2005－NL－170（12）
2005／11／21

確率モデルを利用した照応解析の研究

高橋　慎之介　榑松　理樹　藤田　ハミド
岩手県立大学大学院　ソフトウェア情報学研究科
g231c023@edu.soft.iwate-pu.ac.jp { kure, issam}@soft.iwate-pu.ac.jp

あらまし　本稿では、単語の共起情報と人手で与えられた照応情報により学習した確率モデルを利用することで日本語における前方照応の指示詞に対する先行詞を推定する方法を提案する．確率モデルを利用することにより，指示詞の推定ルール間の整合性などを取る必要がなくなる。本研究では，指示詞より前に出現する語のうち、指示詞が係る語との組合せが不自然で無い語が，指示詞の先行詞であるという考えに基づき，語の係り受け関係の共起情報に着目し、指示詞の候補の妥当性を示す確率モデルを利用することで、先行詞の推定を試みる．

Study of Estimate of Referents using a Probabilistic Model

Shinnosuke Takahashi, Masaki Kurematsu, Hamido Fujita
Iwate Prefectual University.
Graduate school of Software and Information Science.
g231c023@edu.soft.iwate-pu.ac.jp { kure, issam}@soft.iwate-pu.ac.jp

abstract This paper proposes a method of identifying antecedents using co-occurrence of words and a probabilistic model. We make the probabilistic model form documents what have anaphoric relation. The model shows whether a word is suitable as antecedents or not. First, This method extracts demonstratives from a given document using dependency grammar. It extracts independent words as antecedent candidates, too. Next, it evaluates the suitability of independent words as antecedent based on co-occurrence of words and the probabilistic model. Finally, it regards some independent words whose suitability is high as antecedent candidates. This method extracts not one word but some words as antecedents.

1　はじめに

人は通常，会話や文章において，既に登場したものと同一の事物を繰り返し表現する際に，対象となる事物の名称を述べず，別の語句に置き換えて表現することが多い．この置き換えられた語句は，既に登場した事物と同一の内容を指す．このような置き換えられた語句と置き換えた語句は照応関係にあると言う．このとき，置き換えられた語句を先行詞，置き換えた語句を照応詞という．文脈照応は，照応詞の内容となる先行詞が照応詞より前に現れている場合には前方照応と呼び，先行詞が照応詞よりも後に現れている場合には後方照応と呼ぶ．照応詞の中に，実際の先行詞以外の言葉で置き換える照応詞が存在する．それを指示詞と呼ぶ．

コンピュータによる自然言語処理の分野において，照応関係の解析は高品位の対話システム，機械翻訳の実現のために必要とされており，現在までに照応解析の為に様々な手法が提案されている．しかし，決定的とされる手法はまだ提案さ

表１　指示代名詞の種類（長尾，岩波「自然言語処理」より）

指示代名詞	コ系	ソ系	ア系	ド系
名詞形態	これ（ら） ここ（ら） こちら こっち	それ（ら） そこ（ら） そちら そっち	あれ（ら） あそこ あちら あっち	どれ どこ どちら どっち
連体詞形態	この こんな こういう こうした こういった このような	その そんな そういう そうした そういった そのような	あの あんな ああいう ああした ああいった あのような	どの どんな どういう どうした どういった どのような
副詞形態	こう このように こんなに こんなふうに	そう そのように そんなに そんなふうに	ああ あのように あんなに あんなふうに	どう どのように どんなに どんなふうに

れていない．より高度な自然言語処理を行ううえでも、精度の高い照合解析の手法を構築する必要がある．

以上のような背景から本研究では，前方照応における指示詞の照応解析手法について提案し，その有効性を検証する．

以下２章において対象とする指示詞について説明する。３章において本手法について述べ，４章において評価実験について説明する．

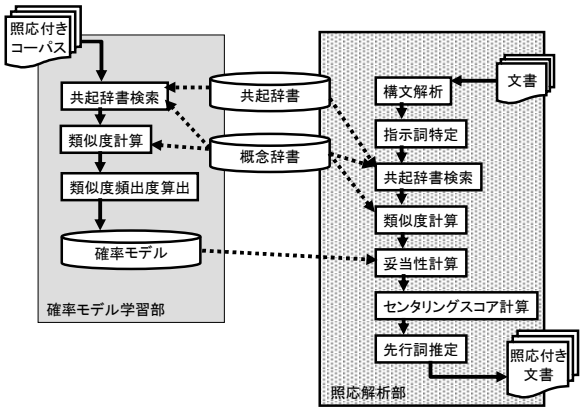

図１　システムの概要

２　指示詞

文献[1]の中では，指示詞は表１のようにまとめられている．本研究では，ド系音から始まる指示詞については照応解析しない．これは，ド系音指示詞は不定称であり，指示対象が一意に同定できないためである．また照応解析の対象としては名詞形態の指示詞を扱う．これは連体詞形態，副詞形態の指示詞は，特定の自立語を指示対象として持つ他に，既に述べられた概念やそれに付随する属性全てを指示対象として持つ場合が存在し，指示対象の特定が困難なためである．

３　提案手法

先行詞を推定するための１つの基準として，先行詞と指示詞を受ける語との組合せの妥当性が考えられる．先行詞と，指示詞を受ける語との関係が適切でなければ，文意が通らなくなり，照応関係が成り立たなくなると考えられる．我々はこの点に着目する．提案手法では，照応解析における，候補となる自立語と指示詞を受ける語との組合せの妥当性を，確率モデルを用いて求める．本手法における確率モデルは，人が解析した照応解析の結果に基づき求めたものである．これは，語の共起関係に対する意味的な類似性の出現傾向を示している．更にセンタリング理論に基づくスコアを指示詞と候補となった自立語に接続する助詞から算出し，それを利用することで，候補の絞込みを行う．

図１に本提案手法に基づくシステムの概要を示す．

本システムは，既に照応付けが行われたコーパスから学習を行う確率モデル学習部と，学習が行われた確率モデルを利用することで照応解析

```
〈レコード番号〉                          J〈レコード番号〉                          JCC7173641
〈見出し情報〉
        〈句見出し〉                      昼食    を      食べ
〈共起句構成要素情報〉
〈要素    〈形態    〈かな表記〉      〈品詞〉 〈慣用句〈概念情報〉
 番号〉    素〉                        フラグ〉
{ 1      昼食     チュウショク     名詞     0       3bec74  lunch  昼食［チュウショク］  ″a meal eaten
at noon″  昼の食事 }
{ 2      を       ヲ               助詞     0         ″″ }
{ 3      食べ     タベ             動詞     0       3bc6f0   ″″  食べる［タベ・ル］  ″to eat something″
食物をとる }
〈構文情報〉
        〈部分構文木〉
                〈受け側要素〉            3/食べ
                〈関係要素〉              2/を/を
                〈係り側要素〉            1/昼食
〈意味情報〉
        〈部分意味フレーム〉
                〈受け側概念要素〉        3/3bc6f0/食べ
                〈概念関係子〉            object
                〈係り側概念要素〉        1/3bec74/昼食
〈共起状況情報〉
        〈頻度〉                          1;1;488;6
        〈例文〉                          {00050003b57d-8-3/〈昼食〉を…(食べ)に帰る}
〈管理情報〉
        〈管理履歴レコード〉                  DATE=″95/3/31″
```

図２　EDR 共起辞書のレコード例（EDR 電子化辞書，日本語共起辞書より）

を行う照応解析部から構成される．

確率モデル学習部では，共起情報と概念辞書を利用することで，辞書に示された語と実際の先行詞との類似度を算出し，類似度の出現確率を学習する．学習が行われたモデルは，照応解析で利用される．

照応解析部では，入力文章を構文解析し，指示詞を特定する．特定された指示詞の係り受け関係と共起情報，概念辞書，さらに確率モデル，センタリングスコアを利用することで指示詞の先行詞を推定する．

以下，それぞれの部分について説明する．

３．１　確率モデル学習部

本節では確率モデルとモデルの学習方法について述べる．

確率モデルは，先行詞の候補となる自立語と指示詞を受ける語との組合せがどの程度妥当であるかを示すものである．確率モデルは，共起関係として示されている係る語と受ける語，それらの関係において，共起関係において示された係る語とその部分に実際に出現する語との意味的類似度の出現頻度をしめしている．すなわち，今出現した語の係る語としての妥当性を過去の出現から判断するものである．

次に確率モデルの学習方法について述べる．確率モデルは，照応付けされた文書（コーパス）を与え，その中の照応関係に対し先行詞と共起関係のかかる語との意味的類似度の出現頻度を元に学習を行う．また共起情報や類似度を求めるために必要な概念辞書としては、現在は EDR 電子化辞書[2]に含まれる共起辞書と概念辞書を利用する．以下，学習方法について説明する．

①　共起辞書検索

指示詞を受ける語と組み合わせるのが適当な語がどの概念に属するかを明らかにするために，EDR 共起辞書を利用する．EDR 共起辞書は日本語コーパスに格納された実例文の解析結果から、係り受けを構成している部分、すなわち共起句を抽出したものである．一例を図２に示す．我々は，この EDR 共起辞書の持つ共起情報を自立語と自立語の組合せが妥当性であるという指標であると捉える．

検索は次の手順で行う．

最初に，学習データの各文章を係り受け文法に基

づく構文解析を行い，係り受け関係を抽出する．次に，共起辞書のレコードのうち，指示詞を含む文節を受ける文節に含まれる自立語 Wi が<受け側要素>，先行詞 Aj の直後に出てくる助詞 Pj が<関係要素>に出現するレコード Rk を取り出す．

自立語 Wi からレコード Rk が発見できない場合，Wi の類義語 Wi*と Pj を使用して辞書の検索を行う．それでも発見できない場合，特殊な照応関係と捉え，学習対象から取り除く．

② 類似度計算

先行詞 Aj と共起辞書に記載されている語 Wl が意味的にどれほど近いのかを類似度として数値化する．類似度は EDR 概念辞書における概念と概念の距離として計算する．これが大きいほど意味が近いと考えられる．概念辞書は多重継承を許した木構造となっており，語の類似度は，文献[4]を参考に式(1)を用いて計算する．

$$Sim(Aj,Wl)=\frac{C(Aj),C(Wl)\text{共通段数}\times 2}{(C(Aj)\text{の段数}+C(Wl)\text{の段数})}\cdots(1)$$

式(1)において，C（Aj）は Aj の概念であり，C（Wl）は Wl の概念である．段数とは最も上位の概念を 1 段目とし，そこからひとつ下位の概念になるごとに 1 を加算したものである．共通段数とは，対象概念の共通の上位概念の段数である．また 1 つの語に対し，複数の概念があることが考えられることから，すべての組合せに対し，式(1)を適用し，最大値を類似度 d として取り出す．

取り出した類似度に対し，出現回数を 1 加算する．

①および②を学習データとして与えた文書に現れるすべての先行詞と指示詞に対して実施する．

③ 確率モデルの生成

①および②によって求まった類似度に対する出現回数を元に確率モデルを生成する。確率モデルは、次の式(2)によって定義する。

$$P(d)=\frac{Freq(d)}{\int_0^1 Freq(x)\,dx}\cdots(2)$$

式(2)において、Freq(d)は，共起情報に記されている語の概念と意味的な類似度が d である語が出現した回数を示しているものであり，①から②によって求める．分母になっている関数は、類似度ごとの出現回数の総和を示す．類似度は離散値をとるが，各値の差が小さいと考えられることから積分で近似する．

更にデータスパースネス問題を想定し，モデルのスムージングを行う．

３．２　照応解析部

次に照応解析部の処理手順を述べる．

① 構文解析

指示詞を含む文章を入力として与え，その文章に対して，構文解析器 Cabocha[3]を適用し，語の品詞と，文節，文節の係り受け関係を得る．

② 指示詞特定

品詞と語の基本形から対象とする指示詞 Ax を特定する．さらに係り受け関係から指示詞を含む文節 Sa と，Sa を受ける文節 Sg を特定する．

③ 共起辞書検索

共起辞書のレコードのうち，指示詞 Ax を含む文節 Sa を受ける文節 Sg に含まれる自立語 Wy が<受け側要素>，指示詞 Ax の直後に出てくる助詞 Pz が<関係要素>に出現するレコード Ri を取り出す．

自立語 Wy からレコード Ri が発見できない場合，Wy の類義語 Wy*と Pz を使用して辞書の検索を行う．それでも発見できない場合，係り受け関係解析不能と判断し，指示詞 Ax を解析対象からはずす．

④ 類似度計算

検索されたレコード Ri から，<意味情報>の<係り側概念要素>に記述されている概念 Ct を取り出す．

Sa を含む文において，その指示詞より前の部分とその直前の文に出現する自立語 Wi を先行詞の候補として取り出す.

次に Wi の持つ概念 C(Wi)をそれぞれ概念辞書から取り出す

次に Ct と C（Wi）から，概念辞書を用いて，類似度 d（Wi）=MAX(Sim(Ct,C(Wi)))を算出する．ここで，MAX(x)は，x の取りうる値の最大値を意味する．これは，複数の概念を持つ語については，類似度の最大値を取ることを意味する.

⑤　妥当性の計算

確率モデルは，共起情報で示されている係り側の語と，その部分に現れる語との意味的類似度の出現確率を表している．⑤までで求めた先行詞の候補となる自立語の類似度に対し，この確率モデルを適応することで，妥当性 Sui(Wi)を求める．Sui(Wi)は、以下の式(3)で示される。

$$Sui(Wi) = P(d(Wi) \cdots (3)$$

このとき、一定の基準，閾値を事前に与えておき，その値を以下である自立語は，先行詞の候補から取り除く．この処理によって，全ての自立語が削除される場合がある．この場合は，「候補無し」と判断する．これにより，先行詞がある前方照応以外の照応関係を選別し，誤った照応関係の推定を回避することを試みる.

⑥　センタリングスコア計算

センタリングとは，照応解析におけるひとつの知見である．文章に既に現れた主題や焦点が話の中心を担っている，ということから指示詞はそれらを先行詞として指示しやすくなる．主題は前の文の焦点を指すことが多く，また焦点は次の文などで主題として参照されることが多いというものである．しかし，これら参照されやすい特性を持つ主題や焦点を特定するのは難しいとされる.

そこで，助詞，句読点等の表層表現から，主題や焦点を近似的に類推する．表層表現とその表現が主題・焦点である可能性の重みを対応付けした表を利用することで，文における主題・焦点の度合いを名詞句に対して付ける．実際に使われる表を表２に示す．本表は文献[1]にまとめられた表を参考にして制作した.

表 2 主題・焦点に与える重み

	表層表現	重み	例
主題	<指示詞>が	20	それがした
焦点	ガ格以外の指示詞，代名詞	0.27	それにした
	<名詞>が/も/だ/なら/こそ	0.25	太郎がした
	<名詞>を/に/，/.	0.24	太郎にした
	<名詞>へ/で/から/より	0.22	学校へ行く

表２の主題の項目を使用し，Ax の持つ助詞 Pz を見て，Ax が文の主題であるかを推定する．Ax が主題であった場合，Wi の直後に登場する助詞 Pi を見て，候補の焦点の度合いを表２の焦点の項目から推定し，重みをセンタリングによるスコア E(Wi)として与える．Ax が主題ではない，または表２には存在しない表現を Wi が持っていた場合，E(Wi)は 0 とする.

センタリングを導入することによって，同じ類似度を持つ語が出現した場合において，候補に差違を付け，指示対象の候補を絞り込むことが可能になると考える.

⑦　推定スコア計算

⑤までで求めた妥当性 Sui(Wi)と，⑥で求めたセンタリングスコア E(Wi)を足し合わせることで，推定スコア V(Wi)を求める.

⑧　先行詞推定

先行詞の候補である自立語 Wi のうち，⑦で求めた推定スコア V(Wi)が一定の閾値をこえたものを取り出す．それらを推定スコアに対し昇順で並べたものを，最終的に先行詞として出力する．本研究では，先行詞の候補をひとつに決定するのではなく，先行詞となりうる可能性のある語を全て出力する．これは，照応は曖昧な現象であり，

一意に決定できるものではないと考えるためである．この処理は本研究の特徴的な部分である．

４　評価実験

本提案手法の有効性を検証するために２種類の評価実験を行う．１つめの評価実験は，照応解析の精度を評価するものであり，２つめの評価実験は，照応解析の有用性を評価するものである．以下にそれらの評価実験の内容について述べる．

４．１　人手による照応との比較

照応関係が存在するコーパスに対して，人手によって同定した先行詞と，本研究で提案する手法によって同定した先行詞を比較する事によって，本照応解析手法の精度の検証を行う．人手による照応関係の付与はきわめて主観的なものであるが，複数人での結果の検討を行うことにより，客観性を高めることを試みる．この実験によって，提案手法の結果が人が行う照応付けとどの程度一致するかを評価する．

評価実験としては，主に新聞記事から作成した照応関係を付与した文書を利用する．これら 32 件について，22 件を学習用，10 件を評価用とし，その組合せを変えデータを５セット作成する．各セットに提案手法による処理を行い，その結果と人手による結果を比較することで，精度を評価する．図 4 に評価に使用する照応関係が付与されたデータの例（一部）を示す．

４．２　検索システムでの利用

照応解析の有用性を評価するために，コーパスを情報資源対象にした質問応答システムへ適用実験を行う．本評価実験は，質問応答システムの情報源として，コーパスをそのまま利用する場合と，コーパスに対し，提案手法により指示詞を先行詞に置き直したものを利用する場合とで，質問に対する正答率を比較する．本評価実験において，照応解析をしたものを利用した場合のほうが高い正答率を得ることができれば，その有用性が評価できると考える．質問応答システムへの問題としては，テストコレクションの１つを利用する予定である．

・・・

アンケートの対象は埼玉、茨城両県警の交通警察官３００人で、郵送による匿名での回答を求めた。

それ<アンケート>によると、交通警察官１人当たりの事故取扱件数は月平均１２・６件。

・・・

図４：データの一例

４．３　実験結果

現在評価実験を行っている段階であり，本稿にはその結果を示すことができない．実験結果については，当日報告する予定である．

５　まとめ

本稿では，単語を対象とした前方照応の照応解析の新しい手法を提案した．本研究では，先行詞として自立語を特定するだけにとどまっているが，実際の照応関係では自立語だけでなく，文全体を照応する照応現象が存在している．今後は，そのような照応現象についても解析が可能な手法の提案が必要になると考える．

参考文献

[1]長尾真　編：“岩波講座ソフトウェア科学 15 自然言語処理”，岩波書店(1996)

[2]“EDR 電子化辞書”
http://www2.nict.go.jp/kk/e416/EDR/J_index.html

[3]“日本語係り受け解析器　Cabocha”
http://chasen.org/~taku/software/cabocha/

[4]川島貴広，石川勉：“言葉の意味に関する類似性判別能力における概念ベースとシソーラスとの性能比較”，情報処理学会第６５回全国大会，2M-1,pp.2-135 – 2-136(2004)