The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

機械翻訳における訳語一貫性評価用データセットの構築

Construction of Coherent Translation Evaluation Dataset in Machine Translation

阿部 香央莉 *1 Kaori Abe　鈴木 潤 *1*2*3 Jun Suzuki　永田 昌明 *3 Masaaki Nagata　乾 健太郎 *1*2 Kentaro Inui

*1東北大学 Tohoku University　*2理化学研究所 AIP センター Center for Advanced Intelligence Project (AIP)

*3NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories

This paper focuses on the topic of coherent translation in the document-level Neural Machine Translation (NMT). To evaluate NMT systems whether they can maintain consistent translations or not, we construct a dataset for evaluating the coherence. We evaluate typical baseline NMT systems on our dataset and discuss a coherent problem in terms of the term selection in the existing systems.

1. はじめに

2014 年に系列変換モデル (sequence-to-sequence model)[1] が提案されて以降，深層ニューラルネットを用いた機械翻訳（以降 NMT と呼ぶ）が盛んに研究されている．特に直近数年の技術発展はめざましく，注意機構 [2, 3]，サブワード [4]，逆翻訳による擬似データ作成 [5]，Transformer モデル [6] といった翻訳品質を大幅に向上させる革新的な方法論が次々に考案されている．これらの技術により，現在では非常に良好な翻訳結果が得られるようになりつつある．

しかし，これらの技術は，一文単位での翻訳を行う方法論としての研究成果である．よって，文章単位での翻訳を考えた場合に，文脈に依存した訳語選択，代名詞／ゼロ照応補完，訳語 *1 の一貫性といった，文章全体として自然な翻訳を実現するという観点では，まだ十分対応できているとは言えない．実際に，今年度開催予定の WMT2019*2 のシェアードタスクでは，文章単位の翻訳（Document-level MT）のデータが新たに追加される．また，文章単位，あるいは，連続する複数文を考慮して翻訳を行う方法論に関する研究成果も徐々に報告されるようになってきた [7, 8] このように，NMT の研究は，一文単位の翻訳から文章単位の翻訳へと研究の焦点が移動する転換期にさしかかっていると考えられる．このような背景から，本稿では，文章単位の翻訳を考えた際に解決すべき課題の一つとなる「訳語の一貫性」に焦点を当てる．

訳語の一貫性には様々なパターンが考えられる．例えば，表 1 中の例 1 のように，英語の文章における “our company” に対して，日本語では「我が社」「当社」「弊社」などの様々な訳語が許容できる場合がある．これは，書き手と読み手の関係性や状況などに応じて適切な訳語を使い分ける必要がある例である．また，表 1 中の例 2 では，日本語の文章中では「時計」と表現される対象について，英語では掛け時計を “clock”，腕時計を “watch” と呼び分けるため，文脈に応じて正しい訳語に統一する必要がある．これは言語間で，単語の内包する概念が一部異なるため，訳し分けが必要な例といえる．

一般論として，同一文章内で「同じもの（名詞・名詞句）」や「同じ行為（動詞・動詞句）」を指すとき,「同じ表現」であることが望ましい．特に，マニュアルや論文などの事実を読み手に正確に伝える必要のある文章では，同一文章内で同じものを異なる表現で表すと，読み手に同一のものを指していることが伝わらず，混乱を招く要因となり得る．よって，NMT の能力向上の一つとして，ある文章中で訳語の一貫性を考慮できる方法論の研究を推進したいと考える．しかし，現実的には訳語の一貫性の方法論の研究を行うには大きな課題が存在する．それは，現状の翻訳の評価用データが一貫性を評価できる作りになっていない点，および，BLEU 等の翻訳で使われる標準的な自動評価指標では，訳語の一貫性を正当に評価することが難しい点である．そこで本研究では，方法論作成の前段階として，まずは訳語の一貫性を評価できるデータセットを構築することに取り組む．

表 1: 一貫性がない翻訳の例

	例 1：日本語側に一貫性がない例
Japanese	我が社は赤字続きだ.
English	**Our company** has been in the red.
Japanese	当社は人材紹介会社である.
English	**Our company** is a staff agency.
Japanese	弊社の西側のオフィス
English	**our company**’s western office
	例 2：英語側に一貫性がない例
Japanese	いい時計ですね.
	この時計は父の形見なんです.
English	it’s a nice **clock**.
	this **watch** is a memento of my father.

表 2: 評価用データセットの統計値

	日英-英	日英-日	英日-英	英日-日
一貫性訳語種類数		31		125
行数		415		418
全語彙数	2947	2962	3137	2796
平均文長数（文字単位）	200.3	97.99	169.5	86.82
平均文長数（トークン単位）	36.93	41．03	28.84	34.29

2. NMT の学習・評価用データセット

2.1 現状

前述したように，NMT では，これまで一文単位で独立に翻訳を行う方法論が主流であったため，文章単位での訳語の一貫性に関しては考慮されてこなかった．そのため，NMT で用いられる学習および評価用データセットも，基本的に訳語の一貫性を評価するようには設計されていない．つまり，NMT で用いられる学習および評価用データセットでは,「同じもの・行為」を表すのが同じ表現で統一されているとは限らない．特に，NMT では，より多くの対訳データを用いるとより翻訳品質が高くなるという性質が観測されているため，基本的に様々な文章から対訳データを収集して学習用データとして用いている．

連絡先: 阿部　香央莉 (Kaori Abe)　東北大学大学院 情報科学研究科　乾・鈴木研究室
〒 980-8579 宮城県仙台市青葉区荒巻字青葉 6-6-05 東北大学工学研究科 電子情報システム・応物系 1 号館 6 階
E-mail: abe-k(at)ecei.tohoku.ac.jp
TEL: 022-795-7091

*1　ここでは「訳語」という用語を用いるが，一単語である場合だけでなく，複数の語の並び（句）の場合も含むこととする．

*2　http://www.statmt.org/wmt19/

よって，学習用データでは訳語の一貫性が担保されるような状況にはなっていないと容易に推測できる．よって，そのデータを用いて学習された NMT モデルでは訳語を統一することは基本的にはあまり考慮されていない．同様に，評価用データに関しても，複数の文章から収集してきたものを一つの評価データとして用いることが多いため，相対的に量は少なくても学習用データと同様に，ある語（または句）に対して複数の訳語が割り当てられている状況となっていることが多い．

2.2 訳語の一貫性評価の要件

本稿では，議論を明確にするために，以降は日英・英日翻訳に限定して議論を進める．

本研究では，前述のように複数の訳語が考えられる状況において，文章中で訳語が一貫しているかを評価できるデータセットの構築が目的である．その要件として，複数の訳語が考えられる翻訳元の語（または句）を訳語との対応付きで抽出し，抽出された語に対して同一の語（または句）に関しては，同じ訳語が割り当てられるかを評価することで訳語の一貫性の評価をすることを考える．例えば，“our company” に対して「我が社」「当社」「弊社」という 3 種類の訳語を含む文が存在する評価データがある場合，全ての “our company” に対して「我が社」「当社」「弊社」のいずれか一つを一貫して訳語として選択できたかで訳語の一貫性を評価する．この時，本研究では「我が社」「当社」「弊社」の 3 種類のどれを選んでも一貫性が保たれていれば正解とみなす．或いは，どれか一つの訳語を利用者が事前に決定し，その訳語を選択できたかを評価する方法でもよい．よって，例えば，対象とする訳語がきちんと翻訳文に出現したかの割合（正解率）で訳語の一貫性を評価することができる．或いは，より詳細に翻訳元の文と翻訳先の文で一貫性を担保したい語と訳語の語句対がきちんと対応がとれて翻訳されているかを評価することで，一貫性の評価を行うことができる．

よって，訳語一貫性評価用データセットとして，ある対訳データのある翻訳方向において，翻訳元の文のある語句の翻訳として十分に尤もらしい複数の翻訳先の訳語が存在するような語句対と文対を抽出する．つまり，訳語一貫性評価用データセットの構成要素は，1:対訳データ，2:翻訳方向，3:語句対（一つの語に複数の訳語を許容する），4: 文対 (入力文と参照訳) となる．このような要件を満たす評価用データの作成方法を次節で述べる．

3. データセット作成手順

本節では，訳語の一貫性を評価する評価用データセットの構築方法を述べる．図 1 にデータセット構築方法の概略を示す．まず，基本的な考え方として，データを最初から人手にて作成するのではなく，既存の評価用データセットを改変して訳語の一貫性を評価できるようにする．

訳語の一貫性を評価する評価用データセットを構築する概略は以下の通りである．

1. 単語アライメントから語句対の辞書取得
2. 語句対を自動或いは人手フィルタリングし，ある翻訳元の語から複数の訳語が存在する語句対を抽出
3. 元データセット中から，2 で抽出された語句対が含まれる文を抽出
4. 抽出された文に対し，自動・手動で 2 で抽出した語句対の訳語候補を人手で最終チェック

本研究では，日英翻訳の評価用データセットとして KFTT（京都翻訳フリータスク）コーパス [9][*3]，英日翻訳の評価用データセットとして ASPEC（Asian Scientific Paper Excerpt Corpus）[10][*4] を取り上げ，これらを基にして訳語の一貫性評価用データセットを構築する．表 2 に，作成した評価用データセットの統計値を示す．表 2 における一貫性訳語種類数と行数は，それぞれデータセットの対訳文中で訳語が統一されている語句対の種類数，データセット中における対訳文数を表す．また，その他の統計値として，各データセットにおける全語彙数，平均文長数をそれぞれ示す．

以下に，それぞれのデータセットの詳細な作成方法を述べる．

3.1 日英翻訳用データセット (KFTT)

KFTT は，京都関連の Wikipedia 記事から作成されたデータセットであり，日本の文化・歴史などが主に記述されている．このデータセット中には日本独自の概念（地名・人名などの固有名詞含む）が多く含まれているため，英語に翻訳する場合にその概念を表現する方法が異なる場合がある．例えば，固有名詞の日本語における読みをそのままローマ字で記述することもあれば，その名詞の概念そのものを英語で説明した句で置き換える例がみられる．具体的には，「源氏」という日本語が「genji」と「minamoto clan」という 2 通りの英語表現に訳されている．日本語話者にとって，この 2 つの英語表現が完全に同じ「源氏」という概念を表すことはほぼ自明であると思われるが，英語話者が日本の歴史を勉強しようとしてこの 2 文に遭遇した場合，これらは違うものではないかと疑問を抱く恐れがある．このように，KFTT は，前述の訳語の一貫性を考慮した評価用データを作成するのに適した要件を満たしている．

まず，前述の 1 の処理として，KFTT のチューニングセットに付加されている人手アライメントを利用し，語句対の辞書を取得する．次に，2 の処理として，一般的な頻出単語ではないが KFTT に頻出する単語を取得する．本研究では，OpenSubtitles[11][*5] という映画字幕コーパスの単語分布を一般的な頻出単語分布とみなし，「KFTT に出現するが Opensubtitles には出現しない単語」を抽出する．先ほどアライメントで得られた単語ペア辞書のうち，KFTT 頻出単語に該当する単語ペアのみを抽出する [*6]．3 の処理として，抽出された語句対が含まれる対訳文対を，元の KFTT のチューニングセットから抽出する．最後に 4 の処理として，人手で語句対と対訳文対の対応確認を行う．またその際に，対応する箇所を句単位に伸ばすことができる場合，手作業で句単位に拡張する．

3.2 英日翻訳用データセット (ASPEC)

ASPEC は，科学技術系論文から作成されたデータセットである．科学技術分野では，専門用語に相当する表現を日本語に訳す際，その訳語は翻訳者あるいは翻訳した時期によって変化することがあると考えられる．具体例として，「Faraday's law」の「Faraday」をそのままアルファベット表記で記述したり，カタカナに置き換えて記述したりする例が見られる．また，この「ファラデーの法則」自体が，元々は「電気分解の法則」と「電磁誘導の法則」の 2 種類を表す句であり，上記の例を厳密に記述する場合は「Faraday の電気分解の法則」と記述される場合もある．上記実例以外にも，訳語の一貫性がとれていない同様の状況が散見される．ASPEC も KFTT 同様に訳語の一貫性を評価するのに適した要件を満たしていると考えられる．

まず，ASPEC の評価データに対し GIZA++で自動アライメントを行い，単語ペア辞書を取得する．この時，GIZA++の学習は ASPEC の整理済み学習データを用いる．次に，得られた単語ペア辞書の中から「ASPEC 学習セット内では訳語の揺れがあるが，ASPEC 評価セット内では訳語がほぼ 1 つに定まる単語ペア」のみを抽出する．訳語揺れの基準は，出現頻度 3 回以上の訳し先が 3 つ以上ある場合とする．ここで ASPEC 評価セットの訳し先が複数ある場合には，アライメントが間違っていないことが判断でき，かつ最も高頻度な訳し先を人手で 1 つ選び単語ペアとして抽出する．抽出された単語ペアが含まれる対訳文対を ASPEC 評価セットから抽出し，人手で単語ペアと対訳文対の対応確認を行う．対応する箇所を句単位に伸ばすことができる場合，句単位に拡張する．また，文

*3 http://www.phontron.com/kftt/index-ja.html
*4 http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
*5 http://opus.nlpl.eu/OpenSubtitles2018.php
*6 抽出の際，訳語が 2 種類以上 10 種類未満，訳語の 1 つが少なくとも 5 回以上出現している，複数の訳語の回数が異なっているといった制約を付加した．

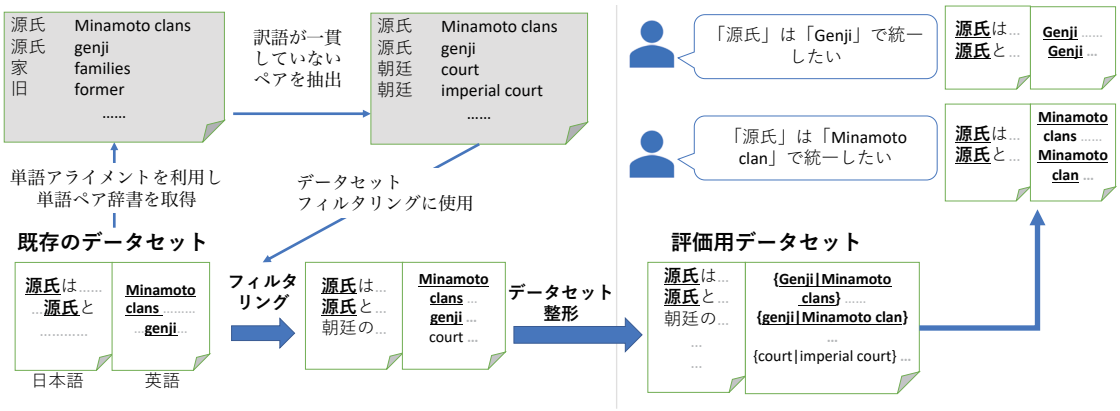

図 1: データセット構築手順（図中の例は KFTT のデータの例.）

表 3: 実験におけるハイパーパラメタ等の設定

前処理	BPE サブワード分割 [4]	8000 マージ
学習時	エポック数	13
	ミニバッチサイズ	64
	単語埋め込み次元数	512
	隠れ層次元数	512
	LSTM レイヤー数	2
	勾配正規化 (grad_norm)	5
	ドロップアウト率	0.3
	初期学習率	1.0
	学習率更新	9 エポック後から毎エポック 0.5 倍
デコード時	ビームサイズ	5

によって訳語が統一されていない場合は，統一する訳語を仮に 1 つ定め，訳語統一を行う．このような手順を経て最終的にできたものを，英日翻訳の評価用データとする．

4. 実験

4.1 実験設定

作成した評価用データセットを用いて，既存の NMT を評価する．本稿での実験の目的は，既存の NMT の翻訳結果が，訳語の一貫性の観点でどの程度妥当かを評価することである．

本研究の実験では，典型的な NMT モデルとして，これまでに広く用いられてきた文献 [12] で用いられている注意機構付き RNN 符号化復号化器を使用する．以下本手法を「ベースライン手法」と呼ぶ．また，訳語の一貫性を検証する上で有効な方法論となりえる制約付きデコーディング [13] も合わせて比較実験を行う．以下，本手法を「制約付きデコーディング手法」と呼ぶ．ただし，制約付きデコーディングに関しては，評価データ中の各文に対して一貫性のある訳語を制約として与えてデコーディングをする．つまり，訳語の一部が事前に与えられることに相当するため，通常の翻訳よりも有利な条件で翻訳していることになる．

翻訳モデル学習の際は，作成した評価用データに合わせて，英日翻訳モデルに対しては ASPEC，日英翻訳モデルに対しては KFTT の学習・開発セットを用いて学習を行った．本実験で用いた具体的なハイパーパラメタ等の設定を表 3 に記す．

評価指標には翻訳タスクにおいて一般的な自動評価尺度とされる BLEU を用いる．ただし，BLEU では「語彙の一貫性」を評価することはできない．そこで，BLEU 以外の評価指標として，暫定的に「訳語を統一した語句が出力文に含まれているどうか」を完全一致で判定し，含まれていれば正解としてその正解率を測る．

4.2 実験結果

表 4 に，日英および英日翻訳評価用データセットでの評価結果を示す．表中の一貫性訳語の出現率とは，翻訳先に一貫性の評価するための訳語が出現した場合に 1，しなかった場合に 0 とし，評価データ全体の出現率を計算した結果である．また，一貫性訳語の正解率とは，翻訳元の文と翻訳先の訳語の対応を求め，対応する翻訳元の単語の訳語となっている場合にのみ正解と考えた場合の正解率である．ただし，翻訳先と翻訳元の単語対応を求めるために TER[14] を用いた．

表 4: 評価用データセットにおける BLEU スコア

	BLEU		一貫性訳語の出現率		一貫性訳語の正解率	
	baseline	制約付き	baseline	制約付き	baseline	制約付き
KFTT	14.32	14.10	67.77	93.77	73.32	91.40
ASPEC	35.11	35.98	63.43	99.58	66.39	74.52

ベースライン手法と制約付きデコーディング手法を比較すると，日英翻訳 (KFTT) においては，ベースライン手法の方が制約付きモデルよりも BLEU スコアは高い．しかし，正解率の観点で見ると，ベースライン手法は 67.77 とだいぶ低い値となっている．先述のように，制約付きデコーディング手法では，一貫性に関する正解の訳語が与えられている状態で翻訳を行なっているため，本来 BLEU スコアも高くなって良いはずである．このように BLEU スコアが低くなる要因として，訳語の一貫性を強制するために，学習した翻訳モデルが破綻をきたし，一貫性に関連しない訳語の選択精度が劣化したことが考えられる．

また，英日翻訳 (ASPEC) においては．想定通り BLEU スコアも正解率も制約付きデコーディング手法が上回った．これは日英翻訳 (KFTT) と比較して，BLEU スコアが相対的にだいぶ高いため，一貫性の訳語を強制的に選択しても他の訳語選択の破綻が起きにくかったことに起因していると考えられる．ただし，この結果をもって制約付きデコーディング手法で訳語の一貫性の問題が解決したわけではない．今回の実験においては，一貫性の評価に用いられる訳語を事前に全て教えるという方法を用いているので，最も有利な状況での結果となっている．今後は，一貫性を評価する訳語も適切に選択する部分も含めて方法論を構築する必要がある．

ここで注意として，制約付きデコーディング手法では，訳語の一貫性ん評価対象の語の事前に与えているため理論上正解率は 100%になるはずである．本実験において，正解率が 100%になっていない理由として，訳語の一貫性を評価する対象の語が，学習データに出現しない未知語（厳密には未知 BPE サブワードを含む語）となるためである *7．

5. 考察

表 5 に，各手法による実際の翻訳例を示す．例 1 および例 2 では日英翻訳の例，例 3 および例 4 では英日翻訳の例をそれぞれ示している．

例 1 では，ベースライン手法のままでは,「南朝」を表す別の表現である “southern court” と翻訳されている．しかし，デコーディングで制約を付加したことにより，制約付きモデルではデータセット中で統一された訳語である “nancho” に正しく翻訳できている．しかし，例 2 においては，制約付きモデルは

*7 未知語以外の語に関しては，全て正解となっていることを確認した．

表 5: 実際の翻訳例

	例 1　一貫性訳語の種類 : (南朝, nancho)
入力文	また 南朝 最大 の 勢力 圏 で あ っ た 九州 に 今川 貞世 （ 了俊 ） ・ 大内 義弘 を 派遣 し て ， [**南朝**] 勢力 を 弱体 化 さ せ 幕府 権力 を 固め る ．
参照文	he also dispatched sadayo (ryoshun) imagawa and yoshihiro ouchi to kyushu , where [**nancho**] dominated , to debilitate its influence and consolidate the power of the bakufu . ”
baseline	in kyushu , he dispatched sadayo (ryoshun) and yoshihiro ouchi to kyushu , and strengthened the power of the **southern court** power . ”
制約付き	in kyushu , he dispatched sadayo (ryoshun) and yoshihiro ouchi to kyushu , and strengthened the power of the [**nancho**] power .
	例 2　一貫性訳語の種類 : (足利, ashikaga)
入力文	足利 尊氏 の 死 から 丁度 100 日 目 の こと で あ る ．
参照文	it fell precisely on the 100th day after the death of takauji [**ashikaga**] .
baseline	it was the 100th day after takauji [**ashikaga**] ’s death .
制約付き	[**ashikaga**] was the 100th day after takauji [**ashikaga**] ’s death .
	例 3 一貫性訳語の種類 : (標記, above)
入力文	”In the paper , an explanation is given on the [**above**] problem .”
参照文	本文 は [**標記**] 問題 の 解説 記事 で あ る 。
baseline	本稿 で は ， この 問題 に つ い て 解説 し た 。
制約付き	本稿 で は ， [**標記**] 問題 に つ い て 解説 し た 。
	例 4 一貫性訳語の種類 : (通信 事業 者, carrier)
入力文	”In the future of IP / IMS (IP multi - media subsystem) , the carrier networks remarkably convert .”
参照文	ＩＰ / ＩＭＳ (ＩＰ マルチメディア サブ システム) の 将来 に お い て ， [通信 事業 者] ネットワーク は 著し く 転換 する 。
baseline	ＩＰ / ＩＭＳ (ＩＰ マルチメディア サブ システム) の 将来 に お い て は ， キャリア ネットワーク が 著し く 転換 し て い る 。
制約付き	ＩＰ / ＩＭＳ (ＩＰ マルチメディア サブ システム) の 将来 に お い て は ， 通信 網 が 著し く 転換 し て い る 。

指定した訳語である “ashikaga” を過剰に出力してしまい，結果として意味が通らないような文を出力してしまっている．このエラーの理由としては，デコーディングの制約により強引に特定の単語を出力させたことで，本来は正しい出力を行うことができる言語モデルに悪影響を及ぼした可能性が高い．このように，単純な制約付きデコーディングを行うと，却って NMT が持つ翻訳能力を損なう場合があると考えられる．

例 3 では，制約付きデコーディングによって “above” に対応する「標記」という単語を正しく出力できている．このような例は英日翻訳では多く見られ，結果として BLEU スコアの向上に繋がったと考えられる．しかし，現状の制約付きデコーディングでは限界もある．例 4 では，デコーディングに制約を加えてもなお句である「通信事業者」が出現しなかった．これは,「通信事業者」という句が，学習データに出現しない未知語を含む句であるためである．このように，現状の NMT による方法では，完全に訳語の一貫性を保つことは難しいということが判明した．

6. おわりに

本稿では，文章単位の NMT を実現する上で課題の一つとなる「訳語の一貫性」に焦点を当て，訳語の一貫性を評価するための日英および英日翻訳評価用データセットを構築した．また，構築した評価用データセットを用いて，現状の典型的な NMT の手法を評価し，現状の翻訳器による出力の問題点を考察した．今後の課題としては，現状の NMT の翻訳品質を保ったまま，訳語の一貫性を実現する方法論を考案する．

参考文献

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. Curran Associates, Inc., 2014.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, Vol. abs/1409.0473, , September 2014.

[3] 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. Association for Computational Linguistics, 2015.

[4] 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 (ACL 2016)*, pp. 1715–1725, 2016.

[5] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500. Association for Computational Linguistics, 2018.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ￥L ukaszKaiser, Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017.

[7] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1304–1313, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[8] 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, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[9] Graham Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[10] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2204–2208, Portoro, Slovenia, may 2016. European Language Resources Association (ELRA).

[11] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association, 2018.

[12] 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, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[13] Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1535–1546. Association for Computational Linguistics, 2017.

[14] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pp. 223–231, 2006.