，実験により，同一法令に対する複数の翻訳に対して，対訳表現の統一性という観点から，優劣が比較できることを確認した．

今後の課題としては，擬似参照訳獲得手法の検討が挙げられる．また，今回の指標では複数の文書に対する相対的な評価であったが，一つの翻訳文書に対して，絶対的な評価を行うことが挙げられる．

[2] うち 7 文は，政府訳と法令出版社訳が同じである．

参考文献

[1] K.Toyama, Y.Ogawa, K.Imai, Y.Matsuura: Application of Word Alignment for Supporting Translation of Japanese Statutes into English, In JURIX'06, pp.141-150, 2006.

[2] 法令外国語訳・専門家会議：法令用語日英標準対訳辞書 (平成 19 年 3 月改訂版): http://www.cas.go.jp/jp/seisaku/hourei /0703dictionary.pdf , 2007.

[3] 内閣官房：法令翻訳データ集 http://www.cas.go.jp/jp/seisaku/hourei/data1.html

[4] K.Papineni, S.Roukos, T.Ward, W.Zhu: “BLEU: a Method for Automatic Evaluation of Machine Translation”, In Proc. ACL'02, pp.311-318, 2002.

[5] Chin-Yew Lin: ROUGE: a Package for Automatic Evaluation of Summaries, In Proc. of Text Summarization Branches Out, Workshop at the ACL'04, pp.74-81, 2004.

[6] 大島稔彦: 法令起案マニュアル, ぎょうせい, 2004.

[7] F. Nakane: “Ehs Law Bulletin Series, Japan”, vol.VIII, pp. EA, Eibun-Horei-Sha, Inc., 2002.

[8] Google 言語ツール http://www.google.co.jp/language_tools/