対話システムのための因果関係知識を含む文からの関連語抽出

藤田基靖* 荒木健治 ジェプカ・ラファウ
(北海道大学大学院情報科学研究科)

1 はじめに

対話の内容が特定されていない非タスク指向型対話システムにおいて，人間が違和感なく対話を続けられるシステムの実現には至っていない[1][2]．これは，現在の自然言語処理技術では，システムが話者の意図を理解した上で応答を返すことが困難なためである．この問題を解決するため，本手法では文中に含まれる因果関係知識に着目する．因果関係にある文は，「原因」と「結果」の関係で結ばれ，一貫した主張で意義深い表現を含むものが多い．この知識を利用できれば，人間が行う発話に近い発話をシステム上で実現することが期待できる．よって本稿では，因果関係にある文からその関連語を取り出し，利用する手法について提案する．

2 因果関係知識を含む文

本手法で扱う因果関係にある文は，文中に接続表現の「ため・から・ので」が含まれるかどうかで判断する．接続表現を「ため・から・ので」としたのは，これらを含む文は２つの事象間を結びつけ，多くの場合において高い必然性を持つ因果関係を表現しているからである[3]．また，「ため・から・ので」が含まれる文であっても，「北から」や「なので」といった表現も「から」や「ので」を含むため，これだけで因果関係知識を含む文と断定することはできない．そのため，Web を日本語でテキスト化したコーパスである JpWaC[4]から，因果関係を持つ「ため」，「から」，「ので」を含む文をそれぞれ 50 文ずつ抜き出し，形態素解析エンジン MeCab[5]によりその品詞について調査を行った．その結果どの接続表現も，因果関係を含む文である場合は特定の品詞に分類されることを確認した．各接続表現について，「ため」は「名詞,非自立,副詞可能」，「から」は「助詞,接続助詞」，「ので」は「助詞,接続助詞」であるものが因果関係知識を含む文となっており，以後本手法ではこれらの品詞に分類される「ため」，「から」，「ので」を含む文を因果関係知識を含む文として扱う．さらに，単文には「原因」と「結果」の２つが含まれないため，単文以外のものを対象とする．以上の条件に合致する文を，本手法では因果関係知識を含む文とする．

表１ 接続表現の品詞分類

接続表現	品詞，品詞細分類 1，品詞細分類 2
ため	名詞，非自立，副詞可能
から	助詞，接続助詞
ので	助詞，接続助詞

3 目指す対話とその具体例

前章までに述べた点を踏まえて，因果関係にある例文を用いて，実際にどのような関連語が得られ，どのような応答が可能になるかについて述べる．

（１）「風邪を引いたので，薬を飲みました」

例文（１）は「風邪」を「薬」で「治療」したこと暗に伝えている．人間は瞬時に理解できる内容であるが，システムがこの内容を理解することは難しい．しかし，「風邪」と「薬」の因果関係知識から，明示されていない関連語である「治療」が取り出せれば，以下の対話例が可能になると考えられる．

ユーザ：「風邪を引いたので，薬を飲みました」
システム：「治療したのですね」

上記の対話例でユーザの発話に対するシステムの応答は，一般的な因果関係に基づく知識により得られたものである．この対話例のように，因果関係を含む文から関連語を取り出して応答ができれば，人間が無意識に理解できることをシステムも理解しているように見せることができる．そうすれば人間らしい意味を理解した対話が，システムとの間でできているように見えるのではないか．これが本研究で目指す対話システムである．

4 関連語抽出処理の概要

本手法では，因果関係知識を含む文から知識を抽出する際に，知識源として Web を用いる．Web 上の情報はジャンルの偏りが少なく，膨大な量の文を容易に入手できるためである．Web 知識源の有効性は，相良らの研究

*fuji_riv@media.eng.hokudai.ac.jp
†札幌市北区北 14 条西 9 丁目北海道大学大学院情報科学研究科

[6]で述べられており，検索エンジン Google[7]のスニペットを知識源として用いる．また，知識を抽出する際の重要語として「名詞」を扱い抽出する．「名詞」は，他の品詞よりも扱いやすく，多くの情報を集めることができるためである．さらに，抽出した「名詞」に加えてシソーラス[8]を用いて上位語を抽出し利用することで，多くの関連語を利用する．

本手法の処理過程について，発話を解析して入力を分割・シソーラスなどを獲得する「解析部」と，獲得した語を基に，Web 検索結果のスニペットを用いて関連語を抽出する「関連語抽出部」の２つに分け，例文「風邪を引いたので，薬を飲んだ」を用いて説明する．

まずは，「解析部」について解説する．最初に，接続表現の前後で文を２つに分ける．この分割により，例文からは「風邪を引いた」，「薬を飲んだ」という２つの表現を得る．また，分割後の「風邪を引いた」から名詞を抽出すると「風邪」が得られ，同じく「薬を飲んだ」からは「薬」を得る．続いて「風邪」と「薬」のシソーラスを解析することで，「風邪」の上位語「病気・体調」と「薬」の上位語「薬剤・薬品」を得る．以上の処理により得られた表現が，関連語抽出に用いるキーワードとなる．この例では，「風邪を引いた」，「薬を飲んだ」，「風邪」，「薬」，「病気・体調」，「薬剤・薬品」の６つがキーワードである．

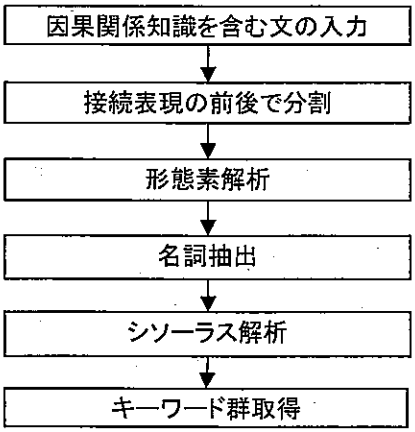

図１　解析部

次に「関連語抽出部」について説明する．まず，「解析部」で得られたキーワードに対して，検索エンジン Google を用いて Web 検索を行う．利用するスニペットは，Google の検索結果上位１００件中のスニペットである．これは１００件から３００件のスニペット中に表れる名詞を，出現頻度順に記載した表２で示すように，出現頻度３００件まで増やした結果を見ても，上位の名詞に大幅な変化はない．このため，Google の１ページに表示できる最大件数である１００件中のスニペットより関連語抽出を行う．このスニペットから，関連語抽出に関係の薄い表現である「http」，「www」，「関連ページ」などをあらかじめ取り除き，このスニペットから，「名詞,一般」，「名詞,サ変接続」である語を抽出する．この時１００種類ほどの「名詞」が抽出されるが，出現回数の多さが上位７位までの「名詞,一般」，「名詞,サ変接続」のみを残す．上位７位までにした理由については，「名詞，一般」の中で出現回数が多い上位３０まで

表２　検索結果スニペット中の出現回数の高い名詞

キーワード	検索結果スニペット中の出現回数の高い名詞上位 30
風邪	**スニペット上位１００件** インフルエンザ, 症状, ウイルス, 湯, かぜ, レシピ, 咽頭, 民間, 療法, 夏, 感冒, ビタミン, 記事, 方法, 薬, 商品, 子供, 動画, 鼻, 症候群, 漢方薬, ツール, 熱, 胃腸, 原因, キーワード, 喉, 漢方, メンバー, 外来, C **スニペット上位２００件** 症状, ウイルス, インフルエンザ, かぜ, 胃腸, おたふく, 鼻, 気, 咽頭, 夏, 鼻水, 猫, 湯, 症候群, 疾患, 原因, 療法, 咳, 薬, ツール, 急性, 民間, 感冒, レシピ, 子供, ウィルス, ビタミン, 頭痛, 効果, キーワード, 記事 **スニペット上位３００件** 症状, インフルエンザ, ウイルス, かぜ, 胃腸, 原因, 気, 咽頭, おたふく, 鼻, 薬, 夏, レシピ, 体, 疾患, 咳, 季節, 症候群, 急性, 感冒, 子供, 湯, 効果, 栄養, 医学, ツール, 喉, ウィルス, 喘息, 熱, 鼻水
薬	**スニペット上位１００件** 医薬品, 医療, 薬剤師, くすり, 株式会社, 精神, 膳, 専門, 病院, 知識, 漢方, 家庭, 症状, 企業, 医学, 薬事, 技術, 全国, 茶, 麻薬, 中心, 臨床, 胃腸, バイオ, 価値, ジェネリック, 食, 食品, 薬品, 副作用, 辞書 **スニペット上位２００件** 膳, 医薬品, 薬剤師, 医療, 専門, 企業, くすり, 株式会社, 食品, 症状, 病院, 知識, 茶, 精神, 漢方, 家庭, 技術, 会社, 医学, 薬事, 会員, 全国, 薬品, 副作用, 漢方薬, 臨床, 胃腸, 皆様, 書籍, 商品, 症候群 **スニペット上位３００件** 膳, 医薬品, 薬剤師, 医療, 漢方, 病院, 食品, 症状, 専門, 企業, 株式会社, くすり, 知識, 精神, 効果, 会社, 家庭, 医学, 商品, 医師, 技術, 茶, 麻薬, 副作用, ツール, 薬局, 風邪, 会員, 全国, 中心, 臨床

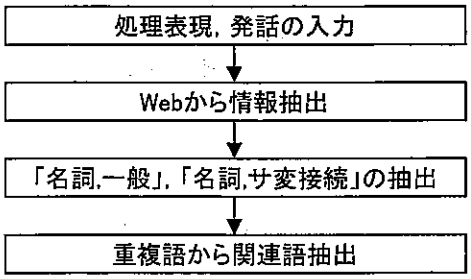

図２　関連語抽出部

5 評価実験

４章で述べたシステムの評価実験を行う．この評価実験の目的は，「因果関係知識を含む文を入力」，「関連語を応答」と仮定した際に，この組み合わせが対話として成り立つかを調べることである．

実験に用いるデータセットの作成には，JpWaC コーパスを用いる．このコーパスから２章で述べた基準を基に，因果関係知識を含む単文以外の文を抽出する．また，通常の発話と似たデータセットにするため，含まれる単語数が 15 以下となる文を抽出し，抽出された文から 50 文をランダムに抽出し，これを実験のデータセットとした．

このデータセットを入力と仮定し，その入力全てに本システムを用いて因果関係知識にかかわる関連語抽出を行う．得られた関連語について「名詞,一般」であるものは語尾に「ですね」を補い，「名詞,サ変接続」であるものの語尾には「するのですね」を補う．以上の処理によって得られたものを，各入力に対する応答と仮定して，入力に関連した応答として適切であるかを評価する．以下に実験結果の一部を示す．

入力：「昨夜は早寝したので，今朝は５時半に目覚める」
応答：「早起きするのですね」

上記結果で，入力は「早寝」と「５時半に目覚める」となっており，一般的には因果関係が成立すると考えられる．それに対して，関連語は「早起き」を獲得しており，入力の意味を読み取って，関連した応答をしていると考えられる．

上記の実験結果のように，全てのデータセットについて応答の生成を行い，精度評価を行う．精度の評価点は，「入力が因果関係知識を含む文であるか」，「応答が入力に関連しているか」の２点である．入力の評価は，「因果関係知識を含む文である」，「因果関係知識を含む文ではない」，「どちらともいえない」の３段階評価とする．また，応答の評価は，「因果関係知識に関連している」，「曖昧な応答である」，「因果関係知識に関連していない」の３段階評価とし，生成された応答全てにこの評価を行い，一番多かった評価をその応答の評価とする．応答がなかった場合はこの評価を行わない．評価者は著者らを含まない自然言語処理の研究に従事する２０代の理系大学生３名である．以下は得られた実験結果の一部である．

入力：山の中なので電波が届かないのです．
応答：電話するのですね

入力は「山の中」と「電波が届かない」といった関係性のため，一般的な因果関係が成立するといえる．応答は，「電話」という関連語から「電話するのですね」となった．山の中で携帯電話が通じないというユーザの意図を理解した応答と考えることができるため，因果関係知識に関連した応答であると考えられる．

入力：チタンレールを使っているので，軽量です．
応答：素材ですね

入力は，「チタンレール」と「重さ」の関係について述べており，得られた関連語が「素材」となっている．物体の軽さにチタンレールという素材が関与していることを読み取っており，隠れた意味を読み取った因果関係知識に関連した応答であると考えられる．

入力：私は眠くなった・・・ので，一眠り.a.
応答：応答なし

入力は「眠くなったため，眠った」という内容であるが，文中に「・・・」がある，語尾に「.a」がついているなど，文法に問題があるといえる．このため，応答も生成されなかった．これは，接続表現のみから因果関係知識の抽出を行ったこと，Web 上の文書を対象としたことの２点が原因として考えられる．誤りがあった１４例の内，この誤りに該当するものは３例あり，誤り全体の２１.４%になるため，この問題は実験の条件を変えない限り今後も起こりうることが予想される．

以上のように実験結果５０例に対して，評価を行い，表５に示す．データセット中で，因果関係にあると認められた文は３６あり，その抽出精度は７２.０%となる．

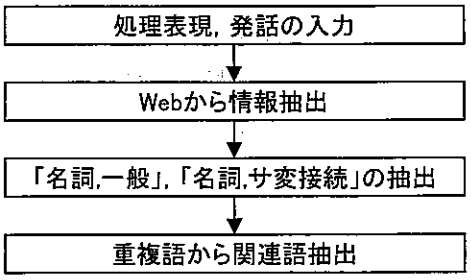

図２　関連語抽出部

5 評価実験

４章で述べたシステムの評価実験を行う．この評価実験の目的は，「因果関係知識を含む文を入力」，「関連語を応答」と仮定した際に，この組み合わせが対話として成り立つかを調べることである．

実験に用いるデータセットの作成には，JpWaC コーパスを用いる．このコーパスから２章で述べた基準を基に，因果関係知識を含む単文以外の文を抽出する．また，通常の発話と似たデータセットにするため，含まれる単語数が 15 以下となる文を抽出し，抽出された文から 50 文をランダムに抽出し，これを実験のデータセットとした．

このデータセットを入力と仮定し，その入力全てに本システムを用いて因果関係知識にかかわる関連語抽出を行う．得られた関連語について「名詞,一般」であるものは語尾に「ですね」を補い，「名詞,サ変接続」であるものの語尾には「するのですね」を補う．以上の処理によって得られたものを，各入力に対する応答と仮定して，入力に関連した応答として適切であるかを評価する．以下に実験結果の一部を示す．

入力：「昨夜は早寝したので，今朝は５時半に目覚める」
応答：「早起きするのですね」

上記結果で，入力は「早寝」と「５時半に目覚める」となっており，一般的には因果関係が成立すると考えられる．それに対して，関連語は「早起き」を獲得しており，入力の意味を読み取って，関連した応答をしていると考えられる．

上記の実験結果のように，全てのデータセットについて応答の生成を行い，精度評価を行う．精度の評価点は，「入力が因果関係知識を含む文であるか」，「応答が入力に関連しているか」の２点である．入力の評価は，「因果関係知識を含む文である」，「因果関係知識を含む文ではない」，「どちらともいえない」の３段階評価とする．また，応答の評価は，「因果関係知識に関連している」，「曖昧な応答である」，「因果関係知識に関連していない」の３段階評価とし，生成された応答全てにこの評価を行い，一番多かった評価をその応答の評価とする．応答がなかった場合はこの評価を行わない．評価者は著者らを含まない自然言語処理の研究に従事する２０代の理系大学生３名である．以下は得られた実験結果の一部である．

入力：山の中なので電波が届かないのです．
応答：電話するのですね

入力は「山の中」と「電波が届かない」といった関係性のため，一般的な因果関係が成立するといえる．応答は，「電話」という関連語から「電話するのですね」となった．山の中で携帯電話が通じないというユーザの意図を理解した応答と考えることができるため，因果関係知識に関連した応答であると考えられる．

入力：チタンレールを使っているので，軽量です．
応答：素材ですね

入力は，「チタンレール」と「重さ」の関係について述べており，得られた関連語が「素材」となっている．物体の軽さにチタンレールという素材が関与していることを読み取っており，隠れた意味を読み取った因果関係知識に関連した応答であると考えられる．

入力：私は眠くなった・・・ので，一眠り.a.
応答：応答なし

入力は「眠くなったため，眠った」という内容であるが，文中に「・・・」がある，語尾に「.a」がついているなど，文法に問題があるといえる．このため，応答も生成されなかった．これは，接続表現のみから因果関係知識の抽出を行ったこと，Web 上の文書を対象としたことの２点が原因として考えられる．誤りがあった１４例の内，この誤りに該当するものは３例あり，誤り全体の２１.４%になるため，この問題は実験の条件を変えない限り今後も起こりうることが予想される．

以上のように実験結果５０例に対して，評価を行い，表５に示す．データセット中で，因果関係にあると認められた文は３６あり，その抽出精度は７２.０%となる．

応答文が生成されたのはデータセット５０文の内３５文からで，その生成率は７０.０％であった．また，因果関係知識を含む３６文から応答文が生成されたものは２０文で，その精度は７５.９％であった．さらにその２０文の中で，因果関係知識に基づく応答をしているものは１４文あり，データセットから因果関係知識に基づく応答を行う精度は２８.０％となった．因果関係知識を含む文のみからで考えると，因果関係知識に基づく応答ができる精度は３８.９％となり，データセットからの生成と比べると精度は向上している．

表５　抽出された入力と生成応答文の精度

抽出された入力，生成される応答文	該当	精度
因果関係知識を含む入力の抽出精度	36	72.0%
入力から応答を生成する精度	35	70.0%
因果関係知識を含む文から関連語を生成する精度	20	55.6%
因果関係知識に関連した応答を生成する精度	16	32.0%
因果関係知識を含む文からその知識に関連した応答を生成する精度	14	38.9%

6 考察

因果関係知識を含む文の抽出精度は 72.0%，因果関係知識を含む文から得られる関連語抽出精度は 38.9%であった．これらの精度が満足できる値であるかは，まず人間同士の対話から因果関係を捉えた応答を行う割合を調査する必要がある．その上で，発話者の意図を対話システムが理解していると感じる精度が，人間同士の対話と同程度の精度であるかを調査する必要がある．

次に，因果関係知識を含む文の抽出精度について述べる．誤りの理由として，「原因」と「結果」が因果関係になかったこと，Web の文表現に由来する誤りがあったことが挙げられる．これらの誤りは，文内の命題要素を抽出しその判定を行い，正しい因果関係が含まれる文のみを抽出することで改善できると考えられる．

関連語抽出の誤りとして，複数の関連語を獲得したことや，関連語を抽出できなかった点が挙げられる．そのため，関連語が複数得られた際にさらに絞り込みを行い，関連語が抽出できなかった際は，逆に絞り込む条件を緩和することで，精度が向上できると考えられる．

7 まとめと今後の課題

本研究では，Web コーパスである JpWaC を用いて因果関係知識を含む文の抽出を行い，検索エンジン Google を用いて，抽出した文から関連語を獲得し，対話の応答として適切であるかの評価を行った．評価点は，因果関係知識を含む文の抽出精度と，その文から抽出した関連語が因果関係知識に基づいているかである．因果関係知識を含む文の抽出精度は７２.０％，因果関係知識に基づいた応答を生成する精度は３８.９％であった．今後は対話システムへの利用を目的とした達成目標となる抽出精度の調査を行い，命題要素を因果関係知識の判定へ利用及び，関連語抽出アルゴリズムの変更により精度向上を目指す．

参考文献

[1] J.Weizenbaum,"ELIZA—A computer program for the study of natural. :language communication between man and machine,"Commun.. :ACM, vol.9, no.1, pp.36-45, 1966.

[2] Wallace, R. The Anatomy of A. L. I. C. E. , http://www.alicebot.org/anatomy.html

[3] 接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得. :乾孝司, 乾健太郎, 松本裕治. :情報処理学会論文誌, Vol.45, No.3, pp.919-933, 2004.

[4] JpWaC(Japanese Web as Corpus), http://kilgarriff.co.uk/

[5] MeCab, http://mecab.sourceforge.net/

[6] 質問応答に対する知識源としての Web 検索エンジンの Snippet の有効性. :相良春樹, 森辰則, 中川裕志. :言語処理学会第 12 回年次大会発表論文集, pp.316-319, 2006.

[7] Google, http://www.google.co.jp/

[8] 分類語彙表 増補改定版. :国立国語研究所編, 大日本図書刊, 2004.

[9] 因果の言語学. :有田節子,月間言語, Vol.25, No.5, pp.20-23, 1996.