# 括弧表現の分割による法令のあらましの翻訳精度向上

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@kl.i.is.nagoya-u.ac.jp

## 1　はじめに

近年、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などのため、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。それに対して、法務省は「日本法令外国語訳データベースシステム」(以下、JLT) を公開し、日本法令の英訳を提供している。しかし、所管府省庁ごとでの翻訳作業や、法令翻訳の難しさなどの理由から、翻訳品質のばらつきや翻訳作業の遅延が生じている。

そのような背景を基に、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の研究が進められてきた [1]。あらましは公布された法令を要約した公的な文書であり、日本法令に関する情報の概要を理解するために有用である。また、あらましは、元の法令より短く、簡潔な文で構成されるため、機械翻訳に適しているといえる。

文献 [1] のあらましの翻訳システムは、翻訳モデルの学習に法令文を使用している。同様に、あらまし文も翻訳モデルの学習に使用できると考えられる。しかし、法令文は一文が長く、また、あらまし文も法令文より短いとはいえ、長文が多く含まれている。長文から翻訳モデルを学習することは、対訳候補の多さから、アライメント精度低下の原因となる。同様に翻訳システムの入力文についても、長文の翻訳に失敗することが多い。しかし、長文を複数の短文に分割することができれば、アライメントや翻訳精度の改善に期待ができる。そこで本研究では、法令やあらましにおける括弧表現に着目した。

法令や法令のあらましでは、括弧を用いた表現（以下、括弧表現）により補足や説明など、さまざまな情報が提示される。括弧表現の使い方によって、開括弧と閉括弧で囲まれている部分は独立した文（以下、括弧文）とみなすことができる。この場合、括弧文が抜き出された元の文（以下、括弧抜き文）と括弧文は、互いに独立している。そのため、法令やあらまし文は、括弧文を抜き出すことにより、主語や目的語を欠くことなく分割できる。そこで本研究では、翻訳精度の向上のため、法令文やあらまし文を括弧表現に基づいて分割し、それらを学習コーパスや入力文とする統計的機械翻訳手法を提案する。本稿では、まず、法令やあらましの括弧表現について調査し、使い方によって括弧文かどうか検討する。その後、その結果を基にして、学習コーパスと入力文を分割する。また、翻訳実験により、提案手法の有効性を明らかにする。

## 2　括弧表現の分類

### 2.1　法令の括弧表現の分類

法令の表現に用いられる括弧には、かぎ括弧「 」と丸括弧（ ）が存在する。文献 [2] では、かぎ括弧と丸括弧の使い方を以下のように分類している。

かぎ括弧は以下の場合に用いられる。

**(i)** 用語を定義する場合で、その用語を示す場合
**(ii)** ある用語について略称を定める場合で、その略称を示す場合
**(iii)** 他の条文を準用する場合で、その準用する条文の読み替えを行う部分を示す場合
**(iv)** 他の条文を読み替えて適用する場合で、その読み替えを行う部分を示す場合
**(v)** 既存の法令の一部を改正する法令等において、字句を改め、加え、又は削る部分を示す場合

一方、丸括弧は以下の場合に用いられる。

**(1)** 目次において章、節等に含まれる条の範囲を示す場合
**(2)** 条文の中で法令等の題名又は件名の次に法令番号等を示す場合
**(3)** 条文見出しを付ける場合
**(4)** 括弧の前の字句について略称を定める場合
**(5)** 括弧の前の字句を定義する場合
**(6)** 括弧の前の字句から特定の範囲のものを除外し、その字句に特定のものを含ませ、又はその字句を特定の範囲に限定する場合

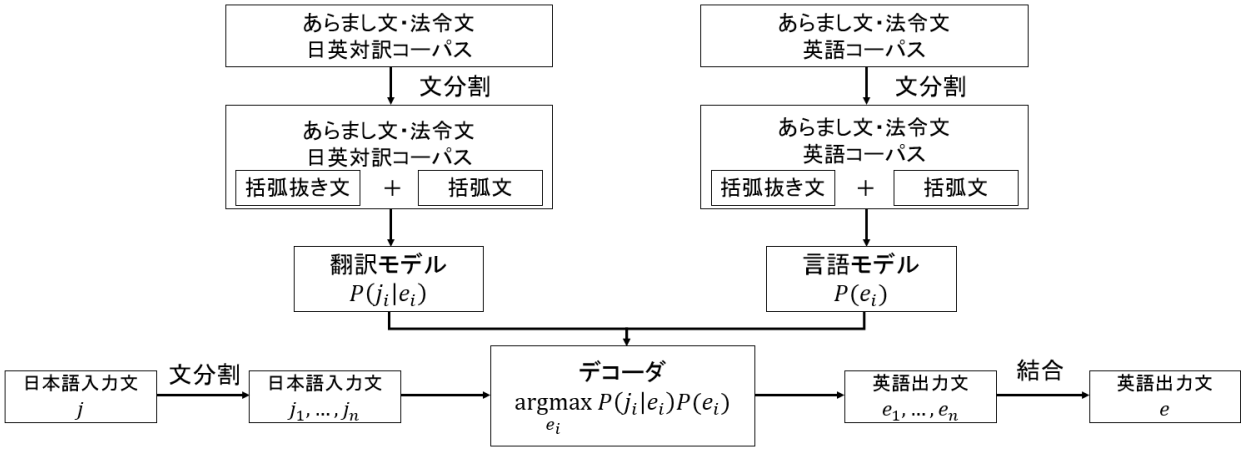


図 1: 提案手法の概略図

**(7)** 括弧の前の字句を特定の場合に別の字句に置き換える場合
**(8)** 他の条項を引用する場合で、その引用条文の要旨を付ける場合
**(9)** 別表または様式等において、本則の規定との関係を明らかにするため、その別表又は様式等について定める本則中の規定を示す場合
**(10)** 左横書きの例規等において、条の中の号である「(1)、(2)、(3)」を示す場合
**(11)** 条の中の号の細分を更に細分する場合で、その区分を示す場合

### 2.2 あらましの括弧表現の分類

あらましにおける括弧表現を調査した。その結果、括弧の使い方は基本的に法令と同じだが、あらましに特有な丸括弧の使い方として、以下の三つを発見した。

**(12)** 題名の中で法令名の次に法令番号や所管府省庁を示す場合
**(13)** 関係規定を付ける場合
**(14)** 箇条書きの番号である「(一)、(二)、(三)」や「(イ)、(ロ)、(ハ)」を示す場合

## 3 提案手法

学習や入力として使用する法令文やあらまし文には長文が多く、翻訳精度を下げる要因となる。そこで、学習コーパスや入力文に対して括弧文を分割し、それらを学習コーパスや入力文とする統計的機械翻訳手法を提案する。今回の提案手法の概略を図 1 に示す。また、括弧文の分割を含めたコーパスの整形は、次の 4 ステップで行う。

1. 括弧文と括弧抜き文の分離
2. 文の形態素への分割
3. 英語の小文字化
4. 翻訳モデルの学習コーパスに対するクリーニング

学習コーパスの分割に関する詳細は 3.2 節、入力文の分割に関する詳細は 3.3 節において示す。

### 3.1 括弧表現に応じた分割

2 節で述べた括弧表現の分類に応じて、括弧表現が括弧文であれば、すなわち、開括弧と閉括弧で囲まれている部分と、その部分を取り除いた文が互いに独立であれば、分割する。かぎ括弧の場合、どの括弧表現も括弧文ではないため、分割は行わない。丸括弧の場合、(1) から (9)、(12)、(13) の括弧表現は括弧文であるため、分割する。(10)、(11)、(14) の括弧表現は括弧文ではないため、かぎ括弧と同様、分割しない。

また、一文に括弧文が複数存在する場合、以下のように対応する。

- 一文内に括弧文が複数ある場合は、すべての括弧文を抜き出す。
- 括弧文内にさらに別の括弧文があり、括弧が入れ子になっている場合は、まず一番外側の括弧文を抜き出す。その後、抜き出した括弧文の中でさらに括弧文を再帰的に抜き出す。

次節以降、学習コーパス及び入力文の分割は、括弧文を含む文に対してのみ行う。

### 3.2 学習コーパスの分割

学習コーパスである日英対訳コーパスや英語コーパスに対して、括弧文の分割を行う。まず、対訳文から

表 1: 翻訳モデルの学習に使用したコーパスの文数

| 手法 | | クリーニング前 | クリーニング後 | 削除された文 |
|---|---|---|---|---|
| ベースライン | | 168,719 | 150,472 | 18,247 |
| 提案手法 | 括弧なし文 | 142,199 | 133,847 | 8,352 |
| | 括弧抜き文 | 26,520 | 20,682 | 5,838 |
| | 括弧文 | 19,281 | 19,139 | 142 |
| | 計 | 188,000 | 173,668 | 14,332 |

括弧文をすべて抜き出した括弧抜き文を、新たな対訳とする。その後、抜き出した括弧文から対訳関係の同定により、括弧文の対訳を獲得する。そして、獲得した括弧文、括弧抜き文、さらにもともと括弧文を含まない括弧なし文を合わせて、新たなコーパスとする。

### 3.3 入力文の分割

学習コーパス同様、入力文に対しても括弧文を分割する。括弧文を含む文は、括弧抜き文と括弧文に分割され、別々に翻訳される。その後、括弧文の翻訳を括弧抜き文の翻訳に挿入し、一文に結合する。このとき、括弧文の翻訳を挿入する位置は、分割前の文において、その括弧文が登場した直前の単語の訳語の直後とする。

## 4 翻訳実験

本節では、提案手法の有効性を検証するため、翻訳実験を行う。本実験では、括弧文を分割しないベースライン手法と、分割する提案手法の二つを比較した。

提案手法では、元の対訳文から抜き出した括弧文を新たな対訳文としてコーパスに追加する。その際、一文から抜き出した括弧文が複数存在する場合、それぞれの括弧文が、どの対訳と対応するか同定する必要がある。しかし、今回は対訳関係の同定まで実装できなかったため、抜き出した括弧文が日英ともに一文である場合のみ、対訳コーパスに追加した。複数の括弧文への対処は今後の課題とする。

### 4.1 実験手順

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[3]（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[4] と GIZA++[5] によって学習した。その際、JLT 法令文対訳コーパス (313 法令、分割前 166,977 文) と平成 22 年公布のあらまし対訳コーパス (72 法律、分割前 1,742 文) を合わせたコーパスを用いた。また、翻訳モデルの学習コーパスに対するクリーニングとして、日英どちらかの文長が 80 単語を超える場合、または、日英対訳の間の単語数の比が 9 倍を超える場合は、その対訳文を削除した。各手法に対して、クリーニングを行ったときの対訳コーパスの文数を表 1 に示す。

表 2: 各手法に対する BLEU, RIBES スコア

| 手法 | BLEU | RIBES |
|---|---|---|
| ベースライン | 37.37 | 71.05 |
| 提案手法 | 39.32 | 72.31 |

デコーダのパラメータは MERT[6] によって最適化した。その際、デベロップメントデータとして、平成 22 年公布のあらまし対訳コーパスからランダムに抽出した 300 文を用いた。

入力文には平成 23 年公布のあらまし (52 法律、分割前 1,371 文)、デコーダには Moses を用いた。出力文は、BLEU[7]、RIBES[8] によって自動評価した。なお、自動評価用の参照訳は 2 セット使用した。

### 4.2 実験結果と考察

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。提案手法は、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 ($p < 0.05$)。

また、RIBES は一文ごとの評価が可能なので、括弧ありの 120 文と括弧なしの 1,251 文に分けて、自動評価を行った [2]。その結果を表 3 に示す。提案手法における括弧文を含む文の RIBES 値は、ベースラインより 2.5 ポイント以上高くなり、括弧なし文の RIBES 値でも 1.5 ポイント以上高い結果となった。括弧なし文に対する RIBES 値の上昇は、クリーニングによって学習コーパスから削除される文が減少したことと、学習に使用したコーパスが文分割により短くなった結果、アライメント精度が向上したことが理由として挙

[1] http://www.statmt.org/moses/

[2] 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。

| | |
|---|---|
| 原文 | 地方公共団体等以外の者は、港湾運営会社の株式について、保有基準割合（**原則として総株主の議決権の一〇〇分の二〇**）以上の数の議決権を取得し、又は保有してはならないこととした。 |
| ベースライン | no person other than the holding ratio threshold **of the voting rights of all shareholders** or more **( 20 / 100ths** , with respect to the shares of the company shall not acquire or hold a number of voting rights in the local government , etc. port operation **principle )** . |
| 提案手法 | local governments , etc. with regard to shares of the company shall not be acquired or held by a person other than the holding ratio **( 100 20 percent of the voting rights of all shareholders in principle )** threshold of the voting rights of the port operation . |
| 参照訳 1 | a person other than local governments or the like shall not obtain or hold a number of voting rights for the shares of the port operation company that is at least equal to the holding ratio threshold **( 20 percent of the voting rights of all shareholders , in principle )** . |
| 参照訳 2 | no person except for local public entities , etc. may acquire or hold voting rights of the port and harbor management company equivalent to or exceeding the holding ratio threshold **( twenty percent of the voting rights of all shareholders in principle )** . |

図 2: ベースラインと提案手法の翻訳例

表 3: 括弧文の有無で分けた RIBES スコア

| 手法 | 括弧あり文 | 括弧なし文 |
|---|---|---|
| ベースライン | 53.14 | 70.49 |
| 提案手法 | 55.74 | 72.12 |

げられる。括弧文を含む文の RIBES 値の上昇は、先述の理由に加えて、入力文が分割され一文が短くなった結果、正しい翻訳が多く得られるようになったためだと考えられる。

また、図 2 にベースラインと提案手法の翻訳例を示す。この図における太字は、括弧文とその翻訳を表している。ベースラインでは、括弧文の内容が三つに分かれて翻訳されているのに対し、提案手法では、一つにまとまって翻訳されている。また、この翻訳では、ベースラインの RIBES 値は 52.98 であるのに対し、提案手法では 67.04 となっている。

以上から、括弧文の分割を行った結果、入力文全体の翻訳精度が向上したことが分かり、提案手法の有効性を示すことができた。

## 5 おわりに

本研究では、法令やあらましの括弧表現に着目した分割手法を提案し、翻訳実験によってその有効性を確認した。提案手法では、翻訳モデルの学習に使用した対訳コーパスに対して分割を行ったが、今回は抜き出した複数の括弧文の対訳関係の同定は行わなかったため、括弧文の対訳をすべて獲得することはできなかった。したがって、今後は括弧文の対訳をより多く獲得し、翻訳精度のさらなる改善を目指す。

## 参考文献


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

[2] 株式会社ぎょうせい法制執務研究会: 図説法制執務入門. ぎょうせい, 2013.

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

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

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

[6] Och, F. J.: Minimum Error Rate Training in Statistical Machine Translation. *Proc. of ACL 2003*, pp. 160–167, 2003.

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

[8] 平尾努, 磯崎秀樹, Duh, K., 須藤克仁, 塚田元, 永田昌明: RIBES: 順位相関に基づく翻訳の自動評価法. 言語処理学会 第 17 回年次大会発表論文集, pp. 1115–1118, 2011.