語の類義度に基づく確率モデルを利用した照応解析手法の提案

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

1　はじめに

人は，会話や文章において，既に登場した事物と同一の事物を繰り返し表現する際に，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．

コンピュータによる自然言語処理の分野において，照応解析は高品位の対話システム等の実現のために必要とされており，現在まで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、精度の高い照応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章において本手法について述べ，３章において評価実験について説明する．

2　提案手法

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，既に人手で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を行う照応解析部から構成される．以下，それぞれの部分について説明する．

2.1　確率モデル学習部

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，及び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コーパス中で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．これは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．

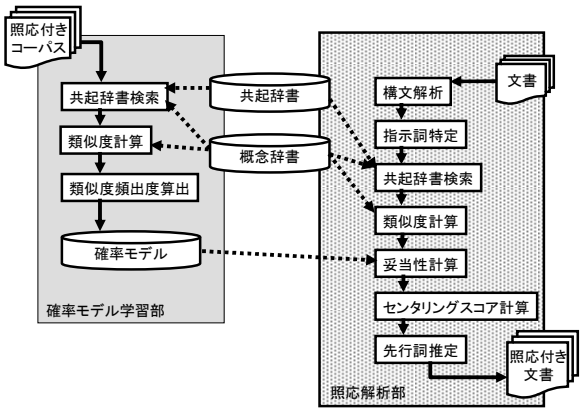

図１：システムの概要

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係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

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，共起辞書の中から，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，概念辞書から得た 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，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 が意味的な近さを類似度として数値化する．

語の類似度は，文献[4]を参考に式(1)を用いて計算する．式(1)において，　C（X）は X の概念である．

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

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．

$$P\{X=d\}=\frac{Freq\ (d)}{\sum_{k=0}^{1}Freq\ (k)}\cdots(2)$$

式(2)において、Freq(d)は，辞書に示されている語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．分母は、類似度ごとの出現回数の総和を示す．

２．２　照応解析部

照応解析は学習部で生成された確率モデルをもとに照応関係を解析する部分である．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，構文解析し，語の品詞と文節，文節間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

共起辞書の中から，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，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，係り側要素の意味を示す概念 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．

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．④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。

$$Sui(Wi)=P\{X=d(Wi)\}\cdots(3)$$

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．

これを利用して助詞，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．名詞句に続く表層表現に着目し，その名詞句が主題・焦点である可能性を重みとして与える．文献[1]を参考にして重みを次のように定義した．「：」以下が重みである．

- 主題…＜指示詞＞が：--
- 焦点…
  ＜名詞＞が/も/だ/なら/こそ：0.25

＜名詞＞を/に/，/．：0.24

<名詞>へ/で/から/より：0.22

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，または定義されていない表現を Wi が持っていた場合，E(Wi)は 0 とする．

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

先行詞の候補を推定スコア V(Wi)に基づいて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．

3　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．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．この実験によって，提案手法の結果が人の行う照応付けとどの程度一致するかを評価する．

1 つめの評価実験は，新聞記事を対象とし，2 つめの評価実験は，文学作品を対象とした．以下にそれらの評価実験の内容について述べる．係り受け関係を得るため，構文解析器として Cabocha[3]を利用した．

3.1　新聞記事を対象とした推定

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 30 件を利用する．これら 30 件について，20 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 2 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．

3.1.1　実験結果

結果をまとめたものをを表 1 に示す．

表 1 ：推定結果と人との一致件数

正解の出現順位	データセット1の出現数	データセット2の出現数	累計
1位	1	1	2
2位	2	0	2
3位	1	1	2
4位	0	0	0
5位	0	0	0
6位以下	1	2	3
出現無し	0	1	1

人手によって付けられた先行詞が 1 位に登場することが望ましい．しかし表 1 では，人手による先行詞と一致した推定結果が 2 件と少なかった．

そこで，推定の際，候補の品詞と共起辞書の品詞を一致させることでスコアを計算する手法を導入した．手法の改良後の結果を表 2 に示す．

表 2 ：品詞を一致させた結果

正解の出現順位	データセット1の出現数	データセット2の出現数	累計
1位	1	2	3
2位	2	1	3
3位	1	0	1
4位	1	0	1
5位	0	0	0
6位以下	0	1	1
出現無し	0	1	1

この手法によって，1 位の出現が 3 件，6 位以下の出現が 4 件から 2 件と，大幅に改善された．

また，出力結果の動詞に着目し，動詞を候補から除外する処理を行った．結果を表 3 に示す．

表 3 ：動詞を除外した結果

正解の出現順位	データセット1の出現数	データセット2の出現数	累計
1位	3	3	6
2位	1	0	1
3位	1	0	1
4位	0	0	0
5位	0	0	0
6位以下	0	1	1
出現無し	0	1	1

動詞を抜くことで 1 位に 6 件，5 位まで含めると 8

件と高い精度を獲得できた．

３．２　文学作品を対象とした推定

次に文学作品を対象とした実験を行った．文学作品は，新聞記事と違い完全な文法を保証されていない．その文学作品に本手法を適用した結果を考察する．対象として青空文庫[5]から宮沢賢治の「注文の多い料理店」「カイロ団長」を選択した．

３．２．１　実験結果

実験結果をまとめたものを表４，表５に示す．

表４：「注文の多い料理店」結果

出現順位	一致した候補	妥当な候補	一致した候補（区別有り）	妥当な候補（区別有り）
1位	0	1	0	1
2位	0	0	1	1
3位	1	1	0	0
4位	0	0	0	0
5位	0	0	0	0
6位以下	0	0	0	0

表５：「カイロ団長」結果

出現順位	一致した候補	妥当な候補	一致した候補（区別有り）	妥当な候補（区別有り）
1位	0	0	0	0
2位	0	1	0	0
3位	0	0	0	0
4位	0	0	1	1
5位	0	0	0	1
6位以下	1	1	0	0

実験２では，人手で同定した先行詞の他に候補中に妥当と思われる語があれば，それも結果に含めて検証することにした．

「注文の多い料理店」に比較して「カイロ団長」の結果が良いとは言えなかった．そこで結果から動詞を除外してまとめたものを表６に示す．

表６：動詞を除外した結果（文学）

出現順位	一致した候補	妥当な候補	一致した候補（区別有り）	妥当な候補（区別有り）
1位	0	1	0	0
2位	0	0	0	0
3位	0	0	1	1
4位	0	0	0	1
5位	1	1	0	0
6位以下	0	0	0	0

新聞記事に対する実験に比べて劣るが，多少の精度向上が見られた．

３．３　考察

確率モデルとセンタリングを利用することで，高い精度の照応解析を行うことができた．しかし，本手法は EDR の概念構造への依存が高く，辞書の内容が変わると解析精度も変わってしまうという点が問題である．さらに概念構造をそのまま利用しているため，確率モデルにスコアがばらついてしまう．それが精度に影響を及ぼしていると考える．

更に動詞を除外することで精度の変化を検証したが，単に動詞を除外するだけでなく，先行詞として妥当なのはどの品詞であるかという点を考慮し，動詞に関しても慎重な扱いが必要であると考える．推定の際，候補と共起情報の品詞を一致させる処理は精度を上昇させる一端を担った．以上のことから，品詞情報の活用は照応解析において，大きなファクタとなると考える．

４　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)

[5]“青空文庫”：http://www.aozora.gr.jp/