Q6-8

LLM を用いた知識補完の統合による 自然言語推論システム lightblue の拡張

富田朝　戸次大介
お茶の水女子大学
tomita.asa@is.ocha.ac.jp

概要

- 形式意味論に基づく自然言語推論システムは厳密な推論が可能だが、語彙知識の不足が課題となっている
- 本研究では、論理推論システムlightblueに大規模言語モデル（LLM）を統合し、推論に必要な語彙知識のみを公理として動的に生成・補完する手法を提案する
- 評価実験の結果、提案手法は公理生成を行わない論理推論システムを上回る推論性能を示した

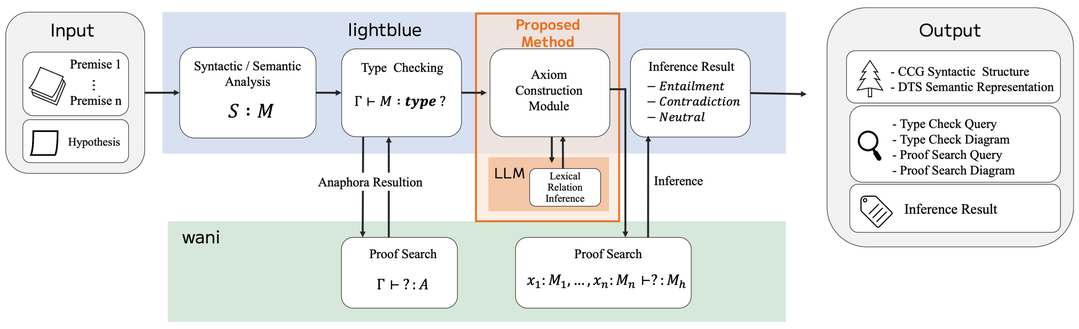
研究背景：自然言語推論のアプローチと課題

大規模言語モデル（LLM）を用いる手法：
- ベンチマークデータセットで高い精度を達成している（Qin et al., 2023）
- 否定などの複雑な論理表現を含むNLIの精度が低い（Truong et al., 2023
- 推論の判断過程が不透明である（Turpin et al., 2023）

理論言語学に基づくモデルを用いる手法：
- 文の意味を論理式として明示的に表現し、論理に基づいて推論を行うことができ、証明可能性に基づく高い適合率（Precision）を有する
- 語彙間の関係性や世界知識に依存する語彙的な推論を扱いづらいという課題がある　例）太郎はりんごを食べた ⊢ 太郎はフルーツを食べた

本研究の目標

論理推論システムlightblue (Tomtia et al., 2025)に、LLMを用いて語彙的な知識を動的に補完する公理生成モジュールを統合し、システムを拡張する



提案手法：公理補完機構の追加

前提文 (premise)：太郎が**りんご**を食べた　⊢　仮説文 (hypothesis)：太郎が**フルーツ**を食べた

1. lightblueで入力文の統語・意味解析、意味表示の型検査を行う

lightblue

$$\left[\begin{array}{c} x:entity \\ u_0:\left[\begin{array}{c} x:entity \\ s:entity \\ りんご(s\ x) \end{array}\right] \\ e:evt \\ \left[\left[\begin{array}{c} u_1:\ 食べる(e, 太郎, \pi_1 u_0) \\ \#タ(e) \end{array}\right]\right] \end{array}\right] \qquad \left[\begin{array}{c} x:entity \\ u_0:\left[\begin{array}{c} x:entity \\ s:entity \\ フルーツ(s,x) \end{array}\right] \\ e:evt \\ \left[\left[\begin{array}{c} u_1:\ 食べる(e, 太郎, \pi_1 u_0) \\ \#タ(e) \end{array}\right]\right] \end{array}\right]$$

Subgoalの特定

2. 定数の型（**signature**）と その語の統語的型（**syntactic type**）を抽出する

前提文
$((太郎, entity), \boldsymbol{NP})$　$((りんご, entity \to entity \to type), \boldsymbol{N})$
$((食べる_{ga/o}, entity \to entity \to entity \to type), \boldsymbol{S \setminus NP_{ga} \setminus NP_o})$

仮説文
$((太郎, entity), \boldsymbol{NP})$　$((フルーツ, entity \to entity \to type), \boldsymbol{N})$
$((食べる_{ga/o}, entity \to entity \to entity \to type), \boldsymbol{S \setminus NP_{ga} \setminus NP_o})$

3. signatureとsyntactic typeが一致する語を抽出する

[（りんご, フルーツ）,（フルーツ, りんご）]

知識の獲得と公理の生成

4. プロンプトの生成

りんごはフルーツに包含されるか
フルーツはりんごに包含されるか

5. LLMを呼び出し語彙的な推論を行う

`[Yes, Unknown]`

6. `Yes/No`の場合は含意の公理を生成する
（`Unknown`の場合は公理の生成を行わない）

含意の公理：$(x:entity) \to (s:entity) \to (u : りんご(s,x)) \to フルーツ(s,x)$

評価実験

データセット

JSICK (Yanaka & Mineshima, 2022) のデータ約5000文のうち、名詞の言い換えのみが行われているデータセットを310文抽出し、ランダムサンプリングでした50文を利用（正解データのチェックには高度に専門的な知識を要するため）

文のペア	ラベル	Subgoal	生成された公理
T: 花子が馬に乗っている **H:** 花子が動物に乗っている	Entailment	(馬, 動物) ✓ (動物, 馬)	$(x : entity) \to (s : entity) \to (u : 馬(s,x)) \to 動物(s,x)$
T: 太郎がジャガイモを切っている **H:** 太郎がトマトを切っている	Neutral	(ジャガイモ, トマト) (トマト, ジャガイモ)	
T: 太郎はインタビューを拒否している **H:** 太郎はインタビューを許可している	Contradiction	(許可, 拒否) × (拒否, 許可) ×	$(x : entity) \to (s : entity) \to (u : 許可(s,x)) \to \neg 拒否(s,x)$ $(x : entity) \to (s : entity) \to (u : 拒否(s,x)) \to \neg 許可(s,x)$

実験1: LLM の語彙的関係推論能力の評価

Subgoal（(りんご, フルーツ)のような語彙ペア）191件に対して、人手で `Yes, No, Unknown` のラベルを振り、LLMの出力を評価

含意のRecallが高くPrecisionが低い
→ 含意を広く検出できる一方で、偽陽性が多く公理の過剰生成が行われる恐れがある
矛盾のF1が低い
→ LLMの語彙の矛盾の判定は不安定

表 2　LLM による語彙間の関係性判定の性能の結果

Label	Precision	Recall	F1
Yes	0.636	0.840	0.724
No	0.361	0.722	0.481
Unknown	0.939	0.764	0.843
Macro avg	0.645	0.776	0.683
Micro avg	0.770	0.770	0.770

結論：LLM を用いた語彙関係判定は含意公理の候補生成に一定の有効性を持つものの、その出力を無条件に公理として採用することは危険であり、特に矛盾公理の生成は制御する必要がある

実験2:公理生成をともなう推論性能の評価

lightblueに自動生成した公理を組み入れて推論を行うことで、提案システム全体としての推論性能を評価

比較対象
- Majority Baseline：全てのラベルを最頻ラベル（Neutral）で判定した際の精度
- Vanilla lightblue：自動公理生成を行わない素のlightblueの精度

結論： Accuracy 0.72 を達成し、 Majority Baseline および Vanilla lightblue を上回る性能を示した。

表 3　自然言語推論性能の比較結果

Method	Precision	Recall	Accuracy
Majority Baseline	–	–	0.580
Vanilla lightblue	0.898	0.262	0.600
提案手法	**0.919**	**0.338**	**0.720**

まとめ

- LLMによる語彙関係の判定には誤りも含まれるが、論理推論システムを定理証明器に統合することで、全体として推論能力（Accuracy）の向上に寄与することが確認された。論理推論と LLM を対立的に捉えるのではなく、相補的に統合することで、語彙的推論を含むより実用的な自然言語推論が可能となることを示し
- 今後の展望としては、名詞間の包含関係にとどまらず、述語間の意味的関係性を扱えるように公理生成手法を拡張することが挙げられる。