

# 統計的翻訳とニューラル翻訳による翻訳候補の文の分散表現に基づくリスコアリングの検討

佐橋 広也<sup>1,a)</sup> 西村 友樹<sup>1</sup> 秋葉 友良<sup>1,b)</sup> 中川 聖一<sup>1,†1,c)</sup>

**概要:** 近年、ニューラル機械翻訳 (NMT) が目覚ましい発展を遂げており、従来の統計的翻訳機械翻訳 (SMT) の性能を上回っている。しかし NMT は SMT に比べ、学習に必要なパラレルコーパスの量が十分でなければ、翻訳性能を向上させることが難しく、翻訳の語彙サイズについても制限を持つ。本稿では同じパラレルコーパスで学習した NMT と SMT の翻訳文を比較し、人為的、または文の分散表現ベクトルを利用して自動的にリスコアリングする手法を報告する。両言語のベクトル空間を同一の意味空間に原言語のベクトル表現と目的言語のベクトル表現を写像する手法を提案し、逆翻訳による手法も行った。ベクトル表現を利用した自動リスコアリングでは、ベースラインを下回ったが、逆翻訳によるリスコアリングではベースラインを上回る結果となった。SMT と NMT の翻訳結果のリスコアリングは、我々 [11] と同時に NICT のグループ [12] が発表しており、共に有効性が示されている。我々の文献 [11] はロイター記事の翻訳タスクで評価したが、経済的な数値表現が多く現れて特殊だったので、本稿では、論文の抄録の翻訳タスク ASPECT でも評価したので報告する。

**キーワード:** 機械翻訳、深層学習、分散表現、リスコアリング

## Consideration of rescoring based on distributed representation and back translation of sentences of translation candidates by Statistical Machine Translation Neural Machine Translation

KOYA SAHASHI<sup>1,a)</sup> TOMOKI NISHIMURA<sup>1</sup> TOMOYOSHI AKIBA<sup>1,b)</sup> SEIICHI NAKAGAWA<sup>1,†1,c)</sup>

### 1. はじめに

近年、ニューラル機械翻訳 (NMT) が目覚ましい発展を遂げており、従来の統計的翻訳機械翻訳 (SMT) の性能を上回っている。しかし NMT は SMT に比べ、学習に必要なパラレルコーパスの量が十分でなければ、翻訳性能を向

上させることが難しく、翻訳の語彙サイズについても制限を持つ。そのため機械翻訳の候補をリスコアリングし、性能を向上させる研究が複数報告されている。SMT ではラティスデコーディングを使用した翻訳候補のリスコアリングが行われている [4]。NMT では翻訳候補を別の NMT で再度スコアを評価しリスコアリングする方法が報告されている [10]。SMT と NMT では同じ文でも翻訳結果が異なることが多く、両手法を相補的に使用する方法が考えられる。そこで、SMT と NMT 両方を併用した翻訳システムとして、SMT のシステムやフレーズテーブルを組み込んだ NMT の報告がされている [8][7]。

<sup>1</sup> 豊橋技術科学大学  
Toyohashi University of Technology

<sup>†1</sup> 現在, 中部大学  
Presently with Chubu University

<sup>a)</sup> sahashi@nlp.cs.tut.ac.jp

<sup>b)</sup> akiba@cs.tut.ac.jp

<sup>c)</sup> nakagawa@cs.tut.ac.jp

本稿では同じパラレルコーパスで学習した NMT と SMT の翻訳文を比較し、人為的、または文の分散表現ベクトルを利用して自動的にリスコアリングする手法を報告する。

## 2. 翻訳システム

### 2.1 SMT

SMT の翻訳モデルは原言語の単語列から目的言語の単語列へ翻訳される確率を計算するモデルである (図 1)。翻訳確率は 2 つの言語間の単語またはフレーズ単位で計算される。フレーズ単位の翻訳確率を言語間の翻訳確率とするために、学習コーパスからフレーズテーブルを学習する。原言語文  $F$  に対応する目的言語文  $E$  の単語アライメントを  $a$  としたとき、計算式は以下のように表すことができる。

$$\begin{aligned} \hat{E} &= \arg \max_E P(E|F) \\ &= \arg \max_E \frac{P(F|E)P(E)}{P(F)} \end{aligned} \quad (1)$$

$$P(F|E) = \sum_a P(F, a|E) \quad (2)$$

ここで、 $a$  は単語のアライメントを示し、 $P(E)$  は目的言語の言語モデル、 $P(F|E)$  は翻訳モデルと呼ばれる。翻訳モデルにおいて、各原言語の単語に対応する目的言語の単語は 1 つであるのに対し、各目的言語の単語に対応する原言語の単語は 0 から  $n$  個であると仮定する。また、原言語の単語に適切な目的言語の単語が対応しない場合、目的言語の先頭に  $e_0$  という空の単語があると仮定し、原言語の単語と対応させる。

以前、我々は、目的言語の言語モデルを種々用意して複数の翻訳文候補をリスコアリングする手法を試みた [5]。

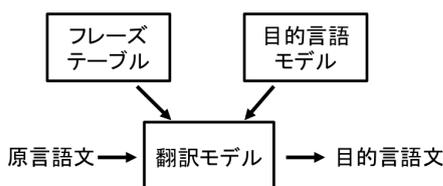


図 1 SMT のブロック図

### 2.2 NMT

NMT の主流であるエンコーダ-デコーダモデルについて説明する。原言語  $F$  の入力文を単語レベルの埋め込みベクトルに変換してエンコーダへ入力する。エンコーダから出力される分散表現は入力文の意味や構造を捉えた文ベクトルとなる。文ベクトルをデコーダに入力した場合、最初の目的言語の単語  $e_1$  を出力確率によって予測する。次の単語を予測するために、出力された単語を入力として与え、終端記号が予測されるまで単語の予測を繰り返し、最終的に目的言語文  $E$  を出力する (図 2)。単語の予測の際にそれ

ぞれどの原単語に対して注目するかを与えるために、エンコーダから出力される単語ベクトルに重みをかけるアテンション機構によって制御する [2]。 $\theta$  をモデルのパラメータとしたとき、デコーダの計算式は、以下のように表すことができる。

$$P(E|F; \theta) = \prod_{j=1}^J P(e_j | F, E < j; \theta) \quad (3)$$

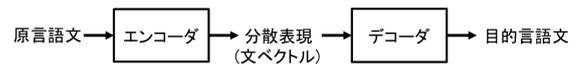


図 2 NMT のブロック図

## 3. 単語と文の分散表現と同一空間への写像

入力文のベクトルと翻訳文のベクトルの類似度によって翻訳候補をリスコアリングする際、両言語のベクトル空間が同一の意味空間になっている必要がある。そこで原言語のベクトル表現と目的言語のベクトル表現を同一の意味空間に写像する手法を提案する。

### 3.1 文レベルの分散表現

エンコーダ-デコーダモデルでは原言語文をベクトルに変換し、ベクトルを入力として目的言語の単語を予測していくモデルである。エンコーダから出力される文ベクトルは原言語の入力文の意味や構造を表現した実数値の連続であると考えられる [6]。そのため 2 つの文の文ベクトルが類似していた場合、2 つの文は意味や構造だけでなく、表層的な単語列の生成も類似していると考えられる。リスコアリングの評価指標として、翻訳候補文と入力文の 2 つの文ベクトルの類似度は有用であると考えられる。

### 3.2 英日分散表現のマッピング

自動的にリスコアリングする場合、原言語 (本稿では英語) の文ベクトルと、翻訳候補である目的言語 (本稿では日本語) の文ベクトルを比較する必要がある。それぞれの文ベクトルはそれぞれの単語ベクトルから生成されたベクトルであり、通常は両ベクトルの意味空間が対応されていないため比較することができない。そのため本研究では、原言語ベクトルを同一意味空間の目的言語ベクトルに変換するために線形変換と非線形変換 (ニューラルネットワーク) を使用する (図 3)。原言語の文ベクトルを入力とし、目的言語の文ベクトルを出力とするために文ベクトルの両言語ペアを学習に使用する。

線形変換では原言語  $X$  とターゲット目的言語  $Z$  のベクトルの空間を文ペアを使い、意味空間で同じマッピングになるよう行列  $W$  を学習する [1]。文献 [1] では単語レベルの線形マッピングを試みている。非線形変換では 3 層の隠れ層を持つニューラルネットワークで実現した。

$$W^* = \arg \min_W \sum_i \sum_j D_{ij} \|X_{i*} W - Z_{j*}\|^2 \quad (4)$$

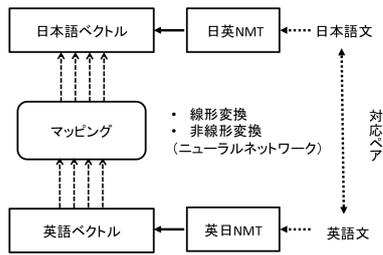


図 3 文ベクトル間のマッピングの構成

### 3.3 リスコアリング

NMT と SMT では翻訳システムの違いから、同じパラレルコーパスを使用した場合でも異なる翻訳文が生成される。そのため本研究では複数の翻訳候補から、最良の翻訳結果を得るために英日翻訳のタスクについてリスコアリングを行う。まず、翻訳精度の指標である BLEU 又は BLEU+1 を文ごとに導出し、NMT と SMT の翻訳文を比較し、それぞれの違いとリスコアリングの有用性について検討する。次に、NMT のエンコーダから出力される文ベクトルを利用してリスコアリングする手法を提案する(図 4)。自動的にリスコアリングする場合は英日方向の英語文ベクトルと日本語の翻訳結果候補の日英方向の日本語文ベクトルを同一空間に写像して比較する。

原言語文を使用したリスコアリングの方法の 1 つとして逆翻訳によるリスコアリングを行う。NMT と SMT の翻訳文を日英方向の翻訳モデルを使用し、原言語方向へ翻訳する。その後原言語文と逆翻訳文の BLEU 又は BLEU+1 を文ごとに導出し、NMT と SMT の翻訳文をリスコアリングする。

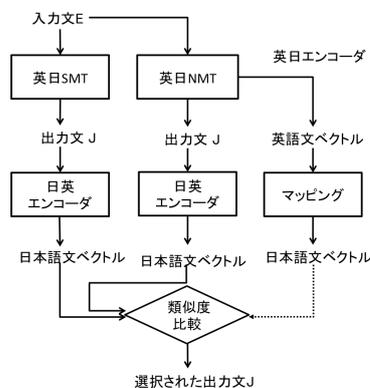


図 4 自動的なリスコアリングの構成

## 4. 評価実験

### 4.1 実験条件

SMT と NMT による英日翻訳と翻訳候補のリスコアリングの評価実験を行った。学習コーパスとテスト文には新聞記事のライター英語-日本語ペアを用意し、パラレルコーパスと評価データ、テストデータに使用する。学習コーパスとして 50,000 文、開発データとして 2,000 文、テストデータとして 2,000 文を使用する。<sup>\*1</sup>

SMT の翻訳モデルの作成ツールには Moses を使い、パラメータの調整には Moses に搭載されている MERT ツールを用いる。学習の語彙サイズは日本語が 20404 語、英語は 32240 語である。SMT は最良の翻訳結果と予測される 1 ベスト翻訳結果以外に、複数の翻訳候補を 1000 ベストまで出力する。

NMT の語彙サイズは両言語 10,000 語に制限し、エンコーダは双方向 LSTM(500+500 次元)、デコーダは翻訳性能の向上のため、アテンション機構を持った LSTM(500 次元)で構成される。学習のエポック数は 50 で、実験には 50 エポックのモデルを使用する。リスコアリングに必要なベクトルのマッピングには、日英-英日の翻訳モデルを上記のパラメータで作成し、学習コーパスから日本語文と英語文のペアをそれぞれのシステムに入力してエンコーダから出力される文ベクトル (500 次元) のペア 50,000 個をペアとし、マッピングの学習に使用する。英日翻訳モデルは逆翻訳のモデルとしても利用される。非線形写像用のネットワークは入力層にドロップアウトを持つ隠れ層 3 層 (512 次元)、活性化関数には Relu 関数を使用する。学習の損失関数は平均二乗誤差を用い、100 エポック学習し、開発データで評価されたモデルを実験に使用する。

リスコアリングの文ベクトルの比較にはコサイン類似度を採用し、入力文 (英語) とコサイン類似度が高い翻訳文 (日本語) を最良の翻訳結果として選択する。人為的にリスコアリングする場合には SMT と NMT の 2 つの翻訳候補で、1 文ごとに BLEU の高い方の文を選択する場合 (オラクル)、正解日本語文の日本語文ベクトルと日本語翻訳結果の 2 つの日本語文の文ベクトルとの高い方を選択する場合 (日本語ベクトルによるオラクル) の 2 つの評価で実験を行う。BLEU の導出には Moses のスクリプト multi-bleu.perl を使用し、1 文ごとの BLEU についても同様に導出する。分単位の BLEU では 4Gram 一致が文中にない場合、計算することができない。その為改良された指標方法である BLEU+1 を導出した場合でも比較する。他のコーパスでも同様な結果が得られるかを確認するため

\*1 講義音声の音声翻訳のタスクとして [9]、話し言葉のドメインとなる講演 TED の書き起こしによる評価実験も予定していたが、NMT の翻訳結果が悪く (BLEU で 4~6 の値) リスコアリングに適した文を生成できないと考えたため、話し言葉では評価していない。

に、ASPEC コーパスを使用した英日翻訳のリスコアリング実験を行う。NMT、SMT それぞれのコーパスとして翻訳品質でソートされた上位 100 万文を使用し、テスト文も用意された 1812 文を使用する。モデルのパラメータはロイターと同様だが、コーパスのサイズから SMT は MOSE チューニングを行っておらず、NMT のは 9 エポック目のモデルを使用していることに注意する。

## 4.2 ロイターコーパスの翻訳実験結果

### 4.2.1 ベースラインの翻訳実験結果

ロイター文で学習した SMT または NMT の翻訳結果の BLEU を表 1 に示す。SMT の 1 ベストの翻訳結果の BLEU20.09 と NMT の翻訳結果の BLEU 値 21.97 をベースとする。SMT と NMT の文の BLEU の差異について、テストデータ 2,000 文に対する NMT と SMT の BLEU の分布を図 5 に、1 文ごとの BLEU の差分 (NMT-SMT) を図 6 に示す。図 5 で 0.0 とは BLEU が 0.0 になった文、10 とは BLEU が 0 から 10.0 未満となった文を意味する (0.0 は除く)。図 5 から SMT と NMT の BLEU の分布に大きな変化は見られなかった。また図 6 から 2,000 文中 1,000 文程度が BLEU において 10 以上の差が見られた。図 7 に SMT と NMT による翻訳例を示す。これらから、SMT の翻訳文と NMT の翻訳文はかなり異なり、互いに相補的になっていることがわかる。つまり、NMT と SMT 両方のシステムで翻訳結果をリスコアリングすることが翻訳結果の改善につながると考えられる。

### 4.2.2 オラクルによるリスコアリング

オラクルにリスコアリングした結果において、SMT の 1 ベストと NMT の 1 ベストを候補にした場合で、SMT と NMT を組み合わせてリスコアリングすることで 1,000 候補の SMT の BLEU を上回る結果を得た (27.05)。更に NMT の 1 ベストと SMT の 1,000 ベストを組み合わせることで、30.28 の BLEU を得た。組み合わせることで BLEU が 0 になる文が減少している結果が得られた。

正解日本語文の日本語ベクトルを利用したリスコアリングにおいても BLEU の向上が見られた。日本語文ベクトルによる選択とオラクルによる選択が一致した割合は 68.2% で、BLEU の差が 30 以上の文では 85.9% であった。このことは日本語文ベクトル空間での比較が可能なこと、及び文ベクトルによるリスコアリングが有用であることを示している。問題は英語文ベクトルと日本語文ベクトルを同一意味空間に写像できるかどうかである。

### 4.2.3 マッピングによる自動リスコアリング

SMT と NMT の翻訳結果をマッピングネットワークによってリスコアリングした場合の翻訳結果の BLEU を表 1 に示す。残念ながら、自動的にリスコアリングした場合、文ベクトル間のマッピングを線形、非線形 (ニューラルネッ

ト) いずれで行ってもベースラインを上回る結果は得られなかった。自動による選択とオラクルによる選択が一致した割合は 65.1% であった。翻訳の候補数が多くなるにつれ BLEU が低下した。これは英語文ベクトルと日本語文ベクトルとのベクトルのマッピングの精度がまだ不十分であるためと考えられる。

### 4.2.4 逆翻訳による自動リスコアリング

SMT と NMT の翻訳結果を逆翻訳によってリスコアリングした場合の翻訳結果の BLEU を表 1 に示す。逆翻訳では SMT の 1 ベストと NMT の 1 ベストを候補にした場合で、ベースラインの NMT の BLEU を上回る結果となった (23.42)。しかし、翻訳候補数の増加に伴い BLEU が上昇するという結果とはならなかった。翻訳候補数が少ない場合に、逆翻訳がリスコアリングに有用であることがわかった。

## 4.3 ASPEC コーパスによる翻訳実験結果

SMT、NMT それぞれ 1 ベストに対してオラクルまたは逆翻訳によってリスコアリングをした場合の翻訳結果の BLEU を表 2 に示す。ベースラインの SMT (25.8) と比較し、オラクルでは 31.04、逆翻訳では 27.48 とロイターと同様に BLEU の上昇が見られた。このため、リスコアリングの有用性はコーパスに依存しないことが考えられる。

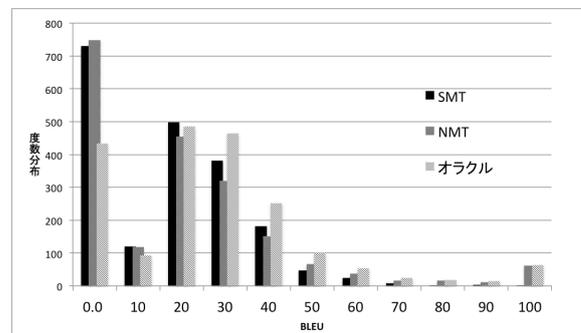


図 5 NMT と SMT のセンテンス BLEU の分布図 (ロイター)

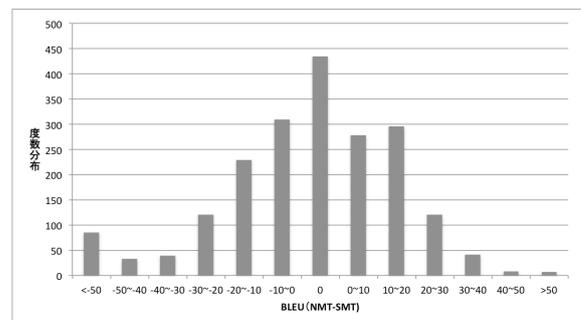


図 6 センテンス BLEU の差分の分布図 (ロイター)

表 1 翻訳のリスコアリング実験結果 (ロイター)

SMT 候補数	0	1	30	100	1000	1	30	100	1000
NMT 候補数	1	0	0	0	0	1	1	1	1
オラクル									
BLEU	21.97	20.09	22.35	23.32	25.20	26.93	28.25	28.86	30.15
BLEU+1	-	-	-	-	-	<b>27.05</b>	28.40	29.01	<b>30.28</b>
COS 類似度	-	-	20.89	21.36	21.90	23.54	24.63	25.10	25.80
自動リスコアリング									
マッピング (ニューラルネット)	-	-	20.06	19.93	19.80	21.99	21.88	21.69	21.29
マッピング (線形変換)	-	-	20.00	19.90	19.57	21.11	20.57	20.03	19.70
逆翻訳 (BLEU)	-	-	20.11	20.04	-	22.34	22.82	22.57	-
逆翻訳 (BLEU+1)	-	-	-	-	-	<b>23.42</b>	<b>23.59</b>	23.25	-

表 2 翻訳のリスコアリング実験結果 (ASPEC)

NMT	SMT	オラクル		逆翻訳	
		(BLEU)	(BLEU+1)	(BLEU)	(BLEU+1)
25.61	25.83	30.78	31.04	26.71	27.48

<p>入力: prime minister ryutaro hashimoto said on monday that japan was not in a situation now to consider a further increase in the consumption tax . SMT: (BLEU 36.2) 橋本首相は、日本が、現在の状況では、消費税引き上げを検討することはない、との見解を示した。 NMT: (BLEU 13.1) 橋本首相は、消費税 (再輸出) を拡大することを今現在のところで、今はないと述べた。</p>
<p>入力: mexico will record a farm trade balance deficit in 1996 of \$ 1.4 billion reversing a surplus in 1995 of \$ 764 million , the national farm council said wednesday in a year end report . SMT: (BLEU 15.1) 過去最高の140億ドルの財政赤字は、96年のメキシコの貿易収支は、95年の100万ドルで、96年の農業委員会の発表によると、年末までには供給過剰 reversing 農業 764 リポートを発表した。 NMT: (BLEU 40.4) メキシコ全国駐在の全国農業評議会 (cna) は、96年のメキシコの貿易収支は、過去最高と13億ドルの黒字となる、との見通しを発表した。</p>

図 7 翻訳結果例 (ロイター)

## 5. おわりに

本研究では英日翻訳において NMT と SMT の複数の翻訳システムを使用し、翻訳結果のリスコアリングを検討した。人為的にリスコアリングした場合において、ベースラインから最大 7.8 の BLEU の上昇が得られることを示した。翻訳候補の文ベクトルとリファレンスの日本語ベクトルを使用したリスコアリングにおいても同様に BLEU の改善を得ることができた。そのため、文ベクトルの意味表現をリスコアリング以外のタスクでも利用できる可能性がある [3]。ただし、日本語と英語の文ベクトルの同一空間へのマッピングによるリスコアリングにおいてはベースラインを下回る結果となった。英語と日本語の同一空間への写像のさらなる工夫が必要である。しかし逆翻訳によるリスコアリングではベースラインを上回る結果となり、逆翻訳の結果がリスコアリングに使用できることを示した。

## 参考文献

- [1] M. Artetxe, G. Labaka, and E. Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proc. ACL2017*, pp. 451–462, 2017.
- [2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [3] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proc. ICML-14*, pp. 1188–1196, 2014.
- [4] E. Matusov, S. Kanthak, and H. Ney. On the integration of speech recognition and statistical machine translation. In *Proc. INTERSPEECH2005*, 2005.
- [5] K. Sahashi, N. Goto, H. Seki, K. Yamamoto, T. Akiba, and S. Nakagawa. Robust lecture speech translation for speech misrecognition and its rescoring effect from multiple candidates. In *Proc. ICAICTA2017*, pp. 1–6, 2017.
- [6] I. Sutskever, O. Vinyals, and Q. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.
- [7] Y. Tang, F. Meng, Z. Lu, and P. LH Li, H. and Yu. Neural machine translation with external phrase memory. *arXiv preprint arXiv:1606.01792*, 2016.
- [8] J. Zhang, Y. Liu, H. Luan, J. Xu, and M. Sun. Prior knowledge integration for neural machine translation using posterior regularization. In *Proc. ACL2017*, pp. 1514–1523, 2017.
- [9] 後藤統興, 山本一公, 中川聖一. 英日講義音声翻訳に対する音声認識誤りを考慮したパラレルコーパスの利用. 情報処理学会研究報告音声言語情報処理 (SLP), Vol. 2016, pp. 1–7, 2016.
- [10] 今村賢治, 隅田英一郎ほか. 双方向ランキングとアンサンブルを併用したニューラル機械翻訳における複数モデルの利用法. 情報処理学会研究報告自然言語処理 (NL), Vol. 2017, No. 9, pp. 1–8, 2017.
- [11] 佐橋広也, 西村友樹, 秋葉友良, 中川聖一. 統計的翻訳とニューラル翻訳による翻訳候補の文の分散表現に基づくリスコアリングの検討. 言語処理学会 第 24 回年次大会 発表論文集, pp. 260–263, 2018.
- [12] Benjamin Marie, 藤田篤. 統計的機械翻訳とニューラル機械翻訳の混合 n ベストランキング. 言語処理学会 第 24 回年次大会 発表論文集, pp. 746–749, 2018.

## 正誤表

下記の箇所に誤りがございました。お詫びして訂正いたします。

訂正箇所	誤	正
1 ページ 題名	統計的翻訳とニューラル翻訳による 翻訳候補の文の分散表現に基づくり スコアリングの検討	統計的翻訳とニューラル翻訳に基づ く翻訳候補文の分散表現と逆翻訳に よるリスコアリングの検討
1 ページ 英題	Consideration of rescoring based on distributed representation and back translation of sentences of translation candidates by Statistical Machine Translation Neural Machine Translation	Consideration of rescoring based on distributed representation and back translation of sentences of translation candidates by Statistical Machine Translation Neural and Machine Translation
1 ページ 概要 12 行目	ASPECT	ASPEC
3 ページ 4.1. 38 行目	分単位、4Gram	文単位、4gram
5 ページ 14 行目		謝辞 本研究は科学研究費(課題番号 25280062)の支援を受けた。