

# 定理証明に基づく対話的な自然言語推論システム

隅田 敦  
東京大学

atsushisumita0421@gmail.com

峯島 宏次  
慶應義塾大学

minesima@abelard.flet.keio.ac.jp

宮尾 祐介  
東京大学

yusuke@is.s.u-tokyo.ac.jp

## 1 概要

自然言語推論 (Natural Language Inference; NLI) とは、ある文から別の文が帰結するか否かを判定するタスクである [1]. 自然言語推論は質問応答, 自動要約, ファクトチェックなどの様々な応用先が考えられ, 近年では自然言語理解のベンチマークとしても重要性を増している [2].

本論文では, 計算機と人間のインタラクションが NLI を解く上で有益であるかを検討する. ここでインタラクションとは, 人間と計算機が互いの弱みを補完しながら協力して NLI タスクを解くという形式を想定している. 一般的に, 人間の方が語彙間の含意関係や常識推論などの柔軟な判断に優れる一方で, 計算機の方が交錯した論理関係の包括的な探索や検証において優位性があると考えられる. 自然言語推論は様々な意味現象が組み合わさった複雑なタスクであり, これら両方の意味現象を処理する必要がある.

本研究では形式意味論と定理証明器を用いるアプローチに着目する. 定理証明器は複雑な数学的命題や形式手法の証明を補佐するためのソフトウェアであり, 人間と計算機が互いの強みを活かしてインタラクティブに複雑な課題を解決している好例である. いくつかの既存研究において, 自然言語の意味を論理式を用いてモデル化する形式意味論 [3, 4] と組み合わせることで, 定理証明器を NLI に応用することが示されている [5, 6, 7]. しかし, 定理証明器の重要な特徴であるインタラクティブな課題解決という側面に目を向けた NLI システムはこれまでに十分に研究が行われて来なかった.

そこで, 本研究では形式意味論と定理証明器に基づいた NLI システムである ccg2lambda [8] に, 証明

に関する知識が不足した際にユーザーが自然言語でインタラクションを行うことが出来るシステムを提案する.

仮説の検証のために, SICK [9] をデータセット, Amazon Mechanical Turk にて募ったクラウドワーカーを被験者として適合率, 再現率, 精度を計測し, エラー分析を行った. 以降で提案手法と実験について詳細に説明する.

## 2 提案手法

本節ではユーザーと定理証明器とのインタラクションを実現するための提案手法について説明する. 重要な点として, インタラクションは基本的に自然言語で行われる. これにより, 形式意味論や定理証明器について知識がないユーザーであってもインタラクションが可能となる.

図 1 に提案手法の大枠を示した. 提案手法の内, セマンティックパーサーと定理証明器については既存手法を用いている. まず, ユーザーが証明したい文の組をシステムに入力し, これがセマンティックパーサーにより論理式に変換される. これが定理証明器に入力される. 定理証明器は証明をいくつかのサブゴールと呼ばれる命題に分解し, 全てのサブゴールが証明された時に全体の証明が完成する. 図 1 の例では,  $person(x0), move(e0)$  がサブゴールである.

本研究ではサブゴールを自然言語に変換しユーザーに提示するインタラクティブな NLI システムを提案する. 以下, サブゴールを自然言語に変換したものを可読なサブゴールと呼ぶことにする. より正確には, あるサブゴールの可読なサブゴールとは, 前提文とそのサブゴールが真の時かつその時のみ真となるような自然言語文のことである. 図の例では,  $person(x0), move(e0)$  に対応する可読なサブゴール

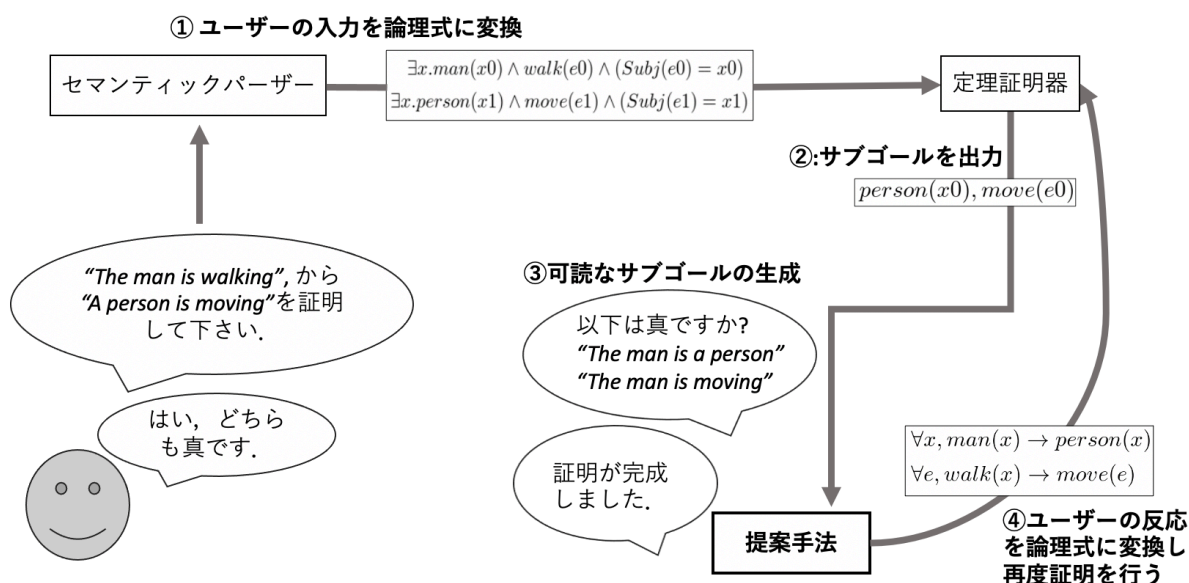


図1 提案手法によるインタラクションの例. ユーザーが証明したい文対を入力すると, セマンティックパーサーにより文対が論理表現に変換され, 定理証明器に入力される. 証明が完成しなかった場合にサブゴールを抽出し, 提案手法により可読なサブゴールが生成されユーザーに提示される. ユーザーの反応に応じて公理が生成され, 証明が続行される.

は *The man is a person*, *The man is moving* となる. ユーザーが可読なサブゴールの真偽を入力すると, それに応じて新たな論理式が公理として生成され, 定理証明器にフィードバックされる. 以下でより詳細なステップについて説明する.

## 2.1 可読なサブゴールの生成

サブゴールを可読なサブゴールに変換する問題を考える. 以下の説明では, 図1中の  $move(e0)$  を例にとる.

サブゴールを自然言語文に変換するためには, サブゴールに含まれていないが自然言語文には必須の要素を識別し正しく補充することが必要となる. 例えば, 適格な(英語の)自然言語文は主語が必ず必要である.  $move(e0)$  の例では, 誰が  $move$  しているのかを識別しなければならない.

まず第一ステップとして, 可読なサブゴールに含めるべき変数の集合を識別する. これは論理式中の意味役割を表す式から識別できる. 意味役割は主語や動詞の必須項などを表現しており, 適格な自然言語文には含める必要がある. よって, 各変数についてその変数が含まれている意味役割の式を抽出し, その式に含まれている全ての変数を集合に追加するという手続きを再帰的に適用すれば良い. 現在の例では,  $Subj(e0) = x0$  から可読なサブゴールに含めるべき変数は  $\{e0, x0\}$  であることが分かる.

次のステップでは, それぞれの変数に語を割り当て, 変数の内容を自然言語で表す. これには, その変数を引数としている述語を選択すれば良い. 現在の手法では *entity* 型の変数に関してはそれを引数にする名詞を, *event* 型の変数に関してはそれを引数にする動詞を選択している. 現在の例では,  $x0$  には *man*,  $e0$  には *move* を選ぶ. 結果として,  $\{e0: "move", x0: "man"\}$  を得る.

最後に, 得られた語の正しい語順と必要な機能語を適宜補充し, 文法的に正しい文を生成する. 可読なサブゴールはごく簡単な宣言文なので, 語の品詞から語順, 時制, 必要な機能語を特定するルールベースのシステムが比較的容易に構築可能である. この例では, *The man is moving* が可読なサブゴールとなる.

表1に実際に生成された可読なサブゴールを示した. 二列目にサブゴール, 三列目に可読なサブゴールが示されている.

## 2.2 ユーザーの反応

可読なサブゴールを受け取ったユーザーは *true*, *false*, *unknown* から一つを選び, これを返答とする. *true* が選択されると, 可読なサブゴールが論理表現に変換され, 公理として定理証明器に追加される. *false* が選ばれた場合には可読なサブゴールを否定した論理表現が追加される. *unknown* が選ばれた場合には公理は追加されず, 証明が完成しないことが

文のペア	サブゴール	可読なサブゴール	公理
<b>T:</b> A group of <i>kids</i> is playing in a yard and an old man is standing in the background. <b>H:</b> A group of <i>boys</i> in a yard is playing and a man is standing in the background	$boy(x0)$	The kid is a boy.	$\forall x, kid(x) \rightarrow boy(x)$
<b>T:</b> The young <i>boys</i> are playing outdoors and the man <i>is smiling nearby</i> . <b>H:</b> The <i>kids</i> are playing outdoors <i>near</i> a man <i>with a smile</i> .	$kid(x0)$ $near(e0, x1)$ $with(e1, ?x2)$ $smile(?x2)$	The boys are kids. The man is with a smile. The boys are near a man.	$\forall x, kid(x) \rightarrow boy(x)$ $\forall x, y, e, boy(x) \rightarrow$ $Subj(x) = e \wedge near(e, y)$ $\wedge man(y)$
<b>T:</b> Two dogs are fighting. <b>H:</b> Two dogs are wrestling and hugging.	$wrestle(e0)$ $hug(e1)$	The dogs are wrestling. The dogs are hugging.	$\forall e, fight(e) \rightarrow hug(e)$ $\forall e, fight(e) \rightarrow wrestle(e)$

表 1 文のペアに SICK の訓練データから抽出した文を示している。T から始まる文が前提文、H から始まる文が帰結文であり、前者から後者への含意関係を示すことがタスクの内容となる。イタリックで示した箇所が非論理的推論が必要となる箇所であり、サブゴールとして定理証明器により出力される。サブゴールには定理証明器が証明出来なかった命題を示している。可読なサブゴールには提案手法が生成しユーザーに提示された文を示した。公理にはユーザーが true で反応した場合に生成される論理式を示した。

	Methods		適合率	再現率	精度
	axiom injection	readable subgoal			
Baseline					61.6
ccg2lambda			99.7	34.7	74.7
ccg2lambda & abduction	✓		98.5	53.8	81.3
提案手法		✓	96.6	59.0	82.1
提案手法 & abduction	✓	✓	96.4	59.2	82.1

表 2 SICK における適合率, 再現率, 精度. チェックマークは axiom injection [5], readable subgoal は提案手法の可読なサブゴールを使用しているか否かを示している。

確定するのでプロセスが停止する。表 1 の四列目に true が選択された場合の公理を示した。

## 3 実験

### 3.1 データ

提案手法の評価には、語彙的な推論と論理的な推論の両方を適度に含み、関連する既存手法 [5, 6] においても使用されている自然言語推論データセットである SICK [9] を用いた。SICK は 10000 件の文対から成り、各文対について一つ目の文から二つ目の文が帰結するか否かを判定する分類タスクである。正解ラベルは *entailment*, *contradiction*, *neutral* の三通りである。

### 3.2 実験設定

セマンティックパーザーには ccg2lambda [8], 定理証明器に Coq [10] を使用した。SICK の訓練データからサンプルした 1000 ペアを depccg と ccg2lambda により論理表現に変換し、定理証明器に渡して証明を行った。定理証明器が証明を完成させた場合はプロセスが終了する。証明が完成しなかった場合はサブゴールを抽出し、提案手法により可読なサブゴール

を生成し、ユーザーがインタラクションを行う。本実験においては Amazon Mechanical Turk により被験者を募り、各ペアについて 3 人が *true*, *false*, *unknown* を付与した。精度を高めるため、3 人が一致した時のみ公理を生成した。

### 3.3 結果

適合率, 再現率, 精度を表 2 に示した。一行目のベースラインは majority class であり, 61.5% を占める *neutral* の割合を示している。二行目は ccg2lambda [8] の結果であり, 99.7% と高い適合率を達成しているが, 再現率が 34.7% と低く, 全体として低い精度となっている。三行目に ccg2lambda を拡張し知識データベースから公理生成を行う word abduction [5] の結果を示した。語彙的推論がカバーされたため再現率が 53.8% に向上している。四行目に提案手法の結果を示した。再現率は 59.0% と最も高く, 偽陽性も少なく適合率の高さも維持されたため, 精度についても比較した手法の中では最も良い結果となった。

### 3.4 エラー分析

提案手法の成功例と失敗例について表 3 にまとめた。

文ペア	正解	予測	可読なサブゴール
T: A man with a helmet painted red is riding a blue motorcycle down the road. H: A motorcyclist with a red helmet is riding a blue motorcycle down the road.	Yes	Yes	The helmet is red. The man is a motorcyclist.
T: A surfer is riding the wave. H: A rider is surfing the wave	Yes	Yes	The surfer is surfing the wave. The surfer is a rider.
T: A man is walking in a yard. H: A man is walking outside.	Yes	Yes	The man is walking outside.
T: A man is speaking on a stage H: A man is speaking on a podium	Unk	Yes	The stage is a podium
T: A man is cleaning a dish H: A man is preparing a dish.	Unk	Yes	The man is preparing a dish.

表 3 本実験中に得られた提案手法の成功例と失敗例. 単純な語彙間の含意関係では捉えられない句間の含意関係が捉えられている (1-2 行目). 一方で, 帰結文との重複が多いケース (3 行目) や蓋然性を高めるだけの関係を含意とアノテーションしてしまうケース (4-5 行目) が多く見られた.

一行目のペアは成功例である. この証明を完成させるためには, 句間に成立する含意関係を認識する必要があり, 単純な語彙間の含意関係に基づいた手法では処理することが出来ない. より形式的には, ここで用いられている知識は以下のようなものである. *The helmet is red* については

$$\begin{aligned} \forall x, y, e, \text{paint}(e) \wedge \text{red}(x) \wedge (\text{Acc}(e) = x) \\ \wedge (\text{Dat}(e) = y) \rightarrow \text{red}(y) \end{aligned} \quad (1)$$

*The man is a motorcyclist* については

$$\begin{aligned} \forall e, x, y, \text{ride}(e) \wedge \text{man}(x) \wedge \text{motorcycle}(y) \\ \rightarrow \text{motorcyclist}(x) \end{aligned} \quad (2)$$

実際この例においては知識ベースにより公理を生成する既存手法 [5] は証明を完成させることが出来ていない. 二行目に示した例も同様に, *ride the wave* から *surf* を導く必要がる.

三行目に示した例は成功しているがやや問題を含んでいる. 可読なサブゴールは正しく生成され予測結果も正しいが, 帰結文とほとんど重複してしまっている. これは帰結文が短すぎるために分解の余地がないためである. したがってこのようなケースでは可読なサブゴールを介する有用性が低く, より長い前提文・帰結文において提案手法が価値を発揮することが示唆される.

四行目に示した例は典型的な失敗例である. この例では *The stage is a podium* という可読なサブゴールに対して *true* とラベルが付与されたが, 全ての *stage* が *podium* である訳ではないので *unknown* が正しい反応となる. ただしこの文脈で *stage* が *podium* であるという蓋然性は高く, このようなケースと他の語彙間の含意関係や常識判断の間に厳密な線を引くのは難しい. 五行目に示した例も同様である.

また, 曖昧性のある場合にも解析誤りが見られた. 例えば *The men are women* はその女性は男性であるともその女性は人間であるとも解釈でき, それぞれの解釈の下で適切な反応が異なってしまう. このような曖昧性の解消が将来的には必要になると考えられる.

他の失敗ケースで多く見られたのは CCG タグ付けやセマンティックパーザーなどの前処理の段階での解析結果の誤りが影響するケースである. 特に複合表現など構成性を満たさない表現は誤った意味表現が付与される例が多く見られた.

## 4 結論と今後の課題

本論文では形式意味論と定理証明器に基づいたインタラクティブな自然言語推論システムを提案した. 実験の結果, 提案手法により再現率と精度が向上しており, ユーザーとのインタラクションが有益であることが示された.

今後の課題としては, 前節で示した典型的な失敗例の解決に加え, より複雑な設定でインタラクションを成立させることが挙げられる. SICK は単文と単文の間の含意関係を認識するタスクだが, 人間が論理関係を包括的に把握するのが困難な量のドキュメントを前提文とする場合の方が応用上は有益である. そのためには共参照などの問題を考慮する必要がある. また, SICK では非論理的推論のほとんどがシンプルな語彙間の推論だが, 常識推論などのより高度かつ計算機には処理が難しい推論を含んだ場合にもインタラクションを成立させることが出来れば, ユーザーとシステムの協働により高い価値を見いだせるので, これらを今後の課題としたい.



## 参考文献

- [1] Omer Levy, Torsten Zesch, Ido Dagan, and Iryna Gurevych. Recognizing partial textual entailment. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 451–455, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.
- [2] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.
- [3] Terence Parsons. *Events in the Semantics of English*. MIT Press, 1990.
- [4] Claudia Maienborn. Event semantics. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics. An International Handbook of Natural Language Meaning*, volume 1, pages 802–829. Mouton de Gruyter, 2011.
- [5] Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. On-demand injection of lexical knowledge for recognising textual entailment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 710–720, Valencia, Spain, April 2017. Association for Computational Linguistics.
- [6] Hitomi Yanaka, Koji Mineshima, Pascual Martínez-Gómez, and Daisuke Bekki. Acquisition of phrase correspondences using natural deduction proofs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 756–766, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [7] Lasha Abzianidze. LangPro: Natural language theorem prover. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 115–120, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [8] Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. ccg2lambda: A compositional semantics system. In *Proceedings of ACL-2016 System Demonstrations*, pages 85–90, Berlin, Germany, August 2016. Association for Computational Linguistics.
- [9] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).
- [10] Yves Bertot and Pierre Castéran. *Interactive Theorem Proving and Program Development: Coq'Art: the Calculus of Inductive Constructions*. Springer, 2013.