テキスト平易化技術を利用した攻撃性緩和システムの提案とコーパス生成

村山 貴志 * 中村 朋生 * 谷合 廣紀 * 入江 英嗣 * 坂井 修一 *

概要. 本研究では，テキストの意味内容を保持しつつ攻撃性のみを緩和する変換を統計的機械翻訳の枠組みで提案し，その中で特に機械翻訳で必要となるパラレルコーパスを自動生成する手法を提案し実装する．機械翻訳には大規模パラレルコーパスが必要だが，本タスクでは入手が難しい．そこで，英語及び日本語の大規模生コーパスを対象に，単語分散表現に基づく文間類似度判定とポライトネス理論に基づく攻撃性判定を組み合わせた品質推定をコーパスから選択した 2 文に対して繰り返すことで，擬似的なパラレルコーパスを生成した．品質推定を行った文対数に対する，パラレルコーパスとして選択された文対数の割合は，英語において 3.13×10^{-6}，日本語において 3.56×10^{-4} であった．

1 はじめに

近年モバイル通信端末の普及と高性能化により一般の人々が不特定多数の人々とコミュニケーションを取る機会が増加したが，コミュニケーション量の増大に伴って，攻撃的文章が伝達されることによる，受け手に対する心理的負担の問題が立ち現れてくることとなる．既存サービスにおいては，このような攻撃的な文章表現により発生するコストに対し，悪意のある表現が含まれていた場合受信者に文章を到達させないといったシステムが構築されている．

しかし，この場合送り手が送信した情報を受け手は全く得られなくなってしまう．攻撃的表現を含む文章は，その中で伝達されている情報そのものは受け手にとって有用である場合がある．従来の対策手法のもとでは受け手は送り手の送信した有用な情報にアクセスする機会を失うこととなる．

以上を踏まえ，攻撃的表現を含むテキストから攻撃性のみを除去，あるいは緩和し，必要な情報のみで構成された文章に再構築する問題を同一言語内の翻訳問題として考える．この同一言語内の翻訳問題という枠組みは，テキスト平易化という研究分野でも同様に用いられている．そこで，テキスト平易化分野における研究成果 [5] を本問題に援用してテキスト攻撃性緩和システムの全体像を提案する．さらにこのシステムにおいて中核的役割を果たすコーパス生成方式を提案および実装する．

2 提案：テキスト攻撃性緩和方式

テキストの攻撃性緩和を同一言語内の統計的機械翻訳問題として実現するシステムを提案する．

本研究の目的を統計的機械翻訳によって実現するためには，攻撃的表現を含む単言語コーパスと，攻撃的表現を含む文と含まない文のパラレルコーパス

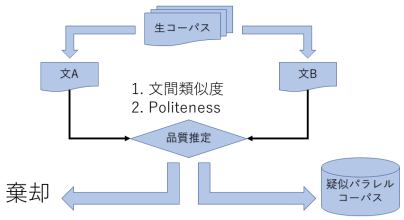

図 1. 擬似パラレルコーパス生成手法の概観

必要となる．前者は一般的な統計的機械翻訳に用いられている単言語コーパスがそのまま利用できるが，後者については現在入手可能な大規模なものは存在しない．そこで，本研究ではこのパラレルコーパスを擬似的に構築する手法を提案する．ここにおいて統計的機械翻訳によるテキスト平易化研究で用いられていたアイデアを援用する．

擬似パラレルコーパス生成手法提案の概観を図 1 に示す．まず，巨大な単言語コーパスから 2 文を選択する．次に，その文対に対して同義性と攻撃性の判定を行う．攻撃的表現の有無のみ異なる同義文という条件を満たしていた場合に擬似パラレルコーパスに対訳文として収録し，そうでない場合には棄却する．これを繰り返し行うことで単言語コーパスから擬似パラレルコーパスを生成する．

文間類似度判定においては，選択された 2 文の類似度を定量的に評価し，その閾値が一定の値を超えたものを類似文とみなす．梶原らの研究で行われた文間類似度判定の実験結果に基づき，提案されているアルゴリズムの中で最高精度を実現した Maximum Alignment を採用する．なお，Maximum Alignment アルゴリズムの提案は Song らによる [4].

文の攻撃性判定には，Brown and Levinson の

* 東京大学大学院情報理工学系研究科

ポライトネス理論 [1] に基づいたポライトネスの定量化ツールである Stanford Politeness API [2] を利用し，そこから出力される Politeness 値を用いる．Stanford Politeness API はある文に登場するポライトネス・ストラテジーを根拠としてその文の Politeness 値を -1 から $+1$ の間の値で定量化する．ポライトネス・ストラテジーは，相手のプラスの状態に置かれたいという欲求を満たすポジティブ・ポライトネスと，相手のマイナスの状態から遠ざかりたいという欲求を満たすネガティブ・ポライトネスからなる．Politeness 値はテキストのポライトネス・ストラテジー表現とそれに逆行する表現それぞれの多寡と強さを定量化したものであるので，攻撃性を示す指標として妥当なものである．

Stanford Politeness API で Politeness 値を推定できるのは英文のみのであるが，今回は日本語文でもテキスト攻撃性緩和を考える．そこで，日本語文を英訳した上で Stanford Politeness API によって Politeness 値を推定する．ポライトネス理論は言語によらない普遍的現象をモデル化した理論であり，言語表現そのものではなく，その背景にある意味内容から人々はポライトネスを判断していることを示している．そのため，完全に意味内容を保持した日本語から英語への翻訳が可能であり，かつそれが実現されているならば，翻訳後の英文から得られる Politeness 値は原文の Politeness 値と理論上一致する．以上の理由から，ポライトネス推定による攻撃性判定では日本語テキストを英訳した上でその Politeness 値を計測するという手法を採用する．

3 実装

英語生コーパスは EC サイト「Amazon.com」のレビューコーパス，日本語生コーパスは日本語質疑応答 Web サービスの「Yahoo!知恵袋」のスクレイピングを利用した．

品質推定においては，文間類似度判定における Maximum Alignment で使用する単語分散表現は Word2Vec を利用し [3]，Politeness 値計算に進むための文間類似度の閾値は 0.6 とした．Politeness 値計算において，日本語文の英訳には Google Cloud Platform の Translation API を利用した．文対を生成する上での Politeness 値は，負の値の時に攻撃的表現を含む文，正の値の時に攻撃的表現を含まない文と判断した．

4 結果と考察

比較した文対数に対する生成された文対数の割合は英語において 3.13×10^{-6}，日本語において 3.56×10^{-4} であった．この割合が常に維持されるときにサイズ 100 万の擬似パラレルコーパスを用意するには，英語において 3200 億回，日本語において 28 億回の比較が必要となる．乱数を用いるアルゴリズムのため並列化は容易であるが，各ユーザが目的に合わせて生コーパスや閾値を調整するという実用を考えると、さらなる比較回数削減や比較自体のの高速化が期待される。特に英語の日本語に対する生成割合が低いので生コーパスや実装の見直しなど改善の余地があると予想される．

5 まとめ

本研究では，テキスト攻撃性緩和という問題に対し，統計的機械翻訳の枠組みで実現する方法を提案した．この機械翻訳を行うにあたって必要となるパラレルコーパスを，文間類似度判定と Politeness 値推定からなる品質推定を用いて，擬似的に構築する方式を提案した．この機械翻訳を行うにあたって必要となるパラレルコーパスを文間類似度判定と Politeness 値推定からなる品質推定を用いて擬似的に構築する方式を提案し実装した．

今後の展望としては，今回の擬似パラレルコーパス生成アルゴリズムを用いて大規模な擬似パラレルコーパスを実際に生成し，翻訳アルゴリズムの選択も含めて統計的機械翻訳システム全体を実現することが期待される．

参考文献

[1] P. Brown, S. C. Levinson, 田中 典子, 斉藤 早智子, 津留崎 毅, 鶴田 庸子, 日野寿憲, 山下 早代子. ポライトネス : 言語使用における、ある普遍現象. 研究社, 2011.

[2] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts. A Computational Approach to Politeness with Application to Social Factors. *CoRR*, abs/1306.6078, 2013.

[3] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781, 2013.

[4] Y. Song and D. Roth. Unsupervised sparse vector densification for short text similarity. In *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1275–1280. Association for Computational Linguistics (ACL), 2015.

[5] 梶原 智之, 小町 守. 平易なコーパスを用いないテキスト平易化. 自然言語処理, 25(2):223–249, 2018.