

DRS意味解析における出現位置を 利用した語彙数削減 (+a)

黒澤 友哉, 谷中 瞳 (東京大学)

March 17, 2023 14:45 – 15:00



THE UNIVERSITY OF TOKYO

自己紹介

黒澤 友哉（くろさわ ともや）

所属

東京大学情報理工学系研究科コンピュータ科学専攻
谷中研究室 M1（4月から M2）

出身

神奈川県秦野市

経歴

- 厚木高校（神奈川）
- 東京大学 理科二類 → 理学部情報科学科

自己紹介

性格など

- 論理学への興味
 - 記号論理学、論理アプローチ（定理証明、モデル検査）
- ブラックボックスに対する忌避感
 - プロービング (probing)
 - 説明可能性
 - 言語モデルとオープンな処理の両立

意味解析とは

- 自然言語から**構造的に解釈可能な意味表現**に変換するタスク
- 計算機による言語理解の実現に向けた**基礎タスク**の一つ

Yuri eats an apple. $\rightarrow$ $yuri(x) \wedge apple(y) \wedge eat(x, y)$

談話表示構造 (DRS, Discourse Representation Structure)

- 次の2つから構成される「箱」をもつ（箱形式 DRS）
 - 談話指示子（変数: **t1**, **e1**）の集合
 - 条件式（命題: **t1** < “now”）の集合

t1	e1	[b1]
t1 < “now”		time.n.08(t1)
Time(e1 , t1)		DayOfMonth(t1 , “31”)
Patient(e1 , “speaker”)		MonthOfYear(t1 , “05”)
bear.v.02(e1)		YearOfCentury(t1 , “1940”)

図: *I was born on the 31st of May in 1940* を表す DRS

談話表示構造 (DRS)

- 一文に対し、複数の箱をもつこともある
- 箱同士で関係性が定義可能（論理結合子など）
- テンスや時間関係など、**様々な言語現象を表せる高度な意味表現形式**

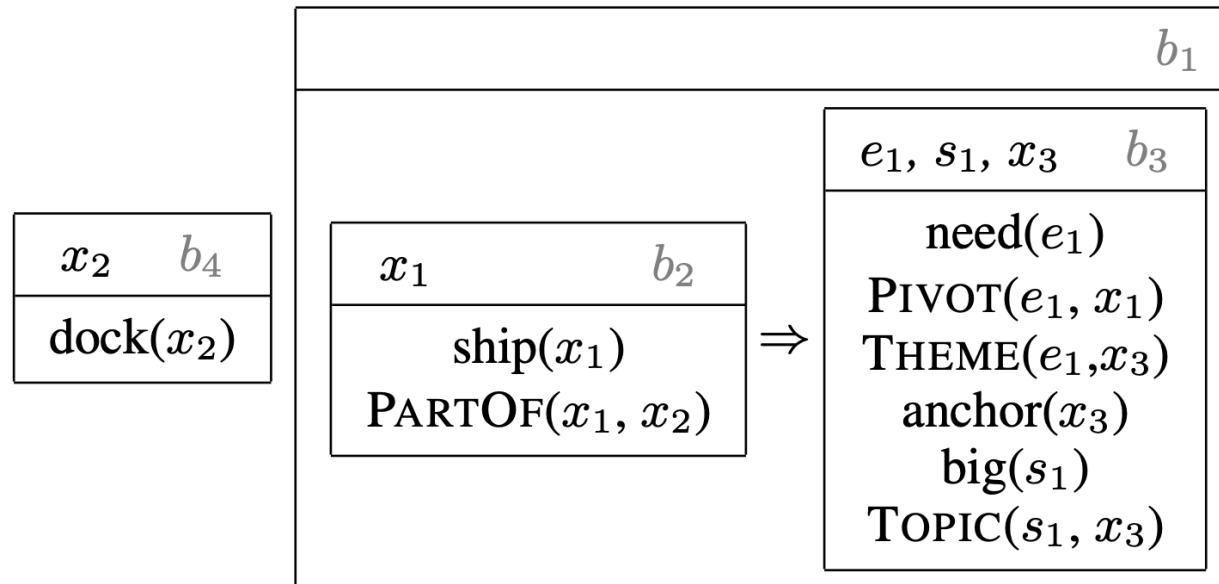


図: *Every ship in the dock needs a big anchor* の箱形式 DRS

DRS の表現形式

相対句形式 DRS

- 各命題や箱間の関係性を1行（句）に表した文字列
- 箱形式 DRS と比較し、**計算機上で扱いやすい**

\$NEW TPR @1 “now”

\$0 Time @2 @1

\$0 REF

\$0 REF

\$0 Patient @0 “speaker”

\$0 bear “v.02” @0

\$0 DayOfMonth @-1 “31”

\$0 time “n.08” @-1

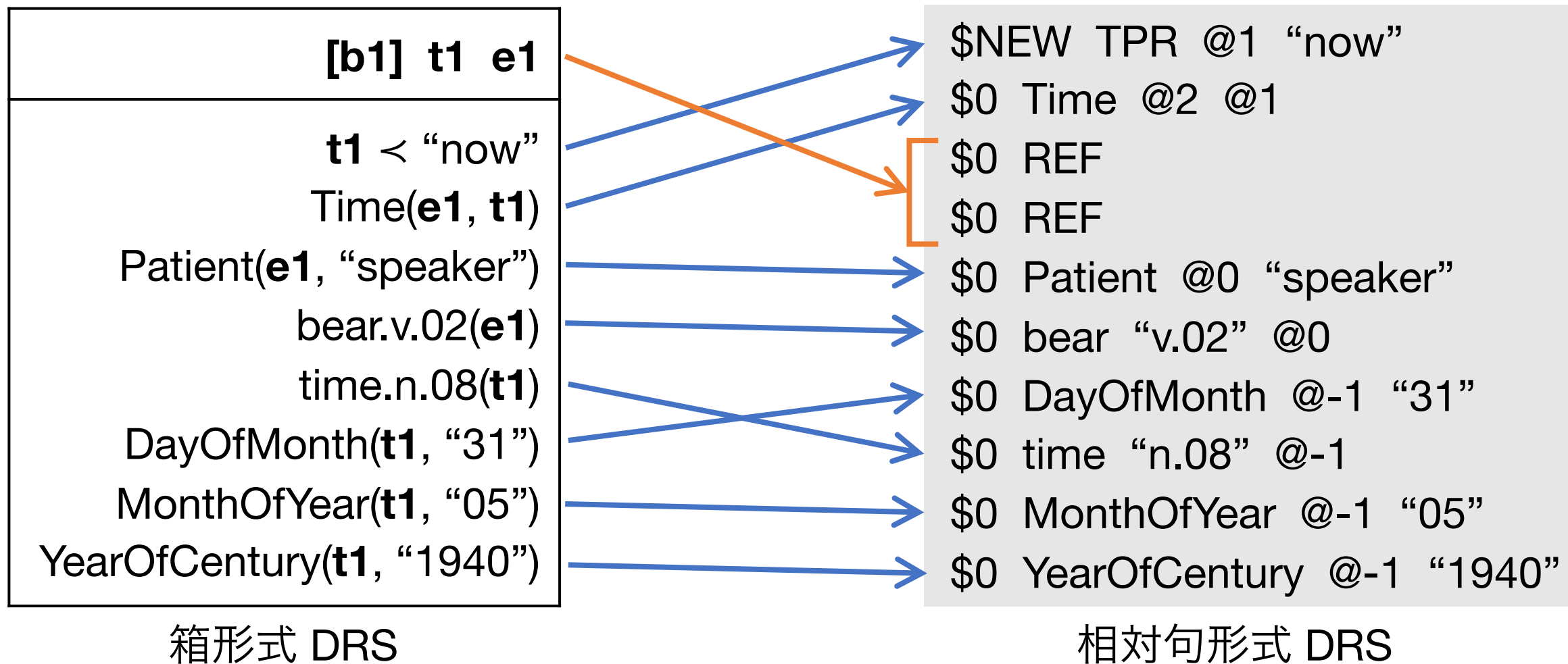
\$0 MonthOfYear @-1 “05”

\$0 YearOfCentury @-1 “1940”

図: *I was born on the 31st of May in 1940* を表す相対句形式 DRS

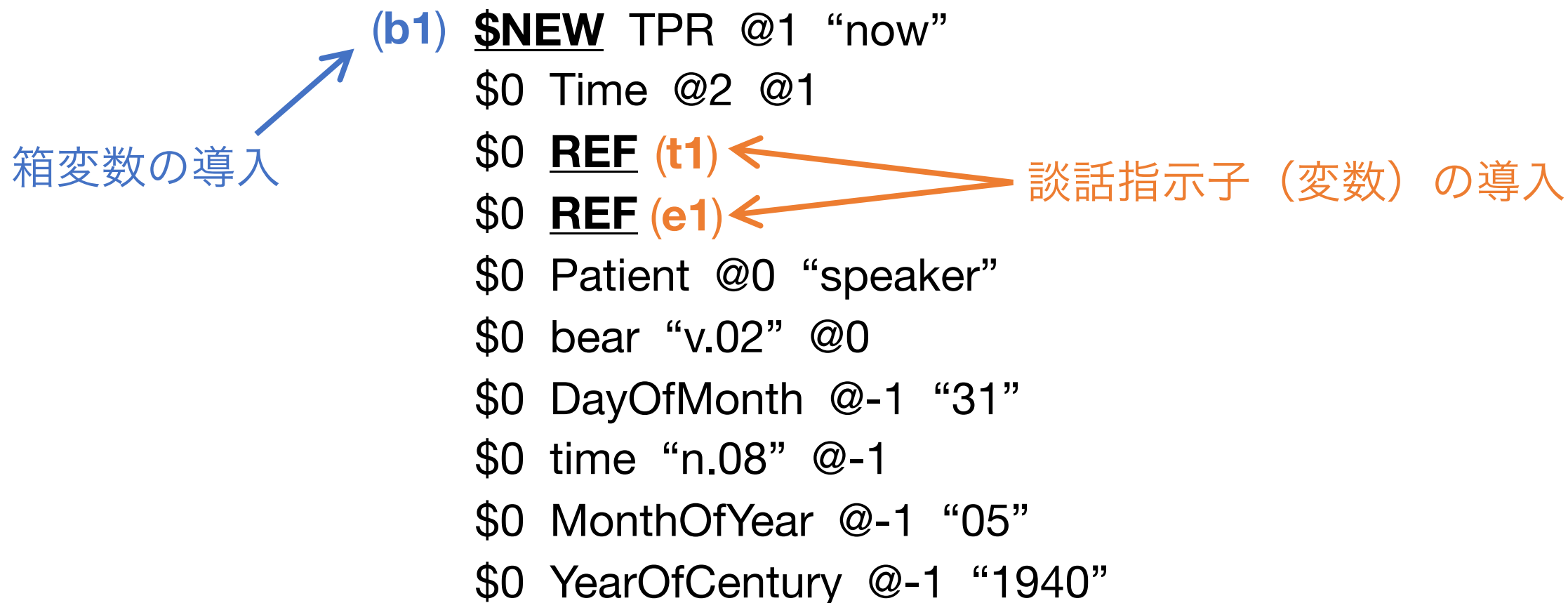
箱形式 DRS と相対句形式 DRS

箱形式 DRS の条件式と相対句形式 DRS の句は一対一対応する



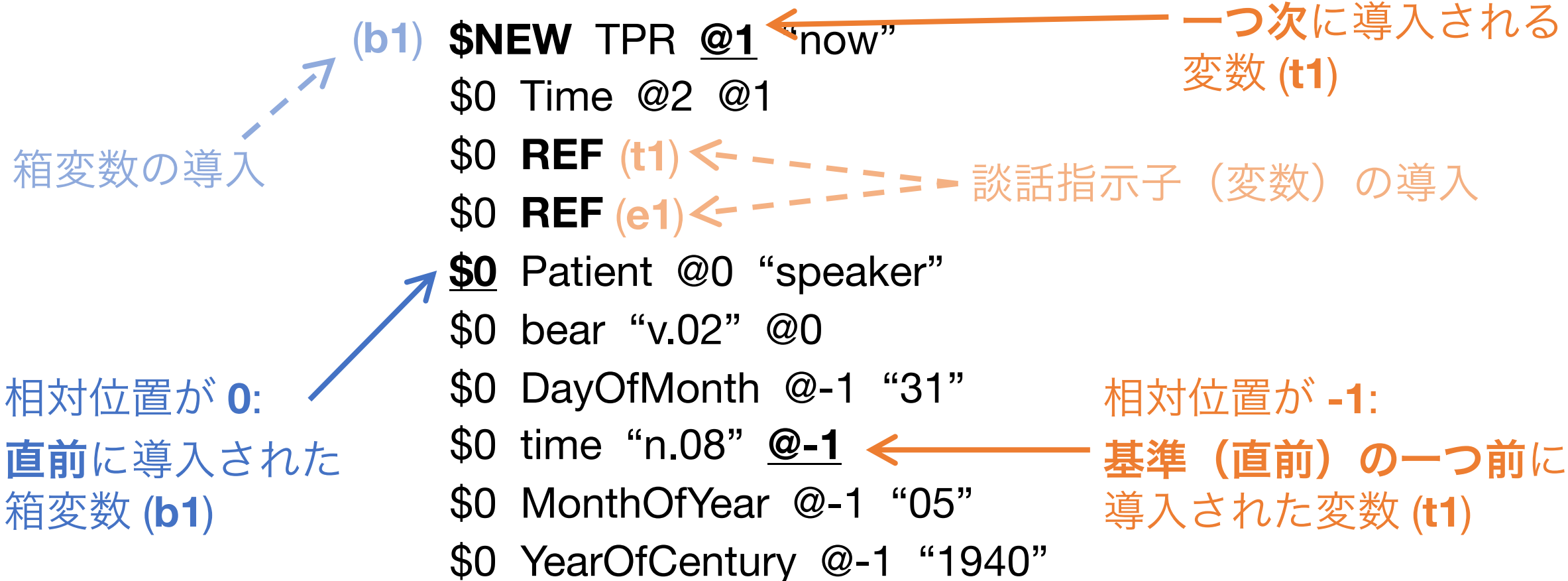
相対句形式 DRS の変数

各変数は、導入を基準とした相対位置で示される



相対句形式 DRS の変数

各変数は、導入を基準とした相対位置で示される



相対句形式 DRS の句の形式

相対句形式 DRS の句は、以下の4つのタイプに限定される

タイプ	項目数	例
変数宣言	2	\$0 REF
箱間の関係	3 or 4	\$0 PRESUPPOSITION \$NEW
意味	4	\$0 bear “v.02” @0 \$0 mobile_phone “n.01” @0
役割	4	\$0 DayOfMonth @-1 “ 31 ” \$0 Name @0 “ cristina ”

様々な単語や値が入り得る

本研究の動機とアプローチ

動機

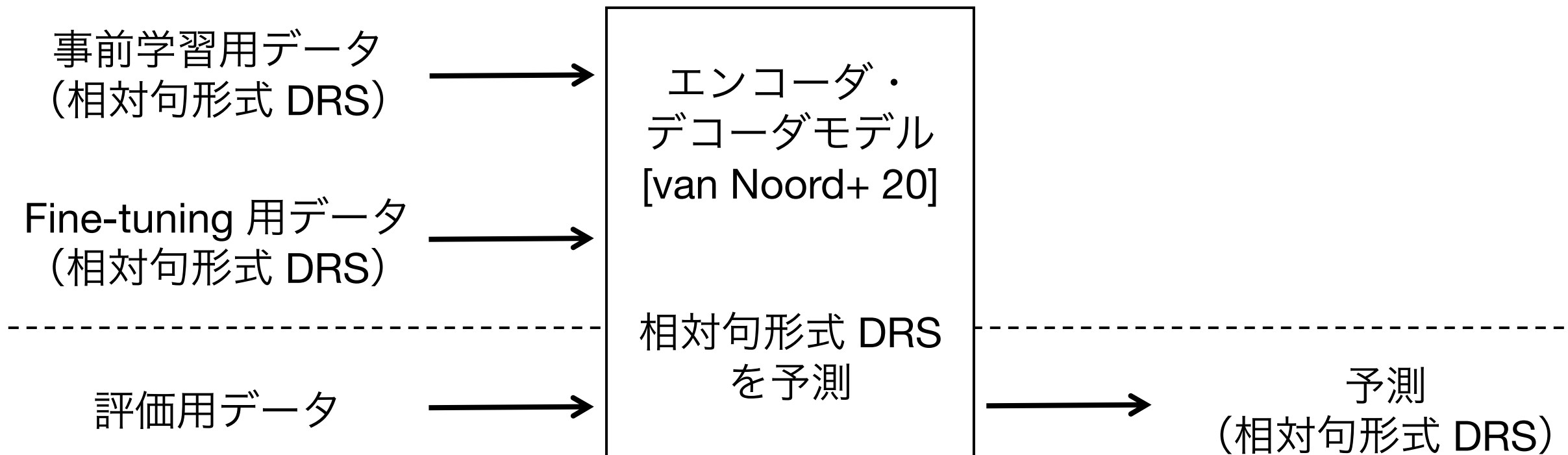
- 語義や役割を表す句について、**出力すべき語彙の数が多い**
 - 低頻度語や未知語の予測でスコアが下がりやすい

アプローチ

- 前処理と後処理を行ない、**言語モデルが出力すべき語彙を減らすこと**でスコア上昇を図る

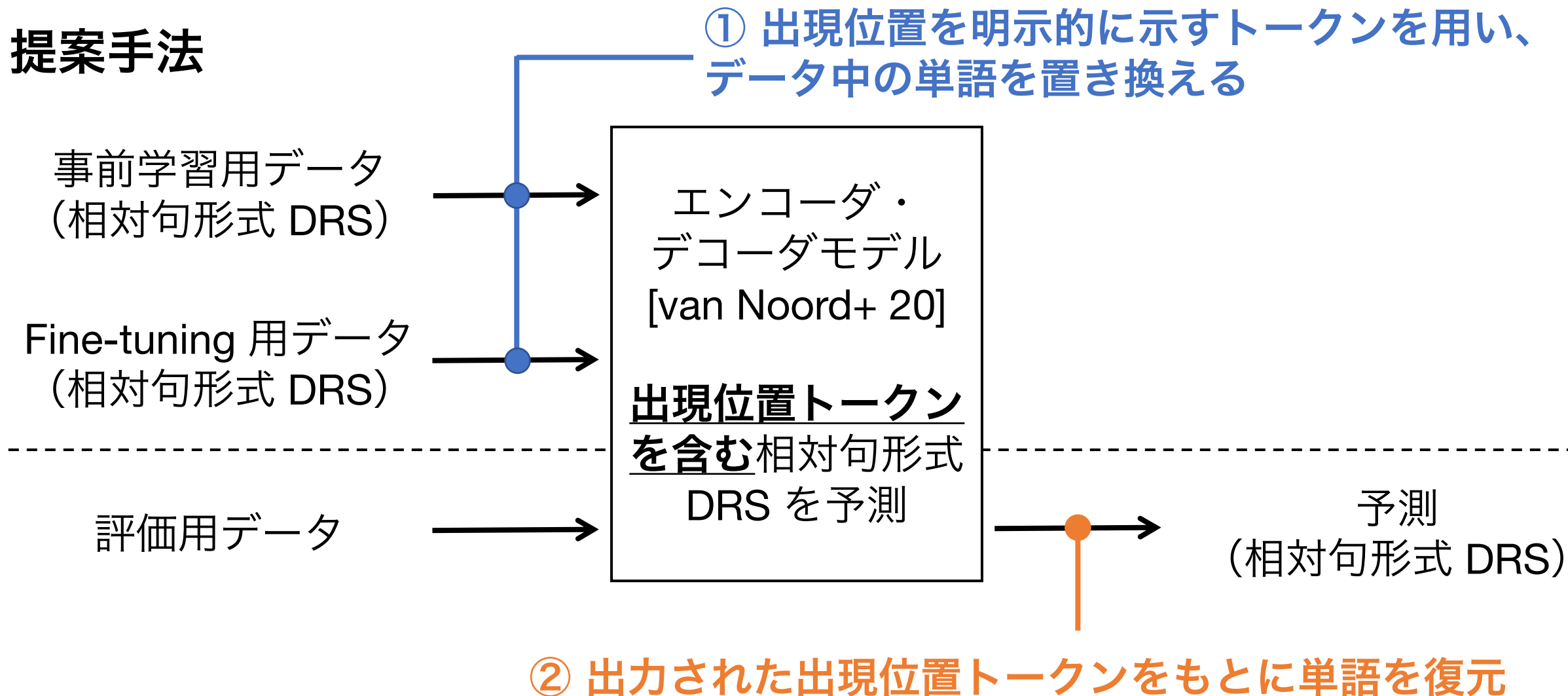
システムの全体図と提案手法

先行研究 [van Noord+ 20]



システムの全体図と提案手法

提案手法



置換対象の単語

DRS に含まれるトークンのうち、解析対象の文にも含まれるトークン

文 I was **born** on the **31st** of **May** in **1940**.

t1 e1	[b1]
t1 < “now”	time.n.08(t1)
Time(e1 , t1)	DayOfMonth(t1 , “31”)
Patient(e1 , “speaker”)	MonthOfYear(t1 , “05”)
bear.v.02 (e1)	YearOfCentury(t1 , “1940”)

置換例

文 I was **born** on the **31st** of **May** in **1940**.

相対句形式 DRS

\$NEW TPR @1 “now”
 \$0 Time @2 @1
 \$0 REF
 \$0 REF
 \$0 Patient @0 “speaker”
 \$0 **bear** “v.02” @0
 \$0 DayOfMonth @-1 “**31**”
 \$0 time “n.08” @-1
 \$0 MonthOfYear @-1 “**05**”
 \$0 YearOfCentury @-1 “**1940**”

提案手法を施した DRS

\$NEW TPR @1 “now”
 \$0 Time @2 @1
 \$0 REF
 \$0 REF
 \$0 Patient @0 “speaker”
 \$0 **#2** “v.02” @0
 \$0 DayOfMonth @-1 “**#5**”
 \$0 time “n.08” @-1
 \$0 MonthOfYear @-1 “**#7**”
 \$0 YearOfCentury @-1 “**#9**”

置換の詳細

2つのタイプの句において、出現位置トークンを利用する

語義タイプ 談話指示子に単語とその synset（語義）を与える

\$0 **bear** “v.02” @0

役割タイプ 談話指示子の役割として具体的な値を与える

\$0 DayOfMonth @-1 “**31**”

語義タイプの句の置換

- (1) 文中の単語を原形に変換
- (2) 原形に変換された単語が DRS のどの句に存在するか判定
- (3) 出現位置に # を付与したトークン（出現位置トークン）で置き換え

I was born on the 31st of May in 1940.

\$0 **bear** “v.02” @0 → \$0 **#2** “v.02” @0

役割タイプの句の置換

I was born on the 31st of May in 1940.

(a) 単語と完全一致するトークンが DRS 中に存在すれば置き換え

\$0 YearOfCentury @-1 “**1940**” → \$0 YearOfCentury @-1 “**#9**”

(b) 単語に対応する値が DRS 中に存在すれば置き換え

\$0 MonthOfYear @-1 “**05**” → \$0 MonthOfYear @-1 “**#7**”

\$0 DayOfMonth @-1 “**31**” → \$0 DayOfMonth @-1 “**#5**”

復元方法

語義タイプ 出力された synset トークンをもとに原形に変換して復元

入力文	I was <u>born</u> on the 31st of May in 1940.
予測した句	\$0 #2 “ <u>v</u> .02” @0
→ 復元結果	\$0 bear “v.02” @0

役割タイプ 学習時に観測した単語と DRS トークンの変換に関して辞書を構成し、その辞書をもとに復元

入力文	I was born on the <u>31st</u> of May in 1940.
予測した句	\$0 DayOfMonth @-1 “#5”
辞書	<u>31st</u> → 31
→ 復元結果	\$0 DayOfMonth @-1 “ 31 ”

実験: データセット

Parallel Meaning Bank (PMB)

- 大規模 DRS コーパスの一種
- 英語、ドイツ語、オランダ語、イタリア語の 4 言語
- アノテーションの質に応じて Gold, Silver, Bronze の 3 つに分割

PMB に含まれる英語のドキュメント数

	Gold			Silver	Bronze
Version	Train	Dev	Test	Train	Train
3.0.0	6,620	885	898	97,598	146,371

実験設定

モデルとエンコーダへの入力

van Noord et al. (2020) で最も F-score が高かった設定を採用

- bi-LSTM エンコーダを 2 つ導入したエンコーダ・デコーダモデル
- 各エンコーダへの入力
 - 一方は**トークン列と意味タグを交互に並べる**
 - 他方は**文字ごとに区切る**

実験設定

データセット

PMB 3.0.0 の英語データ

事前学習

Gold Train + Silver Train
融合セット (104,218)

ファインチューニング

Gold Train (6,620)

評価

Gold Test (898)

評価指標

評価セット全体に対するミクロ F-score の5回平均
(1回の事前学習、5回のファインチューニング)

結果

- 提案手法を施すことで先行研究よりも **F-score が上昇**
- 同時に**計算コストの減少**にも成功

	語彙数 ↓	F-score (%) ↑	事前学習 時間 (分) ↓	消費 GPU メモリ (GB) ↓
提案手法	1,081	89.9	186	8.98
[van Noord+ 20]	11,394	89.3	237	15.30

アブレーション分析

- ・ 語義タイプと役割タイプそれぞれのスコア上昇の度合いを分析
- ・ **X** は対応するタイプに対して**復元を行なわない**

→ 語義タイプは約 **15%**、役割タイプは約 **5%** のスコア上昇に貢献

語義タイプへの適用	役割タイプへの適用	F-score (%)
✓	✓	89.9
✓	X	84.6
X	✓	74.9

予測 DRS の ill-formed な形式に関する分析

相対句形式 DRS が **ill-formed** な形式とは

- どのタイプ (→ 11ページ目) にも属さない句の存在
- 変数参照の失敗 (相対位置で示された変数が存在しない等)

ill-formed な形式の出現数に関する分析

- 学習データ中の**出現頻度が低い言語現象**を含むケースで ill-formed が多かった
- 特に、**接続や量化を示す意味タグ**を有する文について、
予測 DRS の ill-formed の出現数は全 31 ケース中 10 ケース存在

まとめと今後の展望

- 文と共通する DRS トークンを出現位置トークンに置き換える手法を試みた
- 出力語彙数が大幅に減少したことにより、既存手法と比較し有意なスコア上昇が見られ、さらに事前学習時間と GPU メモリの消費量も減少した
- 今後は、PMB に含まれる英語以外の言語に対して、語彙数を抑制するアプローチを試みる予定である

宣伝

LINE株式会社 / NLPチームでの就業型インターンシップの成果報告

**【インターンレポート】 グラフベースで、機械学習を用いない
ニュース記事要約文の hallucination 検出**

<https://engineering.linecorp.com/ja/blog/graph-based-non-machine-learning-hallucination-detection-of-news-article-summary-sentences>

hallucination とは

幻惑。生成においては、入力の内容に対し**不誠実な内容**を言語モデルが**生成**してしまう現象のこと

LINE 株式会社

テキスト入力によるインタラクションの動作生成

3次元アバターやキャラクターを使った臨場感ある動画制作の際に、複数の人物が登場する場面はモデリングが必要です。DVLチームでは、独自のAIシミュレーションである動画を生成し、2人の共同動作を生成する研究を行っています。

大規模な言語モデルを活用したマルチモーダルな対話システムの研究開発

対話ロボットコンペティション2022用のシステム

対話システムライブコンペティション5 オープンラック用のシステム

のの中に写っている猫は至く「It tells silvestris catus」という品種であることがわかりました。

NLP って何の略ですか

❌ Neural Language Processingの略です。

❌ 人間の脳機能を模したニューラルネットワークを用いた機械学習の手法の総称です。

200 FFX NLP HELLO WORLD 路ですか

ご清聴ありがとうございました

Slack 上でのコメントなど頂けると大変うれしいです。