日本語事実性解析課題の経験的分析

成田 和弥†, 水野 淳太‡†, 乾 健太郎†
†東北大学 ‡奈良先端科学技術大学院大学

概要

○事実性に影響する表現を手がかりとした，日本語事実性解析器を構築
○特にリソースの問題に着目して誤り分析を行った

背景・目的

事実性

文中のある事象が実際起こったことなのか，あるいは起こる可能性を述べただけなのかに関する情報

実際に起こった(CT+)
彼はその**発言**を**知ら**ない。
実際には起こっていない(CT-)
彼女は先に**帰っ**たんだろう。
実際に起こった可能性が高い(PR+)

事実性解析

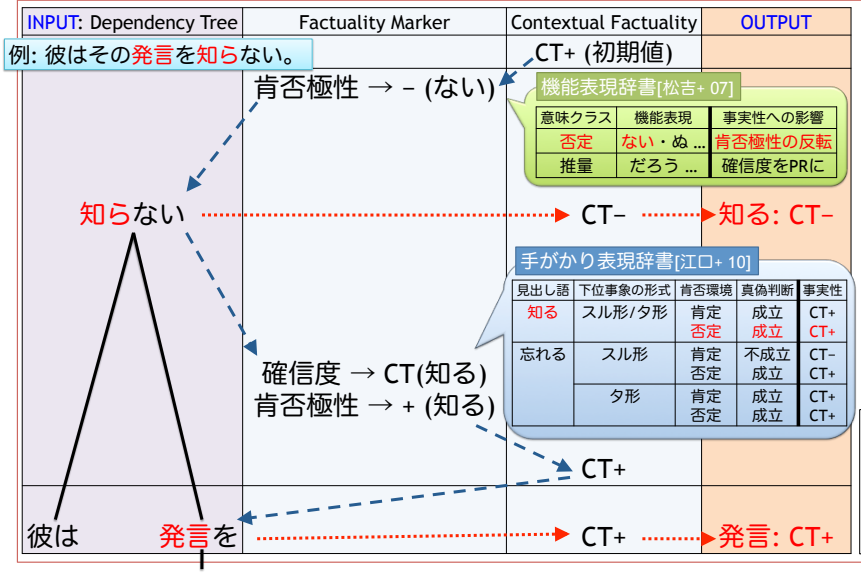
○自然言語処理分野[Inui+ 08][川添+ 11]だけでなく，生物医学分野[Light+ 04][Medlock+ 07]においても研究されている
○機械学習を用いた解析手法[江口+ 10]やパターンに基づく解析手法[Saurí 08]が提案されている
◎リソース(語彙知識)の問題について分析を行いたい

[Saurí 08]のモデル

◎事実性に影響を与える表現を手がかりとして，事実性を依存構造木の根から伝搬させて解析
○辞書が与える情報を組み合わせて事実性解析を行う構成性をもつ
○事実性を，確信度と肯否極性の2軸に分割して分析を行うことが可能

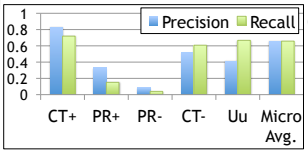
[Saurí 08]のモデルをもとに日本語事実性解析
特にリソースの問題に着目して分析を行った

日本語事実性解析器



事実性		肯否極性		
		+ (positive)	– (negative)	u (underspecified)
確信度	CT (Certain)	CT+	CT–	
	PR (Probable)	PR+	PR–	
	U (Underspecified)			Uu

【実験設定】
○入力：一文の係り受け解析結果
（正解の形態素情報に対して構文解析を行った）
○出力：各事象に対する事実性
○評価データ：拡張モダリティタグ付与コーパス[松吉+ 10]の一部(6,404文/14,917事象)
○辞書：機能表現辞書の一部（5345表現）
手がかり表現辞書（3692表現）



○PR–は確信度と肯否極性の両方を更新する必要があるため，簡単に解くことはできない

誤り分析

辞書項目(語彙知識)の不足 (3割程度)

サーバーは接続を解除しました。
正解：CT–
出力：CT+
手がかり表現辞書になく，正しく更新ができない

おそらく自慢したいだけの見栄っ張りです。
辞書にないだけでなく，アルゴリズムの拡張が必要
正解：PR+
出力：CT+

辞書だけでは難しい (3割程度)

放棄してしまったのが敗因ではない。
正解：CT+
出力：CT–
「ない」の与える肯否極性–が伝搬されてしまう

日本語が入らず英数字も入りません。
正解：CT–
出力：CT+
「ん」の与える肯否極性–が伝搬されてしまう

辞書にはあるが解釈が難しい (1割程度)

初歩的な質問で申し訳ありません。
正解：CT+
出力：CT–
「ん」が否定表現として認識されてしまう

※残り3割程度は構文解析誤りやアノテーション誤りなど

今後の課題

語彙知識の拡充

○副詞などの別のタイプの辞書[川添+ 11]の追加
○現在の辞書をシードとした知識の拡充
スル形/タ形をとるかとらないか
実際に起こっている/いないことが多いか
類義語には同じ情報を付与できるか

節間・事象間の関係の認識

○どのような文脈では事実性を伝搬させ，どのような文脈では伝搬させるべきではないのか，を判断する必要がある
談話解析との連携

更なる分析

○今回の分析では表現の多様性は低いので，全てリストアップして解決できるのではないか？

モダリティ解析へ