単語の分散表現を用いた日本語イベント連鎖の自動構築

瀧下 祥　　Rafal Rzepka　　荒木 健治
北海道大学大学院 情報科学研究科
{takishita_uni, rzepka, araki}@ist.hokudai.ac.jp

1 はじめに

言語理解では，文章で書かれていない知識を背景 (常識) 知識を用いて補完する．例えば,「風呂に入る」という行動には,「服を脱ぐ」や「体を洗う」といったサブイベントが含まれる．このような，イベントの理解は事前知識がないと困難であるため，本稿ではイベントに関する知識 (イベント知識) を構築する．

イベント知識に関して Schank ら [1] はスクリプトの概念構造を提案している．スクリプトとは，イベントと参加者の関係，イベント間の因果関係を構造化して扱う知識表現である．これは，特定のシナリオ間でイベントがどのように展開されるかを推測するために使用できる．また，参加者の関係をモデル化するため，共参照解析や談話解析のようなタスクに適用することができる．しかしながら，Schank らはスクリプト構築を人手で行っているため，大きなコストがかかり，膨大な知識を作成することは困難である．

また，日本語におけるイベント関係獲得に関する研究の多くは,「タネを植える → 土をかける」などの二つのイベントを含む 1 組のイベントを扱う．これらは同じ文に現れるイベントから知識を獲得し，選別・列挙するものが多いが,「風呂に入る」などの行動は 1 組のイベントでなく複数のイベントが一連の流れに従うものである．そのため，より長いイベント知識を構築する必要があると考える．しかしながら，文に 3 個以上のイベントが含まれることは多くないため，既存手法では長いイベント関係を獲得することは困難である．

本稿では，単語の分散表現を用いてイベント知識間の関連度を計算し，連結することで長いイベント知識を構築する．その後，人手で構築したイベント知識の有効性を評価する．提案手法の独創性は，同じ文に現れていないイベント知識を構築すること，2 つ以上の長さを持つ行動連鎖を獲得することである．本研究の最終目的は，ボトムアップにスクリプトのようなイベント知識を構築することである．

2 関連研究

テキストからイベント知識を獲得する研究が多く行われている．特に英語のような情報の省略が少ない言語を対象とした研究は，共参照を用いて物語イベント連鎖を構造化する手法が多く用いられている．Chambers と Jurafsky[2] は，共参照を用いて，特定の主語に関するイベントを連鎖的に獲得し，頻度が高いイベント関係を一般的なイベント連鎖とした．また，Granroth と Stephen[3] は，2 つのイベントの時系列関係を機械学習を用いて計算する手法を提案した．しかし，日本語は主語や目的語を省略する傾向にあるため，これらの手法を用いることは困難である [4].

日本語のイベント知識獲得に関する研究は共起頻度や助詞・接続詞に基づいて行われるもの [5]，知識を使用するもの [6][7] が挙げられる．阿部ら [5] は日本語 Web コーパスから，”X をしても Y ない”などのパターンと共起頻度を用いてイベントを獲得した．また，旭ら [6] は特定のイベント語 (インフルエンザに対して咳など) を使用し，時系列に沿ってイベント知識を獲得した．これらの研究は知識を限られた条件で獲得するため文に大きく依存すること，意味の考慮に事前知識が必要である点が問題である．

本稿では分散表現を用いてイベント間の意味的なつながりを考慮することにより，頻度やフレーズに依存しにくいイベント獲得手法を提案する．また，分散表現を使用するため，辞書などの事前知識を必要としない．加えて，イベント知識間の類似度を計算し，関連度の高い組み合わせを連結することにより，同一文に含まれていない知識や長いイベント知識 (イベント連鎖) を構築することが可能である．

3　イベント連鎖構築手法

本稿では,「意味的なつながりがある事象」「時間の流れに沿った事象の連鎖」に基づいてイベント知識を獲得する.「事象」とは，動詞と動詞にかかる単語をまとめたものを指し，これを時系列に沿って並べたものをイベント連鎖と定義する．ここでは，事象関係獲得とイベント連鎖構築について述べる．

3.1　事象関係獲得

事象関係獲得の流れを以下に示す．NWC[8] と YACIS[9] の Web コーパスのテキストに対して，意味役割付与システム ASA[10] を用いて意味解析を行い，述語論理構造を獲得する．加えて，時系列語や出現順序から事象間の時系列関係を付与する．コーパスの構造が異なるため，別手法で時系列関係を獲得する．以下にコーパスの特徴を述べる．

NWC(文単位)

NWC コーパスは文が独立しており，前後関係から時系列を判断することが困難である．そのため，時系列語「した後に」を用いて時系列関係を獲得する．表 1 に獲得の流れを示す.

YACIS(文章単位)

YACIS コーパスは，投稿ごとに文がまとまっている．ブログの筆者は経験を時系列順に列挙すると仮定し，事象の出現順から時系列関係を獲得する．

表 1: 事象関係獲得の具体例 (NWC の場合)

0. 対象文 シャンプーで洗髪した後にお湯で洗う
1. キーワードによる分割 + 意味解析 シャンプーで洗髪する:した後に:お湯で洗う
2. 時系列処理 シャンプーで, 洗髪する → お湯で, 洗う

3.2　単語の意味関連度計算

イベント知識には，ノイズなどにより意味的なつながりが少ないものがある．そこで，分散表現を用いて関連度を計算し，事象間で意味的に関連性のある単語のみを獲得する．分散表現には，常識的知識を考慮した ConceptNet Numberbatch[10] を用いる．

分散表現を用いた関連度計算を式 (1) に示す．事象 A に含まれる単語の集合 A と単語 W_{Ak}，事象 B に含まれる単語の集合 B と単語 W_{Bn} とする (W_{Ak} は単語集合 A の k 番目の単語，W_{Bn} は単語集合 B の n 番目の単語)．関連度を式 (1) で計算し，実験的に設定した閾値 R より高い場合に単語を獲得する．なお，分散表現に含まれていない単語は関連度 0 として計算する．

$$similarity(W_{Ak}) = \frac{\sum_{n=0}^{N-1} cosine(W_{Ak}, W_{Bn})}{N} \quad (1)$$

3.3　イベント連鎖構築

3.3.1　イベント知識連結

3.2 で作成した意味的なつながりのあるイベント知識を連結し，イベント連鎖を構築する．イベント知識 A「事象 A → 事象 B」とイベント知識 B「事象 C → 事象 D」に対して，事象 B と事象 C の関連度が高い場合，イベント知識 A・B は共通のイベント連鎖に現れやすいと考えられる．例を挙げると,「頭皮を洗う → シャンプーで洗う」と「お湯をかき混ぜる → 入浴する」では，事象 B「シャンプーで洗う」と事象 C「お湯をかき混ぜる」の関連度が高いため，イベント知識 A・B を連結することでイベント連鎖「風呂に入る」として扱うことができる．このように，意味関連度を式 (2) を用いて計算し，関連度が閾値 R より高い場合にイベント知識を連結してイベント連鎖を構築する．表 2 に構築したイベント連鎖の例を示す.

$$similarity(事象_A, 事象_B) = \frac{\sum_{k=0}^{K-1} \sum_{n=0}^{N-1} cosine(W_{Ak}, W_{Bn})}{MN} \quad (2)$$

表 2: 提案手法を用いて構築したイベント連鎖

頭皮をよく洗う → シャンプーで洗う → お湯をかき混ぜる → 入浴する
記事を書く → ブログを読む → 本を読む → 英語を勉強する
お金を払う → お店を借りる → 営業を営む → 宣伝を行う

3.3.2 パラメータ：関連度計算の閾値

連結規則の閾値について述べる．はじめに，異なるイベント知識に含まれる事象 B と事象 C の関連度 R_{bc} を計算して 2 つの事象間に意味的なつながりがあるか判別する．次に，事象 A と事象 C, 事象 D ならびに事象 B と事象 D の関連度を計算する．これは，事象 B と事象 C しか計算しない場合に，他の事象間で関連性の低いイベント連鎖が構築されることを防ぐため用いた条件である．したがって，すべての関連度があらかじめ指定した閾値 R_{bc},R_{ac},R_{ad},R_{bd} より高い場合にイベント連鎖として獲得する．

4 評価実験

4.1 実験設定

獲得したイベント連鎖に対する評価を以下に示す．評価項目は，イベント知識に含まれる事象間の意味的なつながりとイベント連鎖全体の一般性である．以下で各評価項目について説明する．

事象関係の意味的なつながり

獲得したイベント連鎖の中で，2 事象ごとの関連性を評価する．意味的なつながりを“ 関連 ”として扱い，事象 A と事象 B に対して”関連あり”,”文脈依存”,”関連していない”の三択で選択する．これを全ての組み合わせに対して行い，”関連あり”を 3 ポイント,”文脈依存”を 2 ポイント,”関係なし”を 1 ポイントとして計算する．

イベント連鎖の一般性

イベント連鎖に一般性があるかどうかを評価する．一般性を“ 起こりやすさ ”として扱い，”起こりにくい”から”起こりやすい”をリッカート尺度を用いて 1-5 の数値で評価する．

4.2 実験結果

20 代大学院生 3 人 (日本語母語話者女性 1 名と日本語検定 1 級の中国人留学生の男性 2 名) にアンケート形式で評価を行った．関連度計算の各閾値は第一著者が目視で調節しており，それぞれを表 3 に示す．3 回以上出現しているイベント知識で構築した 3 事象，4 事象のイベント連鎖をコーパスごとに 20 件ずつランダムに抽出し，合計 80 件で評価を行う．

獲得したイベント知識は NWC コーパスでは 21 件，YACIS コーパスでは 124 件であり，イベント連鎖は 90 件と 459 件であった．件数が少ない要因は対象文が少ないこと，頻度を加味していること，良質な知識を獲得するために閾値を高めていることが挙げられる．また，表 4 に事象間の意味的なつながり，表 5 に評価者ごとのイベント連鎖の一般性の評価結果を示す．表に示す評価結果は評価結果の平均である．

表 3:条件と閾値の設定

条件 \ 閾値	意味関連度	R_{bc}	$R_{ac,ad,bd}$
YACIS:事象 3 つ	0.15	0.25	0.15
YACIS:事象 4 つ	0.15	0.25	0.15
NWC:事象 3 つ	0.15	0.2	0.1
NWC:事象 4 つ	0.15	0.2	0.1

5 考察

表 4 より，評価では全ての部分で平均が 2 以上であるため，意味的に関連があるイベント知識を獲得できていることが確認された．事象 BC 間などの直接文に現れていないイベント知識も提案手法により，関連のあるつながりを構築できている．また，事象 AB 間と事象 CD 間は，文に現れるイベント関係であるため評価が高いと考える．以上のことより，意味的なつながりのあるイベント知識の構築に関して，提案手法が有効であること確認された．

表 4:事象間関連度の評価結果

条件 \ 部分	AB	AC	AD	BC	BD	CD
YACIS:事象 3 つ	2.8	2.5	-	2.6	-	-
YACIS:事象 4 つ	2.8	2.4	2.5	2.4	2.5	2.7
NWC:事象 3 つ	2.8	2.0	-	2.0	-	-
NWC:事象 4 つ	2.8	1.7	1.9	1.8	2.0	2.6
平均	2.8	2.15	2.2	2.2	2.25	2.65

表 5:評価者ごとのイベント連鎖の一般性

一般性 \ 評価者	A	B	C	平均
YACIS:事象 3 つ	4.3	3.7	3.95	4.0
YACIS:事象 4 つ	4.05	3.7	3.75	3.8
NWC:事象 3 つ	3.4	2.05	3.15	2.9
NWC:事象 4 つ	3.0	2.2	3.2	2.8

次に，コーパスごとに結果を比較する．NWC と YACIS では，NWC の方が全体的に評価が低くなる傾向がある．イベント知識が 21 件と少ないため，閾値を下げてイベント連鎖を構築していることが要因として考えられる．YACIS はイベント知識が 124 件と多いため，閾値を高くしても多くのイベント連鎖を構築されやすい．そのため，関連度の高い組み合わせでイベント連鎖を構築することができ，評価が高くなる傾向にある．加えて，ブログ記事は行動の連鎖をそのまま列挙することが多いため，関連度の高いイベント知識を多く獲得することができる．以上の違いが，コーパスに関して評価に差が生じた要因である．

また，イベント連鎖の一般性では，YACIS と NWC では約 1 ポイントの差が生じている．閾値を 0.05 ポイント高めに設定したことが要因として考えられ，関連度が高いペアが獲得できていることを確認した．また，YACIS は評価値の平均も約 4 であり，起こりやすいイベント連鎖を獲得できていることがわかる．しかし，獲得できる知識のジャンルが狭くなる傾向があるため，多様性を持つ知識を構築するために改善が必要である．

事象数の違いでは 4 事象より 3 事象のものが 0.1 ポイントほど評価が高くなる傾向にある．この要因はイベント連鎖の構築方法にあり，3 事象の連結は事象 BC が一致している場合に連結させている．そのため，事象 AB と事象 BC はどちらもテキストに含まれるため，関連度が高くなる．また，事象 AC も同じ事象の前後事象であるため，同様に関連度が高くなりやすい．

評価が低くなった要因として，文脈に依存するイベント知識が多いことが挙げられる．例えば「本を読む→英語を勉強する→話を聞く」では，話が指すものが明確でないため判断が困難である．また,「コーヒーを飲む→バナナを食べる→お酒を飲む」などは常識に依存するため評価が分かれた．加えて,「本を読む→漫画を読む」などの同義語を含む点も問題である．

6 まとめ

本稿では，日本語の Web コーパスから事象を抽出し，イベント知識を獲得した．また，それらを ConceptNet Numberbatch を用いて意味的な関連性から連結し，従来研究よりも長く意味的なつながりのあるイベント連鎖を構築した．事象間の関連度とイベント連鎖の一般性を人手で評価した結果，どちらも有効性を確認した．

今後の課題として，より幅広いイベント連鎖を構築する必要がある．今回構築したイベント連鎖は同じようなジャンル (本を読むことやブログを書くこと) が多い傾向にあった．ノイズ削減のために出現頻度を用いてイベント知識を限定したことが要因であるが,「風呂に入る」や「学校に行く」などの文に現れにくい常識的な知識を獲得することが困難である．この問題を解決するために，少ない出現頻度でもノイズの少ない知識を獲得できるように改善する必要がある．

参考文献

[1] Roger Schank and Robert Abelson: Scripts, plans, goals, and understanding, An inquiry into human knowledge structures (Artificial Intelligence Series), 1977.

[2] Nathanael Chambers and Dan Jurafsky: Unsupervised learning of narrative event chains. In Proceedings of ACL (Association for Computational Linguistics), pp.789797, 2008.

[3] Mark Granroth-Wilding and Stephen Clark: What happens next? Event prediction using a compositional neural network model. In Proceedings of AAAI, pp.2727-2733, 2016.

[4] Yin Jou Huang and Sadao Kurohashi: Improving Shared Argument Identification in Japanese Event Relation Knowledge Acquisition, Proceedings of the ACL 2017 Events and Stories in the News Workshop, pp.21-30, 2017.

[5] Shuya Abe, Kentaro Inui, and Yuji Matsumoto: Acquiring event relation knowledge by learning cooccurrence patterns and fertilizing cooccurrence samples, Proceedings of the 3rd International Joint Conference on Natural Language Processing, pp.479504, 2008.

[6] 旭直人, 山本岳洋, 中村聡史, 田中克己: 行動連鎖を用いた情報検索支援と Web からの行動連鎖の抽出, A7-2, DEIM Forum, 2009.

[7] 高橋公海, 佐藤進也, 松尾真人: 状況に依存した行動パターン抽出手法の検討, 情報処理学会研究報告 Vol.2013-IFAT-111 No.24, 2013.

[8] Nihongo Web Corpus (NWC コーパス), http://s-yata.jp/corpus/

[9] Michal Ptaszynski, Pawel Dybala,Rafal Rzepka, Kenji Araki and Yoshio Momouchi: YACIS: A Five-Billion-Word Corpus of Japanese Blogs Fully Annotated with Syntactic and Affective Information, In Proceedings of The AISB/IACAP World Congress, pp.4049, 2012.

[10] 竹内 孔一: 意味役割付与システム (ASA), http://www.cl.cs.okayama-u.ac.jp/study/project/asa

[11] Robert Speer and Joshua Chin and Catherine Havasi: ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, AAAI pp.4444-4451, 2017.