テキスト集合からのスクリプト知識自動獲得

藤木 稔明† 難波 英嗣†† 奥村 学††
† 東京工業大学大学院 総合理工学研究科 †† 東京工業大学 精密工学研究所

1 はじめに

スクリプト知識とは, 典型的な事象列を一つの枠の中にまとめた記憶構造であり, 人間が日常行っている定型化された行動を, 時間の流れに沿った事象の連鎖の形式で表現したもの, と定義される. 例えば「レストランでの食事」のスクリプト知識は, 普通「客が店に入り」,「座り」,「メニューはウェイトレスが持ってくるか, テーブルの上に置いてあって」,「客はそれを見て料理を決め」,「注文し」,「できるまで待つ」と表現できる.

この知識は従来, 単語の多義性解消, 文章自動生成, 文章自動要約といった自然言語処理の分野で用いられてきた [1]. しかし, 実際に用いるためには膨大な量のスクリプト知識が必要であり, 人手で記述するには多くのコストがかかる.

そこで, 我々はスクリプト知識を自動獲得するための一手法を提案する. 従来, スクリプト知識自動獲得に関して, 堀らによる研究 [2] が行われている. しかし, 本研究ではスクリプト知識における時間順序関係を考慮に入れている点で従来研究との違いがある.

2 提案手法

本研究は, スクリプト知識の定義における「典型的な事象列」「時間の流れに沿った事象の連鎖」という点に注目する. また, 一つの「事象」を動詞とその動詞の持つ格要素 (ガ格とヲ格) からなる「行為」として表現し, これを時間順に並べたものを「行為の連続」と定義する. そして, このような「行為の連続」の中で典型性を持つものをスクリプト知識とする.

提案手法の概要を図 1 に示す. 図 1 は「殺人事件」に対するスクリプト知識を自動獲得する際の流れを表したものである.

まず, スクリプト知識を獲得する対象となるテキスト集合の生成を行う. 詳細は後述するが, これは図 1 において, 事件毎に分類, 一段落目のみ抽出, という 2 段階で表現されている. 次に, これらのテキスト集合から「行為の連続」を抽出する. そして, 異なるテキスト集合に繰り返し出現する「行為の連続」には典型性があると考え, それらを殺人事件に関するスクリプト知識として自動獲得する.

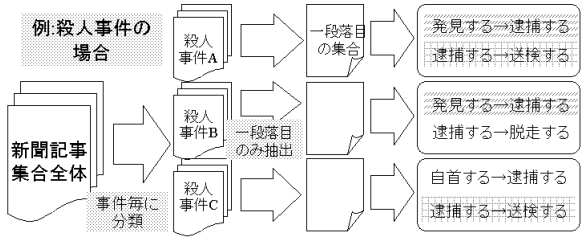

図 1: スクリプト知識自動獲得手法の概要

Automatic acquisition of a script knowledge from a text collection

† Toshiaki FUJIKI (fujiki@lr.pi.titech.ac.jp)
†† Hidetsugu NANBA (nanba@pi.titech.ac.jp)
†† Manabu OKUMURA (oku@pi.titech.ac.jp)
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology(†)
Precision and Intelligence Laboratory, Tokyo Institute of Technology(††)

以下では, 対象テキスト集合の生成手法とスクリプト知識の自動獲得手法について説明する.

2.1 対象テキスト集合の生成

提案手法は次のような性質を持つテキスト集合を対象として, スクリプト知識の獲得を行う. (1) それぞれのテキストは事実のみからなる. (2) それぞれのテキスト集合は単一のトピックからなり, そのトピックは類似している. (3) テキスト集合において, テキストは時間順に整列している.

このような性質を満たすテキスト集合を生成するため, 本研究では新聞記事を利用し, 次のような方法を用いた.

まず, 新聞記事集合全体から同一のトピックを持つ記事を集める. そして, それらの中から事実の記載されることの多い, 報道記事のみを用いる. しかし, 報道記事の中にも事件の経過等, いろいろな情報が含まれるため, 単純にこれらを並べても, 事実のみが時間順に並んだテキスト集合とはならない. そこで, それぞれの報道記事から第一段落のみを抽出し, それらを用いてテキスト集合を生成する. これは, 報道記事の第一段落にはその記事において要点となる事実のみが記述されているためである.

この方法を用いて, 類似するトピックを持つ複数のテキスト集合を生成する. また, 各テキストはあらかじめ構文解析を行っておく.

2.2 スクリプト知識の自動獲得

まず, 各テキストから「行為間の 2 項関係」の抽出を行う. 提案手法では, 以下に示す 3 つの場合に「行為」を表す動詞間における時間順序関係を特定できると考え, これらを用いることで「行為」間に時間順序関係があることを示す「行為間の 2 項関係」を抽出する.

- 異なる文中の動詞が共通の格要素を持つ場合
  「警察は A 容疑者を発見した. 警察は A 容疑者を逮捕した. 」という 2 文があった場合, テキスト集合の性質より, 前の文のほうが過去に起こった出来事を示す文であるはずである. よって,「発見した」「逮捕した」という 2 つの動詞には時間順序関係があることがわかり, これらが共通の格要素を持つ場合に「行為間の 2 項関係」として抽出する.

- 動詞間に連用修飾関係がある場合
  「警察は A 容疑者を発見, 逮捕した」のような連用修飾関係が動詞間にある場合, 修飾動詞「発見」が先に起こり, 被修飾動詞「逮捕した」が後から起こっている. そこで, これらの動詞が表すそれぞれの「行為」を「行為間の 2 項関係」として抽出する.

- 被連体修飾名詞を格要素とする動詞がある場合
  「警視庁は発見した A 容疑者を逮捕した」において, 動詞「発見する」に対する被連体修飾名詞「A 容疑者」は「逮捕した」のヲ格となっている. このような場合, 連体修飾を行う動詞と被連体修飾名詞を格要素とする動詞が表す 2 つの「行為」

はその動詞の出現順と同じ時間順序となる．そこで，それぞれの動詞が表す「行為」に「行為間の2項関係」があると考え，抽出する．

ただし，連体修飾節を構成する動詞は過去形でなければならない，という制約を設けた．これは，「彼は信頼する医者を訪ねた」「彼は訪ねる医者を決めた」というような文において「行為」の時間順序を特定することができないためである．

いずれの場合においても動詞の格要素は省略されることが多く，構文解析の結果からは明らかでないことが多い．そこで本研究では動詞の格フレームに記述される意味素を利用して補完処理を行う．まず，意味素と適合するような語を文内で動詞の前方から探し，そのような語が見つかった場合には，その語を動詞の格要素として補完する．

このとき，格フレームが辞書に書かれていない動詞についてはどのような意味素の格でも持つとして扱う．

また，受動態の文を扱う際にも問題が生じる．そこで本研究では，動詞の能動態格フレーム中のニ格とガ格を入れ替えることで，受動態の格フレームを生成できる場合にのみ，その動詞の受動態を扱う．またその他にも，「～を発見，逮捕．」のようにサ変名詞の直後に句読点がある場合は，そのサ変名詞を動詞として扱う．

次に，このようにして抽出された「行為間の2項関係」が「典型的」であるかどうかの判別を行う．そのためにまず，それぞれの動詞とその格要素の名詞を意味素に変換し，抽象化を行う．このような抽象化を行うことにより，他のテキスト中に出現する「行為間の2項関係」と同一かどうかの比較を，より一般的なレベルで行うことができるようになる．その後，それぞれのテキストから抽出された「行為間の2項関係」について出現頻度を調べ，頻度の高いものに重みをつけていく．

最後に，そのようにして重みづけられた「行為間の2項関係」について，閾値以上の重みを持つものをまとめることで「行為の連続」を形成する．ただし，この「行為の連続」では動詞が意味素になっているため意味が分かりにくい．そこで，同一の意味素を持つ動詞に対しては，その中で最も出現頻度の高い動詞で代表することとする．このようにして全ての意味素を動詞に戻し，それをスクリプト知識として出力する．

3　スクリプト知識獲得実験

3.1　実験

前節で述べた手法の有効性を調べるため「殺人事件」に関するスクリプト知識の獲得実験を行った．この実験では，入力テキストとして，日経全文記事データベース 日経4紙 DVD-ROM 版 1990-1995 年版，1996-2000 年版より，日本経済新聞 10 年分を用い，「殺人事件」という語を含む記事を検索，GETA[3] を用いて 4489 報道記事を 617 トピックに分類したものを用いた．

3.2　結果と課題

実験では全部で 41 個の複数回出現する「行為間の2項関係」を自動獲得することができた．図2に獲得結果から上位8個の「行為間の2項関係」を示す．この図では一つの関係が一つの矢印に対応しており，「[機関] が [人] を発見する」という「行為」の後には「[機関] が [人] を逮捕する」という「行為」が続くことを示している．なお，図中で [] で囲まれる語は意味素を，×印のつけられた矢印は誤って獲得された「行為間の2項関係」を表している．

図2中の「[機関] が [人] を再逮捕する」→「[機関] が [人] を逮捕する」という関係は，異なる容疑者が意味素への変換に際して同一の意味素として扱われてしまったための間違いだが，他にも，構文解析の失敗やトピックへのテキスト分類がうまく行われていないといった，用いたツールによる要因，格要素の補完，受動態の取り扱いといった提案手法による要因により，獲得した知識中にはノイズと考えられるものが含まれている．

また，本手法で用いた入力テキスト集合の生成方法では事件は時間を追って報道される必要がある．そのため，誘拐事件のように変則的な順序で報道される (誘拐事件では犯人が逮捕されてから報道が行われる) ような出来事については獲得することができない．また，記事の第一段落のみを用いているため，2段落目以降に書かれる情報，例えば背景知識のような情報に関するスクリプト知識を獲得することはできない．

この問題を解決するためには，例えば，テキストの構造を利用することが考えられる．修辞構造理論などを用いてテキスト構造が明らかにされている場合，第一段落以外の段落についても時間順序関係を決定できる可能性がある．

また，小さな事件で一度しか報道されない事件などは個々のテキスト集合が小さくなってしまうため，スクリプト知識を獲得することが難しいという問題がある．このような場合も含め，新聞記事で大きく取り上げられない出来事についてのスクリプト知識を獲得するためには，入力テキスト集合の生成に異なる方法を用いる必要がある．

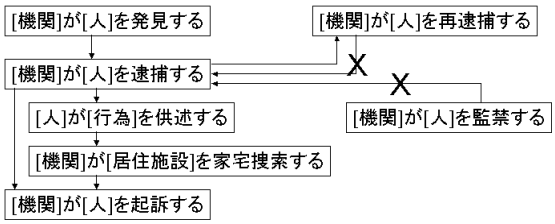

図 2: 「殺人事件」に対するスクリプト知識の一部

4　おわりに

類似するトピックを持つ複数のテキスト集合からのスクリプト知識自動獲得手法を提案し，その手法を用いて実際に自動獲得を行った．

今後の課題としては，本手法で獲得することのできない種類のスクリプト知識を獲得するために手法を拡張することや，獲得したスクリプト知識を外部タスクで用いることによって，精度を客観的に評価することが挙げられる．

参考文献

[1] 内田ユリ子，石崎俊，井佐原均，実際的な知識に基づく文脈表現構造からの英語テキスト生成，電子情報通信学会論文誌 D-II, Vol.J72-D-II, No.9, 電子情報通信学会，1989.

[2] 堀浩一，齋藤忠夫，猪瀬博，スクリプトの文章からの帰納的な学習，自然言語処理研究報告，37-5, 情報処理学会，1983.

[3] 西岡真吾，今一修，汎用連想計算エンジン GETA とそれに基づく連想検索システム，情報処理学会研究報告，2000-NL-137, 情報処理学会，2000.