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

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] 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．

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主な対象であり，機械学習により各種の言語特徴量を統合する手法が提案されてきた．それに対し，羅ら [7] 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，単言語の STS タスクにおいて検討されてきた手法が CL STS タスクにおいても適用可能であることを示した．

一方，最近では，対象の文ペアに対するアライメント情報の利用が STS タスクにおいて有効であることが示された [3]．例えば Sultan[6] らは，アライメントの状況を表すスコアのみを用いて，機械学習に基づいて各種の類似度を統合する従来手法に匹敵する精度が得られることを報告している．

そこで本報告では，異言語文間の意味的類似度の計算におけるアライメント情報の利用法を検討し，その有効性について報告する．

2 タスクの設定と提案手法

2.1 CL STS タスクの設定

本研究の目的は，異言語文間の意味的類似度の計算手法を確立することである．当面の対象言語は，英語 (E)，日本語 (J)，中国語 (C) とする．これまでの STS タスクにおける研究との比較を行うため，本研究の対象データは，STS タスクにおいて公開されているデータを日本語，英語に翻訳したものを用いる．対象データの例文を表 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となっている．ここで，英語における文間の意味的類似度は，日本語，中国語における翻訳文間に引き継がれると仮定している．

前述のように，意味的類似度は双方向的であるので，対象言語の組み合わせから 3 つのタスク (STS-EJ, STS-EC, STS-JC) が存在する．従来の英語に対する STS タスクの研究結果との比較を行うため，以上の 3 つの CL タスクに英語を対象とする単言語タスク STS-EE を加える．

2.2 提案手法

本研究の提案手法の概要を図 1 に示す．本研究においては，意味的類似度の計算の対象となる文ペアの言語は異なるので，まず言語横断を行い，双方の文を特定の言語 (基底言語と呼ぶ) に揃えた後に，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，従来研究と同様に，様々な言語特徴量を用いる．本報告においては，特にアライメント情報を言語特徴量として用いることの有効性を検討する．また，機械学習による各種特徴量の統合を行わずにアライメント情報のみを用いた方法との比較も行う．

3 意味的類似度計算におけるアライメントの利用

3.1 アライメント

本研究におけるアラインメント処理は Sultan らの手法 [5] に基づく．この手法は，完全に一致する系列，固有名詞などを先に対応付けた後に，構文解析 (依存構造解析) の結果を利用し，候補単語の類似度を測り，精度よくアライメントを行う．候補単語の類似度の計算においては，英語の言い換えデータである Paraphrase

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

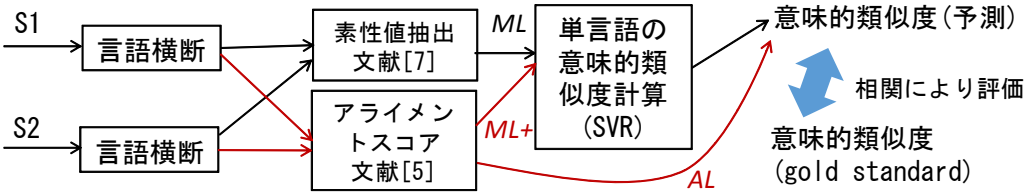

図 1: 提案方式の概要

Database (PPDB) [1]が用いられている．すなわち，候補の単語ペアがこのデータベース中に存在すれば類似度が 1 となるが，存在しなければ類似度は 0 となってしまう．そこで本研究では，PPDB により直接的に与えられる類似度の代わりに，NYT Corpus[2] に対して潜在的意味分析 (Latent Semantic Analysis:LSA) を適用して得られる意味ベクトルを利用し，そのコサイン類似度を測ることにより意味的類似度とした．

3.2 アライメントスコア

アライメントスコアとは，文ペアに対するアライメントの良さを示す指標であり，文間の意味的類似度計算においては，一つの言語特徴量として利用しうる重要な情報である．

アライメントスコアは以下により計算される [5]．ここで，$prop1$ と $prop2$ は，$S1$, $S2$ の中で相手の文に対して対応付けが得られた単語の割合である．定義より，このアライメントスコアは [0,1] の範囲の値を取る．

$$socre(S1, S2) = \frac{2 \times prop_1 \times prop_2}{prop_1 + prop_2}$$

3.3 意味的類似度計算におけるアライメントスコアの利用

意味的類似度計算におけるアライメントスコアの利用に関して，以下の 2 つの手法を比較する．

一つは機械学習に基づく手法である．すでに [7] で報告した手法 (ML 手法と呼ぶ) で用いる以下の言語特徴量 (素性) に加え，上述のアライメントスコアを素性として用いる (ML+手法と呼ぶ).

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. LSA による単語意味ベクトルに基づく言語特徴量
4. LSA に関する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
6. WordNet に基づく言語特徴量

もう一つは，アライメントスコアを線形変換することにより意味的類似度の値域に変換し，直接に意味的類似度とする手法 (AL 手法と呼ぶ) である．この手法は機械学習に基づかないので，事前学習が不要であるという特長がある．

[1] http://paraphrase.org/#/
[2] https://catalog.ldc.upenn.edu/LDC2008T19

4 評価実験と結果評価

4.1 実験データ

本研究では，STS タスクの対象データ[3]から MSRPar と MSRvid の 2 つのデータセット (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．すでに述べたように，STS タスクは英語の単言語タスクであるため，これらのデータ (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)．

4.2 機械学習と評価指標

Python による機械学習のライブラリである scikit-learn[4]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．

評価指標は，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 相関係数を用いる．

4.3 実験結果

表 2 に主要な結果をまとめて示す．本表の各値は，タスクごとに，対象データ (MSRvid，MSRpar) に対して，各手法 (ML，ML+，AL) を適用した場合の Pearson 相関係数であり，太字で表記したものは，タスク・対象データの組み合わせに対して，もっとも良い結果を表す．なお，AL 手法に関しては，単語間の類似度として PPDB ではなく，LSA により導出した意味ベクトルを用いた結果を示している．

総じて，(1) 単言語のタスク (STS-EE) よりも言語横断を要する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 [7] で報告した傾向と符合している．

[3] http://www.cs.york.ac.uk/semeval-2012/task6/
[4] http://scikit-learn.org/stable/

4.4 考察:アライメント情報の有用性

表 2 の結果から以下のことが言える．

- 言語横断を要するタスクにおいては，アライメント情報を他の言語特徴量と併用する ML+手法が有用であることが確認できた．STS-EC タスクに関しては AL 手法の結果が上回っているが，僅差である．
- 一方，単言語のタスク (STS-EE) においては，アライメント情報のみを用いる AL 手法が機械学習を利用した手法 (ML 手法，ML ＋手法) の結果を上回る．

以下，これらの点をさらに考察する．

異言語間タスクによるアライメント情報の利用: 異言語間のタスクでは，アライメント情報だけを用いるより，他の言語特徴量との併用が有用であった．これは良いニュースではあるが，逆に言えば，アライメント情報だけでは不足であることを意味する．異言語間のタスクにおいて言語横断に用いられる機械翻訳の精度は，多くの場合，文意が取れる程度には向上しているが，言語解析が適切に行えるような文でない場合も多くはなく，今回利用したアライメント手法のように，構文解析を要する場合に問題が生じていると推測される．この傾向は，より複雑な長文から構成される MSRpar に対する結果が MSRvid に対する結果に比べて顕著に劣ることからも推測できる．また，今回は英語を基底言語に設定したため，STS-JC のタスクにおいては双方の言語において英語への言語横断が必要となった．これにより，STS-JC に対する結果は他の言語横断タスク (STS-EJ，STS-EC) に比べて低下した．その低下の度合は，ML+手法，AL 手法においてもほぼ同等であり，アライメント情報は有用ではあるが，言語横断に対するロバスト性を改善するものではないことが確認できる．

単言語タスクによるアライメント情報の利用: 単言語のタスク (STS-EE) においては，アライメント情報は驚くほど有用であり，様々な言語特徴量を機械学習により統合する ML 手法を大きく上回っている[5]．特に比較的短く単純な文で構成される MSRvid においては非常に良好な結果を示しており，このような特性の文に対して構文解析を用いることの有用性が示唆され

[5] [6] においても，アライメントスコアだけを用いて STS タスクにおいて上位に相当する結果が得られていることが報告されている．

表 2: 評価実験の結果 (Pearson 相関係数)

	MSRvid			MSRpar		
Task	ML	ML+	AL	ML	ML+	AL
STS-EJ	0.8236	**0.8445**	0.8402	0.6527	**0.6829**	0.6794
STS-EC	0.8249	0.8463	**0.8469**	0.6892	**0.6995**	0.6981
STS-JC	0.7642	**0.7691**	0.7532	0.6316	**0.6326**	0.5928
STS-EE	0.8419	0.8730	**0.9012**	0.6890	0.7308	**0.7429**

る．アライメント情報のみを用いる AL 法は機械学習によらないため，訓練データを用意したり，パラメータをチューニングするなどの過程も必要なく，使い勝手が良い．この方向性を異言語間のタスクにおいても進めていくためには，直接翻訳できる言語ペアを増やすこと，精度良いアライメントツールを各国語において準備することの 2 点が必要になる．前者については，統計的機械翻訳の手法が進展し多言語化が達成されつつあるので，今後は特に後者の開発を行っていくことが必要である．

4.5 考察:アライメントにおける単語の意味的類似度

STS-EE タスクにおいて，単語の意味的類似度として，LSA により導出した意味ベクトルを用いる手法を PPDB を用いるオリジナルの手法と比較した結果を表 3 に示す．

表 3: AL 方法においてアライメント改造の影響

	MSRvid		MSRpar	
Task	LSA	PPDB	LSA	PPDB
STS-EE	0.9012	0.8736	0.7429	0.7125

表 3 に明らかなように，MSRvid，MSRpar の両方において，LSA 意味ベクトルの有効性が確認できた．前節の考察結果と合わせて考えると，対象となる各国語において，単語の意味的類似度を求めるためのリソースを充実させていくことが重要であり，今後は，word2vec[4] のような最近の手法の適用も検討していきたい．

5 おわりに

異言語文間の意味的類似度の計算におけるアライメント情報の有効性を実験的に検討した．その結果，アライメント情報は極めて有用であるが，言語横断を必要とする異言語間の類似度計算においては，他の言語特徴量と併用することが妥当であることが分かった．

今後は，アライメント処理の多言語対応を進め，アライメント情報を素性として機械学習に取り入れる現行のアプローチの限界を極める．また，文間の意味的類似度を考える際にも重要な要素である単語の意味的類似度に関して，word2vec のような最近の手法を取り入れることを検討する．一方で，言語横断を陽に行わない手法 (並行コーパスに基づく多言語の LSA 空間) による精度向上についても検討したい．

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] Mikolov, T., et al. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Proc. of NIPS 2013*.

[5] 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.

[6] Sultan, M.A., et al. 2014. DLS@CU: Sentence Similarity from Word Alignment. *Proc. of SemEval 2014*, pp.241–246.

[7] 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.