The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

意味グラフに基づくテキスト類似性尺度の提案

Semantic Graph-based Simple Semantic Textual Similarity Metrics

田中貴秋 *1 永田昌明 *1 荒瀬由紀 *2 鬼塚真 *2
Takaaki Tanaka Masaaki Nagata Yuki Arase Makoto Onizuka

*1NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

*2大阪大学 大学院情報科学研究科
Osaka University Graduate School of Information Science and Technology

We present a novel method of measuring semantic textual similarity (STS) based on semantic graph. We employed UDepLambda as semantic representation and incorporated the semantic graph-based scores with word based scores. The method improve the correlations between automatic correlation metrics and human judgements in comparison to the other semantic graph-based methods in the evaluations of an STS task and an MT evaluation task.

1. はじめに

機械翻訳，自動要約，剽窃検出など自然言語処理の各タスクにおいて，テキスト内容の意味的類似性 (Semantic Textual Similarity, STS) を測定する重要性が高まっている．頑健性を重視した評価方法として，表層，辞書知識，分散表現を含む語彙，および統語構造に基づく評価尺度が，評価型ワークショップ SemEval 2017 [Daniel 17] などで数多く提案されている．また，異なる言語間の類似性尺度については，SemEval でも実施されており，日本語を含んだ言語対についても研究が行われている [羅 16]．これらの方法では，類似性のあるテキストを比較したときに，類似性，および差異のある部分の根拠を示すことが困難であった．

本稿では，意味グラフをアライメントすることを中心として，テキスト間の意味的類似性を測定する手法について提案する．提案手法では，Universal Dependencies (UD) [Nivre 15] に基づく意味表現 UDepLambda [Reddy 17] を用いる．UDepLambda は論理式による意味表現であるが，各変数をノードとすることにより Abstract Meaning Representation (AMR) [Banarescu 13] と同様の有向グラフと見做すことができる．提案手法では，グラフ間のアライメントを行って類似性評価を行うため，類似している部分，および情報の過不足のある部分を明確にできる．

意味グラフの類似性比較手法の代表的なものとして，AMR を対象とした Smatch [Cai 13] が提案されている．Smatch は，純粋に「意味グラフの重複度合い」を計測するものであるが，提案手法は，典型的な比較評価値である，単語類似性，単語列類似性，アライメントの適合度合い等を組み合わせることで，より頑健で人間の感覚に近似する指標を目指している．

また，意味グラフに基づく STS の評価方法の 1 種として，画像からの captioning により生成されたキャプションと正解キャプション群との類似性を計測する SPICE と呼ばれる方法が提案されている [Anderson 16]．SPICE では，Scene Graph [Schuster 15] と呼ばれる主に画像上のオブジェクト，属性，オブジェクト間の関係を表現した意味グラフを tuple に変換し，その tuple の一致度合いを F-measure により計測する．提案手法でも，同様の tuple の 一致度合いを一要素として採用しているが，述語と項の関係など，人間が類似性を判断する上で重視する関係について，重みを加えることで，人間によるアノテーションのスコアと近いものになるようにしている．

連絡先: 田中貴秋，NTT コミュニケーション科学基礎研究所, 京都府相楽郡精華町光台 2-4, tanaka.takaaki@lab.ntt.co.jp

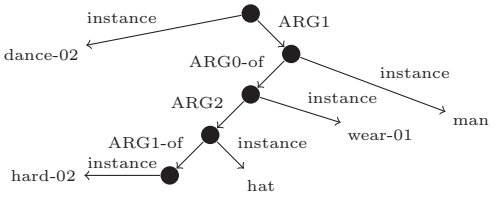

図 1: AMR の例.

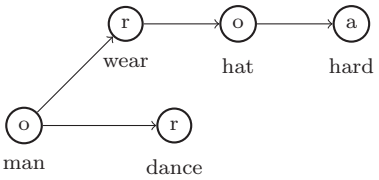

図 2: Scene Graph の例．◯ 内の文字はノードの種類を表している（“o”:object,“a”:attribute，“r”:relation）.

AMR も Scene Graph も基本的に英語を中心として活用されている意味グラフであるが，UDepLambda は ベースとなっている UD とともに多言語処理を念頭において設計されている．本稿では，多言語間の類似性測定への拡張を見据えて，意味表現に UDepLambda を用いた手法を提案し，その有効性を検証する．

STS の測定手法の評価として，2 種類のベンチマークによる実験結果を報告する．一つは，内容的に無関係な文対を含む評価セットに対して，0-5 の絶対評価を行うもので，SemEval の STS benchmark を用いた．もう一つは，翻訳評価手法の妥当性を測るもので，WMT 2015 の Translation Metrics shared task のベンチマークを用いた．提案手法は，STS においては上位システムには及ばないが，翻訳評価タスクでは参加システムと同等の結果を示し，どちらのタスクにおいても，意味グラフに基づく SPICE, Smatch に比較して良い結果が得られたこ

とを報告する．

2. 意味グラフと評価尺度

2.1 AMR

単一の root を持ち，有向グラフで構成される意味表現である (図 1)．AMR の持つ情報には，PropBank [Palmer 05] の意味役割，文内の照応関係，固有表現，モダリティ等がある．AMR の比較手法としては，Smatch[Cai 13] が提案されている．Smatch は，AMR から変換した triple を単位として，2 文間の重なりを調べる．ノードは，entity や event からなり，エッジが relation となる．

2.2 Scene Graph

画像の記述を目的とした有向グラフによる意味表現で Stanford coreNLP [*1] で解析可能な形式 [Schuster 15] である (図 2)．Scene Graph では，object(entity)，attribute, relation がノードになり，エッジには関係ラベルを持たない．Captioning の evaluation を目的とした，SPICE (Semantic Propositional Image Caption Evaluation) [Anderson 16] が提案されている．SPICE では，Graph のノードとエッジをばらして，それぞれの組み合わせを抽出して，2 文間の重なりを F-measure により調べる．各 relation (subject-object などの意味関係) や，color, size などの属性ごとに関係を抽出することも述べられているが，具体的な評価尺度には反映されていない．

2.3 UDepLambda

UDepLambda [Reddy 17] は，多言語共通のアノテーションスキームである UD [Nivre 15] に基づいて提案された論理式タイプの意味表現である．図 3 は，類似した 2 文に関する UD および UDepLambda を示している．ラムダ式による意味表現は，UD による構文解析結果を項の間の照応解析を行って変換している．多言語への適用例として，英語，ドイツ語，スペイン語等のクエリをそれぞれ UDepLambda に変換することにより，同一の枠組みでそれぞれの言語における質問応答システムを実現した結果について報告されているが，意味グラフ間の比較方法については言及されていない．本稿では，UDepLambda の論理式を，AMR タイプの意味グラフと見做して，アライメントを行うことにより類似度を測定する．

3. 提案手法

UDepLamba の 意味グラフに基づくスコアと，単語 (列) に基づくスコアを組み合わせてモデルを構築する．以下，それぞれについて詳細を説明する．

3.1 意味グラフに基づくスコア

SPICE や Smatch で行っているように意味グラフを tuple の形に分解し，対象とする文間の tuple の比較を行う．tuple 間の比較は，Smatch 方式のアライメントスコアに基づくものと，SPICE 方式の F-measure に基づくものを組み合わせる．語彙の比較は，WordNet による類似度スコアあるいは単語分散表現による類似度スコアを用いた．

UDepLambda を用いることで，文法構造の違いを吸収して意味的関係をある程度抽象化できる．その結果，例えば，述語と主格の関係が *nsubj* という関係になっている場合と関係節で *acl* になっている場合を，同一の関係 arg_1 として扱うことができる．これは意味的関係を比較する上で，大きな利点である．また，項の binding を行うことにより，コントロールされている項が解消されている点も意味表現を利用する大きな効用である．例えば，*the children want to swin* から，$(relation, val_1, val_2)$ の tuple を抽出することを考えると，$(arg_1, want, children)$, $(arg_1, swin, children)$ の二つの関係が得られる [*2]．

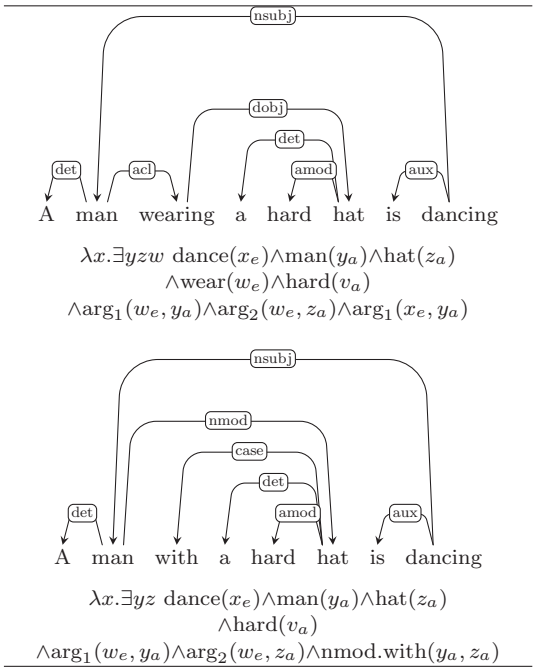

図 3: Universal Dependencies と UDepLambda の例.

ただし，UD, UDepLambda ともに内容語と内容語の間の関係をベースにしているため，述語項構造のように内容語を中心とした関係になっている場合と，前置詞を介した関係になっている場合では，tuple の単位やそれに基づく構造が異なってしまう場合がある．例えば，*a man wearing a hat* と *a man with a hat* における，*man* と *hat* の関係はほぼ同様のものと考えられるが，前者の tuple は，$(arg_1, wear, man)$, $(arg_2, wear, hat)$，後者の tuple は，$(nmod.with, man, hat)$ となりそのまま比較すると類似性を捉えづらい．そこで，同一の内容語を介した複数の tuple に関しては，2 語間の関係を含む tuple に縮約した tuple を新たに追加する．前者の例の場合は，共通の内容語 wear を縮約した tuple $(rel*, man, hat)$ を追加する．$rel*$ は縮約した tuple に付与する関係である．

意味グラフのアライメント Smatch で用いられているように tuple に変換した上でアライメントを行う．ただし，アライメントスコアの算出時に，前述のように内容語ベースの tuple を縮約した tuple を考慮する．

UDepLambda から tuple への変換および，tuple のアライメントは以下のように行う．アライメントは，[Cai 13] で行われているように，初期値から，山登り法によって，アライメントスコアが変化しなくなるまで，探索を行う．局所解に陥る危険性があるが，ここでは計算の簡便性に重きを置いた．

1. 意味グラフのノードから各変数 v を抽出する．比較する 2 文から抽出した変数群をそれぞれ $V_A = \{v_{A1}, ..., v_{Am}\}$, $V_B = \{v_{B1}, ..., v_{Bn}\}$ $(m \leq n)$ とする．
2. 変数 v とインスタンス（entity または event e） を組にした tuple $(inst, v, e)$ を追加する，

*1 https://stanfordnlp.github.io/CoreNLP/

*2 実際の UDepLambda では，変数を介して論理式で得られる．

Genre	Train	Dev	Test	Total	≥ 4	≥ 2	≥ 0
news	3299	500	500	4299	1079	2141	1079
caption	2000	625	525	3250	721	1235	1294
forum	450	375	254	1079	208	432	439
Total	5749	1500	1379	8628	2008	3808	2812

表 1: STS benchmark の統計量 [Daniel 17].

3. 変数 v_i, v_j とその間の関係 r の組を抽出し，tuple $t = (r, v_i, v_j)$ として追加する．比較する 2 文から抽出した tuple 群をそれぞれ T_A, T_B とする
4. head (最初の変数 v_i) が同一の変数を含む複数の tuple $(r_k, v_i, v_j), (r_p, v_i, v_r)$ について，2 変数間の関係を含み縮約した tuple $(rel*, v_j, v_r)$ を新たに追加する．
5. 変数 のアライメントの初期値 $A_0 = (a_1, ..., a_m)$ を定める．ただし，$a_i = j$ は，変数 v_{Ai} と v_{Bj} がアライメントされていることを表している．$a_i = 0$ の場合は，v_{Ai} に対するアライメント先がないことを示す [*3]．
6. アライメントの 1 点を隣接点に変更するか，2 点を入れ替えて作成した A_k を探索候補リストに加える．探索候補リスト中の要素それぞれにアライメントスコア $\sigma_{align}(A_k)$ を求め，最高のスコアを出すもの次の探索候補とする．
7. 前項を，アライメントスコアが向上する候補がなくなるまで繰り返し，探索範囲内でスコアが最高になったアライメントを，アライメント結果とする．

なお，アライメントスコア σ_{align} は，以下のように tuple 間のスコア σ_T の総和で定める．

$$\sigma_{align}(A_k; V_A, V_B, T_A, T_B) = \sum_{\{k,l|a_k=l\}} \sigma_T(t_{Ak}, t_{Bl}) \quad (1)$$

ただし，$t_{Ak} = (r_{Ak}, v_{Ap}, v_{Aq}) \in T_A$，$t_{Bl} = (r_{Bl}, v_{Br}, v_{Bs}) \in T_B$ である．

$$\sigma_T(t_{Ak}, t_{Bl}) = \begin{cases} sim(I(v_{Aq}), I(v_{Bs})) & \text{if } (r_{Ak} = r_{Bl} \text{ or } r_{Ak} = rel* \text{ or } r_{Bl} = rel*) \text{ and } I(v_{Ap}) = I(v_{Br}) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

ここで，$I(\cdot)$ は，変数からインスタンスへの写像を表しており，$I(v_{Ai}) = e$ のとき，$(inst, v_{Ai}, e) \in T_A$ である．なお，インスタンス e_1,e_2 間のマッチングについては，各インスタンスの属する synset c_1,c_2 の WordNet 上のパスと共通の親 synset の深さによるスコア $sim_{wn}(c_1, c_2)$ [Wu 94]，または GloVe による単語ベクトル $\boldsymbol{v_{e1}}$,$\boldsymbol{v_{e2}}$ の cosine 類似度 $sim_{glove}(\boldsymbol{v_{e1}}, \boldsymbol{v_{e2}})$ を用いて行った．

$$sim_{wn}(c_1, c_2) = \frac{2 \cdot \text{depth}(\text{lso}(c_1, c_2))}{\text{len}(c_1, c_2) + 2 \cdot \text{depth}(\text{lso}(c_1, c_2))} \quad (3)$$

ここで，lso(c_1, c_2) は，synset c_1,c_2 の共通の親のうち最も深い synset, len(c_1, c_2) は，c_1 から c_2 へ辿るパスの長さ，depth$(\cdot)$ は，synset の WordNet 上での深さを表す．

意味グラフの重なり SPICE で行うように意味グラフから変換された tuple 集合の重なりのスコア σ_{ov} を F-measure として算出する．以下で，$T_A \oplus T_B$ は，各 tuple の変数をインスタンスに変換した際に一致する tuple の集合を表している．

$$\sigma_{ov} = \frac{2PR}{P+R} \quad \left(P = \frac{|T_A \oplus T_B|}{|T_B|}, R = \frac{|T_A \oplus T_B|}{|T_A|}\right) \quad (4)$$

*3 本稿では，1 対 1 または 1 対 0 のアライメントを仮定している．

System	Description	Test
ECNU⋆	hybrid	.810
BIT⋆	WordNet	.809
Proposed	UDepLambda	.690
SPICE	Scene Graph	.543
Smatch	AMR	.511

表 2: STS benchmark における人手評価値との相関 (Pearson の積率相関係数 r). ⋆ は，[Daniel 17] より抜粋した．

System	Description	de-en	fr-en
DPMFcomb	1st place in de-en	.482	.395
RATATOUILLE	1st place in fr-en	.441	.398
Proposed	UDepLambda	.398	.398
sentBLEU		.360	.358
SPICE	Scene Graph	.211	.211
Smatch	AMR	.303	.285

表 3: WMT2015 の segment-level 評価 (Kendall's τ).

3.2 単語・単語列に基づくスコア

意味グラフに基づくスコアを補完するため，従来手法で用いられている類似性スコア [Šarić 12, 羅 16] から以下のものを用いる．以下，2 文の単語集合をそれぞれ S_A, S_B と表す．

単語の重なり S_A, S_B で共通する単語の割合を算出する (ϕ_1).

$$\phi_1 = sim_{wo}(S_A, S_B) = \frac{2 \cdot |S_A \cap S_B|}{|S_A| + |S_B|} \quad (5)$$

S_A, S_B の包含関係を表すフラグである (ϕ_2).

$$\phi_2 = \begin{matrix} 1 & \text{iff } (S_A \subseteq S_B) \vee (S_A \supseteq S_B) \neq \emptyset \\ 0 & \text{otherwise} \end{matrix} \quad (6)$$

S_A, S_B を WordNet の synset に変換した場合の重なりを算出する [Šarić 12]. 以下の $P_{wn}(S_A, S_B)$, $P_{wn}(S_B, S_A)$ の調和平均の値を用いる (ϕ_3).

$$\phi_3 = P_{wn}(S_A, S_B) = \frac{1}{|S_B|} \sum_{w \in S_A} score(w, S_B) \quad (7)$$

$$score(w, S) = \begin{cases} 1 & \text{if } w \in S \\ \max_{w' \in S} sim_{wnp}(w, w') & \text{otherwise} \end{cases} \quad (8)$$

ただし，$sim_{wnp}(\cdot, \cdot)$ は，WordNet 上の path 類似度 である．

N-gram の重なり N-gram の重なりの度合いとして，unigram, bigram, trigram における共通する n-gram の割合を算出する (ϕ_4,ϕ_5,ϕ_6) [Šarić 12]. 以下において，M_A^N， M_B^N は文 S_A，S_B それぞれにおける N-gram の集合である．S_N，S_B で共通する n-gram に対するそれぞれの n-gram 全体数の割合の調和平均を用いている．

$$sim_{ng}(M_A^N, M_B^N) = 2 \cdot \left(\frac{|M_A^N|}{|M_A^N \cap M_B^N|} + \frac{|M_B^N|}{|M_1^N \cap M_B^N|} \right)^{-1} \quad (9)$$

単語アライメント 統語的な構造を考慮してアライメント可能な単語の割合を算出する (ϕ_7) [Sultan 13].

$$\phi_7 = sim_{align}(S_A, S_B) = \frac{n_c^a(S_A) + n_c^a(S_B)}{n_c(S_A) + n_c(S_B)} \quad (10)$$

3.3 回帰

対象とする 2 文間の類似度スコアが前述した意味グラフに基づくスコア (σ_{align}, σ_{ov}) と単語・単語列に基づくスコア ($\phi_1, ..., \phi_7$) により推定されると仮定する．回帰は，[Liu 17] で用いられているのと同様に SVR を使って行った．RBF カーネルを用い，SVR のパラメータは，開発セットの結果からペナルティ項 $C = 10$, $\gamma = 0.2$, $\epsilon = 0.5$ とした．

4. 実験

提案手法の評価は，意味的関係の極端に低いものを含む対の類似性評価を含む設定と，翻訳システムの翻訳結果のように，類似性が比較的高いものを分別する設定の 2 種類について行った．前者には，STS shared task の評価ベンチマーク，後者には，WMT2015 のデータセットを用いて，自動で算出した類似スコアと人のアノテーションとの相関で評価した．UDepLambda, AMR の解析, Smatch, SPICE の算出は，公開されているツールを使って行った [*4] [*5] [*6]．

4.1 STS benchmark による評価

STS benchmark は SemEval 2017[Daniel 17] で使われたベンチマークと歴代 SemEval の評価セットから作成されたベンチマークである (表 1)．対になった 2 文それぞれの内容の類似性について，0-5 のアノテーションがされている．5 は 2 文の意味内容が完全に一致する，0 は意味内容に関連性がないことを示す．複数のアノテータによる平均値をとっているため，値は整数とは限らない．評価値は，SemEval と同様に，各手法に基づき算出された類似度 ([0,5]) と，人手で付与された類似度の間の Pearson の積率相関係数 r の値を用いる．結果を表 2 に示す．

参加システムの 1,2 位には及ばないが，意味グラフベースの評価尺度の中では，かなり良い結果となっている．ただし，自動解析が難しい文では解析エラーにより情報がかなり欠落しており，簡易な単語に基づくスコアの補完のみでは，困難なものも目立った．

4.2 WMT 2015 評価データに基づく評価

機械翻訳の Evaluation Metrics shared task 用のデータを使用して，評価を行った．評価データは，英語を含む 5 か国語 10 方向のうち．本稿の実験では，独英対訳の 2169 文，仏英対訳の 1500 文を使用した．STS のように正解の評価値データが十分にないため，提案手法では，STS の訓練データで構築した回帰モデルを使って，STS の類似度スコアを使用した．各システムの翻訳結果について参照訳との類似性評価を行い，人手で行った翻訳品質評価との比較を行った．評価方法は，pairwise での優劣を人手評価と自動評価尺度での判定結果を比較し，これらの間の相関を Kendall's τ により算出した．比較する翻訳結果を同点と判断した場合の扱いについては，WMT15 Metrics Shared Task [Stanojevič 15] で用いられた算出方法に準じた．

表 3 に示すように，独英では参加システムのトップに及ばないものの，仏英では同等の結果を残している．これは，提案手法が本データで学習していないことを考えると良好な結果といえる．ベースライン手法である sentBLEU に優位な結果となっていることも本タスクへの適用可能性が高いことを示している．STS と同様，解析エラーによる影響が大きく，解析精度および補完方法の改善が課題である．

5. おわりに

本稿では，意味グラフ UDepLambda に基づいたテキストの意味的類似性尺度について提案した．提案手法は，従来の意味グラフに基づく手法に比較して，人のアノテーションとの相関が高いことが確かめられた．本手法を多言語間の類似性尺度に拡張し，翻訳評価等への適用を検討する予定である．

*4 https://github.com/sivareddyg/UDepLambda
*5 https://github.com/mdtux89/amr-eager,amr-evaluation
*6 https://github.com/peteanderson80/SPICE

参考文献

[Anderson 16] Anderson, P., Fernando, B., Johnson, M., and Gould, S.: SPICE: Semantic Propositional Image Caption Evaluation, in *Proceedings of the 14th European Conference on Computer Vision*, ECCV 2016 (2016)

[Banarescu 13] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N.: Abstract Meaning Representation for Sembanking, in *Proceedings of the Linguistic Annotation Workshop* (2013)

[Cai 13] Cai, S. and Knight, K.: Smatch: an Evaluation Metric for Semantic Feature Structures, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL 2013 (2013)

[Daniel 17] Daniel, C., Diab, M., Agirre, E., Lopez-Gazpio, nigo I., and Specia, L.: SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation, in *Proceedings of the 11th International Workshop on Semantic Evaluations*, SemEval-2017 (2017)

[Liu 17] Liu, W., Sun, C., Lin, L., and Liu, B.: ITNLP-AiKF at SemEval-2017 Task 1: Rich Features Based SVR for Semantic Textual Similarity Computing, in *Proceedings of the 11th International Workshop on Semantic Evaluations*, SemEval-2017, pp. 159–163 (2017)

[Nivre 15] Nivre, J.: Towards a Universal Grammar for Natural Language Processing, in *Proceedings of CICLing 2015*, pp. 3–16 (2015)

[Palmer 05] Palmer, M., Gildea, D., and Kingsbury, P.: The Proposition Bank: An Annotated Corpus of Semantic Roles, *Comput. Linguist.*, Vol. 31, No. 1, pp. 71–106 (2005)

[Reddy 17] Reddy, S., Täckström, O., Petrov, S., and Lapata, M. S. M.: Universal Semantic Parsing, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2017, pp. 89–101 (2017)

[Schuster 15] Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., and Manning, C. D.: Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval, in *Proceedings of the Workshop on Vision and Language (VL15)* (2015)

[Stanojevič 15] Stanojevič, M., Kamran, A., Koehn, P., and Bojar, O.: Results of the WMT15 Metrics Shared Task, in *Proceedings of the 10th on Statistical Machine Translation*, WMT 15 (2015)

[Sultan 13] Sultan, M. A., Bethard, S., and Symner, T.: Back to basics for monolingual aligner: Exploiting word similarity and contextual evidence, *TACL*, Vol. 2, pp. 219–230 (2013)

[Šarić 12] Šarić, F., Glavaš, G., Karan, M., Šnajder, J., and Bašić, B. D.: TakeLab: Systems for Measuring Semantic Text Similarity, in *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, SemEval '12, pp. 441–448 (2012)

[Wu 94] Wu, Z. and Palmer, M. S.: Verb semantics and lexical selection, in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, ACL-1994, pp. 133–138 (1994)

[羅 16] 羅 文涛, 林 良彦 : 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, 言語処理学会第 22 回年次大会予稿集, pp. 883–886 (2016)