化学物質名の翻訳における統計的機械翻訳とニューラル機械翻訳の比較および併用

近藤 修平 [1]　松本 裕治 [2,1]
理研 AIP[1] 奈良先端科学技術大学院大学 情報科学研究科 [2]
{shuhei-k, matsu}@is.naist.jp

1 はじめに

科学技術論文の多くは英語で書かれており，専門分野の大規模データベースもその多くが英語を対象としている．これらのリソースから得られる知識を翻訳することができれば、他言語での知識獲得にとって有用であると考えられる．例えば，化学物質に関する英語の大規模データベースである CAS REGISTRY[1]は 1 億 4400 万件以上の有機および無機物質を収録している．日本語の化学物質データベースとして最大規模のものである日本化学物質辞書 (日化辞) に収録されている有機化合物は約 373 万件であり，CAS REGISTRY とは 40 倍近い差があるが，日化辞には化学物質の日本語名と英語名の両方が登録されている．これを利用して機械翻訳システムを学習することで，英語文献中の化学物質名を翻訳して日本語データベースの拡張に利用することや，日本語の化学物質名を英訳して言語横断検索を助けることが可能になると期待される．

本研究では，日化辞から抽出した日英の化学物質名を元に統計的機械翻訳 (SMT) およニューラル機械翻訳 (NMT) モデルを学習し，それぞれのモデルの誤訳について考察するとともに，両者の出力が一致する場合の精度，および両者が一致するにも関わらず参照訳とは一致しない事例についても分析する．

2 関連研究

固有表現の翻字については多くの研究がなされており，シェアードタスクなども行われている [2]．近年では NMT をベースとした手法が主流となりつつあるが [5, 3]，言語対やデータによっては SMT ベースの手法も遜色のない精度を残している [8]．本研究で取り扱う化学物質名の翻訳は，固有表現の翻字と同様に固有名を対象としており，多少の表記ゆれはあるものの原則的に正解訳との完全一致が求められるという点は両者に共通している．一方で，前者には文字数が極端に多い名前がかなりの割合で含まれる点や，括弧等の記号を含むことが多い点，特に慣用名を含む事例で大幅な語順並び替えが必要となる点などで違いがある．

3 実験

3.1 実験設定

実験データには日化辞パラレルコーパス[2]を用いる．これは NBDC NikkajiRDF[3]から 1 対 1 対応する化学物質名を抽出したもので，訓練用，開発用，テスト用データがそれぞれ 2,818,784 対，5607 対，5705 対含まれる．本実験では前処理として上付きおよび下付き文字を表す XML タグを同コーパス中の化学物質名に用いられていない文字に置き換えた上で，日本語名に句読点を含む翻訳対を除外し，残った 2,829,849 対を用いる．除外された翻訳対の例を図 1 に示す．また，上記の訓練/開発/テスト用の分割には従わず，3 分割交差検証を行って全翻訳対を SMT および NMT のシステムで翻訳する．分割はそれぞれ 1,000,000 対，1,000,000 対，829,849 対とし，訓練用データから 10,000 対を取り除いて開発用データとして用いる．例えば，829,849 対をテスト用データに用いる分割では，訓練用データは 1,990,000 対，開発用データは 10,000 対となる．SMT の誤り率最小学習によるチューニングおよび NMT のモデル選択は文字単位の BLEU で行い，テストデータは参照訳との完全一致率で評価する．

[1] https://www.cas.org/support/documentation/chemical-substances

[2] http://www2.nict.go.jp/astrec-att/member/mutiyama/nikkaji/index.html

[3] https://dbarchive.biosciencedbc.jp/jp/nikkaji/desc.html

Copolymers of alpha-acryloyl-omega-hydroxypoly(oxyethylene)/3,3,4,4,5,5,6,6,7,7,8,8,8-tridecafluorooctan-1-yl acrylate (It is limited that the content of the components having molecular weight less than 1,000 is 1% or less.)
α-アクリロイル-ω-ヒドロキシポリ (オキシエチレン) 3,3,4,4,5,5,6,6,7,7,8,8,8-トリデカフルオロオクタン-1-イル= アクリラート共重合物 (分子量 1,000 未満の成分の含有率が 1%以下であるものに限る。)

図 1: 除外された翻訳対の例

3.2 ツール

SMT 用の単語アラインメントには mgiza++[4]を，モデルの学習およびデコードには Moses[5]を用いる．NMT 用の単語分割には Sentencepiece[6]を，モデルの実装には DyNet[7]を用いる．

3.3 統計的機械翻訳結果

SMT 用のデータは文字単位で分割し，空白も 1 文字として取り扱った．SMT ではモデル学習の過程で単語アラインメントが必要になるが，一般的なツールで扱えるトークン数には限界があるため，日英いずれか一方が 100 文字を超える翻訳対は訓練データから除外した．目的言語側の言語モデルは除外前の訓練データから 6 グラム文字モデルを学習した．3 分割交差検証の各分割における，文字数による除外前後の翻訳対の数を表 1 に示す．

表 1: 文字数による除外前後の訓練データの数

	除外前	除外後
第 1 分割	1,990,000	1,629,327
第 2 分割	1,819,849	1,489,122
第 3 分割	1,819,849	1,490,329

開発データにおける文字 BLEU を表 2 に，テストデータにおける完全一致率を表 3 に示す．エラー率 1.22%前後という高い精度が得られたが，SMT に特有の性質として，慣用名が存在する物質名を音写したことに起因する誤訳や，単語アラインメントの際に null にアラインされた文字が前後の文字に付随してフレーズとして抽出されたことに起因すると思われる誤訳が見られた．これらの誤訳の具体例を図 2 に示す．

[4]https://github.com/moses-smt/mgiza
[5]https://github.com/moses-smt/mosesdecoder
[6]https://github.com/google/sentencepiece
[7]https://github.com/clab/dynet

入力文: Disulfur decafluoride
参照訳: 十ふっ化二硫黄
出力文: ジ硫酸フルオリド

入力文: Diboron trisulfide
参照訳: 三硫化二ほう素
出力文: ジボロン トリスルフィド

(a) 慣用名がある物質を音写した例

入力文: Bishexanoic acid 1,2-phenylene ester
参照訳: ビスヘキサン酸 1,2-フェニレン
出力文: ビスヘキサン酸 1,2-フェニレン (C=3〜8)

入力文: Phenol sulfonic acid zinc salt
参照訳: フェノールスルホン酸亜鉛
出力文: フェノール スルホン酸及びその亜鉛塩

(b) フレーズ抽出に起因する誤訳

図 2: SMT による誤訳の例

表 2: 開発データにおける文字 BLEU

第 1 分割	第 2 分割	第 3 分割
99.84	99.78	99.81

表 3: テストデータにおける完全一致率

第 1 分割	第 2 分割	第 3 分割	合計
98.74	98.92	98.68	98.78

3.4 ニューラル機械翻訳結果

NMT 用のデータは Byte pair encoding アルゴリズム [7] を用いて分割した．文字単位の分割を採用しなかったのは，予備実験の結果から NMT において長い事例を学習データから除外すると SMT に比べて精度への悪影響が大きかったことと，100 文字を超える物質名を 20%以上含むようなデータでは系列長が長くなりすぎる懸念があったことが理由である．語彙数は 1,600 として日英それぞれの訓練データで独立に学習を行った．NMT のモデルは Attention ベースの Encoder-decoder モデル [1, 6] であり，表 4 に示すハイパーパラメータを用いた．最適化は Adam [4] で行い，1 エポック目の終了以降 20 万ペア毎に開発データで文字 BLEU による評価を行い，1 エポックを通じて改善が見られなかった時点で学習を打ち切り，最高の文字 BLEU を記録したモデルを選択した．

表 4: NMT モデルのハイパーパラメータ

エンコーダー BiLSTM のレイヤー数	1
デコーダー LSTM のレイヤー数	2
入力の次元数	512
隠れ層の次元数	512
ドロップアウト率	0.25

開発データにおける文字 BLEU を表 5 に，テストデータにおける完全一致率を表 6 に示す．テストデータでのエラー率はおよそ 1.59%であり，SMT に比べて若干劣る結果となった．NMT に特有の誤訳として，通常の分野でも問題として知られるいわゆる under-generation に起因する誤訳や，音写として全く異なる誤訳などが見られた．これらの誤訳の具体例を図 3 に示す．

表 5: 開発データにおける文字 BLEU

第 1 分割	第 2 分割	第 3 分割
99.75	99.72	99.73

表 6: テストデータにおける完全一致率

第 1 分割	第 2 分割	第 3 分割	合計
98.39	98.49	98.35	98.41

3.5 統計的機械翻訳とニューラル機械翻訳の併用

統計的機械翻訳（SMT）とニューラル機械翻訳 (NMT) はモデルの性質が大きく異なるため，両者の翻訳結果が一致する場合は信頼性が高いことが期待される．全 2,829,849 翻訳対に対する SMT と NMT の翻訳結果を比較したところ，両者が一致したのは 2,771,837 対であり，それらが参照訳とも一致したのは 2,765,195 対であった．これは一致率としては 99.76%，エラー率としては 0.24%となり，SMT のエラー率 1.22%，NMT のエラー率 1.59%と比較して 1/5 以下に抑えられている．

次に，SMT および NMT の出力と参照訳が一致しなかった事例から 100 対をランダムに抽出し，その原因について分類した結果を表 7 に示す．カルバミン酸とカルバミド酸の表記ゆれが 45 対を占めたのを筆頭に，表記ゆれに分類される不一致が 75 対を占めたが，このうち慣用名と音写の表記ゆれは SMT に多く見られるのと同じ種類の誤訳を NMT が出力したとも解釈できるものであり，他の種類の表記ゆれに比べて誤訳である可能性が高いと考えられる．また，単語間の空白の有無による不一致が 14 対，末尾のエステル，塩の有無による不一致が 6 対あり，うち 1 対はこれらが重複していた．翻訳元の英語名または参照訳の日本語名が誤っていると思われる例が 2 対，英語名にないアスタリスクやハイフンなどの記号が日本語名に含まれ，誤訳かどうか判断が難しいものが 3 対あった．最後に，翻訳結果が実際に間違っていると思われる例は 2 対であった．結論として，100 サンプルのうち控えめに見積もっても 70%以上は参照訳からの乖離が許容範囲内に収まっており，SMT と NMT の出力結果が一致した場合の誤訳率は 0.1%未満に抑えられると考えられる．

4 おわりに

日化辞から抽出された日英の化学物質名を元に統計的機械翻訳およびニューラル機械翻訳モデルを学習し，同一の入力に対して両者の翻訳結果が一致した場合の精度を測定したところ，エラー率をそれぞれ単独の場合の 1/5 以下に抑えることができた．今後はニューラル機械翻訳モデル用データの分割単位の見直しや，さらなるエラー分析を進めていきたい．また，本実験では両者が独立に翻訳した結果の一致に依存しており，

入力文: 1-Morpholino-1,3-diphenyl-2-propynyldiethoxyphosphine oxide
参照訳: 1-モルホリノ-1,3-ジフェニル-2-プロピニルジエトキシホスフィンオキシド
出力文: 1-モルホリノホスフィンオキシド

(a) Under-generation による誤訳

入力文: 3,8-Divinylidenesebacic acid diethyl ester
参照訳: 3,8-ジビニリデンセバシン酸ジエチル
出力文: 3,8-ジビニリデンババハ酸ジエチル

(b) 音として全く異なる誤訳

図 3: NMT による誤訳の例

表 7: SMT と NMT の結果が一致したが参照訳とは異なった 100 サンプルの内訳

カルバミン酸/カルバミド酸	45
漢数字/カタカナ	10
慣用名/音写	7
ひらがな/カタカナ	3
その他表記ゆれ	10
空白の有無	14
エステル/塩の有無	6
参照訳の誤り	2
誤訳	2
その他	3

全事例の 2%前後が対象外となっているため，統計的機械翻訳とニューラル機械翻訳のより効率的な組み合わせについて検討し，より多くの事例をカバーできるようにしたいと考えている．

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[2] Nancy Chen, Rafael E. Banchs, Min Zhang, Xiangyu Duan, and Haizhou Li. Report of NEWS 2018 Named Entity Transliteration Shared Task. In *Proceedings of the Seventh Named Entities Workshop*, pp. 55–73, 2018.

[3] Roman Grundkiewicz and Kenneth Heafield. Neural Machine Translation Techniques for Named Entity Transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pp. 89–94, 2018.

[4] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *CoRR*, Vol. abs/1412.6980, , 2014.

[5] Soumyadeep Kundu, Sayantan Paul, and Santanu Pal. A Deep Learning Based Approach to Transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pp. 79–83, 2018.

[6] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, 2016.

[8] Snigdha Singhania, Minh Nguyen, Gia H Ngo, and Nancy Chen. Statistical Machine Transliteration Baselines for NEWS 2018. In *Proceedings of the Seventh Named Entities Workshop*, pp. 74–78, 2018.