

Phased Predicate Template Taxonomy (PPTT) を用いた 意味的關係・スクリプト的知識の大規模獲得

佐野 大樹 鳥澤 健太郎 Julien Kloetzer
István Varga 橋本 力 吳 鍾勳
情報通信研究機構 (NICT) 情報分析研究室
{msano, torisawa, julien, istvan, ch, rovellia}@nict.go.jp

1 はじめに

含意, 矛盾, 因果関係などの意味的關係は, 多くの自然言語処理技術にとって肝要な知識であり, これを自動獲得する研究 (Talukdar et al., 2012; Szpektor et al., 2007; Hashimoto et al., 2012; Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009) が行われてきたが, Web 上にあらわれる多様な言語表現をカバーできるだけの十分な獲得技術の開発にはまだ至っていない. 本研究では, 我々が構築した意味的レキシコン『*Phased Predicate Template Taxonomy (PPTT)*』を利用し, 既存研究とは全く異なるアプローチで意味的關係, 及び, スクリプト的知識を獲得する. PPTT 及び獲得した意味的關係は NLP コミュニティにリリース予定である.

PPTT は, 「X を使う」のような係り元の名詞をあらわす変数 X (本稿で「X」はこの意味で用いる) と係り先の述語の組み合わせからなる述語テンプレート (以下, 「テンプレート」もしくは「*T*」と表記) の分類法で, 以下の仮定のもと設計した.

いかなる事物も, それが存在に至る前に始まり, それが存在することをやめた時に終わる独自の「ストーリー」をもち, ストーリーはこれを構成するいくつかのフェイズの時間的遷移によって展開する. テンプレートは, 当該の事物がそのストーリー中のどのフェイズにあるかを示す.

PPTT は, ストーリーのフェイズに対応するテンプレートのクラスとフェイズ間の時間的遷移関係をあらわすクラス間リンクからなり, 12,023 件のテンプレートが扱われている. 例えば, 「X を生産する」は CREATION, 「X を起動する」は EXECUTION というクラスに分類され, CREATION と EXECUTION クラスの間には「X のストーリーにおいて, EXECUTION が起きたのであれば, CREATION が前に起きていなければならない」という時間的遷移関係 (後述する「Past+ リンク」) が付与される. このクラス間の時間的遷移関係の付与が多様な意味關係獲得を可能にする PPTT 特有の論理的デバイスであり, 言語学, 言語処理分野で従来行われて

きたレキシコンの構築 (Fillmore, 1976; Halliday, 1985; Levin, 1993; Fellbaum, 1998) と異なる特色である.

本稿では, PPTT を用いることで, 多様な意味的關係, 及び, スクリプト的知識の大規模獲得が可能となることを実証的に示す. 具体的には, PPTT を利用し, *happens-before* 関係 (「 \Rightarrow 」であらわす, 「X を入手する \Rightarrow X を食べる」など) を適合率約 80% で 400 万件, また, 本稿にて新しく提案する意味的關係「無効化 (*anomalous obstruction*) 関係」 (「/」であらわす, 「X が売り切れる / X が使えない」など) を適合率約 70% で 100 万件獲得できることを示す. なお, スペースの都合上詳細は示せないが, 適合率約 60% で 57 万件の含意関係, また適合率約 80% で 18 万件の反義対を含む矛盾関係を獲得し, state-of-the-art の含意・矛盾認識システムにまさる性能を得ることに成功していることを述べておく. さらに, 獲得した意味的關係を用いて, 「車をデザインする \Rightarrow 車を生産する \Rightarrow 車を購入する \Rightarrow 車を運転する」のような X の機能が円滑に発動するまでの一連の過程 (以下, *standard Course of Events (CoEs)*) や「セミナーを計画する \Rightarrow セミナーに登録する \Rightarrow セミナーが中止される \Rightarrow セミナーが行われない」のような X の機能が特定の出来事の干渉により発動できなくなる過程 (以下, *anomalous CoEs*) を示す構造化知識など *open information extraction* (Sekine, 2006; Etzioni et al., 2008) やプラン認識 (Carberry, 1990) などのタスクに利用できるスクリプト的知識を適合率約 90% で 160 万件獲得できることを示す.

2 PPTT のデザイン

PPTT は, テンプレートを 3 階層でクラス分けする. まず Level 0 (L0) で, テンプレートを以下の 5 つのクラスに分類する.

Non-existence クラス: X が存在することを含意しないテンプレート (「X を計画する」など)

Existence クラス: X が存在することを含意するが, その機能が発動すること, もしくは, 目的を達成することを含意しないテンプレート (「X を茹でる」など)

Functioning クラス: X の機能が発動すること, もしくは, 目的を達成することを含意するテンプレート (「X を使う」など)

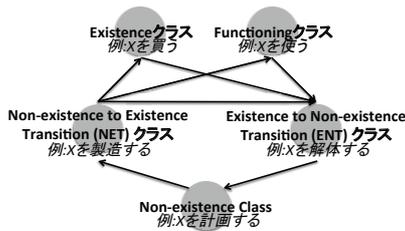


図 1: L0 クラスにおける時間的遷移関係.

Non-existence to Existence Transition (NET) クラス: X が存在しない状態から存在する状態に移行することを含意するテンプレート (「X を製造する」など)

Existence to Non-existence Transition (ENT) クラス: X が存在する状態から存在しない状態に移行することを含意するテンプレート (「X を解体する」など)

このようなクラスを設けることで、図 1 に示すような L0 クラス間の時間的遷移関係 (L0 リンク) を設定でき、例えば、これを用いてテンプレート間の *happens-before* 関係を獲得できる。一例としては、「X を使う」は Functioning クラス、「X が消滅する」は ENT クラスに該当するが、Functioning クラスから ENT クラスに向かって L0 リンクが貼られていることから、「X を使う ⇒ X が消滅する」という *happens-before* 関係を獲得できる。ただし、もちろん、このような形而上学的な概念だけで全てのテンプレートを整理できるわけではない。例えば、L0 クラスに分類することが難しいテンプレートも存在する。一例として、「X を考える」というテンプレートについて考えてみると、当該テンプレートは、X が存在することを必ずしも含意しないため Non-existence クラスに該当するが、X が存在している状態であっても、X が機能している状態であっても、X について「考える」ことは可能である。このようなテンプレートは、L1 クラスの説明にて後述する活性・不活性極性 (Hashimoto et al., 2012) の中立テンプレートに該当するため、PPTT の分類対象から中立テンプレートを除外することとした。また、X が他動詞の主格となるテンプレートも除外した、これは、他動詞主格には生物が該当する場合が多いため、そのような対象の機能や目的を特定することは難しく、Functioning クラスに該当するのかわ、他のクラスに該当するか判定するのが困難なためである。

L0 クラス	活性・不活性極性	
	活性	不活性
Non-existence クラス	POTENTIAL クラス 例:X を計画する	FORECLOSING クラス 例:X を予防する
Existence クラス	ENABLING クラス 例:X を購入する	INCOMMODE クラス 例:X が減る
Functioning クラス	ACTUALIZING クラス 例:X が機能する	DISORDERING クラス 例:X に飽きる
NET クラス	GENERATING クラス 例:X が生まれる	N/A
ENT クラス	N/A	CORRUPTING クラス 例:X は消滅する

表 1: L1 クラスの種類.

PLANNING	X が計画されることを含意するもの。ただし、「X を依頼する」のように計画者以外が X を実行することを含意しないもの。例:「X を企画する」。
CANCELATION	X がキャンセルされること、及び、X がキャンセルされる前に、X の計画や依頼があったことを含意するもの。例:「X を断念する」。
PROHIBIT	X が禁止されることを含意するもの。X が生成される権利や能力が剥奪される。例:「X を禁じる」。
SYMBOLIZATION	X が存在しない状態から存在する状態に移行することを含意するものうち、X が記号の羅列や芸術などに限定されるもの。例:「X を執筆する」。
CREATION	X が存在しない状態から存在する状態に移行することを含意するものうち、X が記号の羅列や芸術などに限定されないもの。例:「X が誕生する」。
COLLAPSE	X が存在する状態から存在しない状態に移行することを含意するものうち、X の変換を前提としないもの。例:「X が死ぬ」。
CONVERSION	X が存在する状態から存在しない状態に移行することを含意するものうち、X の変換を前提とするもの。例:「X を転換する」。
PREPARATION	X が X の機能を発動するための準備として加工、取り付け、登録されることを含意するもの。例:「X をインストールする」。
EXECUTION	X がその機能を発動することを含意するが、X の機能が十分に発揮されたことは含意しないもの。例:「X を起動する」。
WORKING	X がその機能を発動することを含意し、かつ、X の機能が十分に発揮されたことを含意するもの。例:「X で保湿する」。

表 2: L2 クラスの例.

Level1(L1) では、テンプレート間の意味的対立を把握できるよう、活性・不活性極性 (Hashimoto et al., 2012) を用いて L0 クラスを 8 つの L1 クラスに細分化する (表 1 参照)。活性・不活性極性は、テンプレートを活性・不活性・中立の 3 つに分類する (Hashimoto et al., 2012)。活性テンプレートは、「X を使用する」など X の主たる機能、効果、目的、役割、影響が、拡張、準備あるいは活性化されることを含意するもので、不活性テンプレートは、「X を防ぐ」など X の主たる機能、効果、目的、役割、影響が抑制あるいは不活性化されることを含意するものである。中立テンプレートは活性でも不活性でもないものである。この活性・不活性の対立を用いて、同じ L0 クラスに属するが L1 クラスの異なる 2 つのテンプレート、例えば「X が増える」と「X が減る」から矛盾関係を成すテンプレートペアが獲得できる。

Level2 (L2) では、同じ L0・L1 クラスに属するテンプレート間の意味的關係を扱ったり、L0 リンクに比べてより詳細にテンプレート間の時間的遷移関係を把握できるよう、さらに L1 クラスを 41 の L2 クラスに細分化し、表 3 に示す L2 クラス間の時間的遷移/非遷移関係 (L2 リンク) を付与してある。スペースの都合上すべての L2 クラスは列挙できないが、本稿で例示している L2 クラスの定義を表 2 に示す。また、表 3 の「()」内に、PPTT で付与されている当該 L2 リンクの総数を示す。L2 リンクは、先述したように、PPTT の特色のひとつで、意味的關係獲得において重要な役割を担う。例えば、Present+ リンクは、一方の L2 クラスが起きているのであれば、他方の L2 クラスも同時に起きているクラスペアに付与されるもので、「X を救助する ⇨ X が生きる」のような含意關係獲得に役立つ。さらに、3 節で実験について報告する際に、他の

	Positive	Negative
Past	Past⁺ : C ₁ が起きたのであれば, C ₂ が過去に必ず起きている. (55 件)	Past⁻ : C ₁ が起きたのであれば, C ₂ は過去に起きていない. (441 件)
Present	Present⁺ : C ₁ が起きているのであれば C ₂ も同時に起きている. (73 件)	Present⁻ : C ₁ が起きているのであれば C ₂ は同時に起きている. (496 件)
Future	Future⁺ : C ₁ は C ₂ が起きていることを可能にする. (90 件)	Future⁻ : C ₁ は C ₂ が起きていることを不可能にする. (210 件)

表 3: L2 リンク. 「C」は L2 クラス.

L2 リンクも *happens-before* 関係, 及び, 無効化関係などの獲得に貢献することを示す.

3 関係獲得の方法と評価実験

3.1 意味的關係獲得のターゲット

テンプレート間の意味的關係を獲得するために, PPTT に含まれるテンプレート全てを組み合わせてテンプレートペアを生成し, このうちペアを成す両方のテンプレートと 100 回以上共起する名詞が 6 億ウェブページコーパス (以下, *WCorpus*) に少なくとも 1 つあるテンプレートペアを関係獲得の候補とした. 以下, このデータセットを *TP100* とする.

3.2 *happens-before* 関係の獲得

本研究では Regneri et al. (2010) の定義に従い, T_1, T_2 があらず出来事が両方起きたと仮定した場合, T_1 があらず出来事が一般的に T_2 があらず出来事より前に起こる場合に T_1 と T_2 に *happens-before* 関係が成立すると考え, 以下の手法で *happens-before* 関係を獲得, 比較した.

- H1** L2 リンク「Past⁺」でつながれた L2 クラスペア 73 件に属するテンプレートペアを *happens-before* 関係として獲得する手法.
- H2** L2 リンク「Future⁺」でつながれた L2 クラスペア 90 件に属するテンプレートペアを *happens-before* 関係として獲得する手法.
- H3** 図 1 に示した L0 リンクによってリンクされる L2 クラスペア 474 件に属するテンプレートペアを *happens-before* 関係として獲得する手法.
- ALL** H1~H4 の全てで *happens-before* 関係を獲得する手法.
- HB-Ptn** *WCorpus* において「~あとに」「~まえに」など時間関係をあらわす 73 の接続詞, もしくは, 接続辞によって結ばれたテンプレートペアのうち, 同じ名詞を係り元とする, もしくは, 2 つ目のテンプレートの名詞が指示詞のものを *happens-before* 関係として獲得する手法. Chklovski and Pantel(2004) と類似する手法.
- Random** *TP100* から無作為抽出されたテンプレートペアを *happens-before* 関係として獲得する手法.

各手法で獲得したテンプレートペアからそれぞれ無作為抽出した 200 サンプルを, 3 名のアノテータ (筆者以外) がそれぞれ独立して判定した (多数決により正否を判定). 判定においてアノテータは, X に代入した場合に判定対象をターゲット関係の負例とする名詞がないと判定した場合, 当該インスタンスを正例と

手法	適合率 (%)	該当ペア数	再現率 (%)
H1	83.5	1,113,280	18.0
H2	70.5	1,524,557	20.8
H3	67.0	3,837,116	49.7
ALL	79.5	4,387,781	67.5
HB-Ptn	53.0	32,288	0.30
Random	18.0	28,717,454	100.0

表 4: *happens-before* 関係獲得の評価.

し, そのような名詞があると判定した場合は負例とした. κ 値 (Fleiss) は, 0.56 (中程度の一致) であった.

結果を表 4 に示す. 再現率は, **Random** の適合率から求められる *TP100* に含まれる正例数をもとに推定した. 提案手法 **ALL** では, 適合率 79.5%, 再現率 67.5% で図 2 のような約 400 万件の *happens-before* 関係を獲得することができた. 言語や利用可能な言語資源の違いがあるため厳密な比較は難しいが, Chambers et al. が公開している英語の *before* 関係候補 87 万件 (<http://cs.stanford.edu/people/nc/schemas>) と比べて 5 倍以上の *happens-before* 関係を獲得した.

X を煮る ⇒ X を食べる
PREPARATION クラス $\xrightarrow{\text{Future}^+}$ EXECUTION クラス
X を作曲する ⇒ X でリラックスする
SYMBOLIZATION クラス $\xleftarrow{\text{Past}^+}$ WORKING クラス

図 2: 獲得した *happens-before* 関係の例.

3.3 無効化関係の獲得

無効化関係は, (1) T_1 は T_2 が起きていることを不可能にし, (2) T_1 が X の機能が円滑に発動される過程では起きない出来事をあらわし, (3) T_2 が X の機能が発揮, 拡張, 準備される出来事をあらわす場合, T_1 と T_2 は無効化関係にあるものと定義し, 以下の提案手法 **Proposed** の手法で無効化関係を獲得した.

Proposed L2 リンク「Future⁻」でつながれた L2 クラスペアに属するテンプレートペアのうち, T_1 が不活性テンプレート, T_2 が活性テンプレートのものを無効化関係として獲得する手法.

Random *TP100* から無作為抽出されたテンプレートペアを無効化関係として獲得する手法.

アノテーションは *happens-before* 関係と同じスキーマで行った (κ 値 (Fleiss) 0.56, 中程度の一致). 結果を表 5 に示す. 再現率は *happens-before* 関係獲得と同様に, **Random** の適合率から推定した. **Proposed** では, 適合率 73.5%, 再現率 26.4% で図 3 に示すような無効化関係を約 100 万件獲得した. 無効化関係のような新しい関係を獲得できることは, 意味的關係獲得における PPTT の柔軟性を示唆する一例であると考え.

3.4 スクリプト的知識の獲得

スクリプト的知識 (standard/anomalous CoEs) の生成には, 獲得した *happens-before* 関係と無効化関係を用

手法	適合率 (%)	該当ペア数	再現率 (%)
Proposed	73.5	1,081,405	26.4
Random	10.5	28,717,454	100.0

表 5: 無効化関係獲得の評価.

X を禁止する / X を公開できない	
PROHIBIT Class	$\xrightarrow{Future^-}$ EXECUTION Class
X を諦める / X を満喫できない	
CANCELATION Class	$\xrightarrow{Future^-}$ WORKING Class

図 3: 獲得した無効化関係の例.

いた. Standard CoEs の獲得では, まず 3 つの *happens-before* 関係 $T_1 \Rightarrow T_2$, $T_2 \Rightarrow T_3$, $T_3 \Rightarrow T_4$ からテンプレートチェーン $T_1 \Rightarrow T_2 \Rightarrow T_3 \Rightarrow T_4$ を生成した. 次に, テンプレートチェーンが「円滑に機能する過程」をあらわすようチェーンに含まれる全てのテンプレートを活性テンプレートとし, T_4 が X の機能発動を意図する Functioning L0 クラスになるよう制約した.

Anomalous CoEs の獲得では, 2 つの *happens-before* 関係 $T_1 \Rightarrow T_2$, $T_2 \Rightarrow T_3$ と 1 つの無効化関係 T_3/T_4 からテンプレートチェーン $T_1 \Rightarrow T_2 \Rightarrow T_3/T_4$ を生成した. Standard CoEs と同様に T_4 は Functional L0 クラスに限定した (Anomalous CoEs の場合は, X が機能できなくなる出来事をあらわす). また, X の機能が発動してしまう出来事がテンプレートチェーンの途中で起きぬよう T_1 , T_2 は Functioning L0 クラスに属さない活性テンプレートにした.

最後に, CoEs に含まれるテンプレートが類似した文脈にあらわれるものになるよう, チェーンに含まれる全てのテンプレートのペア (T_i, T_j) において, T_1 に対する T_2 , もしくは, T_2 に対する T_1 の Kazama et al.(2010) の文脈類似度スコアが上位 500 位以内に入るという制約を設け, チェーン中の全てのテンプレートと *WCorpus* で 100 回以上共起する名詞ひとつを無作為に選択して X に代入し, CoEs を獲得した.

各 CoEs のサンプルをそれぞれ 200 件無作為抽出し, アノテータ 3 名 (筆者以外) がそれぞれ独立してアノテーションした (多数決で正否を判定). κ 値 (Fleiss) は, 0.79 (かなり高い一致) であった. 結果を表 6 に示し, 例を図 4 に示す. 獲得できた CoEs は両方のタイプでそれぞれ 50 万件をこえており, 適合率も約 90% と高い. 文脈類似度の制約を設けたことで, チェーン生成に用いた *happens-before* 関係や無効化関係獲得よりも高い適合率を得ることができたのではないかと考える. 我々が獲得した CoEs は, standard CoEs と anomalous CoEs を区別できるという点で, これを区別していない Chambers et al. (2008; 2009) の *narrative chains/schemas* と補完的關係にあると考える.

CoE タイプ	適合率 (%)	獲得 CoE 数
Standard CoE	89.5	620,733
Anomalous CoE	93.5	993,470

表 6: CoEs 生成の評価.

Standard CoE	
システムを設計する \Rightarrow システムが開発される \Rightarrow システムが導入される \Rightarrow システムを使う	
PLANNING クラス	$\xrightarrow{Future^+}$ CREATION クラス $\xleftarrow{Past^+}$
PREPARATION クラス	$\xrightarrow{Future^+}$ EXECUTION クラス
Anomalous CoE	
データが生成される \Rightarrow データを読み込ませる \Rightarrow データが上書きされる / データを参照できない	
CREATION クラス	$\xleftarrow{Past^+}$ PREPARATION クラス $\xrightarrow{Future^+}$
CONVERSION クラス	$\xrightarrow{Future^-}$ EXECUTION クラス

図 4: 獲得された standard/anomalous CoEs の例.

4 まとめと今後の展望

本稿では, 我々が構築した PPTT の概要について説明し, さらに, これを用いた意味的關係, 及び, スクリプト的知識の獲得について述べた. 今後は PPTT を拡張し, 構造化知識を獲得するための方法論と資源の開発を進めて行く予定である.

参考文献

- Sandra Carberry. 1990. *Plan Recognition in Natural Language Dialogue*. MIT Press, Cambridge, MA, USA.
- Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio, June. ACL.
- Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of ACL'09*, pages 602–610, Stroudsburg, PA, USA. ACL.
- Timothy Chklovski and Patrick Pantel. 2004. Path analysis for refining verb relations. In *In Proceedings of LinkKDD-04*, Seattle, WA.
- Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74, December.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1):20–32.
- Michael A.K. Halliday. 1985. *An Introduction to Functional Grammar*. Arnold, 1st edition.
- Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. 2012. Excitatory or inhibitory: a new semantic orientation extracts contradiction and causality from the web. In *Proceedings of EMNLP-CoNLL'12*, pages 619–630, Stroudsburg, PA, USA. ACL.
- Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa. 2010. A bayesian method for robust estimation of distributional similarities. In *Proceedings of ACL'10*, pages 247–256, Stroudsburg, PA, USA. ACL.
- Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press, Chicago and London.
- Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of ACL'10*, pages 979–988, Uppsala, Sweden, July. ACL.
- Satoshi Sekine. 2006. On-demand information extraction. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 731–738, Sydney, Australia, July. ACL.
- Idan Szepktor, Eyal Shnarch, and Ido Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In *Proceedings of ACL'07*, pages 456–463, Prague, Czech Republic, June. ACL.
- Partha Pratim Talukdar, Derry Wijaya, and Tom Mitchell. 2012. Acquiring temporal constraints between relations. In *Proceedings of CIKM'12*, pages 992–1001, New York, NY, USA. ACM.