Vol.2016-NL-229 No.20
2016/12/22

単語分散表現を用いた単語アライメントによる日英機械翻訳の自動評価尺度

松尾 潤樹[1,a)]　小町 守[1,b)]　須藤 克仁[2,c)]

概要：
機械翻訳システムの評価においては、システム出力と参照訳の n-gram 一致率に基づく BLEU 等の自動評価尺度がよく用いられている。しかし、BLEU 等の表層の一致で評価する自動評価尺度では、システムが参照訳と同じ意味の異なる表現を出力した際に不当に低い評価がなされる問題点がある。この問題は、機械翻訳の出力をシステム単位ではなく文単位で評価するとき、特に深刻である。また、METEOR などの評価尺度は、類義語を考慮するが、METEOR は品詞が異なるとアライメントを取ることができず、異なる品詞の表現をした際にスコアを与えにくいという問題がある。そこで本稿では、大規模コーパスから学習した単語分散表現を用いた単語アライメントによる意味的文類似度を自動評価尺度に応用する。The 2nd Workshop on Asian Translation と NII Test Collection for IR Systems 8 (NTCIR8) における日本語から英語への翻訳データを用いた実験の結果、全ての単語間のアライメントを考慮する Whole Alignment Similarity が文単位の評価において人手評価との高い相関を得ることを確認した。

1. はじめに

機械翻訳システムの性能を改善するためには、システムを適切に評価することが重要である。機械翻訳の人手評価はコストが高いので、低コストな自動評価尺度が盛んに研究されている [3, 6, 9]。機械翻訳の自動評価のデファクトスタンダードである BLEU [9] は、システムが参照訳と同じ意味の異なる表現を出力した際に不当に低い評価がなされる問題点がある。また、BLEU は主にシステム単位の相関が高い評価尺度として使用がされているが、文単位での BLEU スコアは人手評価の相関がほとんどないことが知られている。特に日英翻訳においては、文法が大きく異なるため、文単位の相関が低いことが問題である。文単位でスコアを計ることによって機械翻訳システムの欠点を直接把握できるため、文単位の評価尺度は重要であると考える。

また、機械翻訳評価と似たタスクとして、意味的文類似度（Semantic Textual Similarity）のタスクがある。意味的文類似度のタスクは、二つの文（適切な文）の類似度を計算するタスクである。意味的文類似度は、単語の言い換えを考慮するため、情報検索などに応用がされているが、まだ機械翻訳の自動評価に用いた研究は見られない。特に、文単位の翻訳の妥当性の評価においては、文同士の関係性を適切に把握することが重要だと考えられ、意味的文類似度の尺度も効果的であることが期待される。意味的文類似度タスクでは、単語の分散表現による単語アライメントを用いた手法が高い性能を示すことが報告されており、機械翻訳の評価にも応用が可能である。

そこで本稿では、単語分散表現を用いた単語アライメントによる意味的文類似度を機械翻訳の評価尺度に応用した。実験には、WAT2015、NTCIR8 の日本語–英語データセットで実験を行った。その結果、我々は文単位の評価において、全ての単語間のアライメントを考慮する Whole Alignment Similarity と人手評価との高い相関を確認した。

2. 関連研究

機械翻訳システムを適切に評価するために、様々な研究がなされている。BLEU は機械翻訳の自動評価尺度におけるデファクトスタンダードである。しかし、Callison-Burch ら [2] はシステム出力と参照訳の n-gram 一致率および文長によるペナルティのみで機械翻訳システムを評価する BLEU は人手評価との相関が低いと示している。そこで、単語の表層の一致だけでなく、同義表現にもスコアを与える METEOR [1] がある。METEOR は表層だけではなく、語幹や同義表現についても考慮する評価尺度である。ま

[1] 首都大学東京 システムデザイン研究科
[2] NTT
[a)] matsuo-junki@ed.tmu.ac.jp
[b)] komachi@tmu.ac.jp
[c)] sudoh.katsuhito@lab.ntt.co.jp

た、言語資源を自動獲得することで言語を問わず使えることを目指した、METEOR-Universal [3] がある。しかしながら、METEOR-Universal で用いられている同義表現は WordNet のような人手で作成する言語資源や stemmer のように対象の言語の知識が必要となるモジュールによって処理され、外部リソースを必要とする。言語依存の処理が不要な知識として PPDB [4] を用いる手法があるが、このデータベースの構築にはパラレルコーパスが必要である。本研究でも、同義表現にもスコアを与えるために、単語分散表現を用いるが、単語分散表現は単言語コーパスだけで学習できる、という利点がある。

機械翻訳の評価は、同義表現を考慮できる仕組みがあることが重要である。例えば、METEOR は表層の一致によらない評価を行うことで、BLEU より高い相関を示した。一方、文同士の同義表現を考慮するタスクとして、Semantic Textual Similarity (STS) のタスクがある。STS タスクは文間の意味的な類似度を計算するタスクで、単語の言い換えを考慮することができる。機械翻訳システムの評価も、流暢性の評価を除けば、機械翻訳のシステムの出力と参照訳を比較することでスコアリングを行う点が共通している。

そこで、本研究では、機械翻訳システムの出力した単語を妥当に評価するために、Song ら [10] の STS タスクにおける意味的文類似度を機械翻訳評価尺度に応用した。Song らの意味的文類似度は STS タスクで高い性能を示しているため、我々は機械翻訳の自動評価尺度にも意味的文類似度が適用可能であると考えた。その結果、Song らが提案している手法の一つが、文単位の相関において、既存の機械翻訳の評価尺度と比較して高い相関を確認した。

3. 意味的文類似度を応用した評価尺度

我々は、意味的文類似度を機械翻訳評価の尺度として応用したため、この節では意味的文類似度を測るための手法について説明する。 3.1 節では One-hot 表現、 3.2 節では文の分散表現に基づく意味的文類似度、 3.3 節では Song ら [10] が提案した単語分散表現のアライメントに基づく意味的文類似度について、それぞれ説明する。

3.1 One-hot 表現に基づく意味的文類似度

単語をベクトルで表現する方法のひとつとして、古くから One-hot 表現が使われている。One-hot 表現は、各単語ベクトルが文書の語彙サイズの次元を持ち、該当する単語を示す次元のみが 1、他の次元は全てが 0 である疎なベクトル表現である。本研究では文のベクトルを文中の単語ベクトルの平均で表現する。そして、システム出力の文ベクトルと参照訳の文ベクトルのコサイン類似度を用いて意味的文類似度を計算する。

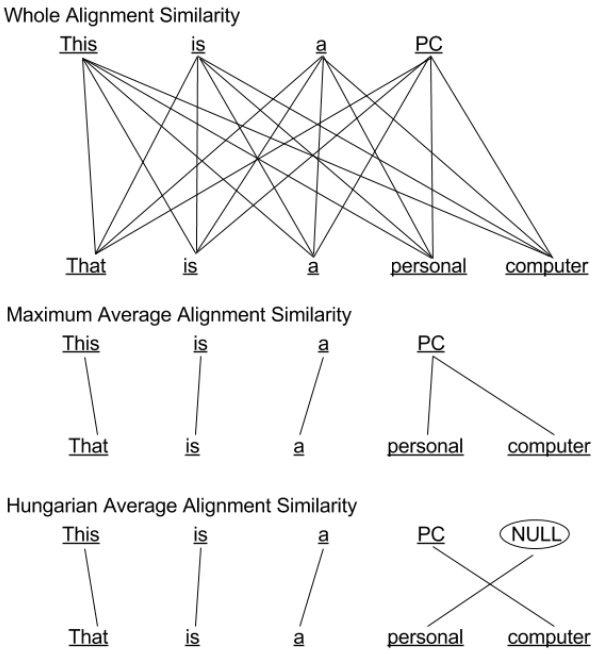

図 1: 単語分散表現のアライメントに基づく意味的文類似度のアライメント

3.2 文の分散表現に基づく意味的文類似度

One-hot 表現では表記が違うだけで異なるベクトルになってしまうため、データスパースネスの問題がある。そこで、表記が異なるような単語も意味的文類似度の計算の際に考慮するために、単語分散表現を用いる。具体的には、分散表現を用いた単語ベクトルの平均で文ベクトルを構成し、文ベクトル間のコサイン類似度によって意味的文類似度を定義する。あらかじめ学習された単語ベクトルを用いた文 x と文 y の文ベクトル SV を以下の式で定義する。

$$\mathrm{SV}(a) = \frac{1}{|a|}\sum_{i=1}^{|a|} a_i \tag{1}$$

a は任意の x、y いずれかの文であり a_i は単語ベクトルである。SV を文ベクトルとし、両文のコサイン類似度で意味的文類似度を定義する。

$$\mathrm{STS}_{\mathrm{SV}}(x,y) = \frac{\mathrm{SV(x)} \cdot \mathrm{SV(y)}}{|\mathrm{SV(x)}||\mathrm{SV(y)}|} \tag{2}$$

3.3 単語分散表現のアライメントに基づく意味的文類似度

本稿では、STS タスクにおいて高い性能を示した Song らの手法を再実装したものを用い、機械翻訳評価尺度に適用する。機械翻訳の評価においては、原言語の単語は目的言語の単語やフレーズに対応するが、単語が文全体に対応することは稀である。しかし、3.1 節と 3.2 節の手法は文中のどの単語がどの単語に対応するかということを全く考慮せず、全ての単語を同じ重みで文全体との類似度を測って

しまう。そこで、単語のアライメントを用い、意味的類似度の計算に関係ない部分の類似度を考慮しないことで、ノイズを減らすことができると考える。

3.3 節で説明される意味的文類似度を簡易に図 1 に示す。実線はアライメントが取れている単語のコサイン類似度である。 3.3 節の手法は、単語アライメントを取る際にノイズを含む可能性が考えられるため、閾値以下の類似度となる単語アライメントを棄却し、人手評価との相関の変化を確認した。閾値を変化させることによって、アラインされる単語が増減する。図 1 は閾値を 1.00 にした時のアライメントの例である。単語の類似度（実線）の値を平均することによって 3.3 のスコアは求められる。

3.3.1 Whole Alignment Similarity

まず我々は、複数のアライメントを取る Whole Alignment Similarity (WAS) を提案する。後述するように、単語間の類似度が一定以上のペアのみを意味的類似度の計算に用いることで、無関係な単語同士の比較を避けることを狙っている。

システム出力 x と参照訳 y の間のすべての単語の組み合わせに対して単語の類似度を計算し、それらの $|x||y|$ 個の単語間類似度を平均して意味的文類似度 $\mathrm{WAS}(x, y)$ を求める。

$$\mathrm{WAS}(x, y) = \frac{1}{|x||y|} \sum_{i=1}^{|x|} \sum_{j=1}^{|y|} \phi(x_i, y_j) \qquad (3)$$

x_i と y_j は、各文に含まれる単語を示す。$\phi(x_i, y_j)$ は単語間類似度を表し、本研究ではコサイン類似度を用いる。

3.3.2 Maximum Alignment Similarity

3.3.1 節の計算は、すべての単語に対してコサイン類似度を取るため、単語間類似度の閾値の設定によってはノイズを含む。閾値以上の単語類似度を持つ単語ペアが複数回カウントされることで、実際は対応している単語は 1 ペアしかなくても、見かけ上多くの単語が対応しているかのように計算されてしまう、という問題がある。

そのため、すべての単語の類似度ではなく、単語間類似度が最大となる値のみの平均をとる Maximum Alignment Similarity (MAS) を考える。また、Maximum Alignment Similarity については、機械翻訳システムの出力 x から参照訳 y の最大値を取る場合と参照訳 y から機械翻訳システムの出力 x の最大値を取る場合と二つのパターンが考えられるため、その二つのパターンの値を平均した値で、$\mathrm{MAS}(x, y)$ を計算する。a と b は任意の出力 x または参照訳 y である。

$$\mathrm{MAS}_{\mathrm{asym}}(a, b) = \frac{1}{|a|} \sum_{i=1}^{|a|} \max_j \phi(a_i, b_j) \qquad (4)$$

$$\mathrm{MAS}(x, y) = \frac{1}{2}(\mathrm{MAS}_{\mathrm{asym}}(x, y) + \mathrm{MAS}_{\mathrm{asym}}(y, x)) \qquad (5)$$

3.3.3 Hungarian Alignment Similarity

3.3.1 節と 3.3.2 節では、Whole Alignment Similarity および Maximum Alignment Similarity を紹介したが、これらの手法は各単語のアライン先に制約を課していないため、同じ単語が複数回対応付けされてしまう可能性がある。

そこで、本節では、Hungarian Alignment Similarity (HAS) の計算方法を定義する。具体的には、システム出力 x および参照訳 y の 2 文を単語をノードとする 2 部グラフとして考える。この 2 部グラフは、単語間類似度 $\phi(x_i, y_j)$ を重みとする辺を持つ完全 2 部グラフである。この完全 2 部グラフの最大マッチングを求めることにより、単語間類似度の総和を最大化する一対一の単語アライメントを得ることができる。2 部グラフの最大マッチング問題は、Hungarian 法 [7] によって解くことができる。システム出力 x に含まれる各単語 x_i に対して、Hungarian 法によって参照訳 y に含まれる単語 $h(x_i)$ を選択し、それらの $\min(|x|, |y|)$ 個の単語の組み合わせについて計算した単語間類似度を平均して、意味的文類似度 $\mathrm{HAS}(x, y)$ を求める。

$$\mathrm{HAS}(x, y) = \frac{1}{\min(|x|, |y|)} \sum_{i=1}^{|x|} \phi(x_i, h(x_i)) \qquad (6)$$

4. 実験

この節では、実験設定と実験結果について報告する。4.1 節では実験設定について述べ、 4.2 節では、 3 節で記述した手法の相関係数をそれぞれ比較する。

4.1 実験設定

実験データは、WAT2015 と NTCIR8 の機械翻訳システムの出力と参照訳を用いた。WAT2015 と NTCIR8 のデータの、特に日本語–英語において、それぞれ 600 文（3 システム出力 ×200 文）、1,200 文（12 システム出力 ×100 文）に対して評価を行った。

機械翻訳システムを評価するために、機械翻訳の出力の妥当性を五段階で評価したスコアと意味的文類似度のスコアの相関を測る。文単位の評価には、ケンドールの順位相関係数を用い、システム単位の評価にはピアソンの積率相関係数を用いた。また、システム単位のスコアは、文のスコアを平均した値を用いた。

3.3 節で用いる単語分散表現は、Google News コーパス [8]（単語数は 30 億）を用いて word2vec によって学習された公開モデルである（`https://code.google.com/archive/p/word2vec/`）。

本実験で使用した意味的文類似度を比較するために、BLEU、METEOR、NIST、RIBES を使用した。また、BLEU、METEOR、NIST は Asiya ツールキット [5] の実装を用いた。RIBES は、NTT コミュニケーション科学基礎研究所のパッケージ の version 1.03.1 を

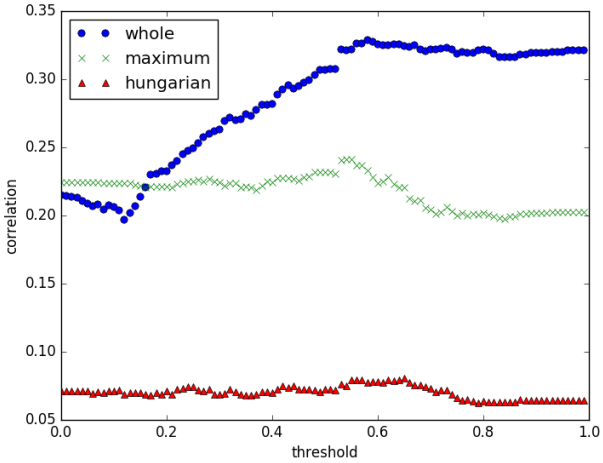

図 2: 単語アライメントに用いる単語類似度の閾値を変化させた時の各意味的文類似度と人手評価との相関（WAT2015）

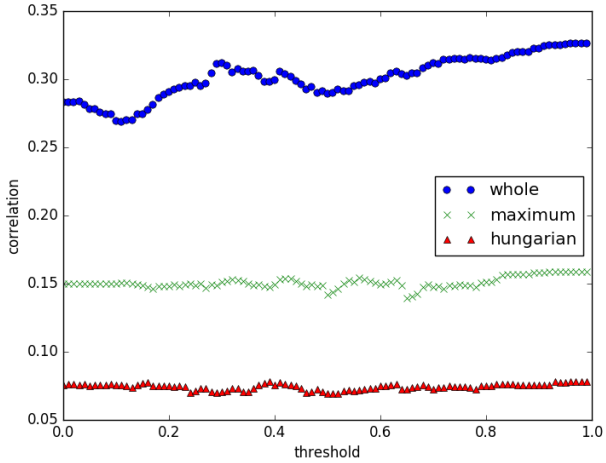

図 3: 単語アライメントに用いる単語類似度の閾値を変化させた時の各意味的文類似度と人手評価との相関（NTCIR8）

用いた（`http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html`）。ただし、それらのパラメーターは標準で実験を行った。

4.2 実験結果

図 2、図 3 は単語の類似度の閾値を変化させた時の人手評価との文単位の相関である。表中の whole は Whole Alignment Similarity、maximum は Maximum Alignment Similarity、hungarian は Hungarian Alignment Similarity と人手評価との相関を示している。WAT2015 のデータでは、一部の閾値を除いて、安定して Whole Alignment Similarity が比較的に強い相関を示した。

表 1 と表 2 は各意味的文類似度もしくは自動評価尺度と人手評価との文単位での相関の値を示している。表 1 と表 2 の結果は文単位で最適化した結果を示している。ベースラインとなる既存の機械翻訳評価尺度と比較しても、文単位の評価においては、Whole Alignment Similarity が比較的高い相関を示すことを確認した。

表 3 と表 4 は各意味的文類似度もしくは自動評価尺度と人手評価とのシステム単位での相関の値を示している。表 3 と表 4 の結果はシステム単位で最適化した結果を示している。システム単位では、比較に行った既存の手法の方が人手評価との相関が高く、文単位の評価ほど意味的文類似度は高い相関を確認することができなかった。

表 5 と表 6 は、システム単位の相関が最大となる閾値にチューニングしたときの各意味的文類似度と人手評価の文単位の相関の値を示している。

意味的文類似度と評価尺度	相関
One-hot 表現に基づく意味的文類似度	0.211
文の分散表現に基づく意味的文類似度	0.097
Whole Alignment Similarity	**0.332**
Maximum Alignment Similarity	0.235
Hungarian Alignment Similarity	0.092
BLEU	0.220
METEOR	0.248
NIST	0.204
RIBES	0.261

表 1: 各意味的文類似度もしくは評価尺度スコアと人手評価との文単位の相関の値（WAT2015）

意味的文類似度と評価尺度	相関
One-hot 表現に基づく意味的文類似度	0.180
文の分散表現に基づく意味的文類似度	0.022
Whole Alignment Similarity	0.343
Maximum Alignment Similarity	0.171
Hungarian Alignment Similarity	0.075
BLEU	0.225
METEOR	0.211
NIST	0.150
RIBES	**0.368**

表 2: 各意味的文類似度もしくは評価尺度スコアと人手評価との文単位の相関の値（NTCIR8）

5. 考察

図 2、図 3 は単語の類似度の閾値を変化させた時の人手との相関であるが、Whole Alignment Similarity (whole) は、閾値を変更することによって相関が左右されていることがわかる。WAT2015 のデータセットでは、閾値を変化させることで相関が大きく変化しているが、NTCIR8 のデータセットでは WAT2015 ほどの違いは見られない。また、NTCIR8 のデータセットでは閾値が 1.00 の時に Whole Alignment Similarity 手法の相関が最大となることを我々

意味的文類似度と評価尺度	相関
One-hot 表現に基づく意味的文類似度	0.983
文の分散表現に基づく意味的文類似度	0.979
Whole Alignment Similarity	0.814
Maximum Alignment Similarity	0.995
Hungarian Alignment Similarity	0.656
BLEU	0.531
METEOR	0.973
NIST	**0.998**
RIBES	0.721

表 3: 各意味的文類似度もしくは評価尺度スコアと人手評価とのシステム単位の相関の値（WAT2015）

意味的文類似度と評価尺度	相関
One-hot 表現に基づく意味的文類似度	0.543
文の分散表現に基づく意味的文類似度	0.645
Whole Alignment Similarity	0.515
Maximum Alignment Similarity	0.657
Hungarian Alignment Similarity	0.097
BLEU	0.075
METEOR	0.482
NIST	0.159
RIBES	**0.861**

表 4: 各意味的文類似度もしくは評価尺度スコアと人手評価とのシステム単位の相関の値（NTCIR8）

意味的文類似度と評価尺度	相関
Whole Alignment Similarity（閾値：0.30）	0.263
Maximum Alignment Similarity （閾値：0.33）	0.223
Hungarian Alignment Similarity（閾値：0.35）	0.068

表 5: システム単位の相関が最大となる閾値にチューニングしたときの各意味的文類似度と人手評価の文単位の相関の値（WAT2015）

意味的文類似度と評価尺度	相関
Whole Alignment Similarity（閾値：0.68）	0.304
Maximum Alignment Similarity （閾値：0.32）	0.152
Hungarian Alignment Similarity（閾値：0.38）	0.073

表 6: システム単位の相関が最大となる閾値にチューニングしたときの各意味的文類似度と人手評価の文単位の相関の値（NTCIR8）

は確認した。これは単語の完全一致のみを許すことを意味し、表層の一致のみで評価する方が相関があることを示している。RIBES などの評価尺度はアラインされない単語は考慮しない評価尺度であるが、RIBES が全ての評価尺度の中で最大の相関となっているため、NTCIR8 のデータでは類義語などを考慮しないほうが高い相関になることが予測される。

また、表 1 と表 2 は意味的文類似度または評価尺度でソートしたリストと人手評価の値でソートしたリストの文単位の相関を測ったものである。Whole Alignment Similarity は BLEU と METEOR を大きく上回っている。また、WAT2015 のデータにおいては、RIBES も凌いでいる。RIBES は WAT2015 の相関係数と NTCIR8 の相関係数の差が大きいが、一方で、Whole Alignment Similarity は安定した値を示している。これは、Whole Alignment Similarity がどのドメインにも安定して動作することを示している。

表 3 と表 4 はシステム単位の相関である。WAT2015、NTCIR8 のデータセットにおいては、Hungarian Alignment Similarity を除き、BLEU、RIBES、METEOR、NIST と比較して意味的文類似度は安定した相関を示している。特に、文の分散表現に基づく意味的文類似度と Maximum Alignment Similarity は、文単位の相関では METEOR や RIBES に及ばないものの、システム単位では METEOR を上回る相関を確認した。しかし、Whole Alignment Similarity は、文単位の相関が比較的に高いにも関わらず、システム単位では先行研究と比較して高い相関を確認することができなかった。これは、意味的文類似度のシステム単位スコアを計算する際に、文長を考慮せず単に文のスコアを平均したことが要因であると考えられる。Whole Alignment Similarity は文長が長くなると、スコアが低くなる特性があるため、BLEU や RIBES のように文長にペナルティをかけるによって、BLEU や RIBES のようにコーパス単位でシステム相関を最適化できる可能性がある。

表 5 と表 6 は、システム単位の相関が最大となる閾値を用いて、文単位の相関を測った結果を示している。文単位に最適化をしていないため、相関は表 1 と表 2 と比べて、値は低くなっている。しかし、システム単位で閾値を決定しても、WAT2015 のドメインでは、どの先行研究よりも高い相関であることが確認できる。

また、意味的文類似度を用いることによって高精度に評価できている例を表 7 で示す。表 7 の例 1 から例 3 までは WAT2015 のデータであり、例 4 は NTCIR8 のデータである。人手評価とそれぞれのスコアは [0,1] に値をとるように正規化している。

例 1 はどの評価尺度も比較的に高いスコアを出している。

例 2 は意味的文類似度が他の評価尺度と比較して、有効性が示されている例である。例 2 では否定表現の言い換えがうまく表現できている翻訳であるにもかかわらず、BLEU と METEOR は比較的に低いスコアリングをしている。これは表層の一致だけでなく、単語アライメントが有効であることを示している。

例 3 は意味的文類似度でもうまく評価ができていない例である。例 3 を考えてみると “that” 節以降の態が異なるため、低い人手評価がなされている。しかし、意味的文類似度はうまくスコアリングがなされていない。これは、意

番号	実例
1	出力: In both cases, the postoperative course was good. 参照: In both the cases, postoperative course were good.
2	出力: In the treatment, side effect was not recognized. 参照: No side effect was noted during treatment.
3	出力: It is found that the deformation is affected by the pair density distribution. 参照: It was found that the deformation gave effects to the pairing density distribution.
4	出力: The above configuration and operation, the water drops, even in the case where the window is not expected operation, it is possible to surely stopped, and the operation switch (dn) is operated and the window is opened, it is possible to a vehicle occupant moves out from possible. 参照: With the above-described construction and operation, even when an automobile falls into water, the windows are surely stopped without performing unexpected operations, and can be surely opened by operating the operation switch (dn), thus enabling occupants to escape from the automobile.

番号	RIBES	METEOR	One-hot	WAS	MAS	人手評価
1	0.83	0.92	0.90	0.89	0.94	1.00
2	0.23	0.37	0.55	0.73	0.71	1.00
3	0.30	0.37	0.64	0.69	0.78	0.20
4	0.75	0.25	0.56	0.11	0.61	0.80

表 7: RIBES、METEOR と意味的文類似度のスコアと人手評価（妥当性）の比較

味的文類似度が単語の依存関係を考慮せず、態の異なりによるペナルティが一切含まれていないことによるものであると考えられる。

例 4 は Whole Alignment Similarity が評価しにくい例である。例 4 は単に単語数が多いために Whole Alignment Similarity は低いスコアを与えている。Whole Alignment Similarity は事実上、文が長くなればなるほどスコアが低くなる評価尺度であるため、単語数が極端に多い出力と参照のペアの評価には向いていないことが懸念される。一方、RIBES や One-hot 表現に基づく意味的文類似度などの表層の一致のみで評価する評価尺度は、比較的低いスコアを与えていない。図 3 から読み取れる通り、NTCIR8 は表層の一致を評価する方が、人手と高い相関を得ることができる。これは、NTCIR8 が持つ特有の性質であり、NTCIR8 のデータセットにおいては表層の一致を評価することがより適切である可能性がある。

6. おわりに

本稿では単語の分散表現を用いた単語アライメントによる意味的文類似度を適用した機械翻訳評価尺度を提案した。本研究で採用した意味的文類似度は、簡易なアルゴリズムで実装ができ、日英翻訳の文単位のスコアリングにおいて有効であることが示された。実験では、表層を評価する先行研究に加えて、単語のアライメントを考慮することが人手評価との相関を上げるために重要な要素であることを示した。Whole Alignment Similarity は態の異なりは考慮できないが、構文木のアライメントによる手法を加えることによって、構文の違いを考慮できると考える。システム単位では、文長にペナルティをかけることによってより高い相関を得ることができると我々は考える。また、最近では既存の教師なしの自動評価尺度を教師あり学習で直接的に最適化する研究がされているため、教師あり学習の手法に意味的文類似度を加えることによって、より高い相関を得ることができる可能性がある。

参考文献

[1] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

[2] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 249–256, 2006.

[3] Michael Denkowski and Alon Lavie. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380, 2014.

[4] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The Paraphrase Database. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 758–764, 2013.

[5] Jesus Gimenez and Lluis Marquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. In *The Prague Bulletin of Mathematical Linguistics*, pp. 77–86, 2010.

[6] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic Evaluation of Translation Quality for Distant Language Pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 994–952, 2010.

[7] Harold W. Kuhn. The Hungarian Method for the assignment problem. In *Naval Research Logistics Quarterly*, pp. 83–97, 1955.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*, 2013.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics*, pp. 311–318, 2002.

[10] Yangqui Song and Dan Roth. Unsupervised Sparse Vector Densification for Short Text Similarity. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL*, pp. 1275–1280, 2015.