# 機械翻訳を利用した異言語文間の意味的類似度計算の評価

羅 文涛 (大阪大学言語文化研究科)，林 良彦 (早稲田大学基幹理工学研究科)

u172761i@ecs.osaka-u.ac.jp，yshk.hayashi@aoni.waseda.jp

## 1 はじめに

意味的類似度 (semantic similarity) とは，言語表現の間の意味的な類似の程度を表す指標である．文間の意味的類似度は対称的 (すなわち，$sim(S_1, S_2) = sim(S_2, S_1)$) であり，類似の程度に応じた実数値を取る．従来は，単語や概念の間，または，文書間の意味的類似度について主に検討されてきたが，近年では，Semantic Textual Similarity (STS) と呼ばれるタスク [1, 2, 3] が設定され，文間の意味的類似度 $sim(S_1, S_2)$ の計算が主要なテーマとして取り上げられている．

これまでの STS タスクにおいては，英語に閉じた単言語の意味的類似度がタスクの対象 [1] であり，機械学習により各種の言語特徴量を統合する手法が議論されてきた．それに対し，羅ら [9, 10] は STS タスクの設定を多言語間のタスク (Cross-lingual STS:CL STS) に展開し，既存の言語横断手段と組み合わせることにより，単言語タスクにおいて検討されてきた手法が CL タスクにおいても適用可能であることを示した．

CL STS においては，対象となる異なる言語の文を同一の言語 (空間) へ写像し，そこで類似度を測ることが行われる．このときの類似度を測る言語を基底言語 (pivot language) と呼ぶ．対象文ペアの一方，または，両方を基底言語へ変換するためには，いわゆる言語横断を行う必要がある．本報告では，機械翻訳を利用した異言語文間の意味的類似度の計算手法に関し，機械翻訳を利用すること自体，翻訳の利用の仕方，翻訳品質がもたらす影響について，評価実験の結果をもとに検討する．

## 2 CL STS タスクの設定

本研究の対象言語は，英語 (en)，日本語 (ja)，中国語 (zh) であり，利用する実験データは，STS タスクにおいて公開されているデータ [2] の一部を人手で日本語，英語に翻訳したものである．対象データからの例文を表 1 に示す．類似度 $sim(S_1, S_2)$ は，STS タスクと同じく，平均して 5 人の評定者による 0 から 5 までの評定値の平均値である．ここで，英語における文間の意味的類似度は，翻訳された日本語，中国語における文間に引き継がれると仮定している．

意味的類似度は対称的であるので，対象言語の組み合わせは 3 つある (英中，英日，日中)．また，基底言語としては，対象言語である英語 (en)，日本語 (ja)，中国語 (zh) の場合を比較する．1 つの CL タスクは言語ペアと基底言語により規定されるが，本報告ではこれらを en-zh:ja (英中タスクで基底言語は日本語) のように書く．後述の評価実験においては，比較のためそれぞれの言語での単言語タスクも行う．これらは，en:en (英語の単言語タスク) のように書く．

## 3 文間の意味的類似度の計算

本報告の類似度計算手法の概要を図 1 に示す．まず既存の機械翻訳エンジン [3] を用いて対象文ペアの双方の文を基底言語に揃えた後に，基底言語における単言語の意味的類似度計算を適用する．意味的類似度計算においては，様々な言語特徴量 (素性) に基づく機械学習ベースの手法を主に用いる．

### 3.1 意味的類似度計算における素性

すでに [9] で報告した#1 から#6 の言語特徴量に加え，[10] で提示したアライメントスコア (#7) を素性として用いる．

1. 単語集合の重なりに基づく言語特徴量
2. 単語 N グラムの重なりに基づく言語特徴量
3. 単語意味ベクトルに基づく言語特徴量
4. 言語特徴量に重み付けしたもの
5. 固有表現の重なりに基づく言語特徴量

[1] SemEval-2016 Task1 で初めて英語とスペイン語の言語横断タスクが設定された．http://alt.qcri.org/semeval2016/task1/

[2] http://www.cs.york.ac.uk/semeval-2012/task6/

[3] 言語グリッド http://langrid.org/jp が提供する翻訳サービスを利用した．

表 1: 対象データにおける文ペア・類似度の例

| $S_1$ | $S_2$ | $sim(S_1, S_2)$ |
|---|---|---|
| A man with a hard hat is dancing.<br>一人のヘルメットをした男がダンスしている.<br>一个頭戴帽子的男人正在跳舞. | A man wearing a hard hat is dancing.<br>一人のヘルメットを被った男がダンスしている.<br>一个戴着帽子的男人正在跳舞. | 5.00 |
| A woman is playing the guitar.<br>一人の女がギターを弾いている.<br>一个頭戴帽子的男人正在跳舞. | A man is playing guitar.<br>一人の男がギターを弾いている.<br>一个戴着帽子的男人正在跳舞. | 2.40 |
| A woman is slicing big pepper.<br>一人の女が大きな胡椒を薄切りにしている.<br>一个女人在切大辣椒. | A dog is moving its mouth.<br>一匹の大きな犬がその口を動かしている.<br>一只狗張着它的嘴. | 0.00 |

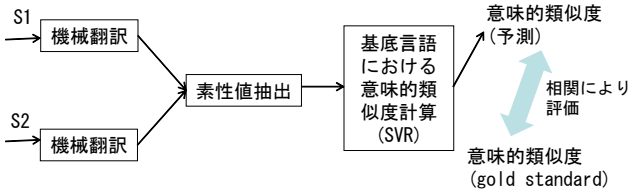


図 1: 意味的類似度計算の枠組み

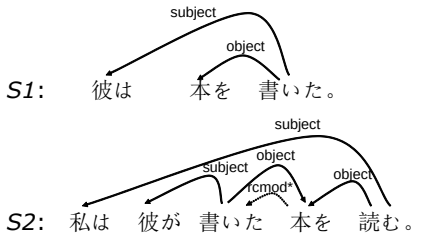


図 2: 日本語の等価係り受け関係

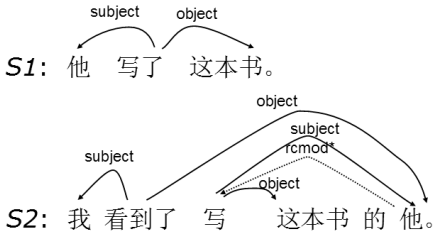


図 3: 中国語の等価係り受け関係

6. WordNet に基づく言語特徴量
7. アライメントに基づく言語特徴量

なお，[9, 10] では，#3 の単語の意味ベクトルとして LSA 手法によるものを用いていたが，本報告では Word2Vec[5] による単語の意味ベクトルを使用した．英語の意味ベクトルは Google の配布する Word2Vec モデル (Continuous Bag-Of-Words に基づく 300 次元のベクトル)[4] を用いたが，日本語，中国語に関しては，Wikipedia のダンプデータから計算した．データ量はそれぞれ約 2GB で，Juman/KNP (日本語)，Stanford CoreNLP (中国語) を用いて前処理を行った．

## 3.2 アライメント手法

本報告におけるアラインメント (文ペアにおける単語の対応付け) 処理は，Sultan らの手法 [8] に基づく．この手法は，完全に一致する単語系列，固有名詞などを先に対応付けた後に，依存構造解析の結果を文脈」として利用し，候補単語の類似度を測り，精度よくアライメントを行う．ただし，Sultan らのオリジナルの手法とは異なり，候補単語の類似度の計算において，各言語の Word2Vec モデルに基づく類似度を用いた．

## 3.3 日本語，中国語対応のアライメント

Sultan らの手法ではアライメントを処理する際には英語の依存構造解析結果から等価受け係り関係を抽出し，対応付けの精度を高めている．本研究では，基底言語として日本語，中国語を考慮するので，これらの言語においても同等の処理を行う必要がある．

日本語のアライメントにおいては，日本語文の構文・格・照応解析を行うシステム KNP[5] によって解析した結果を係り受け関係の補完・縮約を行い，その結果から等価受け係り関係を抽出した．例えば，図 2 においては，object(書く, 本) と rcmod(本, 書く) があるが，これらは，object(verb,noun) と rcmod(noun,verb) を等価な係り受け関係とみることで同一視できる．

中国語では，Stanford CoreNLP[6] を用いて同様の処理を行った．等価係り受け関係の例を図 3[7] に示す．subject(写:書く, 他:彼) と rcmod(他:彼, 写:書く) があるが，subject(verb,noun) と rcmod(noun,verb) の等価性により，これらは同一のものと扱われる．

[4] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing

[5] http://nlp.ist.i.kyoto-u.ac.jp/?KNP

[6] http://stanfordnlp.github.io/CoreNLP/

[7] S1 の日本語訳は：彼はあの本を書いた．S2 の日本語訳は：私があの本を書いた彼を見た．

表 2: 単言語タスクの評価実験の結果

| MSRvid | | | MSRpar | | |
|---|---|---|---|---|---|
| en:en | ja:ja | zh:zh | en:en | ja:ja | zh:zh |
| **0.8712** | 0.8672 | 0.8390 | **0.7303** | 0.7206 | 0.7244 |

# 4 評価実験

## 4.1 実験の設定

**実験データ:** 本研究では，STS タスクの対象データから MSRPar と MSRvid の 2 つのデータセットの一部 (それぞれ 1500 の英語文ペアを含む) を利用する．MSRpar はパラフレーズコーパスからとった長い文ペア群である．一方，MSRvid はビデオ注釈からとった短い文ペア群である．これらの英語データ (の半分．すなわち，各 750 文) を人手により，日本語，中国語に翻訳した．翻訳は完全であると仮定し，英語の文間の gold standard (GS) 類似度は各タスクにおける文間の類似度に引き継がれると仮定する (すなわち，$SimGS(Se_1, Se_2) = SimGS(Se_1, Sj_2)$ など)．

**翻訳品質の評価指標:** 言語横断に用いる機械翻訳の品質が類似度計算に与える影響を評価するための評価指標として RIBES[4] を用いた．これは，RIBES が英語と日本語のように語順が大きく異る言語ペアに対する配慮が行われた評価指標であることによる．

**機械学習:** Python による機械学習のライブラリである scikit-learn[8] が提供するサポートベクトル回帰 (SVR) を利用した．各タスクに対して，グリッドサーチによりパラメータのチューニングを行い，5 分割の交差検定を行った．STS タスクに従い，gold standard として与えられる類似度の系列と類似度計算による出力結果 (予測値) 系列との間の Pearson 相関係数を用いて評価する．

## 4.2 実験結果

表 2 に単言語タスクの結果を，表 3，表 4，表 5 に CL タスクの結果を示す．

各表の第一行めの値は，対象データ (MSRvid, MSRpar) に対して各基底言語 (en,ja,zh) を設定した場合の Pearson 相関係数であり，各表のタスクにおいて最も良い結果を太字で示す．CL タスクに対しては，各表の第二行目に RIBES の値を記入した．*を付した基底言語の場合では，対象文ペアにおける双方 (両側) で翻訳を行った．この場合は，'/' の左側に悪い方の片側翻訳の RIBES 値を示し，' /' の右側には両側の翻訳の RIBES 値の積を表示している．

*SEM のタスクで報告されている MSRvid と MSRpar データに対する最良の結果は，0.8803，0.7343 であり，今回の実験における英語単言語タスクの結果 (0.8712，0.7303) はこれに近いものであった．

総じて，(1) 単言語のタスクよりも機械翻訳を要する CL タスクの結果は悪く，(2) より複雑な長文から構成されている MSRpar の結果は MSRvid の結果より劣る．これは，すでに [9] で報告した傾向と符合している．(3) 単言語のタスクでは英語タスクの場合が他の言語の場合よりもやや良いが，CL タスクにおいては必ずしもそうではない．

## 4.3 機械翻訳の利用に関する評価

**機械翻訳の利用の妥当性:** CL タスクの結果と単言語タスクの結果を比較すると，概ね 9 割程度の精度が得られている．これは，情報検索における言語横断検索の単言語検索に比した精度低下 (80 から 100%)[7] の範囲内であり，機械翻訳を利用する手法として妥当な結果であると言える．

**翻訳自体の影響:** 今回の実験においては，基底言語の設定と翻訳の有無によって，言語横断をしない，片側の文を翻訳，両側の文を翻訳という三つの種類のタスクタイプが存在する．これらのタスクタイプに伴う明らかな傾向として，機械翻訳の処理が重なるたびに精度が下がることが確認できる．例えば，表 2，表 4 によると，MSRpar の場合，英語単言語タスクの結果 (0.7303) が最も良く，en-zh:en の結果 (0.7087) がそれに次ぐ．en-zh:ja の結果 (0.5733) が最も悪いが，これは機械翻訳を両側の文で行うためである．

## 4.4 翻訳品質の影響に関する評価

言語横断に用いる機械翻訳の品質が良ければ，類似度計算の精度も良いことが期待される．ここでは，その傾向を確認する．

例えば，表 3 の MSPpar データにおける en-ja:en と en-ja:ja を比較すると，RIBES の値の差は大きく (0.5492，0.6971)，有意な相関係数の差 (0.6323, 0.7090, p=0.0) がある．一方，MSRvid データにおいては，RIBES の値の差は小さく (0.7762，0.7787)，

[8] http://scikit-learn.org/stable/

表 3: 英日タスクの結果

| | MSRvid | | | MSRpar | | |
|---|---|---|---|---|---|---|
| | en-ja:en | en-ja:ja | en-ja:zh* | en-ja:en | en-ja:ja | en-ja:zh* |
| 相関係数 | 0.8142 | **0.8259** | 0.7108 | 0.6363 | **0.7090** | 0.5949 |
| RIBES | 0.7762 | 0.7787 | 0.7414/0.5579 | 0.5492 | 0.6971 | 0.6368/0.4600 |

表 4: 英中タスクの結果

| | MSRvid | | | MSRpar | | |
|---|---|---|---|---|---|---|
| | en-zh:en | en-ja:ja* | en-zh:zh | en-zh:en | en-zh:ja* | en-zh:zh |
| 相関係数 | **0.8425** | 0.7035 | 0.8295 | 0.7087 | 0.5733 | **0.7219** |
| RIBES | 0.7845 | 0.7114/0.5540 | 0.7525 | 0.6514 | 0.6438/0.4488 | 0.7223 |

表 5: 日中タスクの結果

| | MSRvid | | | MSRpar | | |
|---|---|---|---|---|---|---|
| | ja-zh:en* | ja-zh:ja | zh-ja:zh | ja-zh:en* | ja-zh:ja | ja-zh:zh |
| 相関係数 | 0.7295 | 0.7789 | **0.7982** | 0.6143 | **0.6537** | 0.6456 |
| RIBES | 0.7762/0.6089 | 0.7114 | 0.7414 | 0.5492/0.3577 | 0.6438 | 0.6368 |

相関係数に有意差 (0.8142, 0.8259, p=0.164) はない．より詳しく RIBES 値の差と相関係数の有意差の関係を調べると，表 3 では MSRpar の RIBES の差が約 0.15 で有意差があり (p=0.0)，逆に表 4 では MSRpar の RIBES の差が約 0.07 で有意差がない (p=0.230)．

同様の傾向は，基底言語を固定した場合にも観察される．例えば，MSvid データにおける en-zh:zh と ja-zh:zh の相関係数の有意差は低い (0.8295, 0.7982, p=0.006) のに対し，MSpar データにおいては，相関係数に有意差がある (0.7219, 0.6456, p=0.0)．

## 5 おわりに

機械翻訳を利用した異言語文間の意味的類似度の計算において，機械翻訳の影響について実験的に評価を行った．その結果，機械翻訳の利用における影響は想定される範囲内であること，機械翻訳の品質が類似度計算の精度に強く影響することが示された．このため，可能ならば機械翻訳の適用がなるべく少なくなるように基底言語を選定することが重要である．

今後の課題としては，より良い言語特徴量を開発すること，より良い回帰手段を適用して類似度予測精度の向上を図ることなどが挙げられる．一方で，陽に文の翻訳を行わない手法についても検討を進めたい．例えば，いわゆる単語の分散表現を言語横断的に変換する手法 [6] では，小規模のバイリンガルデータを用いて，原言語における既存の分散表現を目的言語の分散表現へ変換している．あるいは，パラレルコーパスが利用可能であれば，LSA などの手法により，同一の空間において両言語の表現を求めることも可能である．

## 謝辞

本研究は JSPS 科研費#25280117 の助成を受けた．

## 参考文献

[1] Agirre, E., et al. 2012. SemEval-2012 Task 6: A Pilot on semantic textual similarity. *Proc. of STS 2012*, pp.385–393.

[2] Agirre, E., et al. 2013. STS 2013 shared task: Semantic textual similarity. *Proc. of STS 2013*, pp.32–43.

[3] Agirre, E., et al. 2014. SemEval-2014 Task 10: Multilingual semantic textual similarity. *Proc. of SemEval 2014*, pp.81–91.

[4] Isozaki, H., et al. 2010. Automatic evaluation of translation quality for distant language pairs. *Proc. of EMNLP 2010*, pp.944–952.

[5] Mikolov, T., et al. 2013. Distributed representations of words and phrases and their compositionality. *Proc. of NIPS 2013*.

[6] Mikolov, T., et al. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

[7] Nie, J. 2010. *Cross-Language Information Retrieval*. Morgan and Claypool, pp.105.

[8] Sultan, M.A., et al. 2014. Back to basics for monolingual alignment: Exploring word similarity and contextual evidence. *Trans. of the ACL*, Vol.2, pp.219–230.

[9] 羅 文涛, 林 良彦. 2014. 機械学習に基づく異言語文間の意味的類似度の計算. 電子情報通信学会言語理解とコミュニケーション研究会, vol. 114, no. 81, NLC2014-16, pp. 85–90.

[10] 羅 文涛, 林 良彦. 2015. 異言語文間の意味的類似度計算におけるアライメントの利用. 言語処理学会 第 *21* 回年次大会, pp. 63–66.