

単語の分散表現と仮説推論を用いた文の類似度学習

Integrating Distributional Representation Model with Abduction for Text Similarity Learning

谷中 瞳^{*1}
Hitomi Yanaka

峯島 宏次^{*2*4}
Koji Mineshima

Pascual Martínez-Gómez^{*3}
Pascual Martínez-Gómez

戸次 大介^{*2*4}
Daisuke Bekki

^{*1}東京大学
University of Tokyo

^{*2}お茶の水女子大学
Ochanomizu University

^{*3}産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

^{*4}独立行政法人科学技術振興機構, CREST
CREST, Japan Science and Technology Agency

Semantic representations based on logic have a potential to realize deep semantic analysis of natural language. We propose a method of learning semantic textual similarity with features extracted by proving bidirectional entailment relations between a sentence pair. For extracting logic-based features, we use natural deduction proofs with axioms generated from distributional representations of words. In this study, *ccg2lambda*, an integrated system which converts a sentence into a higher-order predicate logic expression by syntactic analysis and semantic composition based on combinatory category grammar (CCG), is applied to automated inference.

1. はじめに

文の類似度計算は情報検索や文書分類、文書要約といった様々な自然言語処理技術の土台となる技術である。類似度計算に用いるベクトル空間モデルとして、ニューラルネットワークを用いて周辺の単語の出現分布から単語をベクトルで表現する Word2Vec モデル [1] がある。Word2Vec モデルは類似した文脈に出現する単語は意味が類似するという分散表現 [2] の考え方に基づく。

他方で、文ペアに対して一方の文を他方の文が内容的に含意できるかどうかを同定する含意関係認識のタスクでは、文の意味を論理式で表現し、推論によって文の意味を分析する研究 [3][4][5][6] が行われてきた。そこで本研究では、単語の分散表現と論理的推論を組み合わせる文の類似度を学習する方法を検討する。具体的には、文の意味を論理式で表現し、自然演繹を用いて 2 文間の含意関係の証明を双方向で行い、証明の過程と結果から類似度学習に必要な特徴量を導出する。このとき、単語の分散表現を語彙知識として推論に用いることによって、より正確な意味推論を目指す。

2. 先行研究

単語の分散表現と論理的推論を組み合わせる文の類似度を学習する先行研究として、Bjerva ら [5] は分散表現から導出した文中の単語間類似度と、推論による含意関係認識の結果を特徴量に用いて類似度を学習している。しかし、この手法では、単語の分散表現は特徴量設計に利用されているが、推論の中では利用されていない。また、Beltagy ら [6] は Probabilistic Soft Logic による類似度学習方法を提案している。Probabilistic Soft Logic では文を一階述語論理に変換し、単語の分散表現を用いて論理式の重みを計算し、重みを特徴量として加法回帰モデルにより類似度を導出している。しかし、この手法では、文中のすべての述語間の重みを計算する必要があり、計算量が多いという問題がある。

本研究では証明論的意味論 [7] の観点から、2 文間の意味のオーバーラップが大きければ、簡潔に含意関係を証明できるはずであるという仮説を立てる。この仮説に基づいて、2 文間の含意関係について自然演繹による証明を試み、証明の過程から類似度学習に必要な特徴量を導出する。このとき、証明不可能と判定されたサブゴールに対して、単語の分散表現を語彙知識として用いて公理を生成することによって、効率的に証明を行う方法を提案する。証明には、組合せ範疇文法 (Combinatory Categorical Grammar, CCG) [8] に基づく統語解析に基づいて文の意味を高階述語論理式で表現し、自動推論を行うシステムである *ccg2lambda* [3][4] を用いる。

3. 提案手法

3.1 CCG に基づく意味表現

まず、CCG パーザによる統語解析によって、自然言語の英文を構文木に変換する。CCG は語彙化文法の一つであり、統語構造からの意味合成に適した文法体系である。CCG は語の意味表示と統語範疇を記述する辞書と、語から句や文を構成する際の統語構造と意味合成の計算方法を指定する組み合わせ規則から成り立つ。CCG では、基本的な統語範疇として、N (名詞)、NP (名詞句)、S (文) 等の基底範疇が定義されている。また、複合範疇は基本範疇と二項演算子 $/, \backslash$ の組み合わせによって定義される。本手法では C&C [9] と EasyCCG [10] という 2 種類の CCG パーザでパーズを行う。意味表示は、パーズに成功した方を採用し、どちらの CCG パーザでも成功した場合は EasyCCG の結果を使用する。

次に、ラムダ計算に基づく意味合成によって、構文木から文の意味表示である高階述語論理式を導出する。語の意味はラムダ項によって表現され、組み合わせ規則によってラムダ項の意味合成を行い、文の論理式を導出する。意味表示は動詞をイベントを項に持つ述語として分析する意味論である Neo-Devidsonian Event Semantics [11] に基づく。Neo-Devidsonian Event Semantics では副詞や前置詞などの修飾表現をイベントを項に持つ述語として扱うため、修飾表現を含む文も簡潔に記述できるという利点がある。

連絡先: 谷中瞳, 東京大学大学院工学系研究科システム創成学専攻, hitomiyana@e.gcc.u-tokyo.ac.jp

3.2 自然演繹による同値関係の証明

本研究では2文間の同値関係について自然演繹による証明を行い、証明の過程と結果から類似度学習に必要な特徴量を導出する。ここで文 A, B 間の同値関係とは、 $A \Rightarrow B, B \Rightarrow A$ という双方向の含意関係のことを示す。2文間の同値関係を証明できる可能性が高いほど、それらの文は同じ意味を持つと考えられる。そこで、2文間の同値関係の証明可能性に基づいて特徴量を設計することで、自然演繹の証明に基づく文の類似度学習を実現する。

高階述語論理による証明には、高階論理・型理論に基づく定理証明器である Coq [12] を用いる。Coq は自然演繹 [13] に基づいて推論規則を適用し、証明を行う。また、Ltac という記述言語を用いて証明探索の手続きを定義することによって、証明の自動化が可能である。cog2lambda は Coq の一階述語論理の部分系に対する自動推論を含む自動証明機能と、高階の公理とを組み合わせることで、自然言語の効率的な推論を可能にしている。

はじめは公理なしで $A \Rightarrow B, B \Rightarrow A$ の証明を試みる。証明に失敗した場合は、結論を否定した $A \Rightarrow \neg B, B \Rightarrow \neg A$ の証明を試みる。これは結論の否定が証明可能な場合は、結論も結論の否定も証明不可能な場合と比較して、論理式が共通の部分を多く含むため、類似度が高くなることを考慮している。

3.3 単語の分散表現を用いた公理の生成

結論と結論の否定の両方の証明に失敗した場合は、語彙知識から前提と結論中の述語間の意味的關係をチェックし、公理を生成して再度 $A \Rightarrow B, B \Rightarrow A$ の証明を試みる。まず、証明の途中で証明不可能と判定されたサブゴールに関して、前提と結論で同じ項をシェアしている述語を生成する公理の候補として絞り込む。さらに、同じ項をシェアしている述語の中でも、項の格が同じである述語について優先して公理を生成する。

次に、語彙知識として WordNet を用いて形態変化、派生形、同義語、反意語、上位語、類似性、下位語の順に前提と結論中の述語間の関係をチェックし、いずれかにマッチした場合は確信度つき公理を生成する。WordNet に述語間の関係が存在しない場合は、Word2Vec モデルを用いて述語間の類似度を計算し、確信度つき公理を生成する。本手法では、Google News Corpus (約 30 億語) で学習済みの Word2Vec モデルを語彙知識に用いる。ここで、公理の確信度は、述語間の類似度を表し、生成された公理の中で確信度が最も高い公理を証明に用いる。WordNet から導出した場合は述語間に共通する上位概念への最短経路の長さを確信度とし、Word2Vec から導出した場合は Word2Vec の \cos 類似度を確信度とする。公理の確信度はいずれも 0.0 から 1.0 の値の範囲をとる。

公理を用いた証明に失敗した場合は結論を否定して、再度公理を用いた証明を試みる。公理を用いても、結論と結論の否定の両方の証明に失敗した場合は、 $A \Rightarrow B, B \Rightarrow A$ の証明の途中で証明不可能と判定されたサブゴールを無視して、強制的に証明を終了させる。

3.4 証明に基づく特徴量の導出

ここまでの自然演繹の証明の過程と結果から、計 9 個の特徴量を導出する。いずれも 0.0 から 1.0 の範囲の値に正規化する。

- 公理の数と公理の確信度

証明に導入した公理の数と確信度を特徴量に用いる。複数の公理を導入した場合は、導入した各公理の確信度の平均を特徴量として採用する。また、公理を導入しなかった場合は特徴量を 1.0 とする。

- 証明可能なサブゴールの数の割合

証明可能なサブゴールの数の割合が高いほど類似度が高くなると仮定して、 n/m (m : 前提のプールに現れる論理式の数, n : 証明で無視しなかったサブゴールの数) を特徴量として導出する。

- 証明不可能なサブゴールの項の格

文の主語や目的語が異なると、修飾語が異なる場合よりも類似度が低くなると仮定して、証明不可能と判定されたサブゴールの項の格 (主格, 直接目的格, 間接目的格) をチェックする。証明不可能と判定されたサブゴールにおけるそれぞれの格の割合を特徴量として導出する。

- 証明のステップ数

証明が簡潔であるほど類似度が高くなると仮定して、自然演繹の証明図の推論ステップ数を特徴量に用いる。

- 証明に適用した推論規則

証明に適用した推論規則の種類が多いほど証明が複雑になり、類似度が低くなると仮定して、証明に適用した推論規則の合計における各推論規則の割合を特徴量に用いる。本手法では $\wedge, \rightarrow, \exists$ の除去規則と導入規則, $=$ の除去規則という計 7 種類の推論規則を対象とする。

- 証明の結果

結論が証明可能か、結論の否定が証明可能か、結論も結論の否定も証明不可能かチェックし、それぞれ該当すれば 1.0, 該当しなければ 0.0 を特徴量とする。

- 前提と結論における述語の一致率

前提と結論との間で述語が一致しているほど、証明できる可能性が高いことを考慮して、前提と結論で述語が一致している割合を特徴量に用いる。

- 前提と結論における型の一致率

前提と結論との間で論理式の型が一致しているほど、証明できる可能性が高いことを考慮して、前提と結論で型が一致している割合を特徴量に用いる。

- 否定演算子

前提と結論のどちらかに否定演算子が含まれる場合と、両方に含まれる/含まれない場合とでは、前者の方が類似度が低いと仮定して、前者の場合は 0.0, 後者の場合は 1.0 として特徴量に用いる。

3.5 表層情報から導出した特徴量の組み合わせと文の類似度学習

本手法では類似度学習の精度を向上させるため、表層情報から導出した特徴量を自然演繹の証明に基づく特徴量と組み合わせ用いる。表層情報から導出した特徴量は名詞の一致率、動詞の一致率、品詞の一致率、WordNet 同義語の一致率、WordNet 同義語の距離、文字列の一致率、文長の差、TF-IDF ベクトル間の \cos 類似度、LSA ベクトル間の \cos 類似度、LDA ベクトル間の \cos 類似度と計 10 個の特徴量を用いる。特徴量はいずれも 0.0 から 1.0 の範囲の値に正規化する。類似度学習は決定木学習を用い、ハイパーパラメータはグリッドサーチを用いて最適化する。予測精度の評価には平均二乗誤差 (MSE) を評価指標とする。

ID	文 1	文 2	含意関係	類似度
1459	A man is exercising	A man is doing physical activity	Yes	4.5
3116	There is no man eating some food	The man is eating	No	3.3
3477	A person is dancing	A man is thinking	Unknown	1.1
3660	A woman is collecting tap water in a mug	A woman is emptying the water from a mug	Unknown	3.8

表 1: SICK データセットに含まれる文ペアの例.

4. 実験

4.1 データセットと評価方法

文の意味的類似度データセットである SemEval2014 Task1 SICK データセット [14] を用いて、本手法の評価を行った。SICK は、2 つの文の意味的類似度を予測するタスクであり、表 1 の通り類似度と含意関係が人手によって付与されている。類似度は 1.0 から 5.0 の範囲の値で付与されている。このデータセットには計 9927 件の文ペアが含まれており、そのうち訓練データ、開発データ、テストデータはそれぞれ 4500 件、500 件、4927 件含まれている。評価指標には、本手法によって計算された類似度スコアと、データセットに付与された正解スコアとの Pearson 相関係数 γ 、Spearman 相関係数 ρ 、MSE という 3 種類の指標を用いる。比較対象は、推論を用いた既存手法として、2 節で紹介した The Meaning Factory[5], UTexas[6] とする。

4.2 実験結果・考察

表 2 に本手法と既存手法の Pearson 相関係数 γ 、Spearman 相関係数 ρ 、MSE の結果を示す。Pearson 相関係数と Spearman 相関係数に関しては、本手法は推論を用いた既存手法よりも高い精度を達成した。一方で、MSE に関しては既存手法よりも精度が低かった。これは 4.3 項で例を示すように、公理の過剰生成によって含意関係が不明とアノテートされているのに証明できてしまい、正解スコアよりも類似度が極端に高くなってしまふ文ペアが存在したからである。

	γ	ρ	MSE
提案法 (WordNet)	0.833	0.791	0.563
提案法 (WordNet+Word2Vec)	0.833	0.790	0.564
提案法 (Word2Vec)	0.830	0.785	0.569
The Meaning Factory	0.827	0.772	0.322
UTexas	0.714	0.674	0.499

表 2: 本手法と既存手法の評価結果.

提案手法の語彙知識に WordNet のみを用いた場合と Word2Vec と WordNet の両方を用いた場合とを比較した結果、精度にほとんど差がなかった。この原因は、WordNet に語彙知識がない場合のみ Word2Vec を用いて公理を生成するが、SICK データセット中の単語はほぼ WordNet に含まれていたからである。今後、他のデータセットを用いて、単語の分散表現の有用性を検証する。語彙知識に WordNet のみを用いた場合と Word2Vec のみを用いた場合とでは、WordNet のみの方が精度が高かった。また、含意関係が不明とアノテートされているテストデータ 2790 件のうち、公理の生成によって証明できてしまった件数は WordNet, WordNet+Word2Vec, Word2Vec のみの順に 1283 件, 1337 件, 1601 件だった。これらの結果は、単語の分散表現を語彙知識に用いることによ

て、語彙知識の不足を解消できる一方で、公理を過剰生成してしまう問題があることを示唆している。

4.3 エラー分析

表 3 に語彙知識に Word2Vec のみを用いた場合の SICK の評価結果の成功例と失敗例を示す。4591 は成功例である。 $\forall x.cut(x) \rightarrow slice(x)$ という公理の生成によって正しく含意関係を証明でき、類似度スコアも正解スコアに近い結果が予測された。3974 は含意関係がある文ペアであり、 $\forall x.awaken(x) \rightarrow wake(x)$, $\forall x.awaken(x) \rightarrow up(x)$ という 2 つの公理の生成によって正しく含意関係を証明できたが、*awaken - up* 間の Word2Vec 類似度が 0.041 と低かったため、類似度スコアが正解スコアよりも低くなってしまった例である。この場合はフレーズに関する語彙知識を追加して、*awaken - wake up* 間の公理を生成する必要があり、今後の課題とする。1941 は含意関係が不明の文ペアだが、複数の公理が生成されることによって含意関係を証明できてしまったため、類似度スコアが正解スコアよりも高くなってしまった例である。このような公理の過剰生成を防ぐためには、確信度に閾値を設けるといった公理の生成条件を追加する必要がある。

5. 結論

本研究では文を高階述語論理式に変換し、単語の分散表現から生成した公理を用いて、2 文間の双方向の含意関係の証明から特徴量を導出し、文の類似度学習を行う手法を提案した。実験の結果から、単語の分散表現が自然演繹の証明における語彙知識として有用であることが示唆された。一方で、単語の分散表現を語彙知識に用いた場合、場合によっては公理を過剰に生成してしまい、証明の精度を下げてしまうという問題が示唆された。今後の課題として、単語の分散表現や外部知識を用いて証明に必要な語彙知識を正しく選択し、正確な証明を実現する方法論を検討する。また、複数のデータセットを用いて提案手法の有用性を検証する。

参考文献

- [1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems, (NIPS'13)*, pp. 3111–3119, 2013.
- [2] Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, Vol. 37, No. 1, pp. 141–188, 2010.
- [3] Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. On-demand injection of

ID	文ペア	類似度		公理
		類似度	(正解)	
4591	Some onions are being cut by a person	4.7	4.9	$\forall x.cut(x) \rightarrow slice(x)$
	A person is slicing some onions			$\forall x.awaken(x) \rightarrow wake(x)$
3974	A girl is awakening	2.8	4.9	$\forall x.awaken(x) \rightarrow up(x)$
	A girl is waking up			$\forall x.woman(x) \rightarrow person(x)$
1941	A woman is putting the baby into a trash can	3.4	1.0	$\forall x.trash(x) \rightarrow skillet(x)$
	A person is putting meat into a skillet			$\forall x.baby(x) \rightarrow meat(x)$

表 3: エラー分析.

lexical knowledge for recognizing textual entailment. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, (EACL'17)*, 2017.

- [4] Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. Higher-order logical inference with compositional semantics. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*, pp. 2055–2061, 2015.
- [5] Johannes Bjerva, Johan Bos, Rob Goot, and Malvina Nissim. The Meaning Factory: Formal semantics for recognizing textual entailment and determining semantic similarity. *SemEval 2014: International Workshop on Semantic Evaluation*, pp. 642–646, 2014.
- [6] Islam Beltagy, Katrin Erk, and Raymond Mooney. Probabilistic soft logic for semantic textual similarity. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, pp. 1210–1219, 2014.
- [7] Daisuke Bekki and Koji Mineshima. Context-passing and underspecification in Dependent Type Semantics. In Stergios Chatzikyriakidis and Zhaohui Luo, editors, *Modern Perspectives in Type Theoretical Semantics*, Studies of Linguistics and Philosophy, pp. 11–41. Springer, 2017.
- [8] Mark Steedman. *The Syntactic Process*. MIT Press, 2000.
- [9] Stephen Clark and James R. Curran. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, Vol. 33, No. 4, pp. 493–552, 2007.
- [10] Mike Lewis and Mark Steedman. A* CCG parsing with a supertag-factored model. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pp. 990–1000, 2014.
- [11] Terence Parsons. *Events in The Semantics of English: a Study in Subatomic Semantics*. MIT Press, 1990.
- [12] Yves Bertot, Pierre Castéran, Gérard Huet, and Christine Paulin-Mohring. *Interactive Theorem Proving and*

Program Development. Coq'Art: The Calculus of Inductive Constructions. Springer, 2004.

- [13] Dag Prawitz. *Natural Deduction – A Proof-Theoretical Study*. Almqvist & Wiksell, Stockholm, 1965.
- [14] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223, 2014.