構造付き Web テキスト翻訳のための高品質多言語データセット

橋本 和真, Raffaella Buschiazzo, James Bradbury,* Teresa Marshall,
Richard Socher, Caiming Xiong
Salesforce
k.hashimoto@salesforce.com

1 はじめに

機械翻訳の研究が進むなかで, ベンチマークデータセットで BLEU を競うだけでなく, どのように実応用していくかという点も重要な研究課題である. 機械翻訳の応用先として, 企業のオンラインヘルプ等の Web テキスト翻訳が挙げられる. サービスを様々な国に提供するためには各言語に対応したテキストが必要となる. ユーザーに提供するテキストであるため完全に機械翻訳のみに頼ることは困難であるが, 人手が関わる時間的・金銭的コスト削減を目指すことが可能である.

Web ページは HTML や XML の構造を有しているため, 本研究では HTML/XML タグがついたテキストを直接翻訳することを研究課題として考える. それにより, Web ページの構造を保持したまま多言語への翻訳補助を目指す. 現実的な手段として Google Translate の利用が非常に有力であるが, 例えば Google Chrome の Web ページ翻訳機能では, 元々のページ構造が崩れてしまうことがある. また, 近年の自然言語処理のデータのほとんどは Web テキストであるため, それらを単に生テキストとして扱うのではなく, 筆者の意図 (フレーズの強調など) を表現するタグに着目することは研究として有意義であると考える.

以上の目的のため本稿では, XML 構造付きテキストの翻訳データセットを構築する.[1] データセットの特徴としては, (i) Salesforce のオンラインヘルプから抽出した XML 付きテキスト, (ii) 人手で整備されているため高品質, (iii) 17 言語間の任意の組み合わせが可能, といったことが挙げられる. 特徴 (iii) に関しては, 英語をピボットとして用いることなく, 「日本語-フィンランド語」といった言語対の直接の翻訳の研究を可能にする. 実験では, ニューラル機械翻訳に基づき, XML 制約付きビームサーチ, コピー機構を取り入れたシステムの評価を行う. ドメインや XML 構造を考慮した自動評価に加え, 「ポストエディットの対象として有用か？」という観点から人手評価も行う.

*現在 Google Brain 所属.
[1]データセットと前処理・学習コードは公開予定である.

- Example (a)
English:
You can use this report on your Community Management Home dashboard or in <ph>Community Workspaces</ph> under <menucascade><uicontrol>Dashboards</uicontrol><uicontrol>Home</uicontrol></menucascade>.
Japanese:
このレポートは、[コミュニティ管理] のホームのダッシュボード、または <ph>コミュニティワークスペース</ph>の <menucascade><uicontrol>[ダッシュボード]</uicontrol> <uicontrol>[ホーム]</uicontrol></menucascade> で使用できます。

- Example (b)
English:
Results with bothbeach and house in the searchable fields of the record.
Japanese:
レコードの検索可能な項目に beach と house の 両方が含まれている結果。

- Example (c)
English:
You can only predefine **this field** to an email address. You can predefine **it** using either T (used to define email addresses) or To Recipients (used to define contact, lead, and user IDs).
Japanese:
この項目はメールアドレスに対してのみ事前に定義できます。
この項目は [宛先] (メールアドレスを定義するために使用) または [宛先受信者] (取引先責任者、リード、ユーザ ID を定義するために使用) のいずれかを使用して事前に定義できます。

図 1: 英日の対訳データの例.

2 データセット構築

オンラインヘルプは各項目に関してまずは英語で作成され, それぞれの対応言語に翻訳されてきた. 長年構築してきた翻訳メモリを利用した人手翻訳で各言語に対応しており, 現在以下の 16 言語に関するデータがある: Brazilian Portuguese, Danish, Dutch, Finnish, French, German, Italian, **Japanese**, Korean, Mexican Spanish, Norwegian, Russian, Simplified Chinese, Spanish, Swedish, Traditional Chinese. 多言語の同一項目の Web ページの構造から XML タグを頼りに対訳データを抽出するため, 任意の言語対を考えることができる. 図 1 に英日の対訳例を示す. 例 (a) では, 英日ともに同じ XML 構造を持つ. 例 (b) では, 英日の文法の違いにより全体の構造が異なる. 例 (c) では, 複数文に関する翻訳の中で共参照の関係がある.

本稿では深く触れないが, 複数文の文脈を考慮した機械翻訳の研究は近年注目されており, 本データセットはそういった研究の機会も提供しうる. その他にも, 単語・フレーズ単位の翻訳例も多く含んでおり, 様々な種類のテキストを扱うことが出来る.

2.1 対訳抽出

ページ対応 本データセットでは 3 つ以上の言語を同時に対応させることも可能であるが, 本稿ではまず従来通り 2 言語の対応から考える. 実際の Web ページを用いるため, 理想的にはクローリングによりデータセットが構築可能であるが, 高品質なデータ提供のために社内の XML ファイルをもとに本データセットを構築する. 例えば, English-Japanese では 7,336 個の XML ファイルが対応付けされ, Finnsih-Japanese では 7,927 個の XML ファイルが対応付けされた.

XML アラインメント 対応付けされた XML ファイルの組を XML パーサで解析し, 「同一のオンラインヘルプ内容を同一の Web ページ構造で表現している」という元々のデータの性質を利用する. 各ファイルを XML 要素の列として線形化し, XML タグのマッチングスコアを用いたペアワイズ系列アラインメント手法により, XML 要素の対応を抽出する. 研究のための言語資源という観点から, このように整備されたデータはデータセット構築を容易にする.

XML タグの分類 次に, 高品質な対訳を抽出するために XML タグを translatable, transparent, untranslatable の 3 種類に人手で分類する. タグごとに扱うテキストの傾向が異なるからであり, `p`, `xref`, `note` などの translatable タグで囲われたテキストを翻訳対象とする. 一般的に translatable に分類されるタグは独立したテキストに対応しており, 系列アラインメントで容易に対応がとれる. それに対して `b`, `ph` などの transparent タグは文中の単語やフレーズに対応することが多く, 文法の違いによって語順が入れ替わることがあるため系列アラインメントで対訳がとれないことが多い. そこで transparent タグは対訳として抽出せず, translatable タグのテキスト中に埋め込む. `sup` などの untranslatable タグは上付き文字などに対応するもので, 今回は翻訳対象から除外する.

対訳抽出 図 2 に対訳の抽出例を示す. ここでは 3 つの translatable タグが対応付けされており, 3 つの対訳例を得ることが出来る. 巨大な対訳例を回避するため, 図中の `note` タグのように translatable タグを個別に分離して対訳例を抽出する. ただし, 図中の

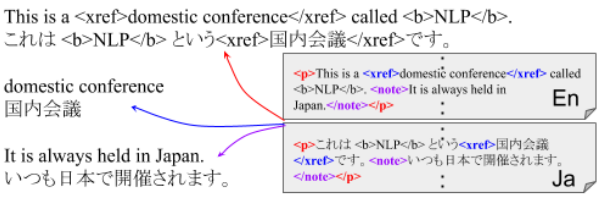

図 2: 対訳の抽出例.

`xref` タグのように文中に埋め込まれている場合には, transparent タグと同様に埋め込んだままにする. ルートタグの対応は自明であるため, 対訳抽出の際に除外する. この過程によって, 図 1 の例 (c) のように複数文対応がとれることがあるが, 必ずしも 2 言語間で文の数が一致しないことを確認したため, 本稿では文分割することを考えない.

フィルタリング 最後に, さらに対訳の質を上げるため, 対訳ペアのうち XML タグの種類・個数が一致するもののみをデータセットに含める. ただし, XML タグの順序, 全体の構造は文法によって異なることがあるので制約として考えない. 対訳例の重複を除外した後に, 開発データ・テストデータ用にそれぞれ 2,000 例ずつ分離し, 残りを学習データとする. 対応する 17 言語すべてのペアにおいて, 学習データは約 10 万の対訳例からなる.

2.2 自動評価尺度

XML タグを除いた生テキストにした状態での BLEU スコアに加えて, IT 分野テキスト特有の製品名などの固有名詞や数値表現 (NE&NUM) に関する Precision と Recall を評価する. また, XML を考慮した評価を行うため, XML 構造の精度 (Acc.) と, 参照訳の XML 構造との一致率 (Match) を評価する. Match に関しては, 参照訳の構造と完全一致したときのみ正解として数える. さらに, Match で正解した場合には翻訳結果を XML タグで分割して, セグメントごとに訳出を比較した BLEU スコアも評価する. この BLEU の場合には, Match で不正解の翻訳結果に関しては出力無しとしてペナルティを与える.

3 XML タグ付き翻訳モデル

構造付きのテキスト翻訳のタスクとして, 構文情報を利用した機械翻訳モデル [2] の利用が考えられるが, 今回はベースラインとして XML タグを含んだ sequence-to-sequence (seq2seq) タスクとして一般的なニューラル機械翻訳モデルを使用する [1]. 単語分割は任意であるが, 対象とする XML タグはそれ以上分割しない制

約を加える. Seq2seq モデルとしては Transformer [5] を使用するが, モデル固有の手法は考えない.[2]

3.1 XML 制約付きビームサーチ

Seq2seq モデルはその利便性から広く使用されているが, 本タスクにおいて正確に XML タグを出力する保証はない. 本タスクではソース言語側の XML 構造がターゲット言語側の出力時に重要な情報になるため, ビームサーチによる出力時に明示的に制約を加える. 最初の制約は, ソース側の XML タグそれぞれ一回ずつだけ出力時にタグを開けることである. 次に, 現在開いているタグのみ閉じることが出来るという制約を課す. 最後に, ソース側の全てのタグを網羅するまで出力を終わらせない制約を加える. 以上により, ソース側に条件づけられた XML 構造を可能な限り正確に出力することが出来る. ただし, XML 構造が正しくとも翻訳が正確であるとは限らない.

3.2 複数のコピー機構

本データセットは企業サービスのオンラインヘルプに基づいているため, 非常に似通ったフレーズが頻繁に出現する. また, 一貫した英語表記の製品名などの固有名詞, または数値表現などが多く扱われている. このような場合には, 単純に seq2seq モデルのみに依存するよりも, そのまま該当箇所をコピーすることが有用であると期待される.

そこで本稿では, コピー機構 [4] を利用する. これは元々は同一言語内のタスクである自動要約のモデルに適用された手法であるが, 本タスクにおいても固有名詞, 数値, XML タグなどをそのままコピーするために有用であると考えられる. また, データセット内の類似した対訳例を翻訳メモリのように活用するため, Gu et al. [3] の手法もコピー機構として利用する. これは See et al. [4] の手法のコピー元がソース側言語でなく, 学習データから取り出した類似対訳例のターゲット側言語をコピー元にした手法である. つまり, 同様のコピー機構のコピー対象が 2 種類になるということになる. 学習時には各タイムステップの seq2seq の単語出力確率 p_g と, 学習データからのコピーによる単語出力確率 p_r と, ソース言語側からのコピーによる単語の出力確率 p_s を個別に学習し, p_r と p_s を学習する際には「コピーしない確率」も学習する. テスト時には単語の出力元を明確にするため, 離散的な選択を以下の式により行う:

$$p = (1 - \delta_s)p_s + \delta_s(\delta_r p_g + (1 - \delta_r)p_r), \qquad (1)$$

[2]モデルの詳細な定式化, 実験設定などに関してはデータセット公開と同時に原稿を公開する.

	BLEU	NE&NUM Precision, Recall	BLEU	NE&NUM Precision, Recall
	English-to-Japanese		English-to-French	
OT	61.61	89.84, 89.84	64.07	88.64, 85.64
X	62.00	92.54, 90.51	63.98	87.48, 86.98
X_{rs}	64.25	91.64, 90.98	63.51	88.42, 85.64
$X_{rs}^{(T)}$	64.34	93.39, 91.75	65.04	88.98, 88.31
	English-to-Finnish		English-to-German	
OT	43.97	87.58, 84.99	50.51	88.40, 86.55
X	42.84	83.17, 85.55	50.96	88.79, 86.43
X_{rs}	45.10	86.41, 86.49	52.91	88.00, 86.78
$X_{rs}^{(T)}$	45.71	87.38, 88.91	52.69	88.22, 88.45

表 1: 開発データにおける XML 無しの評価と, テストデータにおける X_{rs} の評価 ($X_{rs}^{(T)}$).

ここで, δ_r と δ_s はバイナリ値であり, これにより各ソース元からコピーするかどうかの選択を行う. まずは正確性の求められるソース側からのコピーを考え, しない場合には学習データからのコピーを考え, 最後に seq2seq モデル本体からの出力を考える手法になっている.

4 実験

4.1 実験設定

本データセットでは 17×16 言語対を実験対象として考えられるが, 原稿のスペースの都合上, そのうちの 4 言語対 (English-to-{Japanese, French, Finnish, German}) を例として取り上げる. モデルは以下の 3 種類を考える: OT (XML タグを取り除いたデータで学習した Transformer モデル), X (XML を考慮したモデル), X_{rs} (XML とコピー機構を考慮したモデル).

4.2 XML 無しの評価

表 1 に, XML を考慮しない評価結果を示す. まず OT と X の比較により, XML タグと共に学習することで翻訳精度が上がる傾向が, 本稿に載せていない他の 4 言語対に関しても確認できた. これは XML タグによりフレーズの対応情報が暗に学習時に利用できるためだからであると考えられる. しかし English-to-Finnish では BLEU が大きく下がってしまったが, これは後述する人手評価の結果と関連している. BLEU に関しては, 翻訳メモリを模した学習データからのコピーにより大幅な向上が確認できた. しかし BLEU が上がる半面, NE&NUM に関するスコアが悪化し, ソースからのコピーによってその悪化分が改善されるという傾向が確認できた. つまり, BLEU を上げる努力をすることで他の重要な指標で精度が下がってしまうことに注意する必要がある.

ドメイン適用の可能性 本データセットは Salesforce のオンラインヘルプのドメインに特化しているが, 基

学習データ	開発データ	newstest2014
本データセット	64.07	7.35
w/ 10K news	63.66	14.02
w/ 20K news	64.31	16.30
10K news のみ	0.90	2.66
20K news のみ	2.35	6.72

表 2: English-to-French のドメイン適用結果 (BLEU).

	BLEU	XML Acc., Match	BLEU	XML Acc., Match
	English-to-Japanese		English-to-French	
X	59.77	99.80, 99.55	61.81	99.60, 99.30
X_{rs}	62.06	99.80, 99.40	61.87	99.80, 99.50
$X_{rs}^{(T)}$	62.27	99.95, 99.60	63.19	99.80, 99.35
	English-to-Finnish		English-to-German	
X	41.98	99.65, 99.25	48.91	99.85, 99.25
X_{rs}	43.57	99.50, 99.25	51.16	99.75, 99.30
$X_{rs}^{(T)}$	44.22	99.90, 99.65	50.47	99.80, 99.20

表 3: XML を考慮した評価結果.

本的な文法の変換が学習できる程度のデータセットになっていれば, 対象とするすべての言語対におけるドメイン適用のソースデータとして用いることができる. これを示すため, English-to-French 翻訳において語彙や文体が大きく異なるニュースドメインにおける評価を行った. 表 2 に結果を示す通り, ニュースドメインの学習データを 10,000 または 20,000 例加えることで, ニュースドメインの評価データである newstest2014 における BLEU が大きく向上した.

4.3 XML 有りの評価

表 3 に, XML を考慮した自動評価結果を示す. XML 制約付きビームサーチにより, 高い精度でソースに条件付けされた XML 構造が正確に出力されていることがわかる. 例えば English-to-Japanese の X モデルの設定で制約なしのビームサーチを行うと, Acc. と Match が 98.70% と 98.10% まで下がり, BLEU も 59.77 から 58.02 に下がる. XML 無しの評価の場合と同様に, ここでもコピー機構により全体的に翻訳精度が向上することが確認できた.

4.4 人手評価

図 3 に人手評価の結果を示す. スコアは 1 から 4 の 4 段階で, 修正なしですぐ使える完璧な翻訳結果なら 4, 少しポストエディットが必要なら 3, もっとポストエディットが必要であるが無いより良いなら 2, 一から翻訳するほうが良いほど悪い翻訳なら 1, というスコアづけである. XML 有りのコピー機構付きのモデルで, テストデータからランダムに 500 事例取り出して, それぞれのプロの翻訳者による評価を行った.

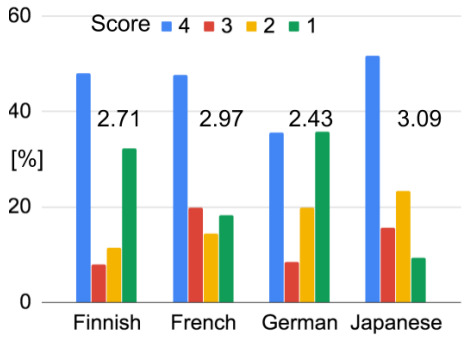

図 3: 人手評価.

人手評価の結果はおおまかに BLEU の結果と相関があることが見て取れ, 日本語の結果が最も良いことがわかった. 逆に評価の低いフィンランド語やドイツ語は複合名詞の扱いが難しい言語であり, ここまで BLEU が高く出るドメインにおける評価においてもまだまだ実用を考えると改善の余地があることがわかる.

この人手評価ではスコアが 4 で無い翻訳例に関してはどういったエラーが存在するのかをアノテーションされており, 文法ミス, 語彙ミス, ニューラルモデル特融の繰り返しミスなどが報告された. 一般的に繰り返しや欠落といったエラーが期待通り多かったものの, フィンランド語についてのみ, XML 構造に関する不適切な出力が多いとの報告があった. この結果は, XML と同時に学習するとフィンランド語のみ BLEU が大きく下がってしまったことと関係しており, 複雑な複合名詞を扱う言語等においては固有の難しさがあると考えられる. これからは様々な対象言語ごとに固有の問題を解決していくことが重要である.

5 おわりに

本稿では XML 構造を有するテキストの翻訳データセットを構築し, システムの評価を行った. 今後は実際のポストエディットの効果も確かめていきたい.

参考文献

[1] R. Aharoni and Y. Goldberg. Towards String-To-Tree Neural Machine Translation. In *ACL*, 2017.

[2] A. Eriguchi, K. Hashimoto, and Y. Tsuruoka. Tree-to-Sequence Attentional Neural Machine Translation. In *ACL*, 2016.

[3] J. Gu, Y. Wang, K. Cho, and V. O. K. Li. Search Engine Guided Neural Machine Translation. In *AAAI*, 2018.

[4] A. See, P. J. Liu, and C. D. Manning. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL*, 2017.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, and I. Polosukhin. Attention is All you Need. In *NIPS*. 2017.