法令対訳コーパスからの複単語表現抽出

坂本 聡美[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
satomi@kl.i.is.nagoya-u.ac.jp

1　はじめに

現在、日本の法情報の国際的発信が進められている。主要な法令は既に英訳され、法務省の日本法令外国語訳データベースシステム (JLT)[1] で公開されている。しかし、法令文特有の用語や、日常語とは意味が異なる用語、法令以外の様々な分野の専門用語が法令理解の妨げになっている。法令で用いられる用語の意味は、特殊なものを除き、日常生活の中で用いられる通常の用語の意味と解するのが原則である [1] が、特殊か否かの判定は日本法令の専門家でなければできない。しかし、JLT の利用者は必ずしも法令に精通していない。したがって、真に国際的な法情報発信の達成には、法令の英訳だけではなく、法令文における用語の概念定義を体系的に示した多言語法令ターミノロジーも提供する必要がある。実際、欧州連合では、公用語を含む 26 言語 (約 800 万語) を収録した多言語ターミノロジー IATE[2] を提供し、域内の円滑な情報共有に貢献している。本研究の目的は、日本語を含む多言語法令ターミノロジーの構築である。その一環として、法令用語とその対訳を収集している。

法令ターミノロジーへ収録すべき用語として、複数の単語から構成される表現がある。これは「複単語表現 (MWE)」や「複合辞」と呼ばれている。MWE の中には対訳が構成的ではないものがあり、表現を構成する各単語の対訳を単に組合せるだけでは全体の対訳はできない。例えば、「民事訴訟法」を形態素解析すると、「民事」「訴訟」「法」の 3 単語に分割される。しかし、「民事訴訟法」の対訳 “Code of Civil Procedure” には、「訴訟」の対訳 “litigation” が含まれておらず、対訳は構成的でない。

また、複数の単語からなる定型表現も MWE である。法令文には、「〜に違反する場合」や「するものとする」などの機能的な定形表現が多く出現する。このような定型表現には、対訳が構成的ではなく、特別な意味を持つものもある。例えば、「するものとする」は、義務を示す「しなければならない」とは異なり、ものごとの原則を示す場合に念のため規定するために用いられることがある。さらに、定型表現であるにも関わらず、JLT では辞書に対訳が無いために訳語が統一されていないことが多い。そのため、MWE は積極的にターミノロジーへ収録することが望ましい。

一般的な日本語 MWE コーパスである日本語フレーズ辞書[3] は既に公開されているが、専門用語は収録されていない。既存の MWE の自動判別方法には、依存解析結果の自動修正 [3] や、YamCha による機能的 MWE の検出 [4]、文節クラスの共起情報を用いた長い名詞句表現の自動抽出 [5] がある。しかし、本研究の対象である法令文は構文構造が複雑であることが多いため、依存関係や文節情報の利用は容易ではない。また、言語資源に乏しく、教師データを用いた手法の適用も容易ではない。

そこで、本稿では法令対訳コーパスからの MWE 抽出手法を提案し、その有効性を実験により明らかにする。提案手法は、Tsvetkov らの教師なし手法 [6] を改良したものである。法令文の特徴に対応するため、MWE のフィルタリング尺度を PMI^k から重複条件付き文書頻度 [7] へ変更する。

2　小規模対訳コーパスを用いた MWE 抽出手法

Tsvetkov らの MWE 抽出手法 [6] は、小規模の対訳コーパスから MWE を獲得するために提案された。対訳テキストにおいて、1 単語対 1 単語のアライメントがされない表現は、すべて MWE の候補であるという考えに基づいている。文献 [6] では、ヘブライ語の MWE を抽出するため、ヘブライ語と英語の対訳コーパス (主として新聞記事) を対象に抽出実験を行っている。この手法を順に説明する。

(1) **前処理**　コーパスに前処理を施すことで、言語特有の違いや自動単語アライメントの誤りを低減させる。ここで行う処理は、トークン化、レンマ化、句読点の除去、言語間において直接対応する単語が存在しない語の除去である。

(2) **単語アライメント**　MWE 候補を特定するため、対

[1] http://www.japaneselawtranslation.go.jp/
[2] http://iate.europa.eu/
[3] http://jefi.info/

訳コーパスの単語アライメントを計算する。GIZA++[4]を用いて双方向の単語アライメント取得し、これらをマージして多単語対多単語の対応を許したアライメントを得る。アライメントのマージは、MWE 候補を増やすため、対応の和集合を取る規則 union を用いる。次に、1 単語対 1 単語のアライメントであるものを対訳辞書で確認する。もし、対訳が辞書に既に存在する場合は、対訳テキストから取り除き、記号「*」に置換する。この置換処理により、対訳が構成的な単語を MWE の候補から外せる。

(3) **MWE 候補のランキングとフィルタリング**　この時点で、対訳テキストは「*」により区切られた単語列となっている。これらの単語列は、対訳が 1 単語対 1 単語に対応していないため、単語列の任意の部分を MWE 候補と見なせる。単語列中のどの部分が抽出すべき MWE であるかを判別するため、単語列の任意のバイグラムに対して自己相互情報量 PMI^k を式 (1) により計算する。

$$PMI^k(x,y) = \frac{P(x,y)^k}{P(x)P(y)} \tag{1}$$

ここで、$P(x)$ はコーパス中のユニグラム x の出現回数、$P(x,y)$ はバイグラム xy の出現回数である。重み k は任意に設定する。PMI^k が閾値以上の場合は連接する表現として認め、閾値を下回る場合は MWE の切れ目であるとする。最後に、前処理により変形した部分を本文中で使用されている形に戻し、2 単語以上の単語列を MWE として抽出する。

3　提案手法

3.1　マージ規則の変更

Tsuvekov らは、双方向の単語アライメントをマージするために和集合を用いている。単方向のアライメントでは、ある言語の単語それぞれが、他方の言語の 1 個以上の単語へ必ず対応付けされる。そのため、双方向のアライメントを考えると、ある方向では 1 単語対 1 単語に対応付けされていても、逆方向では 1 単語対多単語の対応である場合がある。このとき、和集合をマージ規則として用いると、積集合を用いる場合よりも 1 単語対 1 単語の対応が減り、辞書引きの対象となる対応が減るため、「*」への置換数が減少する可能性がある。結果として、MWE 候補は増えるが、対訳テキストを「*」で区切ることが十分にできず、後の MWE 抽出への悪影響が懸念される。

一方で、積集合をマージ規則として用いると、原言語のすべての単語が対象言語の単語にアライメントされることが必ずしも保証されない。この場合も、辞書引きの対象となる単語の数が不当に少なくなり、「*」への置換率が下がる可能性がある。

法令は構文構造が複雑であることが多いため、一般文書よりも GIZA++のアライメントが誤りやすいと考えられる。そのため、多単語対多単語の対応が多くなりやすく、「*」への置換率が下がりやすい可能性がある。そこで、アライメントのマージ規則による「*」への置換率の違いを調査し、置換率の最も高いものを選んで MWE 抽出に用いることとする。GIZA++には和集合と積集合の間をとるマージ規則も用意されているため、これらを含めて調査する。

3.2　フィルタリング尺度の変更

Tsuvekov らの使用した PMI^k は、2 単語の共起性を測る尺度である。各構成語の単体での出現数が大きいと PMI^k の値は一般に小さくなるため、高頻度語で構成されている低頻度な MWE の抽出は容易ではない。例えば、「許可申請書」について考える。この表現を形態素解析すると、「許可」「申請」「書」の 3 単語に分割される。また、この対訳は “license application form” “written application for permission” など複数存在し、元の用語に対して非構成的な対訳となるものがある。そのため、MWE として抽出すべきである。「許可」「申請」「書」の単言語コーパス中での出現数は、それぞれ 17,595 個、19,400 個、63,692 個であり、法令文に比較的出現しやすい語である。一方、「許可申請」と「申請書」の出現数はそれぞれ 151 個、3,821 個で、PMI^k 値はそれぞれ 0.00223、3.80 となる。そのため、閾値 1 の場合に抽出できるものは「申請書」だけとなる。法令文に比較的出現しにくい「許可申請」を伴う「許可申請書」は抽出できない。また、「認可申請書」や「登録申請書」などの似た表現も同様に、構成語と比較して出現数が少なくなりがちである。そのため、PMI^k 値が小さくなり、連接する表現として判定される可能性が低くなる。つまり、PMI^k は、構成語の一部が共通する表現のバリエーションを抽出するためには適切でない。

このような表現のバリエーションは、法令文には多く存在すると考えられる。そこで、提案手法では「*」で区切られた単語列から MWE を抽出するフィルタリング尺度に**重複条件付き文書頻度** [7] を用いる。重複条件付き文書頻度 (df_k) とは、コーパス中で、ある文字列を k 回以上含む文書の数である。武田ら [7] は、

[4] http://www.statmt.org/moses/giza/GIZA++.html

特徴量 df_2/df_1 が自立語境界を判定する基準となることを示した。df_2/df_1 による用語抽出は、はじめに入力された単語列を全体でスコアが最大になるように分割し、分割後の各部分単語列のうち df_1/N が一定値以内のものを抽出する。ここで、N はコーパスの文書数である。部分単語列 x_i に対するスコアは式 (2) により計算する。

$$Score(x_i) = \begin{cases} -\infty & (df_2 < 3) \\ \log 0.5 & (df_2 \geq 3, \quad df_1/N > 0.5) \\ \log(df_2(x_i)/df_1(x_i)) & \\ \qquad (df_2 \geq 3, \quad df_1/N \leq 0.5) \end{cases} \tag{2}$$

文献 [7] では、文書全体を 1 文につなげたものを入力の単語列としている。しかし、法令文は 1 文あたりの単語数が数個から千個以上のものまであり、1 文書 (1 法令) あたりの大きさにもばらつきがある。ある単語が 1 文書あたりに出現する回数は、文書の大きさにも依存する。コーパス内で文書の大きさにばらつきがあると、文書特有の用語間で重複条件付き文書頻度の大きさに差が生まれやすい。一方で、df_2/df_1 による分割は相対的なスコアの差が重要となる。そこで、「*」への置換により抽出範囲をあらかじめ限定することで、ノイズを抑えて抽出できると期待される。

提案手法では、「*」で区切られた部分文字列をさらに df_2/df_1 で分割し、df_1/N が一定値以内であるものを MWE として抽出する。ただし、他の候補と出現数が同じで、かつ、その部分文字列であるものは除く。

4 実験 1: 適切なマージ規則の決定

MWE の抽出実験を行うため、適切なマージ規則をあらかじめ決定する必要がある。そこで、各規則を用いた場合の「*」への置換率を調べ、最大となるものを実際の抽出に使う。

4.1 実験概要

本研究の目的は法令文からの用語獲得であるため、JLT 掲載の法令日英対訳データ 313 本 (166,977 文) を MWE の抽出元とする。また、官報情報検索サービス[5] から収集した日本法令 9,915 本 (1,627,045 文) を単言語コーパスとして用いる。対訳辞書は「英辞郎 (第五版)」と、人手により作成した漢数字・記号の対訳データを用いる。対訳辞書のうち実際に使用したのは、対訳コーパスに出現する単語で、かつ、1 単語対 1 単語の対訳 27,096 個である。

[5] https://search.npb.go.jp/kanpou/

表 1: 「*」への置換率

マージ規則	置換箇所数	置換率 (%)
union	1,099,240	15.0
grow-diag-final	1,142,848	15.6
grow-diag-final-and	**1,158,524**	**15.8**
grow-diag	1,078,594	14.8
grow	1,036,032	14.2
intersection	1,109,478	15.2

手順は、はじめに単言語コーパスと対訳コーパスに対して、分かち書きと単語のレンマ化を行う。日本語文には MeCab[6](IPA 辞書使用) を用いる。英語文の分かち書きには Moses[7] の tokenizer.perl を、レンマ化には Ruby のライブラリ lemmatizer を用いる。次に、対訳コーパスから、英語か日本語のどちらかが 80 語を越える文を削除する。これは、語数の多過ぎる文が GIZA++のエラーの原因になることを防ぐためである。また、単言語コーパスから単語ユニグラムと単語バイグラムの出現数を、対訳コーパスから重複条件付き文書頻度をそれぞれ求める。

次に、GIZA++を用いて、多単語対多単語の対応を認めた単語アライメントを対訳コーパスから得る。比較のため、アライメントのマージ規則は、union、grow-diag-final、grow-diag-final-and、grow-diag、grow、intersection の 6 種類を用いる。得られた対訳のうち、1 単語対 1 単語の対訳について、辞書中にその対訳が存在するかどうかを確認する。もし辞書に存在した場合は、対訳を「*」で置換する。

4.2 実験結果

長すぎる文の削除により、対訳文は 166,977 個から 148,912 個になった。各マージ規則に対する「*」への置換率を表 1 に示す。置換率は、対訳コーパスの単語のべ数 7,310,804 個に対する「*」の個数の割合である。「*」への置換率が最も高かったのは grow-diag-final-and であったため、これを用いて次節で MWE を抽出する。

5 実験 2: MWE 抽出

提案手法の有効性を確認するため、既存手法と提案手法それぞれを用いて、「*」で区切られた単語列から MWE を抽出し、抽出数と精度を比較する。

[6] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

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

表 2: 抽出結果の比較

抽出方法	抽出数	正解数	精度 (%)
PMI^k(閾値 100)	251	102	40.6
PMI^k(閾値 10)	5,928	917	15.5
PMI^k(閾値 1)	23,117	5,892	25.5
重複条件付き文書頻度	39,544	32,869	83.1

5.1 実験概要

実験 1 により、「*」で区切られた単語列が得られる。これを元に MWE を抽出する。既存手法による抽出は 2 節で述べた方法を用いる。PMI^k の k は Tsvetkov ら [6] と同じ 2.7 とし、閾値は 100、10、1 の 3 種類で実験する。提案手法による抽出は 3.2 節で述べた方法を用いる。閾値は、武田ら [7] と同様に、df_1/N が 0.00005 より大きく 0.1 より小さいもので、かつ、2 単語以上のものとする。

5.2 実験結果と考察

MWE 候補は、「*」で区切られた単語列中の任意の部分 (2 単語以上) であり、13,692,940 個が得られた。各手法の抽出数と精度を表 2 に示す。正解は、記号を含まず、かつ、数や番号・条項に関係する語を含まないものとした。

既存手法 PMI^k のどの閾値を設定した場合と比べても、提案手法の重複条件付き文書頻度による抽出精度は高くなった。また、既存手法よりも多くの MWE を抽出できた。これにより、提案手法は有効であるといえる。

閾値 1 の既存手法で抽出した正解 MWE 5,892 個のうち、提案手法でも抽出できたのは 931 個 (15.8%) であった。例えば、「いずれかに該当する場合」は既存手法のみで抽出され、「いずれかに該当する事由」は提案手法のみで抽出された。また、「いずれかに該当する場合を除く」「いずれかに該当する場合における」「いずれかに該当する場合において」は両方の手法で抽出された。このような違いが生じた原因は、フィルタリング尺度の特徴の違いである。単言語コーパス中での「する事由」の出現数 723 個は、「する」の 659,146 個、「事由」の 9,254 個に対して低い。そのため、PMI^k は 0.0086 で閾値 1 を越えず、既存手法では抽出されなかった。一方で、「いずれかに該当する場合」の df_1 は 267 で、df_1/N が 0.1 以上となるため、提案手法では抽出されなかった。既存手法でのみ抽出された MWE のうち、df_1/N が 0.1 以上となるものは 39.4%であり、コーパス中の文書に広く出現する表現が多い。このように、提案手法は文書に広く出現する MWE の抽出は容易ではないが、比較的低頻度の表現を抽出できることが分かる。

また、既存手法でのみ抽出された MWE の 41.2%は、df_2 が 3 未満であった。例えば、「臨床修練」は対訳コーパス中で 1 つの文書にのみ出現する用語であるため、df_2 が 1 である。この MWE は、既存手法の閾値 1 の場合は抽出できたが、提案手法では抽出できなかった。式 (2) で df_2 が 3 未満のものにマイナス無限大のスコアを付けることが原因だと考えられる。ある法令に固有の MWE のうち、キーワードとなるものは TF-IDF が高いことが期待される。そのため、TF-IDF を組み合わせたスコアを用いることで更なる改良が見込める。

6 おわりに

法令対訳コーパスからの MWE 抽出を目的として、対訳コーパスのアライメント誤りを利用した教師なし手法を改良した手法を提案した。提案手法は、コーパス内での単語の出現分布にばらつきがある文書に対して、比較的低頻度の MWE を抽出することができる。実験の結果、提案手法では 80%を越える精度が得られた。さらに、既存手法よりも多くの MWE を抽出できており、その有効性を確認した。

今後の課題としては、提案手法では抽出できなかった MWE を抽出するため、抽出条件を改良する。また、多言語法令ターミノロジーの設計と、必要な用語の選定と収集を行う計画である。さらに、他分野への応用を視野に入れた抽出手法の一般化も検討する。

参考文献

[1] 田島信威. 最新法令の読解法 : やさしい法令の読み方. ぎょうせい, 1996.

[2] 首藤公昭, 田辺利文. 日本語の複単語表現辞書 : JDMWE. 自然言語処理, Vol. 17, No. 5, pp. 51–74, 2010.

[3] 塩田嶺明, 中澤敏明, 黒橋禎夫. 単語間結合度に基づく複単語表現のアライメントの改善. 言語処理学会 第 20 回年次大会, pp. 376–379, 2014.

[4] 注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史. 日本語機能表現の自動検出と統計的係り受け解析への応用. 自然言語処理, Vol. 14, No. 5, pp. 167–197, 2007.

[5] 潮田明. 連体形複合辞に修飾された名詞句の係り受け解析. 言語処理学会 第 18 回年次大会, pp. 967–970, 2012.

[6] Yulia Tsvetkov and Shuly Wintner. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, Vol. 18, No. 04, pp. 549–573, 2010.

[7] 武田善行, 梅村恭司. キーワード抽出を実現する文書頻度分析. 計量国語学, Vol. 23, No. 2, pp. 65–90, 2001.