Q9-18 ツリーバンクの言語学的妥当性の自動評価

富田朝[1] 谷中瞳[2] 戸次大介[1]
お茶の水女子大学[1] 東京大学[2]
tomita.asa@is.ocha.ac.jp

はじめに

コーパス自体の質や妥当性を評価するのは難しい
専門的な知識が必要となる言語資源は人手の評価が一般的で、自動評価する方法は確立されていない
データ数の多いCCGツリーバンクの「妥当性」を自動で評価するための手法を2つ提案する

背景：なぜ妥当なツリーバンクが必要なのか

課題

誤りを含むツリーバンクで学習された統語解析器は
誤りを含む統語構造を出力する
→ 誤った推論へとつながる

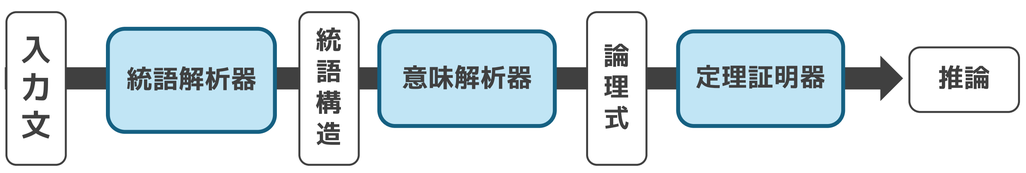


図1：形式文法に基づいた推論パイプライン

LLMでの推論との違い

LLM：誤りが生じた際に原因の特定が難しい
形式文法：誤りの原因の特定が可能
→ 信頼性につながる

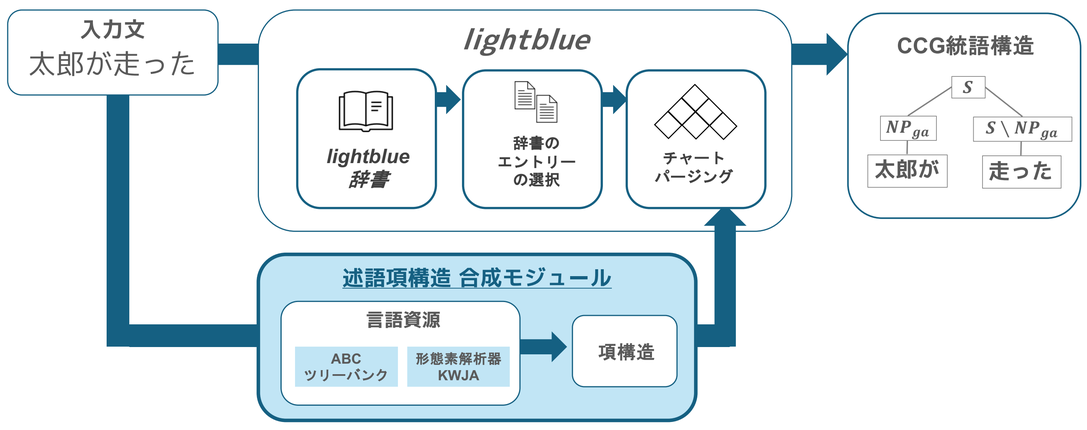
背景：ツリーバンクの構築 （Tomita et al., EACL 2024）

日本語CCG統語解析器lightblue [1)] (Bekki and Kawazoe 2016) を用いて
統語構造の自動アノテーションを行う

課題

*lightblue*が参照する辞書に、述語項構造に関する
誤りが含まれている

述語項構造合成モジュールを開発し、他の言語資源から
項構造情報を抽出し合成する手法を提案



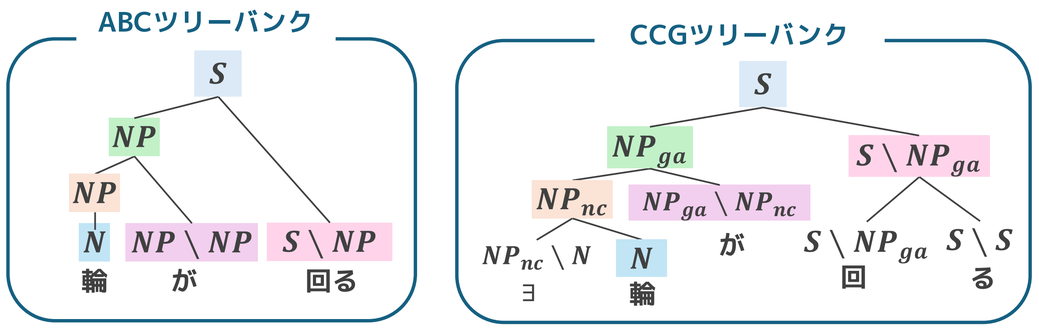
[1)] https://github.com/DaisukeBekki/lightblue

提案手法1：他コーパスとの比較

他の範疇文法コーパスとの一致度をスコア化することで、
統語構造としての信頼性を評価する
本研究では、CCGへの変換が容易な**ABC文法に**基づいた
ABCツリーバンク (Kubota et al. 2019)**を用いる**

スコアの付け方

ABCツリーバンクの部分構造が
CCGツリーバンクの部分構造に含まれる割合



利点 - 量化表現を含むunary規則に対応できる
- 述語の分析の差異にも対応できる

提案手法2：型検査での評価

型検査(type-check)：CCGツリーバンクに含まれる
依存型意味論(Bekki and Mineshima 2017)に基づく**意味表示**が**型**として
整合であるかを確かめることができる

型検査が成功：型レベルでの整合性を担保できる

$$\frac{\vdash \begin{bmatrix} x_0 : entity \\ \begin{bmatrix} x_1 : entity \\ 輪/わ\ (x_1, x_0) \end{bmatrix} \end{bmatrix} : type \qquad s_0 : \begin{bmatrix} x_0 : entity \\ \begin{bmatrix} x_1 : entity \\ 輪/わ\ (x_1, x_0) \end{bmatrix} \end{bmatrix} \vdash \begin{bmatrix} x_2 : entity \\ 回る/ガ(x_2, \pi_1(s_0)) \end{bmatrix} : type}{\vdash \begin{bmatrix} u_0 : \begin{bmatrix} x_o : entity \\ \begin{bmatrix} x_1 : entity \\ 輪\ /\ わ\ (x_1 . x_0) \end{bmatrix} \end{bmatrix} \\ \begin{bmatrix} x_2 : entity \\ 回る\ /\ ガ(x_2, \pi_1(u_0)) \end{bmatrix} \end{bmatrix} : type} (\Sigma F)$$

意味表示が**type**という型を持つことが証明できれば成功

型検査が失敗：統語構造または意味表示に誤りがある

利点 意味表示のレベルで型理論に基づいた評価ができる

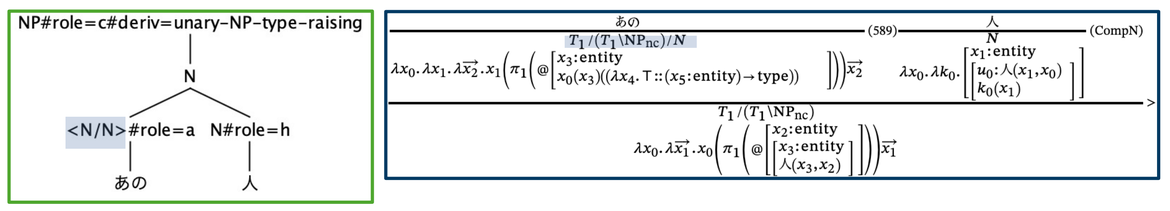
評価実験：ツリーバンクの言語学的妥当性の評価

		サンプリング数	統語的評価		意味的評価	総合評価
			スコア平均	スコア＞50の文	型検査通過数（率）	スコア>50 & 型検査通過数（率）
全体		**456**	**47.0**	**219 (48.0%)**	**306 (67.1%)**	**161 (26.1%)**
ジャンル	青空文庫	75	41.4	26 (34.7%)	48 (64.0%)	15 (16.7%)
	聖書	24	50.8	15 (62.5%)	14 (58.3%)	9 (27.3%)
	書籍	6	56.2	5 (83.3%)	4 (66.7%)	4 (40.0%)
	会議録	60	55.7	43 (71.7%)	33 (55.0%)	25 (29.4%)
	フィクション	21	40.7	8 (38.1%)	16 (76.2%)	7 (25.0%)
	辞書	18	55.4	11 (61.1%)	12 (66.7%)	8 (30.8%)
	法律	6	26.8	2 (33.3%)	4 (66.7%)	2 (25.0%)
	その他	30	52.3	13 (43.3%)	17 (56.7%)	10 (25.0%)
	ニュース	30	41.2	10 (33.3%)	24 (80.0%)	9 (23.1%)
	ノンフィクション	6	53.2	3 (50.0%)	6 (100%)	3 (33.3%)
	話し言葉	30	38.0	10 (33.3%)	28 (93.3%)	10 (25.0%)
	テッドトーク	15	40.9	5 (33.3%)	9 (60.0%)	4 (21.1%)
	教科書	120	50.3	60 (50.0%)	78 (65.0%)	47 (28.1%)
	ウィキペディア	15	50.1	8 (53.3%)	13 (86.7%)	8 (34.8%)

統語的評価：提案手法1に関する平均スコアとスコア>50の文の数を算出
意味的評価：提案手法2に関する型検査通過数と通過率を算出
総合評価　：スコア > 50かつ型検査に通過した文の数と割合を算出

考察

スコアが低い文：ABCツリーバンクとCCGツリーバンクの分析の差異



→ ABCツリーバンクの量化表現の分析の誤りが一致度の低下につながっている

型検査に通らない文：辞書のエントリーの語彙レベルの誤り・エントリーの不足

まとめ

日本語CCGツリーバンクの言語学的な妥当性の自動
評価の手法を提案し、ツリーバンクの評価を行った

今後の展望

1. 統語的評価の改善：詳細な統語情報の評価手法の考案
2. 意味的評価の改善：False positiveを抑えた評価手法の考案
3. ツリーバンクの構築過程へのフィードバック機構の導入