日本語から英語への文脈翻訳テストの提案

永田 昌明　　森下 睦
NTT コミュニケーション科学基礎研究所
{nagata.masaaki, morishita.makoto}@lab.ntt.co.jp

1 はじめに

ニューラル機械翻訳の登場により文単位の翻訳精度は大きく向上し、さらに翻訳精度を上げるために文脈を利用する研究が始まった [15, 1, 17] 。文脈を利用する最も簡単な方法は、原言語および目的言語において直前の文と現在の文を <concat> のような特別な区切り記号で連結し、通常の文単位の翻訳を行うことでである [15]。この方法は 2-to-2 と呼ばれ、最先端の文脈翻訳手法と大差ない精度が得られることが知られている [1, 17]。これに対して通常の文単位の翻訳は 1-to-1 と呼ばれる。また入力文と文脈に別々の encoder を使用する方法も提案されており、encoder として RNN[1] や transformer[17] が使われている。

文脈の利用法を関する研究の問題点の一つは、BLEU のような従来の自動評価尺度では、文脈の何が問題でそれが提案手法によりどう改善されたのかよくわからないことである。文献 [1] では、対照テストセット (contrastive test set)[11] の枠組みに基づいて、英語からフランス語への翻訳において機械翻訳が文脈を理解して適切な訳文を生成する能力を評価する文脈テストセットを提案した

本稿では日本語から英語への翻訳の文脈テストの作成法を提案する。我々は当初、文献 [1] の方法に基づいて共参照と一貫性に関する文脈テストを作成しようとしたが、文脈処理の対象となる問題は言語対および翻訳方法に大きく依存するため、試行錯誤の結果、全く別の方法に辿り着いた。以下では、まず日英文脈翻訳テストセットの概要を述べ、これを用いて 2-to-2 による日英文脈翻訳を分析した結果を報告する。

2 文脈翻訳の対照テスト

日本語から英語への文脈翻訳テストとして、日本語の第 1 文に依存して日本語の第 2 文の英語訳が決定されるような連続する 2 つの日本語文、および、その翻訳となる連続する 2 つの英語文を作成した。さらに、正解と対照可能な不正解として、英語の第 2 文に対して文脈を考慮しないことにより生じる必要最小限の誤りを加えた英語文を作成した。この際、できるだけ正解率が 50%になるよう工夫した。ここで考慮する文脈情報は共参照 (coreference) と一貫性 (coherence/cohesion) とし、その総数は約 1000 件である。

2.1 共参照

Source:	
Context:	申し訳ありませんが、先生は午後少し遅れているんです。
Input:	(*pro* が)診察するまでに 20 分ほどかかると思います。
Target:	
Context:	I'm afraid that the doctor is running a bit late this afternoon.
Correct:	It might be about 20 minutes before he can see you.
Incorrect:	It might be about 20 minutes before we can see you.

(a) ゼロ代名詞

Source:	
Context:	ソファのそばには木製の椅子がある。
Input:	レンズとピンセットが椅子に乗っている。
Target:	
Context:	Beside the couch was a wooden chair.
Correct:	A lens and a forceps was lying upon the seat.
Incorrect:	A lens and a forceps was lying upon a seat.

(b) 冠詞

図 1: 共参照に関するテスト

共参照については、日本語のゼロ代名詞と英語の冠詞に関するテストを作成した。これらは目的言語に対応する原言語の言語要素が存在しないので、基本的に文脈に依存して生成される。

日本語では文脈から了解できる主語や目的語は省略

される。それに対して英語は構文上の制約から主語や目的語が必須である。そのため英語側の代名詞の人称、数、格を正しく決定しなければならない。図 1(a) の例では、文脈がなければ「思う」のような思考動詞や「診察する」のような意思を伴う行為動詞のデフォールトの主語は一人称であるが、ここでは前の文に登場する人物を指していることから正解は三人称になる。

文献 [1] では、OpenSubtitles にある実例を参考にして文脈テストを作成している。当初、我々も OpenSubtitles や Ted Talk のコーパスからテストを作成しようとしたが、日本語のゼロ代名詞のデフォールト訳が一人称または二人称であることが多いという視点が欠け、不自然になることが多かった。また日本人の作業者でも、ゼロ代名詞を正しく分析するのは難しく、作業誤りが目立った。

そこでまず信頼できるゼロ代名詞の分析結果として Keyaki Treebank[2] を用いることにした。具体的には、まず日本語のゼロ代名詞 (*pro*) を含む文を選択し、これを Google 翻訳等の機械翻訳にかけて、デフォールトのゼロ代名詞の英語訳 (正解訳) を確認する。次に不正解となる英語の代名詞 (不正解訳) を選び、正解訳が正解となり不正解訳が不正解となるような先行文脈と、正解訳が不正解になり不正解訳が正解になる先行文脈を作成した。同時に 2 つのテストを作成することで正解率を 50%にすることができる。

図 1(b) に冠詞のテストの例を示す。日本語側で第 1 文と第 2 文で「椅子」に言及しているので、英語側の第 2 文では定冠詞 the が使われている。日本語側の「椅子」に対して英語側は表層形が異なり (chair と seat)、橋渡し参照 (bridge reference) になっている。我々は冠詞の定/不定に関して先行文に (橋渡しも含めて) 先行詞がある場合と先行文に指示対象がないダミーの先行文が半々になるようにテストを作成した。

対訳コーパスから冠詞の定/不定の例を探すのは効率が悪いので、文法誤り訂正のテストセット HOO-2012[4] と共参照解析の注釈つきのコーパス OnteNotes5.0[5] を使用した。自然性が高いダミーの先行文を作成するのは非常に難しいので、文法誤り訂正の正解データにおいて the から a へ訂正されている以下の例のような文は、照応関係が成立しない文を作成するのに有益であった。

> In our country, there are rules that everyone has to follow, and recently a new rule was added.
>
> We aren't allowed to use a (*the) mobile phone in class.

照応関係があり、一方が a で他方が the であるような英語文のデータ OnteNotes の coreference chain から容易に見つけることができる。

2.2 一貫性

Source:
Context: 昨日、渋谷へ行った。
Input: すごい人だった。
Target:
Context: I went to Shibuya yesterday.
Correct: There are a lot of people.
Incorrect: He is a great man.

(a) 曖昧性解消

Source:
Context: いい時計ですね。
Input: この時計は父の形見なんです。
Target:
Context: It's a nice clock.
Correct: This clock is a memento of my father.
Incorrect: This watch is a memento of my father.

(b) 対応

図 2: 一貫性に関するテスト

一貫性に関しては言語に依存する部分が少ないので、文献 [1] に従って曖昧性解消 (disambiguation) と対応 (alignment/repetition) のテストを作成した。

図 2(a) に曖昧性解消のテストの例を示す。「すごい」には「多く」と「偉大な」という 2 つの語義があり、訳し分けに文脈が必要である。一般に曖昧性に関するテストは以下の 3 つの条件を満たす。

- 原言語文は多義 (曖昧性) を持つ語句を含む。
- その多義は目的言語の異なる語句に翻訳される。
- 原言語または目的言語の先行文によってどちらの訳を使うかが決まる。

当初、対訳データから上記の条件にあてはまるものを探そうしたが、日本語は同じ発音でも意味が異なる場合には違う漢字を使うので、英訳が複数ある日本語の単語は少ない。そこで我々は複数の日英辞書から上記の条件を満たす単語を選定し、ゼロから人手でテストを作成した。ちなみに英日翻訳の場合には、日本語が多義になるものが多いので、このテストの作り易さは言語対だけではなく翻訳方向に依存する。

図 2(b) に対応に関するテストの例を示す。時計の英訳は clock または watch であるが、同一文脈で同一対象を指示する際に 2 つの訳が混在してはいけない。一般に一貫性に関するテストは以下の 3 つの条件を満たす。

- 原言語文は多義を持つ語句を含む。
- その多義は目的言語の異なる語句に翻訳される。目的言語ではこれらはほとんど別概念なので、相互に入れ替えることは不可能である。
- 原言語の先行文でも同じ語句単語を含み、当該語句の意味の違いに応じて、目的言語では現在の文と同じ多義が生じる。

目的言語ではほぼ別概念なので、先行文と同じ語句を使う必要があるという制約は、言語対や翻訳方向によらない一般的な問題である。ここでどの語句が正解訳かは言語表現と実世界との対応 (grounding) で決まるが、これは本テストの対象外とする。

3 実験

3.1 データとツール

文脈翻訳の実験に使用した日英対訳データを表 1 に示す。訓練データは約 340 万文、開発データとテストデータはそれぞれ約 3 万文である。このうち OpenSubtitles2018[7], IWSLT-2017[3] は話し言葉、Global Voices[10], Wikipedia 日英京都関連文書 (KFTT)[1], 日英対訳文対応付けデータ [16], ひらがなタイムズ, 読売新聞社説は書き言葉である。

英語文は Moses toolkit を用いて字句解析および小文字化し、日本語文は NFKC 正規化して MeCab UniDic で単語分割した。さらに英語文と日本語文はバイト対符号化 [12] により部分単語に分割した (32K 共通併合操作)。翻訳には、注意付き符号器復号器モデル [8] を実装した OpenNMT-lua [2]をデフォールト設定で使用し、翻訳精度 BLEU[9] は Moses toolkit の multi-bleu.perl で測定した。

Dataset	Split	sents	len(ja)	len(en)
IWSLT2017	train	218,174	22.3	20.6
(TED Talks)	dev	2,577	21.8	19.1
	test	2,357	22.5	19.5
OpenSubtitles2018	train	2,077,430	7.6	8.5
(movie subtitles)	dev	3,245	9.0	7.7
	test	2,901	6.9	8.9
GlobalVoices	train	29,508	27.8	21.6
(blog)	dev	9,426	28.4	22.0
	test	8,148	28.1	21.8
HiraganaTimes	train	16,472	24.6	22.0
Books	dev	2,792	22.4	20.3
(book)	test	2,537	23.4	21.3
HiraganaTimes	train	189,925	24.7	21.1
(magazine)	dev	5,385	21.3	19.8
	test	5,004	21.1	19.9
NICT_align	train	103,417	20.5	14.9
(book)	dev	4,279	21.0	15.1
	test	3,212	17.8	13.1
Wikipedia_Kyoto	train	480,778	23.4	24.9
(Wikipedia)	dev	1,257	20.2	19.9
	test	1,287	21.3	21.7
Yomiuri_editorial	train	283,710	27.2	28.2
(newspaper	dev	3,002	23.3	24.8
editorials)	test	3,014	24.3	26.4
All	train	3,399,414	14.0	14.3
	dev	31,963	22.4	19.1
	test	28,460	22.0	19.4

表 1: 日英対訳データに関する統計量

Dataset	1-to-1	2-to-2
IWSLT2017	12.19	12.26
OpenSubtitles2018	12.23	12.67
GlobalVoices	10.66	10.80
HiraganaTimes_books	12.91	13.23
HiraganaTimes	13.23	13.32
Wikipedia_Kyoto	9.36	9.64
NICT_align	23.09	23.12
Yomiuri_Editorial	14.53	15.26
All	12.77	13.02

表 2: 各データに対する 1-to-1 と 2-to-2 の翻訳精度

3.2 翻訳精度と文脈翻訳テスト

訓練データから 1-to-1 と 2-to-2 のモデルを作成し、各データのテストセットで翻訳精度 (BLEU) を測定した結果を表 2 に示す。2-to-2 の翻訳精度は 1-to-1 に比べて一貫して良いがその差は小さい。

OpenNMT-lua の -tgt オプションを使って、正解文対と不正解文対を強制翻訳 (forced decoding) し、正解文対の対数確率が不正解文対より高いテストの割合を文脈翻訳テストの正解率とする。表 3 に各テストに対する 1-to-1 と 2-to-2 の正解率を示す。

テスト	カテゴリ	tests	1-to-1	2-to-2
共参照	冠詞	330	0.76	0.75
	代名詞	220	0.53	<u>0.70</u>
一貫性	曖昧性	378	0.50	0.52
	対応	73	0.49	0.50

表 3: 各テストに対する 1-to-1 と 2-to-2 の正解率

代名詞翻訳に関して 2-to-2 は 1-to-1 より大幅に正解率が高いが、その他のカテゴリでは両者に大差はない。冠詞の正解率 (約 75%) は設計上のベースライン (約 50%) より大幅に高い。これは、代名詞や曖昧性においては、多義のそれぞれについてそれが正解となる先行文脈を作成し、正解と不正解を入れ替えると、正解率が 50%になるようにテストを作成できるが、冠詞に関してはこのような先行文脈を複数用意することが難しいことと、一文の中だけで言語モデルが正解を出してしまうことが原因だと思われる。

4 おわりに

ルールベース翻訳や統計的機械翻訳では、日本語から英語への翻訳において日本語のゼロ代名詞は最も難しい問題とされ、様々な手法が提案された [13, 6, 14]。本研究で提案した日英文脈翻訳テストにより、ニューラル機械翻訳では、一般的な文脈翻訳の枠組みを用いて日本語のゼロ代名詞を英語へ翻訳する精度を大きく改善できる見通しが得られた。

今後は、英語から日本語への文脈翻訳テストの作成法、および、文脈翻訳において一貫性を改善する手法を検討したい。

[1] https://alaginrc.nict.go.jp/WikiCorpus/index_E.html

[2] http://opennmt.net/

参考文献

[1] Rachel Bawden, Rico Sennrich, , Alexandra Birch, and Barry Haddow. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the NAACL-2018*, pp. 1304–1313, 2018.

[2] Alastair Butler, Tomoko Hotta, Ruriko Otomo, Kei Yoshimoto, Zhen Zhou, and Hong Zhu. Keyaki treebank: Phrase structure with functional information for japanese. In *Proceedings of Text Annotation Workshop*, 2012.

[3] Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the IWSLT-2017*, pp. 2–14, 2017.

[4] Robert Dale, Ilya Anisimoff, and George Narroway. Hoo 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pp. 54–62, 2012.

[5] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: The 90% solution. In *Proceedings of the NAACL-2006*, pp. 57–60, 2006.

[6] Taku Kudo, Hiroshi Ichikawa, and Hideto Kazawa. A joint inference of deep case analysis and zero subject generation for japanese-to-english statistical machine translation. In *Proceedings of the ACL-2014*, pp. 557–562, 2014.

[7] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the LREC-2018*, pp. 1742–1748, 2018.

[8] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the EMNLP-2015*, pp. 1412–1421, 2015.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the ACL-2002*, pp. 311–318, 2002.

[10] Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. Parallel global voices: a collection of multilingual corpora with citizen media stories. In *Proceedings of the LREC-2016*, pp. 900–905, 2016.

[11] Rico Sennrich. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. In *Proceedings of the EACL-2017*, pp. 376–382, 2017.

[12] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the ACL-2016*, pp. 1715–1725, 2016.

[13] Hirotoshi Taira, Katsuhito Sudoh, and Masaaki Nagata. Zero pronoun resolution can improve the quality of j-e translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-2012)*, pp. 111–118, 2012.

[14] Shunsuke Takeno, Masaaki Nagata, and Kazuhide Yamamoto. Integrating empty category detection into preordering machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT-2016)*, pp. 157–165, 2016.

[15] Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pp. 82–92, 2017.

[16] Masao Utiyama and Mayumi Takahashi. English-japanese translation alignment data.

[17] Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the ACL-2018*, pp. 1264–1274, 2018.