多言語対訳サイトの特徴分析に基づくウェブ上の対訳データのフィルタリング*

三好 健悟† Yizhen Wei† 田村 拓也‡ 宇津呂 武仁† 永田 昌明§
†筑波大学大学院 システム情報工学研究科・群 ‡筑波大学 理工学群 工学システム学類
§NTT コミュニケーション科学基礎研究所

1 はじめに

近年，機械翻訳分野においては，ウェブから対訳サイトを大量に収集し，対訳コーパスを作成する技術の研究が進められている [3, 1, 5, 6, 10]．しかし，現時点において，ウェブを情報源として対訳コーパスを作成する技術の問題点として，収集される対訳サイト集合の中に，自動翻訳による対訳サイトが多く含まれる点が挙げられる．自動翻訳による対訳文対が多く含まれる対訳コーパスは，そのままでは機械翻訳モデルの訓練・評価データとしては適さない．そのため，収集される対訳サイト集合のうち人手によって翻訳された対訳サイトのみを抽出することにより，ウェブから収集された対訳コーパスを高品質化する必要がある．

本論文では，ウェブから収集された日英対訳文対集合である JParaCrawl V2 を対象に，その中から人手翻訳による対訳サイトを自動抽出する手法を提案する．提案手法においては，自動翻訳された対訳サイトに特有の現象として，(1) 自動翻訳によって生成された文の特性と，最新の言語モデル [2] に則った文の特性とを比較し，文中における語彙の選択基準において両者の間に差異が見られる傾向があること，および，(2) 文中の一部の固有名詞等は異なるものの，文中の他のフレーズが共有されるテンプレート文が多く含まれること，の二種類の現象に着目する．(1) の特徴に対しては，日本語文に対して，最新の言語モデルを用いて文中の単語の生成確率を予測し，確率一位で予測された単語の含有率を用いるフィルタを適用することによって，自動翻訳による対訳サイト[1] のうちの約 8 割を除外できることを示す．一方，(2) の特徴に対しては，二文間の語の一致度を測る尺度である BLEU-1 [7] を用いたフィルタを適用して自動翻訳による対訳サイトを除外することにより，79%の適合率および 88%の F 値で人手翻訳による対訳サイトを同定できることを示す．さらに，本論文では，(1) の特徴に対する言語モデルフィルタ，および，(2) の特徴に対する BLEU-1 フィルタを併用して，どちらかいずれかのフィルタによって自動翻訳による対訳サイトを除外するアプローチを用いることにより，結果的に，人手翻訳による対訳サイトの同定において 87∼88%前後の適合率・F 値が達成できることを示す．

2 JParaCrawl V2: ウェブから収集された日英対訳文対集合

本論文で高品質化の対象とする JParaCrawl V2 [6] は，ウェブ空間から日本語と英語の対訳文対を収集するシステムによって，NTT コミュニケーション科学基礎研究所が作成した対訳コーパスであり，一般に公開されている[2]．JParaCrawl V2 の日英対訳データ収集過程 [3, 5, 6] は，以下の四段階に分けられる．

(i) ウェブ空間に存在する大規模なドメイン集合のうち，日本語ページおよび英語ページが含まれるドメインにおいて，各ページのテキストデータを収集する．

(ii) 収集されたドメインにおいて，対訳ページの候補と考えられる日本語ページと英語ページ間のページ間対応付けを行う．

(iii) 日英間で対応付けられたページ対において，日英対訳辞書および Bleualign [9] を用いて文対応付けを行う．

(iv) 対応付けられた全ての文対集合に対して，言語モデルおよび句読点情報に基づいて不適切な対訳文

*Filtering Parallel Sentences collected from the Web based on Analyzing Characteristics of Translated Web Sites

†Kengo Miyoshi, Yizhen Wei, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba

‡Takuya Tamura, College of Engineering Systems, University of Tsukuba

§Masaaki Nagata, NTT Communication Science Laboratories, NTT Corporation, Japan

[1] 実際には，日本語側が統計的機械翻訳 (SMT) によって翻訳された対訳サイトである (3.2 節参照).

[2] http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

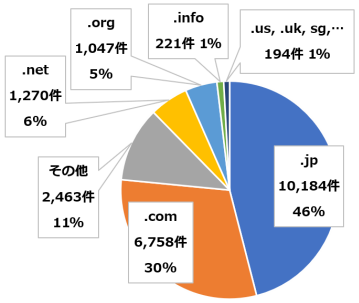


図 1: URL の分布 (サイト数単位)

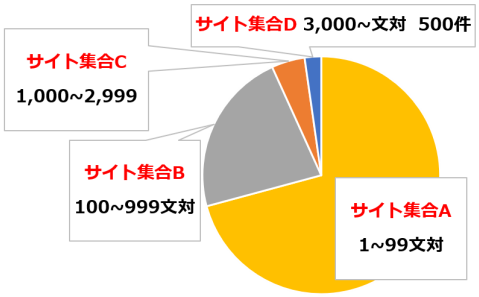


図 2: サイトごとの含有文数の分布 (サイト数単位)

対候補を除外するツール Bicleaner[3]により，対訳文対候補のクリーニングを行う．

本論文では，上記のように収集された日英対訳データ集合に対して，集合内の重複文対排除を行うとともに，英語・日本語以外の言語の文を削除した[4]

3 データセット

本節では，JParaCrawl V2 内のサイト集合における URL，および，サイトごとの文数の分布に基づいて，自動翻訳によるサイトを識別・除外する手法の訓練・評価用サイト集合のデータセットを作成する．

3.1 サイト URL の分布

JParaCrawl V2 における URL の分布 (サイト数単位) を図 1 に示す．さらに，JParaCrawl V2 におけるサイトごとの含有文数[5]の分布 (サイト数単位) を図 2 に示す．ここで，サイトごとの含有文数別に，サイト集合を部分集合 A(1~99 文対)，B(100~999 文対)，C(1,000~2,999 文対)，D(3,000~ 文対) に分割する．そして，これらの部分集合別に，含有文数の文数単位の分布を求めた結果を図 3 に示す．ただし，サイト集合 D については，人手翻訳によるサイトの割合が高いことが分かっ

[3]https://github.com/bitextor/bicleaner

[4]言語識別においては，polyglot (https://github.com/aboSamoor/polyglot) を用いた．

[5]各サイトから収集された文のうち，JParaCrawl V2 データセットに含まれる文の数．

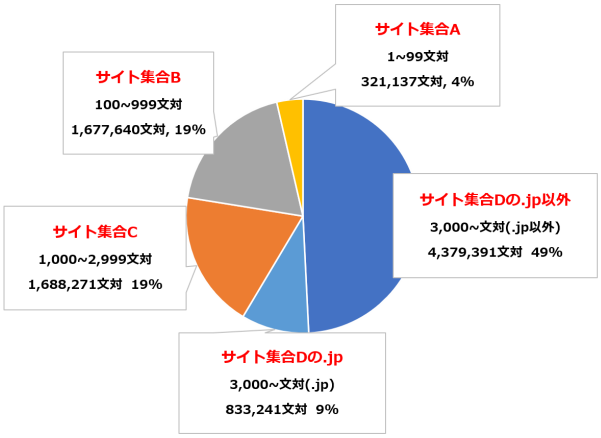


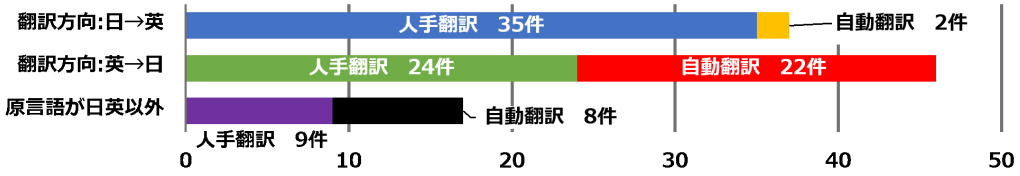
図 3: サイト集合 A・B・C・D ごとの含有文数の分布 (文数単位)

ている“.jp”ドメインと，“.jp”ドメイン以外とに分けて分布を示す．

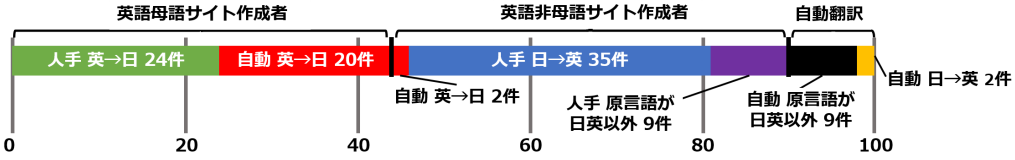
3.2 パラメータ調整用・評価用サイト集合

図 3 に示すように，「サイト集合 D の“.jp”ドメイン以外」の総文数がデータセット全体の約半数を占める．そこで，この「サイト集合 D の“.jp”ドメイン以外」中のサイト集合を母集団として，本論文の手法のパラメータ調整用として 100 サイトを無作為に選定し，各サイトの人手翻訳・自動翻訳の別を分析した．その結果，図 4(a) に示すように，人手翻訳によるサイトが 68 サイト，自動翻訳によるサイトが 32 サイトとなった．そこで，本論文では，自動翻訳によるサイトを識別・除外する手法のパラメータ調整，および，評価の対象を「サイト集合 D の“.jp”ドメイン以外」に設定する．そして，上述のパラメータ調整用 100 サイトを用いてパラメータの調整を行う．また，パラメータ調整用 100 サイトとは別に，評価用 100 サイトを無作為に選定し，本論文の手法を適用し評価を行う．

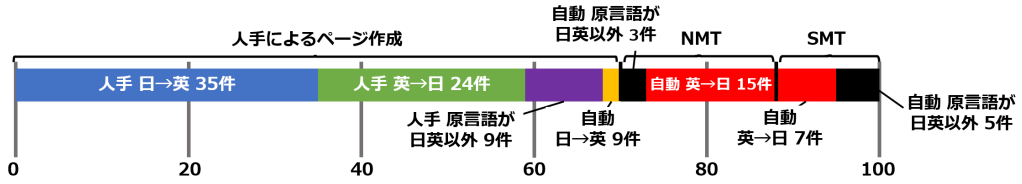
パラメータ調整用 100 サイトに対して，翻訳方向別の人手翻訳・自動翻訳の分布，英語文側の自動翻訳・英語母語サイト作成者・英語非母語サイト作成者の分布，および，日本語文側の SMT による自動翻訳・NMT による自動翻訳・人手によるページ作成の分布を図 4 に示す．このうち，本論文では，日本語言語モデルを用いる手法により，自動翻訳によって生成された日本語ページを含むサイト (図 4(c) 中の NMT の 3+15 サイトおよび SMT の 7+5 サイトの合計 30 サイト) を除外することを目的としてパラメータの調整を行う (4 節参照)．同様に，BLEU-1 フィルタを用いたテンプレート文の検出手法によって，図 4(a) に示す「自動翻訳によるサイト」(32 サイト) を排除することを目的として



(a) 翻訳方向別の人手翻訳・自動翻訳の分布



(b) 英語文側の自動翻訳・英語母語サイト作成者・英語非母語サイト作成者の分布



(c) 日本語文側の SMT・NMT・人手によるページ作成の分布

図 4: パラメータ調整用 100 サイトの分析結果

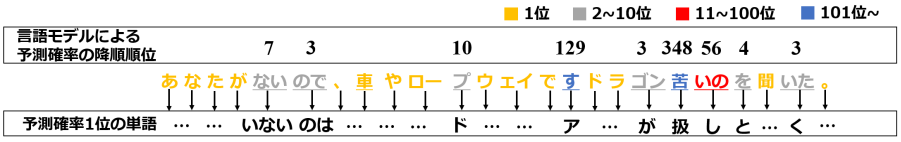


図 5: 「日本語言語モデルを用いたフィルタ」によって除外される自動翻訳文の例 (SMT による自動翻訳文．対象: www.cots.com.cn)

パラメータの調整を行う (5 節参照).

4 言語モデルを用いたフィルタ

本論文では，言語モデルを用いたフィルタによって「自動翻訳による非流暢な文が多く出現するか否か」を識別し，非流暢な文を含むサイトを除外する[6]．具体的には，日本語を含む多言語言語モデルとして，大規模な事前訓練済みニューラル言語モデルである Multilingual BERT (mBERT) [2] を用いて，GLTR [4] によって日本語文中の単語の予測確率の順位分布を求める[7][8]．そして，パラメータ調整用 100 サイトを用いたパラメータ最適化の結果として，「言語モデルによる予測確率が一位となる単語位置の割合の下限値 55.0%」を満たす場合に，人手翻訳によるサイトであると判定する．

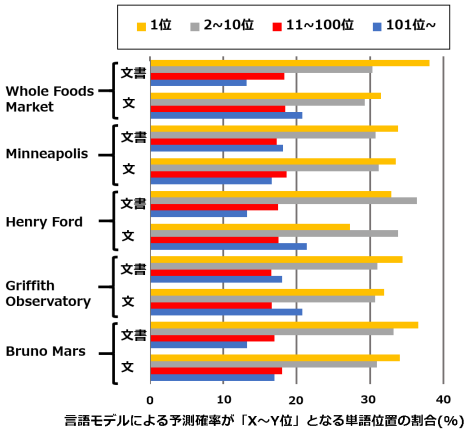


図 6: 英語言語モデル GPT-2 [8] による予測確率の順位の分布の文書単位・文単位間比較 (対象: 英語版 Wikipedia 中の 5 サンプルページ)

[6]日本語側においては，SMT によって生成された日本語文 (図 5) を含む対訳サイトの除外に対しては有効であったが，NMT によって自動翻訳された対訳サイトの除外に対しては有効ではなかった．一方，英語側において，英語言語モデル GPT-2 [8] を用いることにより，自動翻訳によるサイト全般の除外における一定の有効性を確認したが，評価実験の都合上，本論文では，日本語側における評価結果のみを示す．

[7]一サイトあたり 300 文を無作為に抽出し用いる．

[8]本来，GLTR [4] は，「文書単位テキストを対象とした自動・人手生成の識別」において用いる．本論文では，これを「文を対象とした自動・人手生成の識別」において用いた場合の妥当性を検証するために，「GLTR [4] で用いられた英語言語モデル GPT-2 [8] による予測確率の順位の分布の文書単位・文単位間比較」(図 6) を行い，大きな違いがないことを確認した．

表 1: 「テンプレート文検出のための BLEU-1 フィルタ」によって検出されるテンプレート文 (自動翻訳文) の例 (対象: droidchart.com)

テンプレートの原型	Li-Po 【$数字列】 mAh, 取り外し不可能の電池を搭載します。
文例	Li-Po **4000** mAh, 取り外し不可能の電池を搭載します。
	Li-Po **4010** mAh, 取り外し不可能の電池を搭載します。
	Li-Po **4020** mAh, 取り外し不可能の電池を搭載します。

5 テンプレート文検出のための BLEU-1 フィルタ

本論文では，テンプレート文を検出するための BLEU-1 [7] フィルタによって，「自動翻訳によるサイトにおいては，原言語文・目的言語文とも，文中の一部の固有名詞等は異なるものの，文中の他のフレーズが共有されるテンプレート文が多く含まれる」という特徴を検出する．テンプレート文 (自動翻訳文) の典型例を表 1 に示す．BLEU-1 フィルタにおいては，日本語側サイト内から無作為に選定した 1,000 文に対して形態素解析[9]を行った後，任意の二文間の単語重複率を求め，パラメータ調整用 100 サイトを用いたパラメータ最適化の結果として，「二文間の BLEU-1≦70 となる二文の割合の下限値 98.29%」を満たす場合に，人手翻訳によるサイトであると判定する．

6 評価

4 節，および，5 節で述べたフィルタ[10] を用いて評価用 100 サイトの評価を行った結果を図 7 に示す．このうち，日本語言語モデルを用いたフィルタによって実際に除外されたサイトの内訳 (図 8) からは，SMT によって自動翻訳された全 11 サイトのうち，約 8 割の 9 サイトが除外できたことが分かる[11]．また，図 7 の結果から，人手翻訳による対訳サイトの同定の性能は，BLEU-1 [7] を用いたフィルタを単独では 79%の適合率および 88%の F 値であるのに対して，言語モデルフィルタ，および，BLEU-1 フィルタの併用により 87~88%前後の適合率・F 値が達成したことが分かる．

7 関連研究

ウェブから収集された対訳文対対集合から高品質な対訳文対集合を抽出する手法の関連研究として，Zhang

[9]IPA 辞書版 MeCab (https://taku910.github.io/mecab/) を用いた．

[10]パラメータ調整用 100 サイトにおいて，人手翻訳によるサイト検出の再現率・適合率・F 値がそれぞれ最大値付近となるように下限値等のパラメータを調整した．

[11]NMT によって自動翻訳された全 12 サイトについては，2 割弱の 2 サイトしか除外できていない．

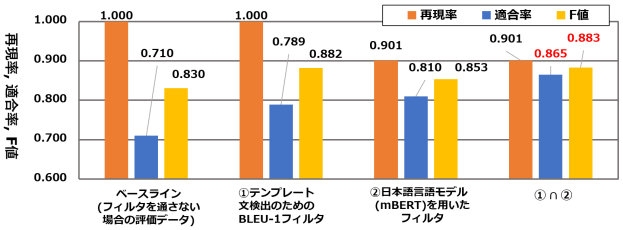


図 7: 「人手翻訳によるサイト」の同定における再現率・適合率・F 値

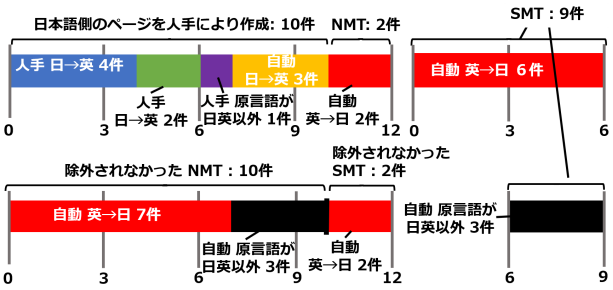


図 8: 「日本語言語モデルを用いたフィルタ」によって除外されたサイトの内訳 (除外されなかった 10 件の NMT サイト，および，2 件の SMT サイトを含む)

ら [10] は，Wikipedia から収集された 104 言語の単言語コーパスを用いて事前訓練された mBERT [2] 等の言語モデルを用いた対訳文対フィルタ手法を提案し，1) 英独 ParaCrawl (3,700 万文対) [1]，および，2) ウェブから独自に収集した日中対訳文対対集合 (2,000 万文対) の高品質化を行っている．Zhang ら [10] が文対の特徴分析に基づいて高品質な対訳文対集合を収集するのに対して，本論文では，対訳サイトを単位とする特徴分析に基づいて高品質な対訳文集合を収集する点で大きく異なる．また，Zhang らの手法では，mBERT 等の言語モデルを訓練するために何種類かの言語資源を必要とする．例えば，mBERT を対訳文対フィルタタスクに適応する際の fine-tuning においては，高品質な対訳文対集合が必要である[12]．したがって，人手を介して整備する必要のある対訳文対の規模の点において，本論文の手法の方が有利であると言える．

8 おわりに

本論文では，JParaCrawl V2 を対象として人手翻訳によるサイトを高性能に抽出することを目的として，言語モデルを用いたフィルタ，および，テンプレート文検出のための BLEU-1 フィルタを提案し，一定以上の適合率・F 値が達成できることを示した．

[12]一例として，日中対訳文対対集合のフィルタタスクでは，ソフトウェアドキュメントに関する既存の対訳コーパスから収集した 30 万対訳文対を用いて mBERT の fine-tuning を行っている．

参考文献

[1] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, and J. Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proc. 58th ACL*, pp. 4555–4567, 2020.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[3] M. Esplà, M. Forcada, G. Ramírez-Sánchez, and H. Hoang. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proc. MT Summit XVII Volume 2: Translator, Project and User Tracks*, pp. 118–119, 2019.

[4] S. Gehrmann, H. Strobelt, and A. Rush. GLTR: Statistical detection and visualization of generated text. In *Proc. 57th ACL*, pp. 111–116, 2019.

[5] M. Morishita, J. Suzuki, and M. Nagata. JParaCrawl: A large scale Web-based Japanese-English parallel corpus. In *Proc. 12th LREC*, pp. 3603–3609, 2020.

[6] 森下睦, 鈴木潤, 永田昌明. JParaCrawl: 大規模 Web ベース日英対訳コーパス. 言語処理学会第 26 回年次大会論文集, pp. 469–472, 2020.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th ACL*, pp. 311–318, 2002.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[9] R. Sennrich and M. Volk. Iterative, MT-based sentence alignment of parallel texts. In *Proc. 18th NODALIDA*, pp. 175–182, 2011.

[10] B. Zhang, A. Nagesh, and K. Knight. Parallel corpus filtering via pre-trained language models. In *Proc. 58th ACL*, pp. 8545–8554, 2020.