

関係代数に基づく推論の含意関係認識への応用

Inference based on Relational Algebra with Applications to Recognizing Textual Entailment

田 然*¹ 宮尾 祐介*²
Ran Tian Yusuke Miyao

*¹国立情報学研究所 *²国立情報学研究所
National Institute of Informatics National Institute of Informatics

In this paper we propose a novel semantic representation framework used for recognizing textual entailment. The framework uses semantically annotated dependency parses as logical representation, of which the semantics can be modeled by algebraic forms. This modelling provides not only faster automated reasoning, but also closer mappings between natural language and logic. Experiments show positive results on NTCIR RITE datasets.

1. はじめに

含意関係認識とは、文 H が文 T から推論されるかどうかを判別する問題であり、自然言語処理における難しい課題の一つとされる。例えば下の例では、文 T から文 H を推論することができるため、「含意関係あり」と判別する。

- T 川端康成は「雪国」などの作品でノーベル文学賞を受賞した。
- H 川端康成は「雪国」の著者である。

文の意味理解に関わる課題であるため、その実現手法として、文の意味を述語論理式で記述し、文の間の含意関係を定理とみなして証明を試みる方法や、これとは対照的に文の表層的な情報 (T と H において一致している単語の数など) しか使わない手法が提案されている。いくつかの評価型ワークショップ [Bentivogli 10][Shima 11] では、証明によるアプローチは表層的アプローチの成果をなかなか越えられなかった。この事実は、自然言語文に対して精密かつ堅牢な意味解析を行うことの難しさを示している。しかし、試験問題を解く場合のように正確な理解・判断が求められる課題では、文の表層的情報に止まらずより精密な意味解析を行う必要がある。例えば世界史の次のような選択問題を考えよう (正しい選択肢は H_2):

- H_1 賠償金取立てを強化するドーズ案によって、ドイツ経済が打撃を受けた。
- H_2 シュトレゼマンが、インフレーション対策のために改革を行った。

教科書には、それぞれに関連して次のような記述がある:

- T_1 翌 24 年には、アメリカの提案したドーズ案により賠償金支払いの条件がゆるめられ、さらにアメリカ資本の導入によりドイツ経済が復興してくると...
- T_2 1923 年に成立したシュトレゼマン内閣は、ルールでの抵抗を中止し、土地財産などを担保に新紙幣 (レンテンマルク) を発行して、インフレを収束させた。

間違った選択肢 1 に対しても、教科書からは「賠償金」「ドーズ案」「ドイツ経済」など一致する単語の多い文が見つかる

連絡先: {tianran, yusuke}@nii.ac.jp

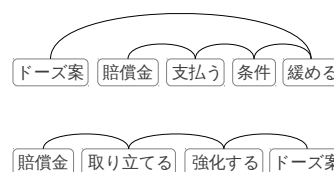


図 1: 係り受け関係

とがわかる。一致している単語数など、表面的な情報だけでは正解は選べないであろう。この例において正しい答えを得るためには、まず単語間のつながりを理解し、例えば T_1 において、「賠償金」を「支払う」ことに対する「条件」が「ゆるめられ」、このことは「ドーズ案」にかかることを認識し、一方 H_1 では「賠償金」を「取り立てる」ことを「強化する」のが「ドーズ案」だと主張していることを理解しなければいけないであろう (図 1)。次に、「支払う条件をゆるめる」と「取り立てを強化する」とは対立していることだと判断する必要がある。更に、図 1 で与えられた単語間のつながりをもって、「支払う条件をゆるめる」と「取り立てを強化する」ことが対立しているのであれば、全体の文「ドーズ案により賠償金支払いの条件がゆるめられ」と「賠償金取立てを強化するドーズ案」も対立していることを推論しなければならない。

このような推論を行うためには何らかの論理体系が必要となってくるが、従来の述語論理はこの課題に対して必ずしも適しているとは言えない [MacCartney 07]。本稿において我々は、含意関係認識に適する新しい意味表現と論理体系を提案し、それを応用したシステムが NTCIR10 RITE2 [Watanabe 13] の大学入試サブタスクでポジティブな結果を得たことを報告する。

我々の提案のアイデアは、DCS (Dependency-based Compositional Semantics) 木 [Liang 11] に似たような形で自然言語文の意味を表現することから始まる。DCS 木は係り受け木に近い構造を持ちながら厳密な論理の意味を表現する。それは一階述語論理式の一部しか表現できないが、本稿において、DCS 木が表現できる意味の全体はある種の代数系を成していることを示す。意味の空間を一階述語論理系からこの代数系に制限することで、自動推論が高速になるだけでなく、DCS 木を経由して我々は論理的な意味空間と自然言語との間のより直接的なマッピングを得る。論理的証明に基づく含意関係認識シ

システムにおいて、知識不足による脆弱性が精度低下の大きな原因の一つとされる [Bos 09] が、意味空間と自然言語との間の直接的なマッピングを用いて、我々は推論を進めるために必要な知識を言語的直感から動的に生成するコンポーネントを作成し、知識不足による脆弱性を克服した。

2. 関連研究

含意関係認識システムは情報検索・テキスト要約・質問応答・機械翻訳など幅広い分野に應用が期待されることから、近年注目を集めている。英語圏では 2005 年から続いた評価型ワークショップ RTE [Bentivogli 10]、そして日本語と中国語を対象に 2011 年から RITE [Shima 11] が開催された。

含意関係認識に対するアプローチは、大別して T と H との間の「類似度」を測る手法と、ある種の「変換」に基づく手法とに分けることができる [Androutsopoulos 10]。「類似度」を測る手法は、表層的な単語一致率から、関係抽出や述語項構造解析の結果を比較する手法、さらに構造的な木カーネル [Zanzotto 10] を用いる方法がある。一方「変換」に基づく手法は、文を論理式で表現し、推論規則を應用して T の表す論理式から H の表す論理式に演繹する方法 [Bos 09] や、構文木の交換規則を直接与えて T の形を書き換えながら H に近づける方法 [Stern 12] などが研究されている。

これらの手法を比較する尺度の一つに、言語や世界知識の取り入れ方に対する自由度がある。「類似度」を測る手法では、おもに単語レベルの知識（同義語・上位語・分布類似度など）を用いて文の類似度をアドホックに計算する方法が採られている。木カーネルは、含意関係認識の訓練データからより構造的な知識を学習できるが、データスパースが深刻な問題である [Zanzotto 10]。一方「変換」に基づく手法は、言語や世界知識を変換規則として自由に導入できる利点を持つ。その中でも論理的推論を用いる方法は、試験問題を解く際に必要であろう構造的な知識 [川添 12] を整合的に取り入れ、例えば「現実的にありえない」という判断を可能にすると期待できる。

文の意味を論理形式で表現する手法として、主に述語論理が使われ、自然言語文を述語論理式に変換するシステムとして英語圏では Boxer [Bos 08] などが広く使われているが、日本語に対して実用的な実装はまだ存在していない [稲田 13]。一方 Natural Logic は、自然言語それ自体に論理的構造を持たせる手法で、含意関係認識における有効性が確認されているが、この手法は必ずしも自明でない文のアライメントに強く依存していることや、時間関係などより高度な構造的知識を表現できないことなどの欠点がある [MacCartney 07]。

DCS は本来、自然言語文を関係データベースのクエリに変換する質問応答システムの枠組みである。DCS 木はデータベース・クエリに変換・実行され答えを出す点において（そして文を DCS 木に変換するパーザは潜在変数として質問と答えのペアから学習される点において）、この枠組みは関係データベースを必要とする。しかしデータベース・クエリ自体を一種の論理的意味と見なせば、これは DCS 木を用いた意味表現の枠組みととらえられる。我々はこの枠組みを述語論理の言葉で再定式化し、DCS 木が表現できる意味の範囲を代数系の形で明確にした。これは、DCS 木による意味表現を含意関係認識に應用するために必要不可欠な理論的な仕事である。

3. 関係代数に基づく意味表現

DCS 木はルート付きの木で、各ノードに述語、各辺に意味役割のペアがラベル付けされている（図 2）。また、DCS 木

Government reported urgent news to citizens.



図 2: 基本 DCS 木

の辺に全称量化マーク、ノードに選択子が付けられる場合もある。全称量化マークを含まない DCS 木を基本 DCS 木と呼ぶ。ここでまず、DCS 木の定める一階述語論理的な意味を定義し、次にその意味が代数形式で表現できることを説明する*1。

3.1 制約充足問題

形式的に、DCS 木 $\mathcal{T} = (\mathcal{N}, \mathcal{E})$ のノード $\sigma \in \mathcal{N}$ に対応する述語を $p(\sigma)$ とおく。辺 $e = (\sigma, \sigma') \in \mathcal{E} \subset \mathcal{N} \times \mathcal{N}$ に意味役割のペア (r, r') が対応する時、 σ をルートに近いノードとし、 r を σ の入力役割、 r' を σ' の出力役割と呼ぶ。また、DCS 木 \mathcal{T} に対してノード σ と $p(\sigma)$ の取りうる意味役割 (σ の入力役割や出力役割など) r のペア (σ, r) を \mathcal{T} の芽と呼ぶ。

基本 DCS 木は、芽を変数とする制約充足問題を定める。芽 (σ, r) に対応する変数を x_{σ_r} とすると、制約は

1. 辺 (σ, σ') に役割 (r, r') が付与されたら、 $x_{\sigma_r} = x_{\sigma'_r}$ 。
2. $p(\sigma)$ の取りうる意味役割の集合を $\{r_1, r_2, \dots\}$ とすると $(x_{\sigma_{r_1}}, x_{\sigma_{r_2}}, \dots) \in \mathcal{I}_{p(\sigma)}$ 。

と定義される。ここで $\mathcal{I}_{p(\sigma)}$ は、元々の DCS 枠組みにおいては関係データベースにおける表に相当し、 $\mathcal{I}_{p(\sigma)}$ の各元は表にある値の組に相当する。我々の枠組みではこの制限を外し、 $\mathcal{I}_{p(\sigma)}$ を単に述語 $p(\sigma)$ に対応する集合とみなす。

図 2 の DCS 木が定める制約充足問題をその右に示した。基本 DCS 木 \mathcal{T} の定める制約充足問題を $CSP(\mathcal{T})$ と書き、 $CSP(\mathcal{T})$ 中の自由変数を y_1, \dots, y_m とおくと、 \mathcal{T} の述語論理の意味を $\exists y_1, \dots, \exists y_m, CSP(\mathcal{T})$ として定義する。また、変数 y_i に対してその外延 $D(y_i; \mathcal{T})$ を

$$D(y_i; \mathcal{T}) = \{y_i | \exists y_1, \dots, \hat{y}_i, \dots, \exists y_m, CSP(\mathcal{T})\}$$

と定義すれば*2、 \mathcal{T} の意味 $\exists y_1, \dots, \exists y_m, CSP(\mathcal{T})$ は、すべて（または何れか一つ）の y_i に対して $D(y_i; \mathcal{T}) \neq \emptyset$ であることと論理的に同値である。

3.2 選択子

DCS 木のノード σ に対して、選択子は $x_{\sigma_{r_1}}, \dots, x_{\sigma_{r_n}}$ を自由変数とする論理式を定める。この論理式を制約充足問題の制約に加えることで、DCS 木の表現力を高めることができる。たとえば「二個」、「半分」のような数量表現や「一番高い」のような最上級表現は選択子によって表現できる。

3.3 全称量子

全称量子を含む文の意味を表現したい時、図 3 のように、DCS 木の辺に q_{all} というマークをつける。辺 (σ, σ') に役割 (r, r') と全称量子マーク q が付与されたら、変数 x_{σ_r} に全称量子が付き論理式が得られる。具体的には、 σ' をルートとする部分木を $\mathcal{T}_{\sigma'}$ とおくと、 x_{σ_r} を自由変数として含む論理式 F に対して、以下 3 種類の全称量子マークを定義する：

*1 理論の再定式化によって我々の DCS 木の定義は元の枠組みと少し異なるが、「DCS 木」という用語を援用する。

*2 \hat{y}_i で i 番目の添字が抜けていることを表す。

Every owner of a siamese cat loves a therapist.

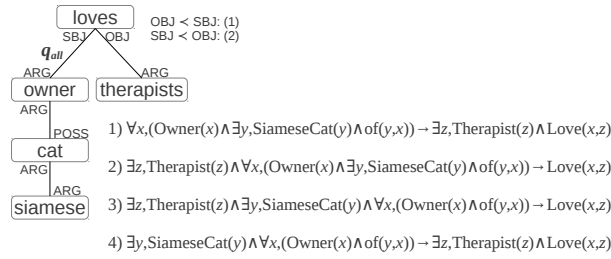


図 3: 全称量子子を含む DCS 木

1. $q_{all}(F): \forall x_{\sigma_r}, x_{\sigma_r} \in D(\sigma'_r; \mathcal{T}_{\sigma'}) \rightarrow F$
2. $q_=(F): \forall x_{\sigma_r}, x_{\sigma_r} \in D(\sigma'_r; \mathcal{T}_{\sigma'}) \leftrightarrow F$
3. $q_{no}(F): \forall x_{\sigma_r}, x_{\sigma_r} \in D(\sigma'_r; \mathcal{T}_{\sigma'}) \rightarrow \neg F$

ここで $D(\sigma'_r; \mathcal{T}_{\sigma'})$ は外延を表す。

基本 DCS 木に対しては、制約充足問題を存在量化した論理式が文の意味となるが、一般には DCS 木の各ノードが一つの論理式を定め、文の意味はこれらの論理積として定義する。具体的には、ノード σ に対して、 $p(\sigma)$ の取れる意味役割の集合に順序を定め、 $r_1 \prec \dots \prec r_n$ とする。 σ を含む最大の基本 DCS 部分木を \mathcal{T}_σ とおくと、 σ が定める論理式は $F_\sigma = CSP(\mathcal{T}_\sigma)$ から出発して $r_1 \prec \dots \prec r_n$ に沿って次のように定義する：

1. F_σ の中で $x_{\sigma_{r_1}}, \dots, x_{\sigma_{r_n}}$ 以外の自由変数を y_1, \dots, y_m とおく。各 $1 \leq i \leq n$ に対して、 r_i が σ の出力役割であれば $F'_{i-1} = \exists y_1, \dots, \exists y_m, F_{i-1}$ 、そうでなければ $F'_{i-1} = F_{i-1}$ とおく。
2. σ の入力役割が r_i となるような辺を集める。辺 e に対して次のように論理式を定義する：
 - (a) e に全称量子化マークが付いていなければ $\exists x_{\sigma_{r_i}}, F'_{i-1}$
 - (b) e に全称量子化マーク q が付いていれば $q(F'_{i-1})$
3. 上で得られたすべての論理式の論理積を F_i とおく

最後に得られた F_n がノード σ に対応する論理式である。

3.4 表現力について

DCS 木は限られた一階述語論理式しか表現できないが、全称量子化などが複雑な入れ子構造になっていない「通常わかりやすい」自然言語文に対して、この枠組みは十分な表現力を持つと思われる。例えば図 3 の文に対して、量子化スコープの違いによる 4 種類の読みが考えられるが、「loves」ノードにおける意味役割の順序の違いで (1)(2) の読みに相当する論理式が生成される。(3)(4) の読みは DCS 木から生成できないが、これは木構造によって「cat」における存在量子子のスコープは自動的に「owner」における全称量子子のスコープの中に置かれたからであり、実際 (3)(4) は不自然な読みと考えられる。

3.5 代数形式

DCS 木は、単に表現できる一階述語論理式に対して制限を設けたわけではなく、この枠組みの大きな特徴の一つは、DCS 木の定める制約充足問題は関係代数の演算を用いて明示的に解けることにある。

DCS 木の定める制約の定義において我々は集合 $\mathcal{J}_{p(\sigma)}$ を用いたが、この集合は記号として与えられたと考え、すべての述



図 4: システム構成

語 p に対して記号 \mathcal{J}_p を用意する。また、変数が取りうるすべての値の集合を W とおく。すると、述語 p が取りうる意味役割を r_1, \dots, r_n とすれば、 $\mathcal{J}_p \subset W_{r_1} \times \dots \times W_{r_n}$ となる^{*3}。 \mathcal{J}_p や W に関係代数の演算を施すことによって、我々は外延 $D(\sigma_r; \mathcal{T})$ の明示的な表現を得ることができる。

我々が用いる関係代数の演算を以下で示す。

- \times, \cap, π : 直積、交わり、射影。
- 各タイプの選択子 s に対する選択演算 s : 選択演算は任意の集合 A をその部分集合 $s(A) \subset A$ に写す写像だと考える。
- 全称量子化マークに対する商演算: 集合 $A \subset W^n, B \subset W$ に対して $q_{all}^i(A, B) \subset W^{n-1}$ を次のように定義する：

$$\{(x_1, \dots, \hat{x}_i, \dots, x_n) \mid \forall x_i, x_i \in B \rightarrow (x_1, \dots, x_n) \in A\}$$

同様に $q_=(A, B)$ を定義する。

(定義) \mathcal{J}_p や W に上記の演算を施してできた表示を代数形式と呼ぶ。

(定理 1) 任意の DCS 木 \mathcal{T} とその芽 (σ, r) に対して、 $D(\sigma_r; \mathcal{T})$ は代数形式で表現することができる。また、この表現は \mathcal{T} のノード数に比例する計算量で計算できる。

(定理 2) 任意の DCS 木の述語論理の意味は、代数形式に関する三つの基本関係: (i) $A \neq \emptyset$, (ii) $A \subset B$, (iii) $A \parallel B \stackrel{def}{=} A \cap B = \emptyset$ で記述できる。

定理 2 によって、 \mathcal{J}_p と W を定数、関係代数演算子を関数、三つの基本関係を原子述語として公理系を構築すれば、我々は任意の DCS 木の意味を表現できる。このような意味表現枠組みは以下の利点を持つ：

1. DCS 木は、全称量子子が入る文も表現でき、これが代数形式を使うとまったく量子子が入らない原子論理式に変換される。言い換えれば、伝統的に一階述語論理式で表現されていた意味がこの枠組みでは命題論理だけで表現できるということであり、命題論理の自動推論は一階述語論理に比べ非常に高速である。
2. DCS 木の意味は、その芽に対応する外延の代数形式の間の関係で表される。表層構造 (DCS 木の芽) と論理構造 (外延の代数形式) との間のこうした明確な対応を利用することで、我々は論理的推論の結果を表層構造にフィードバックする手掛かりを得る。このような手掛かりを利用した動的知識生成コンポーネントは、我々の含意関係認識システムの精度向上に貢献している。

4. 実験

我々の含意関係認識システムの構成を図 4 に示す。DCS 木は、日本語の係り受け解析器 Cabocha[Kudo 02] と述語項構

^{*3} W_{r_1}, \dots, W_{r_n} は W のコピーで、直積の順序による違いは無視する。

表 1: 精度

	RITE-EXAM		RITE2-EXAMBC	
	dev	test	dev	test
baseline	66.73	65.16	-	56.47
shallow	73.15	72.17	71.57	68.75
inference	73.95	73.53	72.16	70.31

造解析器 Syncha[Iida 11] の結果を、日本語語彙大系^{*4}における文型ボタンと照合してルールベースの変換で得られる。

言語知識として、時間表現認識器^{*5}、及び日本語語彙大系や分類語彙表^{*6}、ウィキペディア、広辞苑^{*7}、日本語ワードネット [Bond 12] から抽出した同義語・上位語・対義語を用いた。

推論エンジンの適用によって我々は時間の前後関係など論理的な知識と同義語のような言語知識を合わせて利用することが可能になるが、これだけの知識を用いても H を T から厳密に証明できることは稀である。特に自然言語のフレーズ言い換えに関する知識は少ないが、動的知識生成コンポーネントは推論の表層へのフィードバックを手掛かりに、T と H の DCS 木構造を比較して尤もらしいフレーズ言い換え知識を自動生成する。動的に生成された知識は分類器によって評価され、最終的な含意関係の判断が得られる。この時に分類器が用いる特徴量(単語間の類似度など)は通常の表層のアプローチで用いられるものと大差はないが、推論過程において重要と判断された動的知識に対してだけ類似度を評価することで、我々のシステムは推論的手法と統計的手法をうまく組み合わせたとと言える。

RITE-EXAM と RITE2-EXAMBC データセットで我々のシステムを評価した結果を表 1 に示す。ここで baseline は T と H の一致する文字数を数える単純なシステムによる精度、shallow は通常の表層の特徴量(単語一致率、文の長さ、固有名詞一致数など)しか使わない時の我々のシステムの精度、inference は推論エンジンの出力と生成された動的知識を評価する特徴量を分類器に入れた後の我々のシステムの精度である。shallow は強いシステムであり、これが RITE-EXAM の test データに対して達成した 72.17% の精度は、前回の RITE ワークショップで報告されたトップ成績に匹敵する。推論エンジンの結果と動的知識を評価する特徴量によって、強い shallow システムより有意に高い精度を達成することが示された。

5. おわりに

含意関係認識のために我々は関係度数に基づく新しい意味表現の枠組みを提案した。この枠組みは DCS 木という、係り受け木の形に近い厳密な論理表現を用いる。DCS 木の意味は代数形式で表現され、この表現によって特に表層構造と論理構造との間のより緊密な連携が得られる。この意味表現枠組みを応用した含意関係認識システムは実験でポジティブな結果を示した。将来的にこのシステムは、世界史オントロジーのようなより構造的な知識を取り入れることも可能である。

参考文献

[Androutsopoulos 10] Androutsopoulos, I. and Malakasiotis, P.: A Survey of Paraphrasing and Textual En-

*4 <http://www.kecl.ntt.co.jp/icl/lirg/resources/GoiTaikei/>

*5 <http://www.cl.ecei.tohoku.ac.jp/~katsuma/software/normalizeNumexp/>

*6 <http://www.ninjal.ac.jp/archives/goihyo/>

*7 <http://www.iwanami.co.jp/kojien/>

tailment Methods, *Journal of Artificial Intelligence Research*, Vol. 38, (2010)

[Bentivogli 10] Bentivogli, L., Clark, P., Dagan, I., Dang, H., and Giampiccolo, D.: The Sixth PASCAL Recognizing Textual Entailment Challenge, in *TAC 2010* (2010)

[Bond 12] Bond, F., Baldwin, T., Fothergill, R., and Uchimoto, K.: Japanese SemCor: A Sense-tagged Corpus of Japanese, in *GWC 2012* (2012)

[Bos 08] Bos, J.: Wide-Coverage Semantic Analysis with Boxer, in *STEP 2008* (2008)

[Bos 09] Bos, J., Zanzotto, F. M., and Pennacchiotti, M.: Textual Entailment at EVALITA 2009, in *EVALITA 2009* (2009)

[Iida 11] Iida, R. and Poesio, M.: A Cross-Lingual ILP Solution to Zero Anaphora Resolution, in *ACL-HLT 2011* (2011)

[Kudo 02] Kudo, T. and Matsumoto, Y.: Japanese Dependency Analysis using Cascaded Chunking, in *CoNLL 2002* (2002)

[Liang 11] Liang, P., Jordan, M. I., and Klein, D.: Learning Dependency-Based Compositional Semantics, in *ACL 2011* (2011)

[MacCartney 07] MacCartney, B. and Manning, C. D.: Natural Logic for Textual Inference, in *Workshop on Textual Entailment and Paraphrasing* (2007)

[Shima 11] Shima, H., Kanayama, H., Lee, C.-W., Lin, C.-J., Mitamura, T., Miyao, Y., Shi, S., and Takeda, K.: Overview of NTCIR-9 RITE: Recognizing Inference in TExt, in *NTCIR-9* (2011)

[Stern 12] Stern, A., Stern, R., Dagan, I., and Felner, A.: Efficient Search for Transformation-based Inference, in *ACL 2012* (2012)

[Watanabe 13] Watanabe, Y., Miyao, Y., Mizuno, J., Shibata, T., Kanayama, H., Lee, C.-W., Lin, C.-J., Shi, S., Mitamura, T., Kando, N., Shima, H., and Takeda, K.: Overview of the Recognizing Inference in Text (RITE-2) at the NTCIR-10 Workshop, in *NTCIR-10* (2013)

[Zanzotto 10] Zanzotto, F. M. and Dell'Arciprete, L.: Efficient kernels for sentence pair classification, in *EMNLP 2010* (2010)

[稲田 13] 稲田 和明, 松林 優一郎, 井之上 直也, 乾 健太郎: 効率的な推論処理のための日本語文の論理式変換に向けて, 言語処理学会第 19 次年次大会発表論文集 (2013)

[川添 12] 川添 愛, 宮尾 祐介, 松崎 拓也, 横野 光, 新井 紀子: 計算機による大学入試問題への解答に向けた世界史オントロジーの設計, 人工知能学会第 26 次年次大会発表論文集 (2012)