■2 群（画像・音・言語）- 10 編（自然言語処理）

5 章 意味解析

【本章の構成】

本章では，格解析（5–1 節），省略照応解析（5–2 節），語義曖昧性解消（5–3 節）について述べる．

■2 群 - 10 編 - 5 章

5-1 格解析

■2 群 - 10 編 - 5 章

5-2 省略照応解析

（執筆者：飯田　龍）[2009 年 8 月 受領]

談話中のある表現が他の表現を指す機能を照応という．照応の関係にある 2 つの表現のうち，指す側の表現を照応詞といい，指される側の表現を先行詞という．照応関係のうち，先行詞が照応詞より前に出現する場合を前方照応といい，逆に先行詞が照応詞より後に出現する場合を後方照応という．また，先行詞が話し手や外界の対象であるため明示的に談話中に先行詞を持たない場合を外界照応という．また，照応関係の類似概念として，談話中の複数の表現が現実世界（もしくは仮想世界）の同一の実体を指す関係を共参照（同一指示）という．

(1) 太郎は今日オレンジを買った。彼は昨日もそれを買ったのに。

例えば，(1) の “彼” は “太郎” を指しており，同一実体であるため，照応かつ共参照の関係にある．一方，“それ” は “オレンジ” を指しているが，同じ実体ではない．前者の照応かつ共参照の関係を参照の同一性に基づく照応（Identity-of-Reference Anaphora）といい，後者を語義の同一性に基づく照応（Identity-of-Sense Anaphora）という．また，日本語では，照応詞が文脈から推定可能な場合には頻繁に省略される．この省略された照応詞をゼロ代名詞という．

(2) 太郎は毎日スペイン産のオレンジを食べている。そのオレンジが好みだからだ。

(3) 太郎は毎日スペイン産のオレンジを食べている。その味が好みだからだ。

また，(2) では “スペイン産のオレンジ” と “そのオレンジ” が照応関係にあるのに対し，(3) では “その味” の直接の先行詞はなく，代わりに “スペイン産のオレンジ” の属性を指している．(2) の直接的な照応関係を直接照応と呼ぶのに対し，(3) のような照応関係を間接照応という．この間接照応は橋渡し照応（Bridging Reference）や関連照応とも呼ばれる．

自然言語処理における照応解析の問題設定は談話中のある表現が照応詞となるか否かを分類する問題と，与えられた照応詞について先行詞を候補集合から探索する問題に分けて考えられることが多い．前者の問題を照応性判定，後者を先行詞同定と呼ぶ．例えば，(1) では照応性判定の問題として “彼” や “それ” を照応詞に分類し，先行詞同定の際には “太郎” や “オレンジ” をそれぞれの照応詞の先行詞として同定する必要がある．

照応解析の問題のうち，先行詞同定については人手により作成した規則に基づく解析手法とタグ付きコーパスを利用した手法に分類できる．前者は特に中心化理論[1)]と呼ばれる談話の理論に基づいて研究が進められてきた．一方，後者については照応解析の問題を表現の対が照応関係となるか否かの 2 値分類問題として扱う．本節ではこれら 2 つを説明し，最後に照応（もしくは共参照）解析の課題設計についても紹介する．

5-2-1 中心化理論

中心化理論[1)]は談話の首尾一貫性についての理論で，談話の各発話が中心（Center）と呼ばれる顕現性の最も高い談話要素によって特徴付けられるという考えに基づいている．この理論では，連続する発話間の中心の移り変わりを扱い，前向き中心（*Cf*），後ろ向き中心（*Cb*），

表 5･1　局所焦点の遷移

	$\boldsymbol{Cb}(U_i) = \boldsymbol{Cb}(U_{i-1})$ or $\boldsymbol{Cb}(U_{i-1}) = [?]$	$\boldsymbol{Cb}(U_i) \neq \boldsymbol{Cb}(U_{i-1})$
$\boldsymbol{Cb}(U_i) = \boldsymbol{Cp}(U_i)$	CONTINUE	SMOOTH-SHIFT
$\boldsymbol{Cb}(U_i) \neq \boldsymbol{Cp}(U_i)$	RETAIN	ROUGH-SHIFT

優先中心（*Cp*）という 3 つの特徴を用いて照応の現象を説明する．*Cf* には現行発話内の談話要素すべてが保持され，*Cf* 内の談話要素は “主題 > 主語 > 間接目的語 > 直接目的語 > その他” のような優先度でランキングされる．このうち最上位の談話要素が *Cp* となる．また，*Cb* は 1 つ前の発話に出現した談話要素のうち現行の発話で再度言及されたなかで最上位の談話要素である．この理論では *Cp* と *Cb* を用い，隣接する発話のつながりの良さを表 5･1 に示めす遷移のいずれかに決定する．遷移は CONTINUE，RETAIN，SMOOTH-SHIFT，ROUGH-SHIFT の順で望ましいとされている．

(4) S_1: 太郎は学校から帰ってきた。S_2: 彼は次郎といっしょに食事をした。S_3: その後、彼は夜の散歩に出かけた。

例えば，(4) の 3 文目の “彼” は “太郎” と “次郎” の 2 通りの解釈があり，自然な読みは “ 太郎” であるが，これを *Cb* と *Cp* を使って説明すると，(5) に示すように “太郎” の読みの場合は遷移が CONTINUE であるのに対し，“次郎” の場合は SMOOTH-SHIFT であるため，より自然な遷移である CONTINUE の場合の “太郎” が望ましいことがわかる．

(5)
S_1	*Cb*: ?	*Cp*: 太郎	*Cf*: [太郎, 学校]
S_2	*Cb*: 太郎	*Cp*: 太郎	*Cf*: [太郎, 次郎, 食事]
S_3 (先行詞=太郎)	*Cb*: 太郎	*Cp*: 太郎	*Cf*: [太郎, 夜, 散歩] (CONTINUE)
S_3 (先行詞=次郎)	*Cb*: 次郎	*Cp*: 次郎	*Cf*: [次郎, 夜, 散歩] (SMOOTH-SHIFT)

このように中心化理論では同一談話セグメント内の照応関係を簡潔にモデル化しているが，隣接発話に先行詞を含んでいない場合もしくは一つの発話に照応詞が複数出現する場合にどのように解釈すべきかについては言及されていない．この理論の詳細とその拡張については文献[2]に詳しい．

5-2-2 機械学習に基づく照応解析

典型的な機械学習に基づく照応解析の手法[3]では，照応解析の問題をある表現が他の表現と照応関係になるか否かの 2 値分類問題として扱う．訓練の際には，照応関係となる先行詞と照応詞の対を正例，それ以外の対を負例として学習する．学習には語彙・統語・意味的な情報に加えて，照応詞と候補の相対的な位置の情報などを利用する．実際に先行詞を同定する際には，学習した分類モデルを利用して，照応詞候補に対して前方の先行詞候補集合のうちどの候補が照応関係の対となるかを決定する．ただし，表現の対をそれぞれ独立に分類する問題として扱うと，先行詞候補が他の談話要素と共参照の関係にある場合にその情報を有効に利用することができない．

(6) 山田花子外相は訪米中の英外相と会談した。山田首相は○○条約に署名した。彼女は

「お互いに政治的な信頼を深め、主要な利益に関するテーマについて相互に理解し、協力したい」とコメントした。

例えば，(6) の “山田首相” はこの談話要素の性別がわからないため，“山田首相” と “彼女” が共参照の関係となるか判断できないが，前方文脈の “山田花子首相” と “山田首相” が共参照関係にあるため “山田首相” が女性であるという情報を推論することで正しい関係を導くことができる可能性が高くなる．このような談話要素の集合の関係を捉えるために，談話全体での解釈の最適解を求める試みがなされている[4]．

5-2-3 省略解析

代名詞や定名詞の照応の現象とは異なり，照応詞がゼロ代名詞の場合には談話のどの箇所に省略があるかを検出する処理が必要となる．この処理には述語がどのような格要素をとるかを記述した格フレーム辞書や大規模な共起用例を用いて構築した選択選好のモデルなどが利用される．例えば，(7) の 2 文目では “買う” のヲ格が省略されているが，このゼロ代名詞は “〈 有生物 〉ガ〈 食物 〉ヲ 買う” のような格フレームの情報と実際の談話中の表現を照合することで検出可能である．ただし，動詞によっては語義によって異なる格パターンで出現したり，格要素に異なる意味カテゴリの語をとるため，頑健な格解析の処理が必要となる．

(7) 太郎は携帯電話を買った。花子も（ϕ ヲ）買った。

5-2-4 照応（共参照）の課題設計

機械学習に基づく解析を実現するには照応関係の情報（タグ）が付与されたコーパスが必要となる．ただし，タグ付与する照応もしくは共参照の関係を談話中のすべての表現について厳密に規定しようとした場合には，例えば，今日の私と昨日の私は厳密に同一実体を指していると考えてよいのか，また，“太郎の思い” のような抽象名詞の共参照関係をどのように判定するかといった問題を考える必要がある．このため，自然言語処理における照応解析の問題は特定の応用処理を想定して課題設計されることが多い．例えば，Message Understanding Conference や Automatic Content Extraction プログラムで提供される代表的な共参照解析のデータセットでは情報抽出の精度の向上に関連する箇所にのみタグ付与されている．これにより人手タグ付与の品質は向上するが，タグ付与の対象が限定されているため包括的な照応（共参照）現象の研究には適していない．また，間接照応の問題設計については，どのような場合に間接照応の関係にあるかの定義が困難であり，今後更なる議論が必要となる．

■参考文献

1) B. J. Grosz, A. K. Joshi, and S. Weinstein：“Centering: A framework for modeling the local coherence of discourse,” Computational Linguistics, vol.21, no.2, pp.203-226, 1995.
2) M. Walker, A. K. Joshi, and E. Prince (eds.)：“Centering Theory in Discourse,” Oxford Univ. Press, 1997.
3) W. M. Soon, H. T. Ng, and D. C. Y. Lim：“A Machine Learning Approach to Coreference Resolution of Noun Phrases,” Computational Linguistics, vol.27, no.4, pp.521-544, 2001.

4) H. Poon and P. Domingos：“Joint unsupervised coreference resolution with Markov Logic,” in Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp.650-659, 2008

■2 群 - 10 編 - 5 章

5-3 語義曖昧性解消

（執筆者：白井清昭）［2009 年 7 月 受領］

5-3-1 語義曖昧性解消とは

語義曖昧性解消（Word Sense Disambiguation，以下 WSD）とは，文中に多義の単語があるとき，その文における正しい語義を決定する処理である．単語の多義性解消と呼ばれることもある．ここで単語が多義であるとは，その単語が複数の意味（語義）を持つことを表す．例えば，以下の 2 つの例文には “plant” という多義の単語が含まれる．

1. It is easiest to train the *plant* when the new shoots are very young.
2. The company runs its *plant* day and night.

“plant” は「植物」という語義と「工場」という語義を持つ．ここでは，これらを *plant* $_{植物}$，*plant* $_{工場}$ と表す．これら 2 つの語義のうち，例文 1. における “plant” の語義が *plant* $_{植物}$ であり，例文 2. では *plant* $_{工場}$ であることを決める処理が WSD である．

WSD は意味解析の一つとして位置付けられる．文を構成する個々の単語の正しい語義を決めることは，文の意味を理解する意味解析において必要不可欠な処理である．また，WSD は様々な自然言語処理システムに必要な基礎技術でもある．例えば，機械翻訳システムが例文 1. や 2. を日本語に訳す場合，“plant” の意味が *plant* $_{植物}$ か *plant* $_{工場}$ なのかを正しく決めなければ正しい日本語訳を生成することはできない．

語義の曖昧性を解消するためにこれまでに様々な方法が提案されてきたが，ここでは代表的な 3 つの手法，(1) 格フレーム辞書を利用する手法，(2) 辞書定義文を利用する手法，(3) 機械学習に基づく手法について述べる．

5-3-2 格フレーム辞書を利用する手法

格フレーム辞書の選択制限の情報を使うと語義の曖昧性を解消できる．選択制限（または選択制約）とは，動詞の項に出現する名詞に関する意味的な制約である．例えば，動詞 “train” には *train* $_{訓練}$(訓練する) と *train* $_{曲げる}$(好みの形に仕立てる) という 2 つの語義があり，それぞれの格フレームは以下の通りとする．

train $_{訓練}$　主格: [人間 | 組織]，目的格: [人間 | 動物]
train $_{曲げる}$　主格: [人間]，目的格: [植物]

赤字は選択制限を表す意味素であり，*train* $_{訓練}$ の格フレームは「人間」または「組織」を意味素とする名詞を主格に，「人間」または「動物」を意味素とする名詞を目的格にとることを表す．*train* $_{曲げる}$ の格フレームも同様である．一方，語義 *plant* $_{植物}$ の意味素は「植物」，*plant* $_{工場}$ の意味素は「具体物」であるとする．例文 1. において，“plant” 及び “train” の語義の組み合わせは 2×2 = 4 通りあるが，上記の格フレームの選択制限を満たすものは *plant* $_{植物}$ と *train* $_{曲げる}$ の組のみである．したがって，“plant” の語義は *plant* $_{植物}$，“train” の語義は *train* $_{曲げる}$ であると正しく決定することができる．一般に，動詞の選択制限は名詞または動詞の語義を決定

するための有力な手がかりとなることが知られている．

5-3-3 辞書定義文を利用する手法

語義の曖昧性を解消するために辞書の定義文を利用することができる [1]．例えば，Longman Dictionary of Contemporary English における *plant* 植物，*plant* 工場，*train* 訓練，*train* 曲げる の定義文は以下の通りである．

plant 植物	a living thing that has leaves and roots and grows in earth, especially one that is smaller than a tree.
plant 工場	a factory or building where an industrial process happens.
train 訓練	to teach someone the skills of a particular job or activity, or to be taught these skills.
train 曲げる	to make a plant grow in a particular direction by bending, cutting, or tying it.

ここで，*plant* 植物 と *train* 曲げる の定義文にはともに “grow” という単語が含まれるので，これらの語義は互いに関連があると考えられる．一方，他の語義の組については，“a”, “in” などの機能語以外では共通して現われる単語はない．これにより例文 1. の “plant” 及び “train” の語義はそれぞれ *plant* 植物 と *train* 曲げる であると決定できる．このように，同一文中における多義語の全ての語義の組み合わせについて，辞書の定義文中に共通して現われる単語の数を調べ，その数が最も多い語義の組を正しい語義として選択する．ただし，冠詞，前置詞などの付属語は除いたり，活用形を基本形に戻すなどの前処理が必要である．

5-3-4 機械学習に基づく手法

機械学習に基づく手法は，コーパスを利用して語義の曖昧性を解消するモデルを自動的に学習する方法である．これらの手法は教師あり学習と教師なし学習に大きく分けられる．教師あり学習は正しい語義の付与されたコーパスを，教師なし学習は語義の付与されていないコーパスを訓練データとする．近年では教師あり学習に基づく手法が WSD の主流なアプローチとなっている．ここではその一例として，Naive Bayes モデルを取り上げる [2]．

まず，式 (5･1) のような確率モデルを学習する．

$$P(s|C) = \frac{P(s)P(C|s)}{P(C)} \simeq \frac{P(s)\prod_{w\in C} P(w|s)}{P(C)} \qquad (5\cdot1)$$

$P(s|C)$ は，文 C の中に出現する曖昧語 (語義を決める対象とする単語) の語義が s であるという確率であり，これを Bayes の定理にしたがって変形すると式 (5･1) の 2 番目の項を得る．さらに，文 C に含まれる単語 w は互いに独立に生成されると仮定すると，$P(C|s)$ を $\prod_{w\in C} P(w|s)$ で近似することができ，式 (5･1) の最終項を得る．全ての語義の候補について式 (5･1) の確率を計算し，最も確率の高い語義 s を選択することで語義の曖昧性を解消する．ただし，分母の $P(C)$ は全ての語義について等しいので計算を省略できる（分子が最大となる語義を選択すればよい）．

式 (5･1) のパラメタ $P(s)$ ならびに $P(w|s)$ は語義タグ付きコーパスから最尤推定などで比較的容易に学習できる．$P(s)$ は語義の出現頻度の統計情報を学習する働きをする．一方，

$P(w|s)$ は語義 s と単語 w の共起に関する統計情報を学習する．例えば，“company” という単語は $plant_{植物}$ よりも $plant_{工場}$ の周辺によく出現するので，$P(company|plant_{工場})$ は $P(company|plant_{植物})$ よりも高く推定される．

ここでは Naive Bayes モデルを取り挙げたが，語義タグ付きコーパスを訓練データとして用意できれば，任意の機械学習アルゴリズムを用いて WSD のモデルを学習することが可能である．例えば，学習アルゴリズムとして決定リスト，最大エントロピー法，Support Vector Machine，条件付き確率場（Conditional Randam Field）などを用いた研究がある．また，機械学習を行う上で重要なのは，学習素性としてどのような情報を用いるかという点である．WSD でよく用いられる学習素性を以下にまとめる．

語義の出現頻度 語義の出現頻度の違いは WSD の有力な手がかりとなる．Naive Bayes モデルのパラメータ $P(s)$ は学習素性として語義の出現頻度を用いた例と言える．

周辺の単語 曖昧語の周辺に出現する単語もまた正しい語義の選択にとって有用な情報である．一般に，学習素性として使われる周辺語は，曖昧語と同じ文または段落に出現する，あるいは曖昧語の前後 k 語の範囲内に出現する単語（ただし付属語は除く）とする場合が多い．Naive Bayes モデルのパラメータ $P(w|s)$ は学習素性として周辺の単語を用いた例といえる．

コロケーション コロケーションとは曖昧語を含む連続した単語または品詞の列である．例えば “manufacturing plant” というコロケーションは plant の語義が $plant_{植物}$ であることを示唆する．一般に曖昧語の前後 1〜3 語の範囲のコロケーションが学習素性として使われる．

統語的関係にある語 5-3-2 で述べたように選択制限は WSD の有力な手がかりとなる．そのため，動詞と主語，動詞と目的語の関係など，曖昧語と統語的関係にある単語が学習素性としてよく使われる．

文書のトピック 文書のトピックもまた WSD の有用な手がかりとなる．例えば，同じ “plant” という単語でも，園芸に関する文書では $plant_{植物}$ の語義が，工業に関する文書では $plant_{工場}$ の語義がよく使われる．

語義タグ付きコーパスは正しい語義を人手で与える必要があり，作成コストが高いことから，訓練データとしてプレインテキストを用いる教師なし学習を行う手法もいくつか報告されている．例えば，式 (5･1) の Naive Bayes モデルのパラメータを EM アルゴリズムによって教師なし学習する手法が提案されている[3)]．

■参考文献

1) Michael Lesk：“ Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone,” In *Proceedings of the 5th annual international conference on Systems documentation*, pp.24-26, 1986.
2) William A. Gale, Kenneth W. Church, and David Yarowsky：“ A method for disambiguating word senses in a large corpus,” *Computers and the Humanities*, vol.26, no.5-6, pp.415-439, 1992.

3) Christopher D. Manning and Hinrich Schütze：“ Unsupervised disambiguation,” In *Foundations of Statistical Natural Language Processing*, chapter 7.4, pp.252-256, The MIT press, 1999.