

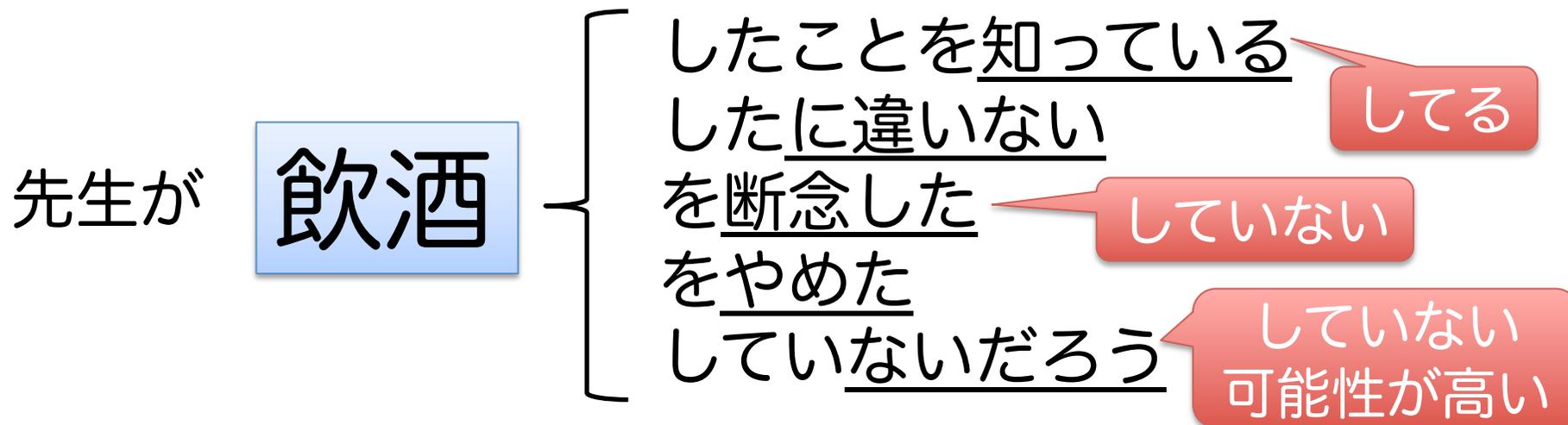
事実性解析のその後

成田 和弥, 乾 健太郎
(東北大学)

Introduction: What is Factuality?

- 事実性 (Factuality) :

文中のある事象が実際に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

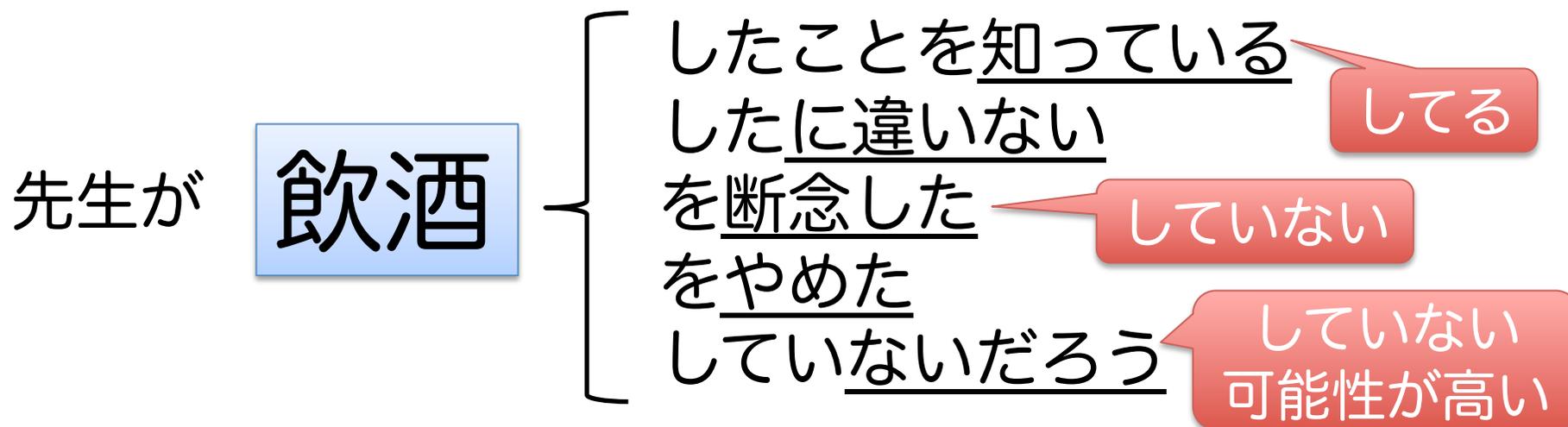


[松吉+ 10]における真偽判断のモダリティ,
[川添+ 11]における確実性 に対応

Introduction: Motivation

- 最終目的：高精度な事実性解析

事実性に影響を与える語彙知識に焦点を当てる

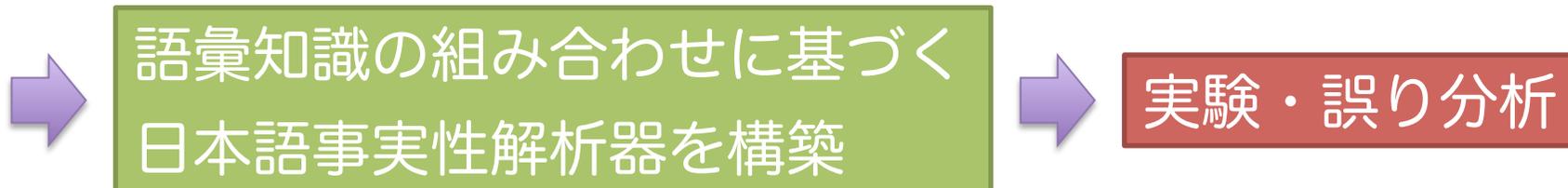


Introduction: Motivation

- 最終目的：高精度な事実性解析

事実性に影響を与える語彙知識に焦点を当てる

- **語彙知識の組み合わせ**による問題
 - 現在利用可能な語彙知識のカバレッジ
 - 語彙知識の拡充・拡張の必要性
 - 語彙知識の曖昧性の有無
 - 局所的な語彙知識の組み合わせだけで解析不可能な問題の有無



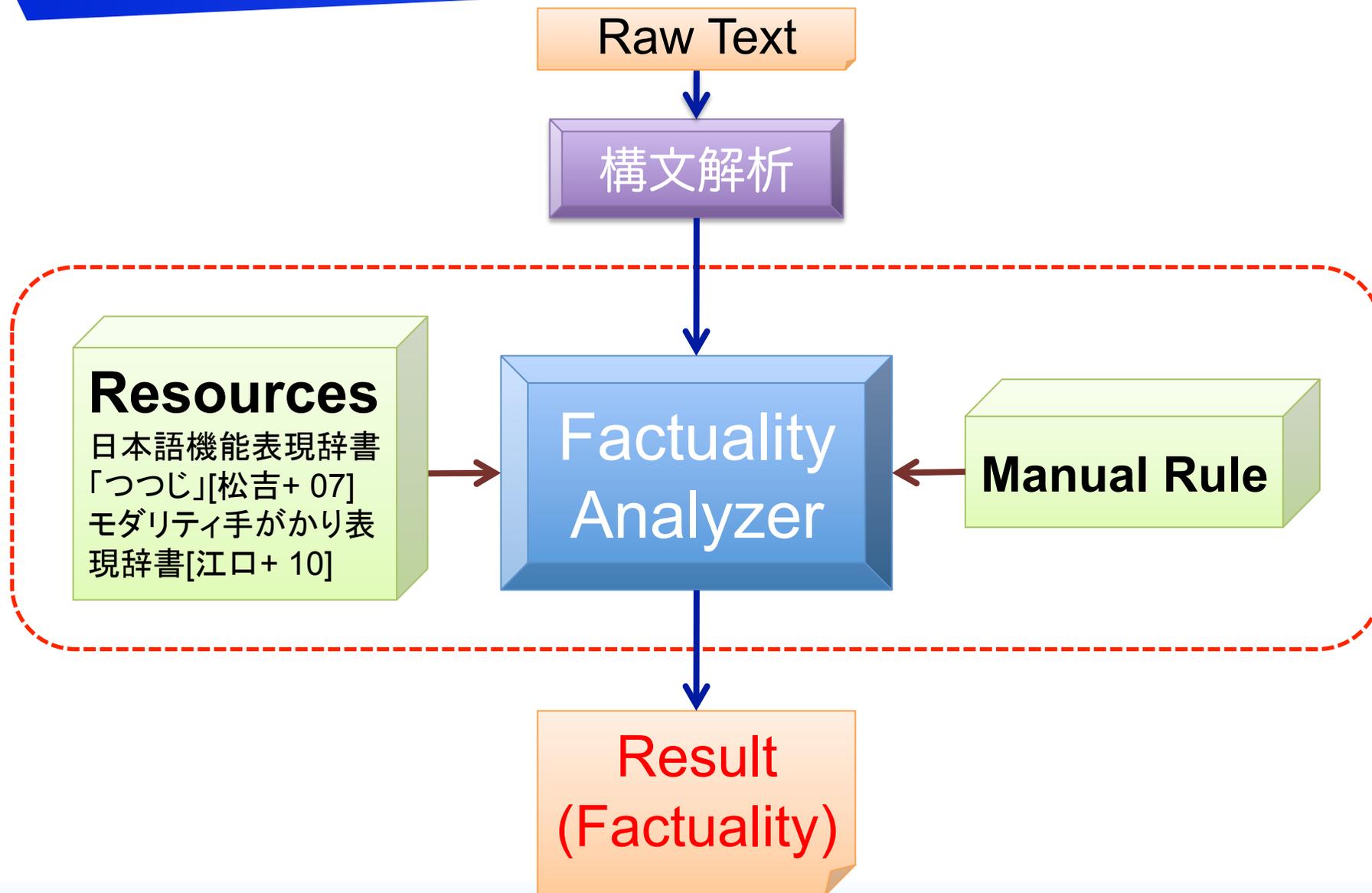
Contents

- Introduction
 - What is Factuality?
 - Motivation
- **Japanese Factuality Analysis**
 - **Defining Factuality**
 - **Lexical Knowledge**
 - **Algorithm**
- Experiment
- Error Analysis
- Conclusion and Future Work

Japanese Factuality Analysis

- [Saurí 08] の手法をベースに実装
 - 事実性を，確信度と肯否極性との組で定義
 - 事象の成立に影響を与える手がかり表現を利用
 - 事実性を依存構造木の根から伝搬させて解析

Japanese Factuality Analysis: Overview



Defining Factuality

- [Saurí 08] では, Factuality Valueを以下のように, 確信度と肯否極性の組として定義

肯否極性

確信度

	Positive (+)	Negative (-)	Underspecified (u)
Certain (CT)	Fact: <CT,+>	Counterfact: <CT,->	Certain but unknown output: <CT,u>
Probable (PR)	Probable: <PR,+>	Not probable: <PR,->	(NA)
Possible (PS)	Possible: <PS,+>	Not certain: <PS,->	(NA)
Underspecified (U)	(NA)	(NA)	Unknown or uncommitted: <U,u>

- 確信度 : ProbableとPossibleの境界が判別しづらい
 - Certain, Probable, Underspecifiedの3段階として考慮

Defining Factuality: Example

- 彼はさきほど部屋を出た。
 - 著者から見て、実際に起こったこと
 - CT+
- もう遅いから、彼は先に帰ったんだらう。
 - 著者から見て起こった可能性が高いこと
 - PR+
- 問題が発生するのを防いだ。
 - 著者から見て実際には起こっていないこと
 - CT-

Defining Factuality: Relation to Related Work

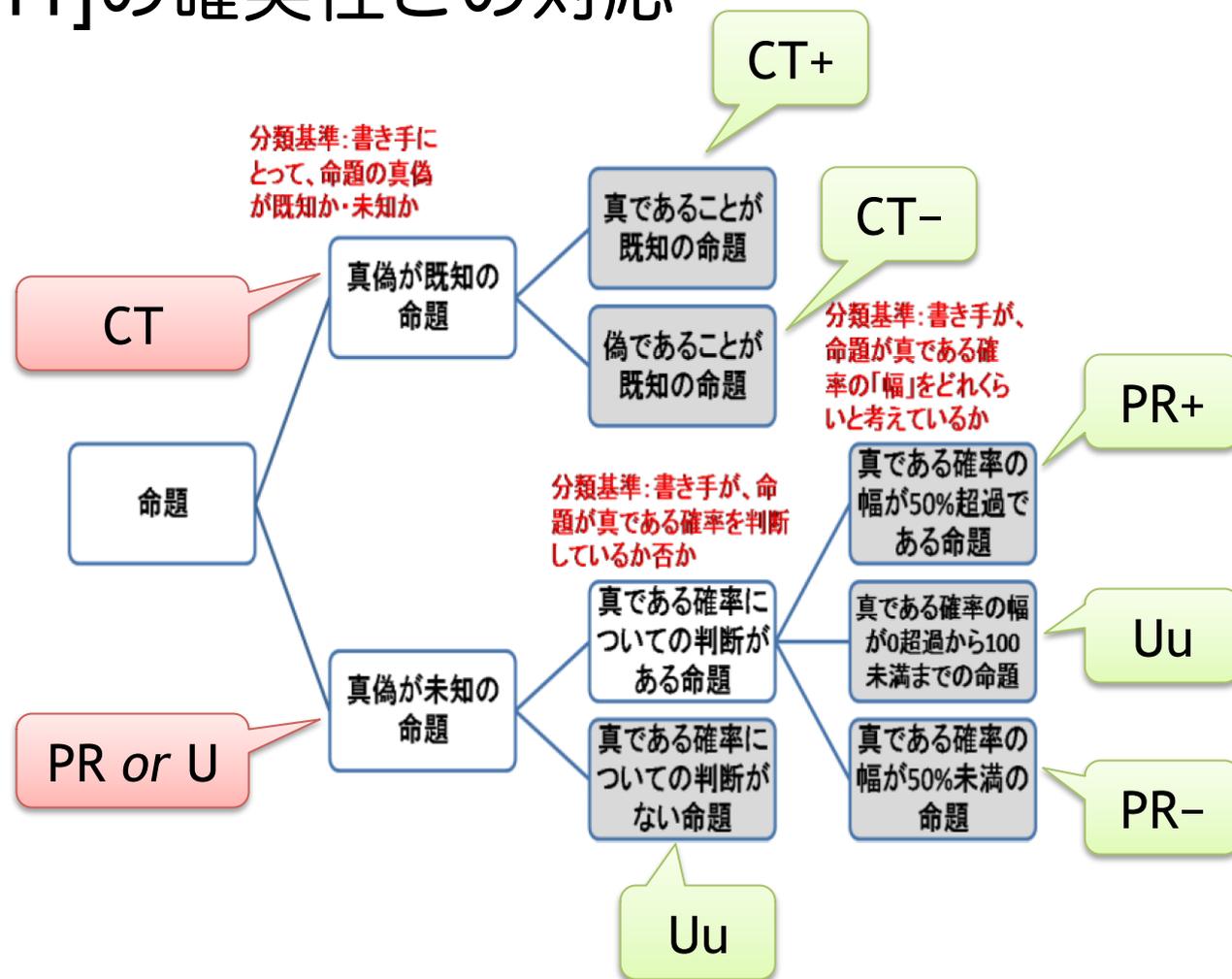
- [松吉+ 10]の真偽判断ラベルとの対応

表 3 真偽判断ラベルと事実性との対応

確信度 \ 肯否極性	+	-	u
CT	成立 不成立から成立 (CT+)	不成立 成立から不成立 (CT-)	0 (Uu)
PR	高確率 低確率から高確率 (PR+)	低確率 高確率から低確率 (PR-)	
U			

Defining Factuality: Relation to Related Work

- [川添+ 11]の確実性との対応



Lexical Knowledge

- 事実性に影響を与える手がかり表現の語彙知識
 - 機能語：日本語機能表現辞書「つつじ」[松吉+ 07]
 - 内容語：モダリティ解析手がかり表現辞書[江口+ 10]
- 例
 - もう遅いから、彼は先に帰ったんだらう。
 - 「だらう」：推量を表す機能語
→ 「帰る」 = PR+
 - 問題が発生するのを防いだ。
 - 「防いだ」：下位事象の肯否極性を反転させる内容語
→ 「発生する」 = CT-

Lexical Knowledge

- 日本語機能表現辞書「つつじ」[松吉+ 07]

表 1 機能表現辞書の意味クラス例と事実性への影響

機能表現辞書中の意味クラス	該当する機能表現	事実性への影響
否定	ない ぬ わけでない	肯否極性: $+\rightarrow -$, $-\rightarrow +$
推量	だろう かもしれない らしい	確信度: $CT\rightarrow PR$
疑問	か かな ではないか	確信度: $CT\rightarrow U$, $PR\rightarrow U$ 肯否極性: $+\rightarrow u$, $-\rightarrow u$

Lexical Knowledge

- モダリティ解析手がかり表現辞書[江口+ 10]

表 2 手がかり表現辞書の見出し語例と対応する事実性

見出し語	下位事象の事象形式	上位事象の肯否環境	上位事象の真偽判断	上位事象に対する事実性
防ぐ	スル形	肯定	不成立	CT-
		否定	成立	CT+
	タ形	肯定	-	-
		否定	-	-
忘れる	スル形	肯定	不成立	CT-
		否定	成立	CT+
	タ形	肯定	成立	CT+
		否定	成立	CT+

Algorithm

- 例：彼が出場を断念したことを相手は知らない。

Algorithm lines	依存構造木	手がかり表現	contextual factuality	event factuality
2			CT+	
(4-7) 8-11 12-15	知らない	polarity: - (ない)	CT-	知る: CT-
(4-7) (8-11) (12-15)	相手は ことを	modality: CT (知る) polarity: + (知る)	CT+	
(4-7) 8-11 12-15	断念した	polarity: - (断念する)	CT+	断念する: CT+
(4-7) 8-11 (12-15)	彼が 出場を		CT-	出場: CT-

Contents

- Introduction
 - What is Factuality?
 - Motivation
- Japanese Factuality Analysis
 - Defining Factuality
 - Lexical Knowledge
 - Algorithm
- **Experiment**
- **Error Analysis**
- **Conclusion and Future Work**

Experiment

- 拡張モダリティタグ付与コーパス[松吉+ 10] のOC (Yahoo!知恵袋) 6,404文に対して適用
 - 入力：
 - 正解の形態素情報→構文解析をした結果
 - 事実性を付与すべき事象の情報
 - 出力：
 - 各事象に対する事実性
 - 評価：
 - それぞれのラベル (CT+, CT-, PR+, PR-, Uu) に対する二値分類
 - Precision, Recall, F1-scoreを計算

Experiment: Result

表 4 それぞれのラベルに対する評価

	CT+	PR+	PR-	CT-	Uu	Micro-Average
TP	7,140	141	7	631	1,893	(9,812)
FP	1,467	286	71	583	2,698	(5,105)
FN	2,791	794	168	408	944	(5,105)
Precision	0.83	0.33	0.09	0.52	0.41	0.66
Recall	0.72	0.15	0.04	0.61	0.67	0.66
F_1	0.77	0.21	0.06	0.56	0.51	0.66

表 5 Confusion Matrix

正解 \ 出力	CT+	PR+	PR-	CT-	Uu	Total
CT+	7,140	237	25	408	2,121	9,931
PR+	528	141	8	23	235	935
PR-	47	6	7	67	48	175
CT-	97	1	16	631	294	1,039
Uu	795	42	22	85	1,893	2,837
Total	8,607	427	78	1,214	4,591	14,917

語彙知識の組み合わせでそれなりに解けた
(PR-はほとんど解けていない)

Contents

- Introduction
 - What is Factuality?
 - Motivation
- Japanese Factuality Analysis
 - Defining Factuality
 - Lexical Knowledge
 - Algorithm
- Experiment
- **Error Analysis**
- Conclusion and Future Work

Error Analysis

- 論点：
 - 現在利用可能な語彙知識のカバレッジ
 - 語彙知識の拡充・拡張の必要性
 - 語彙知識の曖昧性の有無
 - 局所的な語彙知識の組み合わせだけで解析不可能な問題の有無

Error Analysis

- 論点：
 - 現在利用可能な語彙知識のカバレッジ
 - 語彙知識の拡充・拡張の必要性
 - 語彙知識の曖昧性の有無
 - 局所的な語彙知識の組み合わせだけで解析不可能な問題の有無
- 肯否極性に関する200の誤り事例を分類：
 - i. 語彙知識のカバレッジの問題（内容語16%，機能語8%）
 - ii. 語彙知識としては存在しているが，表記揺れなど，辞書引きが難しい問題（2%）
 - iii. 語彙的手がかりの多義性の問題（9%）
 - iv. 手がかりとなる表現のスコープの問題（37%）

Error Analysis

- i. 語彙知識のカバレッジの問題（内容語16%，機能語8%）
 - サーバーは接続を解除しました。
 - 「接続」正解：CT-，出力：CT+

 - 説明しにくいかもしれませんが、教えてください。
 - 「説明する」正解：PR-，出力：PR+

語彙知識の拡充が必要

Error Analysis

- ii. 語彙知識としては存在しているが，表記揺れなど，辞書引きが難しい問題（2%）
 - 関東だから違うかもしれない。
 - 「違う」正解：PR+，出力：PR-

機能表現辞書により，
大抵の表現はカバーできている！

Error Analysis

iii. 語彙的手がかりの多義性の問題 (9%)

- 歩けば交通費はかからないか、いい運動にもなるしね！
 - 「かかる」正解：CT-, 出力：Uu

局所的な文脈だけを用いて、どこまでの問題が解け、そこから先の問題はどの程度残るのかを考える必要がある

Error Analysis

iv. 手がかりとなる表現のスコープの問題 (37%)

- 車が見えたがぶつかるのを避け切れなかった。
 - 「見える」正解：CT+, 出力：CT-

述語の項に関するメーカーなどの知識,
項の範囲を同定するための世界知識
などが必要

Error Analysis

v. その他

- 構文解析誤り
- アノテーション誤り

- 限定修飾などの連体修飾に関する誤り
 - 信用できない人には貸しません。
 - 正解：CT-, 出力：CT+

限定修飾の場合の事実性はどう考えるべきか

Contents

- Introduction
 - What is Factuality?
 - Motivation
- Japanese Factuality Analysis
 - Defining Factuality
 - Lexical Knowledge
 - Algorithm
- Experiment
- Error Analysis
- **Conclusion and Future Work**

Conclusion and Future Work

- 語彙知識の組み合わせに基づいた事実性解析
 - i. 語彙知識のカバレッジの問題
 - ii. 語彙知識としては存在しているが、表記揺れなど、辞書引きが難しい問題
 - iii. 語彙的手がかりの多義性の問題
 - iv. 手がかりとなる表現のスコープの問題
- 今後の課題
 - スコープの問題の解決に向けて
 - 限定修飾の扱い
 - 語彙知識の拡充

References

- 松吉俊, 江口萌, 佐尾ちとせ, 村上浩司, 乾健太郎, 松本裕治. テキスト情報分析のための判断情報アノテーション. 電子情報通信学会論文誌D, Vol. J93-D, No.6, pp. 705-713, 2010.
- 川添愛, 齊藤学, 片岡喜代子, 崔栄殊, 戸次大介. 确实性判断のためのアノテーション済みコーパスの構築. 言語処理学会第17回年次大会発表論文集, pp. 143-146, 2011.
- Roser Saurí. A factuality profiler for eventualities in text. PhD thesis, Brandeis University, 2008.
- 松吉俊, 佐藤理史, 宇津呂武仁. 日本語機能表現辞書の編纂. 自然言語処理, Vol. 14, pp.123-146, 2007.
- 江口萌, 松吉俊, 佐尾ちとせ, 乾健太郎, 松本裕治. モダリティ、真偽情報、価値情報を統合した拡張モダリティ解析. 言語処理学会第16回年次大会発表論文集, pp. 852-855, 2010.