E-002

演劇理解に関する常識知識ベースの構築方式

The construction method of commonsense knowledge base for understanding the conversation of play

荻原 寛† Hiroshi Ogihara　渡部 広一† Hirokazu Watabe　河岡 司† Tsukasa Kawaoka

1．はじめに

人間にとって便利な道具であるコンピュータは，今後，人間との双方向会話によるコミュニケーション機能を備え，また，「常識」を理解することが期待される．我々のプロジェクトでは，感覚や感情をはじめ，歴史や地理，音楽など常識に関する幅広い研究を進めているが，本稿では様々な日常会話の中から「演劇」に関する会話を成立させるための基礎知識の構築を目的とする．

本研究は，演劇の内容に特化した演劇概念ベースと演劇シソーラスの構築を行う．またこの知識のことを「演劇知識」と呼ぶ．本研究で構築した知識の評価は，演劇についての常識的な会話が行えるかどうかで判断する．常識的な会話とは，一つの質問に対して一つの答えを返すことと言えるので，一問一答形式の会話を可能にした，演劇常識判断システムを構築し評価を取る．

2．連想メカニズム

2.1 概念ベース

ある語Aをその語と関連の強いと考えられる語a_iと重みw_iの対の集合として定義する．

$$A = \{(a_1, w_1), (a_2, w_2), \ldots, (a_m, w_m)\} \quad (1)$$

ここで，a_iを 1 次属性と呼ぶ．また便宜上，Aを概念表記と呼ぶ．このような属性の定義された語（概念）を大量に集めたものを概念ベースと呼ぶ．ただし，任意の 1 次属性a_iは，その概念ベース中の概念表記の集合に含まれているものとする．すなわち，属性を表す語もまた概念として定義されている．したがって，1 次属性は必ずある概念表記に一致するので，さらにその 1 次属性を抽出することができる．これを 2 次属性と呼ぶ．概念ベースにおいて，「概念」はn次までの属性の連鎖集合により定義されている．

本研究では，電子化国語辞書から，各見出し語を概念表記，その見出し語の説明文中の自立語を 1 次属性として抽出し，出現頻度に基づく重みを付加した約 4 万の概念からなる概念ベースを基に，新聞などから抽出した概念表記や属性を加え，質の向上を目的にした精錬操作（属性の追加・修正など）を施し，更に，概念間に成り立つ一般的なルールに基づく適切な重みを付加した約 9 万の概念からなる概念ベース[1]を構築し利用している．

2.2 関連度計算方式

関連度とは，概念と概念の関連の強さの度合いであり，関連度計算とは，その度合いを定量的に評価するものである．その具体例を表 1 に示した．「果物」と「夕日」より，「果物」と「林檎」の方が，関連が近く，関連度が高いということになる．具体的には概念連鎖により概念を 2 次属性まで展開したところで，最も対応の良い一次属性同士を対応付け，それらの一致する属性個数を評価することにより算出するものである．

表 1 関連度計算の例

概念 A	概念 B	概念 A と概念 B の関連度
果物	林檎	0.26
果物	夕日	0.01

3. 演劇知識の定義

「演劇，歌劇，舞踊，映画」を演劇と定義し，演劇に関する会話（質問返答）を可能にするための知識を演劇知識と定義する．この演劇知識は，演劇概念ベースと演劇シソーラスの二つで構成されている．演劇知識は，演劇に関する用語の概念や意味などの基礎知識と，俳優の名前や映画のタイトルなど流行に影響される表層的かつ流動的な知識の 2 つに大きく分かれる．前者を演劇知識（下階層），後者を演劇知識（上階層）とした．（図 1）本稿では演劇知識（下階層）についての構築を目的とする．（尚，これ以降『演劇知識』と書いている際は，演劇知識（下階層）のことである．）

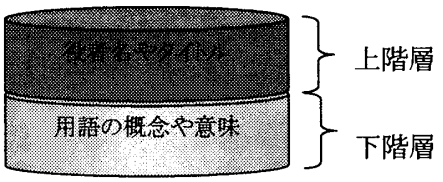

図 1．演劇知識の大まかな分類

4. 演劇知識の構築

演劇知識は，演劇概念ベースと演劇シソーラスの二つで構築されている．

4.1 演劇概念ベース

演劇概念ベースは，演劇に関する知識を概念と属性の組にした知識形態である．概念ベースの構築には Yahoo！や goo の辞書，演劇用語を説明している Web サイトから知識を収集した後，茶筌による形態素解析を行い，自立語を抜き出した．見出し語（説明文のタイトル）を概念，自立語を属性として概念ベースを構築する．しかし初期の演劇概念ベースでは概念に対して不適な属性（雑音）があるため，目視による判断で確実に 1 つずつ除去していき精錬した．

更にオートフィードバック[2]を用いて Web から属性を

†同志社大学大学院　工学研究科
Graduate School of Engineering,Doshisha University

自動に取得し，演劇概念ベースに属性を追加した．オートフィードバックにより新たな概念と属性を取得することができ，概念ベースの拡張が出来た．しかし，オートフィードバックにおいても不適切な属性も取得されるため，ここでも目視で確認しながら雑音を除去している．表２は演劇概念ベースの一部である．

表２　演劇概念ベース（一部）

概念	属性
映画	フィルム、映像、映写機、…
演劇	俳優、舞台、身ぶり、芝居、…
歌舞伎	舞踊、伝統演劇、出雲阿国、…

4.2　演劇シソーラス

演劇シソーラスは，演劇に関する知識を体系的に整えたものである．体系的に整えることにより，概念同士の上位，下位の関係などを容易に知ることが出来る．演劇シソーラスは，映画に関する知識をまとめた「映画ツリー」（図２）と，演劇，歌劇，舞踊に関する知識をまとめた「舞台芸術ツリー」の２つで構成されている．

まず，NTT シソーラス[3]から演劇に関するノードを抜き出し，新しいツリー構造を構築した．新しいツリー構造を構築した後，NTT シソーラスのリーフに登録してあった知識を，新しく構築したツリーの各ノードにリーフとして追加した．さらに演劇概念ベースの知識をシソーラスに追加し拡張をした．

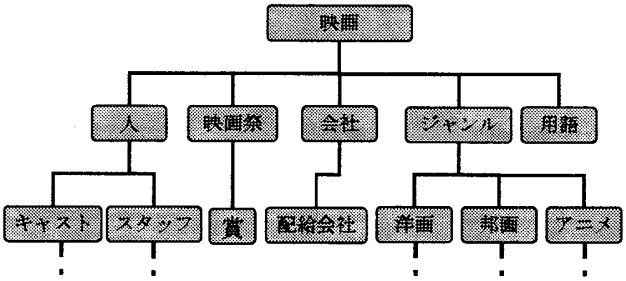

図２．演劇シソーラス－映画ツリー（一部）

5. 演劇問題の定義

演劇に関する一問一答形式の問題を演劇問題と定義する．尚，映画のタイトルや俳優の名前など演劇知識（上階層）の知識を必要とする質問は扱わないことにする．

6. システムの構成

演劇常識判断システムは，入力された質問が演劇に関する内容であるかを判定する判定部と，質問に対する返答をする返答部で構成されている．判定部は演劇シソーラスを用いて，判定する語が演劇シソーラスに存在しているかで判断する．演劇に関する質問であると判断されれば，返答部に処理が移り演劇概念ベースを用いて質問に対する応答を行う．

6.1　演劇問題の判定

判断システム（判定部）は自然言語の文章が演劇に関する質問であるのかを判断する．この判断が出来るということは，「歌舞伎というものはどのようなものか？」などのように演劇に関しての自然言語による入力をコンピュータが理解できているといえる．

雑多な文章の中から演劇に関する文章のみを判定し，抽出することが可能となれば，会話処理を実現する上で，全検索を行う必要が無くなり，会話処理の応答時間の短縮につながることとなる．

6.2　演劇問題の回答

演劇問題の回答に関しては演劇概念ベースを用いて，演劇に関する一問一答形式の問題の意味理解を行う．このシステムを用いる際に，演劇に関する情報文とそれに対応する知識（見出し語）をセットにした演劇知識ベースを用いる．しかし，日常会話にはさまざまな知識があり，それをすべて知識ベース（表３）で登録させるのは困難であり，効率が非常に悪い．そこで，知識ベースには代表的な情報文のみを登録し，演劇概念ベースや演劇シソーラスにより構築した連想システムを用いて，知識の拡張を行う．

表３　演劇知識ベース（一部）

見出し語	情報文
演劇①	観客を前に、俳優が舞台で身ぶりやせりふで物語や人物などを形象化し、演じて見せる芸術
演劇②	舞台装置・照明・音楽など視覚・聴覚上の効果を伴う総合芸術

演劇問題の回答の流れとして，質問文を自立語の単語列に切り分ける．質問文を$\underline{X}$とすると，次式のように表される．

$$\underline{X}=\{(x_1, w_1),\ (x_2, w_2),\ \cdots\ (x_n, w_n)\}$$

（x：自立語，w：重み）

演劇概念ベース中の句読点区切り列の属性列をコンマ区切りの属性列にする．

$$A_1=\{(a_{11}, w_{11}),\ (a_{12}, w_{12}),\ \cdots\ (a_{1n}, w_{1n})\}$$
$$A_2=\{(a_{21}, w_{21}),\ (a_{22}, w_{22}),\ \cdots\ (a_{2m}, w_{2m})\}$$
$$\vdots$$

質問文$\underline{X}$と最も関係の深い見出し語を探すために演劇概念ベースを用いた関連度計算[4]を用いる．

7. 評価と考察

演劇知識の評価を取る．質問判定の評価の方法は，アンケートなどで集めた演劇に関する質問 200 文と，演劇とは関係ない質問 200 文（地理，歴史，音楽などに関係する質問）を演劇常識判断システム（判定部）にかけ，どの程度の正解率を得られるかで判断する．また，返答部の評価に関しては，演劇に関する質問 200 文を「表記一致，関連度計算（重みを一律に付与）」の２手法で処理し，どの程度の正解率を得られるかで判断する．表記一致の評価と比較し，表記一致に対して重みを一律に付与した関連度計算は正解率にどのような差が出るかを検討する．尚，システムにかけた質問文の例を次に示す．

質問文の例
- 映画で演技をする人で女の人は
- 最低の映画に贈られる賞は？
- バッハはどこの国の作曲家ですか？
- 日本一広い平野は何平野ですか？

まず判定部の評価結果は表４のようになった．演劇に関する質問の場合はシステムの返答が関係有りで○，関係なしで×とする．演劇に関係ない場合は，その逆とする．精度は『○の数/質問文総数』とする．

表４　質問判定の結果

試験データ	○	×	精度
演劇に関する質問	179	21	89%
演劇に関しない質問	192	8	96%

結果より，質問判定に関しては演劇に関係ある場合，関係ない場合の両方で高い精度を出すことができた．誤答の主な理由としては，索引語が演劇シソーラスに存在しない場合と，索引語がほかの分野で使われる場合の二つがあげられる．間違えた例は次のようになった．

演劇シソーラスに存在しない場合
- 二人で滑稽な問答を中心に演じる寄席演芸を何という?
- アメリカを代表するアニメ製作会社は？

間違えた理由は，演劇シソーラスには「演芸，アニメ，製作会社」という言葉は存在しているが，質問文を形態素解析にかけて自立語として扱う際に「寄席演芸，アニメ製作会社」と解析してしまったためである．これは，映画祭などのように映画＋祭のような「名詞＋名詞の言葉は結合させる」というルールを今回用いているためである．

索引語がほかの分野で使われる場合
- バッハはどこの国の作曲家ですか？
- 「新世界より」で知られる作曲家は？

これは作曲家という索引語で間違ってしまっている．作曲家という言葉は，音楽の世界で使われる言葉であるが，映画の世界でも映画の BGM を作曲する人のことを作曲家という．このように演劇の世界とそのほかの世界で使われる言葉については間違ってしまっていた．

演劇問題返答部の評価結果は表５のようになった．これは表記一致と関連度計算の二つの手法の結果を比較したものである．質問に関して適切な回答の場合を○，適切な回答ではないが間違いではない場合を△，誤答や複数の回答候補が出てきた場合を×とする．尚，精度は『○の数/質問文総数』とする．また,質問回答の例をいくつか示す．

表5　二つの手法の比較

	表記一致	関連度計算
○	78	111
△	9	12
×	113	77
精度	39%	55%

質問回答の例

表記一致○，関連度計算○
- 映画の撮影現場などでカメラを使って撮影する人をなんと言う？

表記一致×，関連度計算○
- 歌と踊りで物語を表現したものをなんと言う？

表５より，表記一致より重みを一律に付与した関連度計算の方が高い正解率を得ることができた．しかし，まだまだ改善するところもあり，演劇用語の処理に重きを置いているが，重みを一律に付与しているために，一般的な単語の重要性と演劇用語の重要性にあまり差が出ていない．これを改善するためには，演劇用語の重みを重くするなどの，演劇知識に特化した新たな重み付け手法などの考案が必要となる．

8. おわりに

本研究では演劇概念ベースと，演劇シソーラスを用いて演劇に関する知識を構築した．構築する際に，オートフィードバックを用いて Web から自動的に属性を取得するという手法も試みた．しかし，構築した演劇概念ベースには不適切な属性である雑音が多かったため，目視による雑音除去を行った．しかし，目視による除去は効率的ではないため，今後効率的な手法があれば採用をしていく．

また，構成した知識を演劇常識判断システムにかけることにより，高い精度での質問文判定を可能とした．質問返答に関しては 55%の正答率で質問文に対し適切な応答を返すことが可能となった．しかし，今回は演劇概念ベースに重みを一律に付与して関連度計算を行った．そのため，演劇に関する言葉の重要性が薄れてしまった．

今後，演劇に関する言葉の重みを重くするなど属性としての重要性を考慮したうえで，適切な重みを付与することにより，さらなる演劇知識を構築することが期待される．

謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行ったものである．

参考文献

[1] 広瀬 幹規，渡部 広一，河岡 司：概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精錬手法，信学技報，NLC2001-93，pp.109-116，2002
[2] 辻 泰希，渡部 広一，河岡 司：www を用いた概念ベースにない新概念およびその属性獲得手法，人工知能学会全国大会論文集，2D1-01，2004
[3] NTT コミュニケーション科学研究所監修，「日本語語彙体系」，岩波書店，東京，1997
[4] 井筒 大志，東村 貴裕，渡部 広一，河岡 司：概念ベースを用いた連想機能実現のための関連度計算方式，情報科学技術フォーラム FIT2002，E-39，pp.159-160，2002