文長に応じた事前並び替えによる法令のあらましの翻訳

岡田 浩平[1]　小川 泰弘[1,2]　大野 誠寛[1,2]　中村 誠[3]　外山 勝彦[1,2]
[1]名古屋大学 大学院情報科学研究科　[2]同 情報基盤センター　[3]同 大学院法学研究科
{k-okada,yasuhiro}@kl.i.is.nagoya-u.ac.jp

1　はじめに

社会や経済のグローバル化に伴い、対日投資の促進や国際取引の円滑化等のため、日本の法令を外国語へ翻訳することへの要求が高まっている。そのような要求に対して、法令翻訳に関する研究 [1] や、法令そのものではなく、法令のあらまし（以下、あらまし）を翻訳対象とする統計的機械翻訳 (SMT) の研究 [2][3] が進められてきた。

日英翻訳においては、日本語は SOV 型言語、英語は SVO 型言語であるため、翻訳元と翻訳先で語順が大きく異なる。言語間の大きな語順の違いは従来の SMT では対応が容易でなく、翻訳精度低下の原因となっている。そこで、日本語の語順を英語の語順に並び替えてから翻訳する事前並び替え手法を、あらましの SMT に適用した。

今回は学習データから並び替え規則を自動で獲得する事前並び替え手法 lader[4] を用いた。しかし、lader を適用してもあらましの翻訳精度は向上しなかった。その原因を調査したところ、あらましに多く含まれる長文において事前並び替えに失敗し、結果としてそれらの文の翻訳精度が低下したことが判明した。

そこで本研究では、入力文の長さに応じて事前並び替えの有無を分ける翻訳手法を新たに提案し、翻訳精度向上を達成した。

2　法令のあらまし

あらましは公布された法令の内容を簡明かつ正確に要約した文書であり、昭和 48 年度から法令の公布と同時に官報に掲載されている。官報に法令を公布することは、国政の運営や国民の福祉等にとって必要不可欠なものである。しかし、公布される法令は、正確さを金科玉条としており、内容の理解が容易ではなかった。そのため、法令の内容を直ちに理解できるように配慮した形で、法令のあらましが登場した [5]。

図 1 に法令文、図 2 に図 1 の法令文に対応する法令のあらましの例を示す。法令文は三つの項から成る長い文章である一方、そのあらましはわずか一文にその内容をまとめている。このように、あらましは元の法令よりも短く簡潔な文で構成されているため、元の法令文より容易かつ迅速な翻訳が期待できる。また、あらましは法令の概略を把握することに適した文書であるといえる。

> 第四条　人の行為能力は、その本国法によって定める。
> 2　法律行為をした者がその本国法によれば行為能力の制限を受けた者となるときであっても行為地法によれば行為能力者となるべきときは、当該法律行為の当時そのすべての当事者が法を同じくする地に在った場合に限り、当該法律行為をした者は、前項の規定にかかわらず、行為能力者とみなす。
> 3　前項の規定は、親族法又は相続法の規定によるべき法律行為及び行為地と法を異にする地に在る不動産に関する法律行為については、適用しない。

図 1: 法の適用に関する通則法（平成 18 年法律第 78 号）

> 人の行為能力は、その本国法によって定めることとしたほか、日本国外における取引についても、行為地法による取引保護を図ることとした。

図 2: 法の適用に関する通則法のあらまし

3　事前並び替え手法

機械翻訳における事前並び替え手法は大きくルールベースと統計ベースの二つに分けることができる。ルールベースは人手で作成した並び替えルールを用いる手法である。星野ら [6] は、構文情報を利用した並び替えルールを作成している。しかし、これらの並び替えルールは構文情報が正しいことを前提としているため、構文解析器の解析誤りが並び替えに直接悪影響を及ぼすという問題点がある。一方、Katz-Brown ら [7] は構文情報を利用しないルールベース手法を提案して

いる。この手法では、文字列を句読点と助詞「は」で区切り、区切られた各形態素列の並びを逆順にすることで、事前並び替えを実現している。しかし、並列関係を無視した並び替えが行われるという問題点がある。

一方、統計ベースは学習データを用いて並び替え規則を自動で獲得する手法であり、Neubig ら [4] による lader が代表的である。lader ではアライメントがついた対訳データから並べ替えの精度が最大化するように、括弧反転トランスダクション文法 (Bracketing Transduction Grammar, BTG) を教師なしで学習し、事前並び替えを実現する。この手法では学習に用いるアライメントの精度が並べ替えに影響するため、高精度なアライメントを作成することが重要である。

ここで、どの手法があらましの SMT に適しているか検討する。星野らの手法では、高精度の構文解析器が必要である。日本語の代表的な構文解析器として CaboCha[8] がある。しかし、法令文に対する CaboCha の解析精度は十分でなく、後処理や人手による修正が必要であることが山田ら [9] によって報告されている。したがって、これはあらましの SMT に適した手法ではないと考えられる。一方、Katz-Brown らの手法では構文解析なしで並び替えが可能である。しかし、この手法は並列関係を無視した並び替えを行うため、並列構造が頻出する法令文とは相性が悪いと考えられる。Neubig らの手法も同様に構文解析なしで並び替えが可能である。この手法では法令文のコーパスを並び替えの学習に使用することで、法分野に特化した並び替えが可能になると考えられる。

以上から、本研究では lader があらましの SMT に適した手法であると考え、導入を試みた。

4 実験

本節では、事前並び替えなしで翻訳するベースライン手法と、lader による事前並び替えを行い翻訳する lader あり手法を、翻訳実験によって比較し考察した。

4.1 実験設定

各コーパスに対し、文を形態素に分割するツールとして、日本語には MeCab[10]（IPA 辞書使用）を、英語には Moses[1] 付属のトークナイザをそれぞれ用いた。各手法の言語モデルと翻訳モデルは、それぞれ SRILM[11] と GIZA++[12] によって学習した。その際、JLT[2] 法令文対訳コーパス (620 法令、364,936 文) と平成 23 年公布法令のあらまし対訳コーパス (82 法律、2,552 文) を合わせたコーパスを用いた。ただし、翻訳モデルの学習コーパスは、GIZA++が取り扱えない長文や不適切な対訳を取り除くため、日英どちらかの文長が 80 単語を超える対訳文、または、日英対訳の間の単語数の比が 9 倍を超える対訳文を削除した 316,315 文を用いた。lader の学習には前述の平成 23 年公布法令のあらまし対訳コーパスを用い、学習に必要なアライメントは GIZA++から取得した。また、GIZA++から得られる単語クラスを学習の素性として使用した。パラメータチューニングに使用するデベロップメントデータには、平成 23 年公布法令のあらまし対訳コーパス (2 法律、200 文) を用い、BLEU 値が最大になるようにパラメータチューニングを行った。入力文には平成 22 年公布法令のあらまし (72 法律、1,742 文)、デコーダには Moses を用いた。自動評価には、BLEU[13] と RIBES[14] を用いた。自動評価用の参照訳は 2 セット使用した。

なお、Moses のパラメータの一つである distortion-limit の値は、予備実験によって最適な値を求めており、ベースライン手法においては 24、lader ありの手法においては 12 にそれぞれ設定した。lader ありの手法は事前並び替えによって日英の語順の差が小さくなったため、ベースライン手法よりも distortion-limit の最適値が小さくなった。

4.2 結果と考察

はじめに、入力文の並び替え例と、その翻訳結果を表 1 に示す。lader によって日本語文が英語文の語順に近い状態で並び替えられていることがわかる。しかし、「平成 二 四 年」を「年 四 二 平成」、「三 一 日」を「日 一 三」と並び替えているように、人間から見て明らかに不必要、又は誤った解釈をしてしまうような並び替え結果があることがわかった。それでも、これらの並び替えからでも正しい翻訳を得ていることがわかる。これは、学習データの日本語文も同様に並び替えているため、「四 二 平成」と “2012” のような対訳を学習したためであると考えられる。また、この例では並び替えによって、ベースラインよりも正しい翻訳が得られている。

次に、各あらましの出力文に対する自動評価のスコアを平均した結果を表 2 に示す。lader ありの手法は、BLEU においてベースラインより有意にスコアが上昇した ($p < 0.05$)。一方、RIBES の場合、lader ありの手法はベースラインよりも低いスコアを示す結果となった。

それに対して、lader ありの手法における RIBES のスコア低下の要因を調査したところ、あらましに多く含まれる長文が事前並び替えに失敗し、その後の翻訳

[1] http://www.statmt.org/moses/
[2] http://www.japaneselawtranslation.go.jp/

表 1: lader による並び替え例とその翻訳結果

原文	この 法律 は 、 平成 二 四 年 三月 三 一 日 限り 、 その 効力 を 失う もの と する こと と し た 。
並び替え後	この 法律 は 失う を 効力 、 その 限り 日 一 三 三月 年 四 二 平成 、 もの と する こと た 。 し と
ベースライン	the purpose of this act is , 2012 march 31 , cease to be made .
lader あり	this act shall cease to be effective only by march 31 , 2012 .
参照訳 1	this act shall cease to be effective as of march 31, 2012.
参照訳 2	this act shall become ineffective from march 31st, 2012.

表 2: 各手法に対する BLEU, RIBES スコア

手法	BLEU	RIBES
ベースライン	37.72	73.40
lader あり	38.57	73.02

表 3: 形態素数別の RIBES スコア

	ベースライン	lader あり	文数
1-10	75.43	77.93	422
11-20	79.26	80.90	310
21-30	69.77	71.32	223
31-40	64.85	67.27	164
41-50	60.76	61.80	153
51-60	61.00	58.35	134
61-70	62.42	56.07	107
71-80	59.91	49.26	70
81-	60.53	49.54	159

に影響を与えていることが判明した。

4.3 入力文の長さごとの比較

4.2 節の結果をふまえ、入力文の長さが各手法の翻訳精度にどのように影響を与えるか調査した。入力文を形態素数で分類し、それぞれの RIBES 値を求めた結果を表 3 と図 3 に示す [3]。 入力文が 50 形態素以下の場合、lader ありの方が高い RIBES スコアを示している。したがって、短文やそれほど長くない文に対しては、事前並び替えが翻訳精度向上につながっていることがわかる。一方、50 形態素を超える長文になると、ベースラインの方が高い RIBES スコアを示している。これは、あまりに長い文になると事前並び替えの精度が低下し、結果としてその後の翻訳に悪影響を与えていると考えられる。あらまし文は元の法令文より比較的短い文で構成されているが、それでも 50 形態素を超えるような長文はあらまし全体の約 3 割を占めている。つまり、あらましの SMT における長文の取り扱いは、翻訳精度向上における重要な要素であるといえる。

[3]BLEU は一文ごとの評価には適さないため、ここでは使用しなかった。

表 4: lader-hybrid を含めた各手法の翻訳精度の比較

手法	BLEU	RIBES
ベースライン	37.72	73.40
lader あり	38.57	73.02
lader-hybrid	38.49	75.13

5 ハイブリッド手法

4 節で示したように、lader は長文に弱く、それがあらまし全体の翻訳精度を下げる要因となっていた。そこで、二つのデコーダを用意し、50 形態素以下の文は事前並び替えをしてから lader ありのデコーダで翻訳、50 形態素を超える文は事前並び替えせずにベースラインのデコーダで翻訳するハイブリッドな手法 (lader-hybrid) を新たに提案する。

本手法の有効性を検証するため、4 節と同様の翻訳実験を行い、各手法の翻訳精度を比較した。その結果を表 4 に示す。lader-hybrid は、BLEU と RIBES の両方で、ベースラインより有意にスコアが上昇した ($p < 0.05$)。この結果から、入力文の長さに応じて lader の有無を分ける手法は、翻訳精度向上に有効であるといえる。

6 おわりに

本稿では、あらましの SMT に事前並び替え手法を適用した。単純な適用では翻訳精度が向上したとは必ずしもいえなかったが、入力文の長さに応じて事前並び替えの有無を分ける翻訳手法を提案することで、二つの自動評価指標において有意な翻訳精度向上がみられた。

今後の課題として、法分野以外の文書に lader を適用することを考えている。他分野においても長文の並び替えに失敗しやすいのか、入力文の長さに応じて事

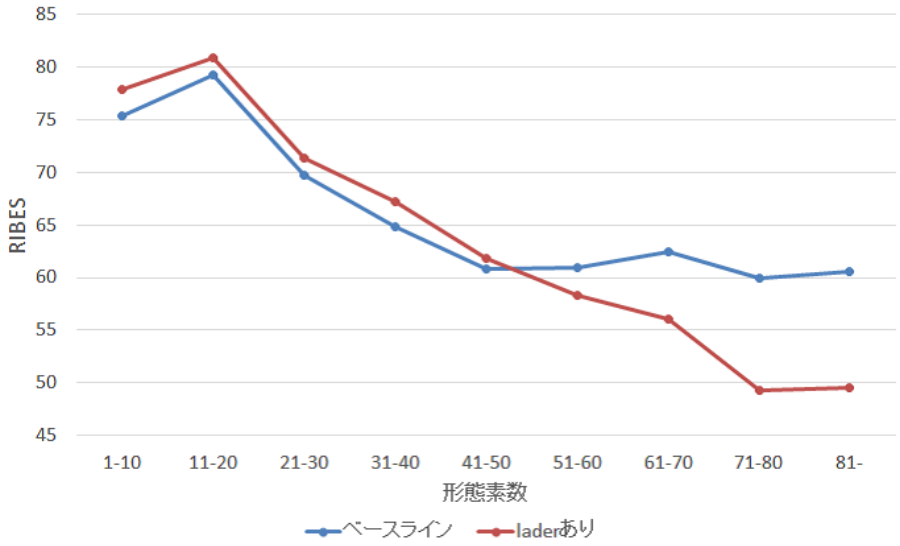

図 3: 形態素数別の RIBES スコア

前並び替えの有無を分ける手法は有効であるかどうか検証する。また、他の事前並び替え手法を適用した場合、長文の並び替えが翻訳精度にどのように影響するか調査することも考えている。

参考文献

[1] Toyama, K., Saito, D., Sekine, Y., Ogawa, Y., Kakuta, T., Kimura, T., and Matsuura, Y. *Design and Development of Japanese Law Translation Database System*, Law via the Internet 2011, 12 pages, 2011.

[2] Inagi, D., Ogawa, Y., Nakamura, M., Ohno, T., and Toyama, K. *Statistical Machine Translation for Outlines of Japanese Statutes*, Proc. of JURISIN 2013, pp. 37–49, 2013.

[3] Okada, K., Ogawa, Y., Nakamura, M., Ohno, T., and Toyama, K. *Improvement of Translation Accuracy for the Outlines of Japanese Statutes by Splitting Parenthesized Expressions*, Proc. of KSE 2015, pp. 67–72, 2015.

[4] Neubig, G., Watanabe, T., and Mori, S. *Inducing a discriminative parser to optimize machine translation reordering*, Proc. of EMNLP 2012, pp. 843–853, 2012.

[5] 井上政文 官報登載の「法令のあらまし」について, 官報百年のあゆみ, pp. 179–184, 1983.

[6] Hoshino, S., Miyao, Y., Sudoh, K., and Nagata, M. *Two-Stage Pre-ordering for Japanese-to-English Statistical Machine Translation*, Proc. of IJCNLP 2013, pp. 1062–1066, 2013.

[7] Katz-Brown, J., and Collins, M. *Syntactic reordering in preprocessing for Japanese → English translation: MIT system description for NTCIR-7 patent translation task*, Proc. of the NTCIR-7 Workshop Meeting, 2008.

[8] 工藤拓, 松本裕治 チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol.43, No.6, pp. 1834–1842, 2002.

[9] 山田将之, 小川泰弘, 外山勝彦 構文情報付き法律文コーパスの設計と構築, 言語処理学会第 14 回年次大会講演論文集, pp. 604–607, 2008.

[10] Kudo, T., Yamamoto, K., and Matsumoto, Y. *Applying Conditional Random Fields to Japanese Morphological Analysis*, Proc. of EMNLP 2004, pp. 230–237, 2004.

[11] Stolcke, A. *SRILM - An Extensible Language Modeling Toolkit*, Proc. of ICSLP 2002, pp. 901–904, 2002.

[12] Och, F. J., and Ney, H. *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, Vol. 29, No. 1, pp. 19–51, 2005.

[13] Papineni, K., Roukos, S., Ward, T., and Zhu, W. *BLEU: A Method for Automatic Evaluation of Machine Translation*, Proc. of ACL 2002, pp. 138–145, 2002.

[14] Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. *Automatic Evaluation of Translation Quality for Distant Language Pairs*, Proc. of EMNLP 2010, pp. 944–952, 2010.