様式 C-19

科学研究費補助金研究成果報告書

平成　22 年　3 月　19 日現在

研究種目：若手研究(B)
研究期間：2008～2009
課題番号：20700138
研究課題名（和文）　照応解析技術を用いた意見の帰結・根拠情報の構造化

研究課題名（英文）　Detecting Conclusion-Evidence Relations in Web Texts Using Anaphora Resolution Technique

研究代表者
飯田 龍（IIDA RYU）
東京工業大学・大学院情報理工学研究科・助教
研究者番号：40464276

研究成果の概要（和文）：
本研究課題では，文章中に記述された書き手の意見情報の抽出問題を，「よい」「すばらしい」などの評価表現や「～と思う」「～らしい」といった表現を伴って出現する意見の“帰結”部分と帰結部分を修飾する“根拠”が記述された談話単位を構造化する問題として扱う．我々はこれまでに談話中において同一実体を指す表現を同定する照応解析の技術構築に取り組んできており，そこで得た知見をこの根拠帰結関係同定に援用することで高品質の関係同定モデルを実現する．モデルの有効性を調査するための評価実験の結果，我々のモデルが Web 文書や新聞記事に出現する根拠帰結関係の同定に関してベースラインとなる既存研究のモデルより高精度で関係同定を実現できることを示した．

研究成果の概要（英文）：
This research focuses on the task of detecting Conclusion-Evidence relations in texts, which consist of evaluative expressions such as ``good'' and ``excellent'' and their evidence expressions. We have developed a technique of resolving anaphoric relations in texts, which is one of clues to capture discourse relations occurring in texts. In order to incorporate such anaphora resolution technique into Conclusion-Evidence relation detection, we refer to the ideas exploited in the works by Iida et al. (2004) and Iida et al. (2005). As we conduct empirical evaluation of Conclusion-Evidence relation detection, we reported that our model achieved better performance of the relation detections appearing in both Web texts and newspaper articles in comparison to a baseline model.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	2,100,000	630,000	2,730,000
2009 年度	1,300,000	390,000	1,690,000
年度			
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：自然言語処理
科研費の分科・細目：情報学・知能情報学
キーワード：照応解析，根拠帰結関係同定

1. 研究開始当初の背景

blogなどのConsumer Generated Mediaから消費者の意見情報を獲得する問題が新しい技術課題として定着しつつある. 研究開発当初の時期には, 意見情報抽出の課題はおおきく(1)文章全体が肯定的な意見を述べているか否定的な意見を述べているかの分類, (2)文章中から〈ipod, 操作性, 良い〉といった情報抽出的な課題として研究が進められていた. 前者は記事全体を分類するため粒度が荒いという問題があり, 後者は文章中から個別に意見断片を収集するため, それらの間の関連性やその意見断片の根拠を知るためにはユーザがあらためて当該文章を読む必要があるといった問題が存在した.

2. 研究の目的

1.で述べた問題点を解決するため, 本研究ではユーザの意見の根拠箇所同定する問題を考える. 具体的には, 文章から根拠を抽出する課題を設計し, それを自動同定する技術を開発する. これにより, ユーザが Web 上に蓄積された意見を閲覧する際に「どのような意見がどういった根拠で述べられているか」という情報を提示する新たな情報サービスを行うための基盤を提供する.

例えば, 以下の例では, 書き手の意見だけを抽出した場合, ipod touch について「操作性などが申し分ない」ことがわかるだけだが, この意見の根拠として, 「他の商品との比較を行っている」「特殊な状況下での意見を述べている」「(個人の経験として) 実際に購入している」といった意見の根拠を参照することで, この意見の読み手が情報の品質を吟味する際の有益な情報の一つとすることができる.

例――――――――――――――――――――――――――
しばらく忙しくて買い物にも行けなかったんですが,
1.今週やっとipod touchを購入しました. 少し触ってみた感想としては
2.これまでのipodと比較しても,
3.操作性や質感・物理的な構造については申し分なし.
4.とくに, Safariで1文字の高さが2〜3ミリくらいの時にリンクをタップしても
5.的確にターゲットを拾ってくれるのはなかなかに感動的.
6.非常に満足です.
――――――――――――――――――――――――――

このような談話セグメント間の関係を同定する問題に関して, まずセグメント間に根拠帰結の関係があるか否かを同定する問題に取り組む.

3. 研究の方法

「根拠帰結」と一言で言っても, 具体的にどのような関係を同定の対象とするか, またどのような談話の範囲の間に関係を認めるかは自明ではない. そこで, 本研究では既存研究のうち最も関連するPenn Discourse Tree Bank(Miltsakaki, 2004)という同一文内もしくは隣接する文の間について接続表現を手がかりに談話関係がアノテーションされたコーパスを参考に関係認定の範囲を決定し, また根拠帰結関係の定義については文献(田窪ら,1992)などに示された接続表現とその用例を参照し, 同定対象とすべき根拠帰結関係を暫定的に定義する.

この後, 定義した関係の妥当性, 関係の自動同定を吟味するための材料とするための根拠帰結関係タグ付きコーパスを構築する. まずWeb記事を対象にした同定のために, 河原ら(2006)の5億文コーパス中のWeb記事のうち「ドラフト精度」「ステロイド」など, 特定のキーワードを含む文とその前後2文を抽出し, その中に出現する根拠帰結関係を人手でタグ付与する. また, Web記事と対比させるために, 新聞記事に対しても根拠帰結関係を人手で付与し, 自動同定の精度の違いを調べる.

根拠帰結関係の自動同定を実現する際には, この関係同定の問題を機械学習に基づく2値分類とみなして扱う. つまり, 与えられた任意の根拠帰結関係の候補に対し, 最終的に同定したい関係か否かを訓練事例をもとに学習し, その結果得られた分類モデルを適用することで未知の問題を解析する. この際, 問題を照応関係のアナロジーとみなし, 照応関係の同定に関して我々が以前提案したトーナメントモデル(飯田ら,2004)と探索先行分類型モデル(飯田ら,2005)を利用する. トーナメントモデルでは与えられた照応詞に対し, 先行詞候補集合に中の2つの候補を比較しより先行詞らしい候補を残す勝ち抜き戦を行い, 最尤先行詞候補を決定する. 次に, 探索先行分類型モデルではトーナメントモデルが決定した最尤先行詞候補と照応詞が照応関係にあるか否かを判別する. このような2段階の解析を行うことで, 照応解析に関しては劇的に解析精度が向上しており, この手順を根拠帰結関係の同定にも導入することで解析精度の向上が期待できる. ただし, 根拠帰結関係の場合には照応関係のように照応詞から先行詞を解くという解析の方向は自明ではないため, 根拠

側から帰結側を決定する，もしくは帰結側から根拠側を決定するという 2 つの方向を比較する必要がある．

また，4 で後述するように同定対象となる根拠帰結の関係は同一文内に頻出し，「ので」「ため」「から」といった接続表現を手がかりとすることで，ほぼ確実に同定可能である．しかし，逆に手がかり表現が出現していない場合には関係同定が非常に困難であることがわかった．接続表現以外の手がかりを導入することでこの問題を解決する必要があるが，本研究では特に「体調を崩す->病気になる」といったある種の因果関係に相当する表現を明示的に捉えることで，接続表現が出現していない場合についてもこのような手がかりが無い状況と比べてより高い精度で関係同定が可能になると考えられる．この事態間の関係知識を導入するために，近年含意関係認識のための資源獲得で用いられている手法を導入することを考える．具体的には，ある特定の共起パタンで出現する事態対を大規模コーパスから獲得し，その収集された頻度に基づいて相互情報量のような共起尺度で順序付けすることで最終的に抽出したい関係にある述語対を獲得する．この考えを根拠帰結関係に当てはめて考えた場合，特定のパタンは「ため」「から」「ので」を伴って出現する事態対となると考えられる．そこで，予備実験としてこれら 3 種の接続表現を伴い，かつ係り受け関係にある述語対を収集し，それらを相互情報量で順序付けしたところ，根拠帰結の関係にある述語対を上位に位置付けることができなかった．なぜこのようなことが起こるかをその後調査したところ，問題は事態の情報をその(意味的な)主辞である述語で代替していることが問題であることがわかった．つまり，「体調を崩す」という事態に関しては「崩す」という述語でその事態を代表させてしまっているため，述語の多義性を考慮できていない．「体調を崩す」という動詞句の粒度で頻度を計上したいが，現存する大規模なコーパスから頻度を求めても疎になるという問題がある．また，語を近傍語で表現したベクトル表現に対し，和や積などの演算を行うことで，句の意味を生成的に計算しようという試みもあるが，この試みも完全に成功しているわけではないため，そのまま利用することができない．このような背景から本研究では述語とその項の共起行列をもとに pLSI(Hoffman, 1999)を用いて次元圧縮を行い，その結果を利用して根拠側と帰結側に出現する動詞句を表現する．述語 v_i が隠れクラス z を生成する確率をそのまま素性として利用することで，述語単体の頻度情報を n 次元の隠れクラスへの帰属確率として表現する．この際，n は述語の総数より数が少ないため，より少ない次元で述語を特徴付けることが可能になる．述語に係るガ格，ヲ格，ニ格についても同様に表現することで，根拠側の事態を 4n 次元，帰結側の事態を 4n 次元，合計 8n 次元で表現することが可能になる．項となる名詞は述語に比べて出現の異なりがおおきいためこの次元圧縮の効果が期待できる．評価実験では，事態の情報の導入に加え，この次元圧縮の効果についても評価を行う．

4. 研究成果

3 で述べた内容について調査した結果をまとめる．まず，根拠帰結関係の定義に関してはさまざまな候補を比較吟味した結果，動機・目的・理由・根拠・原因などの関係を区別せずに根拠側の関係とし，これに対応する箇所を帰結部分として定義した．これは既存研究では談話の論理的関係と修辞的関係が混在した形で定義されており，これをそのまま採用しても応用処理に役立つ見込みが低いと判断したため，まずは単純に根拠と帰結という最も荒い粒度で問題を考えることとした．この定義でどの程度の規模の関係が付与できるかを調査するために，まず Web 記事の抜粋 2,954 事例に対し，根拠帰結関係のタグ付与を人手で行ったところ，4,350 の関係が付与された．この関係を 3 で述べた同定モデルを用いることでどの程度自動同定できるかを調査したところ，提案する 2 段階の処理で関係を同定する手法が単純に 1 段階で同定するベースラインと比較して劇的に精度が向上することがわかり，特に帰結側を与えて根拠側を同定する手順の場合により精度が向上した．これは，根拠となる談話セグメント集合の中から適切な根拠箇所を同定する場合には「ので」などの接続表現が出現するため，関係同定が容易であり，その結果最終的に精度が良くなったと考えられる．また，評価実験の結果，Web 記事を対象にしたため既存の解析ツールでは形態素解析・係り受け解析が適切に行うことができず，その結果関係同定を誤った事例が誤り全体の 25%を占めており，これから Web 記事にも対応できる形態素・係り受け解析器の必要性が再確認できた．これらの誤りを除いた誤り事例のうち，全体の約 35%が「ので」「ため」のような接続表現が明示的に出現していないために同定を誤っていることがわかった．

以後，この接続表現を伴わない場合の同定精度向上の評価実験を行うために，形態素・係り受け関係のような基盤となる解析が誤らない状況で評価を行う．具体的には新聞記事中に出現する談話セグメントのうち主辞となる述語が係り受け関係にある場合のみを対象として評価用データを作成する．根拠帰結関係の定義は Web 記事の場合と同じ定義を採用し，京都テキストコーパスに出現する全ての述語対(合計 12,911 対)に根拠帰結関係のタグを付与した．この結果，3,683 の根拠帰結関係がタグ付与された．このうち，「ので」「から」「ため」などの手がかりとなる接続表現を伴った場合が 906 事例，それ以外が 2,777 事例であり，接続表現が明示的に出現している場合に比べ接続表現を伴わない場合が

非常に多い. この数値からも接続表現が明示的に出現していない場合を適切に検出することが本研究で同定対象としている根拠帰結関係の本質的な問題であることがわかる. 3 でも述べたように, この接続表現を伴わない場合の根拠帰結関係(以後, 非明示的な関係)の同定精度を向上させるために, 根拠や帰結に該当するセグメントの主辞となる述語とその項を同定の手がかりとすることで同定精度に貢献するかを調査した. この結果, 単純にそれぞれの情報を素性として加えただけでも精度が向上し, さらに出現している見出し語そのものではなく, 次元圧縮した情報を加えることで最も良い結果を得た. ただし, 本研究で採用した方法でも動詞句の意味を適切な粒度で扱うための最適な方法ではないため, 最終的な同定精度は再現率約 6 割, 精度 6 割であった(詳細は文献(林ら,2010)を参照されたい). さらなる同定精度向上のためには, (動詞)句などのより細かい粒度の言語表現をスパースネスを解消しつつどのように表現するかを考える必要があり, この点については今後の課題としたい.

5. 主な発表論文等
(研究代表者、研究分担者及び連携研究者には下線)

〔学会発表〕(計 4 件)

① 林賢吾、飯田龍、徳永健伸「述語と項の分布類似度を利用した非明示的な根拠帰結関係の同定」 言語処理学会第 16 回年次大会、2010 年 3 月 8 -11 日、東京大学

② Ryu Iida, Kentaro Inui and Yuji Matsumoto. Capturing Salience with a Trainable Cache Model for Zero-anaphora Resolution. *The Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009)*, pp. 647-655. 2-7 August 2009, Singapore

③ 飯田龍、乾健太郎、松本裕治「根拠情報抽出の課題設計と予備実験」*言語処理学会第 15 回年次大会*、pp.817-820、2009 年 3 月 2- 5 日,鳥取大学

④ 飯田龍、乾健太郎、松本裕治「結束性と首尾一貫性から見たゼロ照応解析」*情報処理学会自然言語処理研究会* 予稿集, NL-178-7. pp.45-52. 2008 年 1 月 21・22 日、国立情報学研究所

〔その他〕
ホームページ等
http://www.cl.cs.titech.ac.jp/~ryu-i/CED/

6. 研究組織
(1)研究代表者
飯田 龍(IIDA RYU)
東京工業大学・大学院情報理工学研究科・助教
研究者番号:40464276

(2)研究分担者
なし

(3)連携研究者
なし