

修 士 論 文

照応詞と先行詞の文脈的な置換可能性
を考慮した共参照解析

Coreference Resolution Considering
Contextual Substitutability
between an Anaphor and its Antecedent

指導教員 鶴岡 慶雅 准教授

東京大学工学系研究科
電気系工学専攻

氏 名 37-126443 佐藤 美沙

提 出 日 平成 26 年 2 月 6 日

概要

文章中にある語句（言及）が同じ実体を指し示す共参照の現象を発見する共参照解析の問題は、機械翻訳や自動要約といった自然言語処理分野の他の応用においても推定結果が利用される重要な問題だが、現在の共参照解析システムの性能は最高レベルのものでも F 値にして 60% 程度と低水準にとどまっている。しかし、共参照には様々な言語現象が関与しており、何が課題であるのかは明らかでない。

本研究でははじめに共参照解析システムの解析結果に生じているエラーの解析を行い、そこで発見された課題点を解消するような手法を提案する。エラー解析では、誤りを原因によって人手で分類した結果、言及の周辺に存在する文脈の情報を考慮できていないことが大きな課題であるとわかった。そこで、文脈の情報を考慮するために、1) 識別的言語モデル、2) 代名詞概念クラス分類それぞれによって推定した文脈的な置換可能性を共参照解析に用いる手法を提案する。文脈的な置換可能性とは、置換した後の文が文脈的に自然な文のまま保たれるという性質のことである。共参照関係にある言及である先行詞と照応詞は同じ実体のことを表しているため、文脈的に置換可能であると考えられる。すなわち、ある照応詞に対して文脈的な置換可能性の高い先行詞候補は、置換可能性の低い先行詞候補よりも先行詞である可能性が高いことが期待される。1) 識別的言語モデルでは、照応詞を先行詞候補に置換したセンテンスに対して、言語モデルにより言語としての自然さを推定した結果を置換可能性とする。2) 代名詞概念クラス分類では、代名詞を「組織」や「人間」といった概念クラスに分類し、概念クラスが一致するならば、置換可能性も高いとする。

提案手法の効果を確認するため、既存の共参照解析システムに文脈的な置換可能性の推定結果を適用する実験を行った。1) 識別的言語モデルの手法では、特に代名詞に対して置換可能性の情報を適用することで、従来システムの性能が向上することを確認した。2) 代名詞概念クラス分類結果を利用する手法では、文脈情報を利用した分類器によって共参照の解析結果が改善した。

文脈の情報を考慮するために文脈的な置換可能性を考慮した共参照解析を行うことで、特に表層の情報が少なく文脈が解析に重要であると考えられる代名詞に関わる部分に対して、有効であることがわかった。

目次

第 1 章	序論	1
1.1	背景	1
1.2	本研究の目的	2
1.3	本研究の貢献	2
1.4	本論文の構成	3
第 2 章	共参照解析	4
2.1	共参照解析とは	4
2.1.1	問題設定	4
2.1.2	評価指標	5
2.1.3	共参照解析の流れ	7
2.2	ルールベースの共参照解析システム	8
2.2.1	ルールベースの共参照解析	8
2.2.2	Stanford Deterministic Coreference Resolution System [11]	8
2.3	教師あり学習ベースの共参照解析システム	10
2.3.1	教師あり学習ベースの共参照解析	10
2.3.2	Reconcile [28]	11
第 3 章	ルールベースの共参照解析システムのエラー解析	13
3.1	設定	13
3.2	エラー分類の結果	13
3.3	課題の発見	16
第 4 章	識別的言語モデルによる 置換可能性を利用した共参照解析	17
4.1	言語モデルによる文脈的な置換可能性の推定	17
4.1.1	文脈的な置換可能性と共参照	17
4.1.2	言語モデリング手法	18
4.1.3	置換可能性と言語モデル	19
4.2	共参照解析のための識別的言語モデルの提案	19
4.2.1	擬似負例の作成	19
4.2.2	識別器の学習	20

4.2.3	識別的言語モデルの表現力測定	21
4.3	共参照解析システムへの置換可能性の適用	24
4.3.1	言語モデルによる置換可能性の推定	24
4.3.2	ルールベースのシステムへの適用	24
4.3.3	教師あり学習ベースのシステムへの適用	24
4.4	共参照解析の性能評価	25
4.4.1	実験設定	25
4.4.2	評価	26
第5章	代名詞の概念クラス分類による 置換可能性を利用した共参照解析	30
5.1	概念クラスと共参照	30
5.1.1	概念クラスとは	30
5.1.2	共参照解析における概念クラスの利用	31
5.1.3	関連研究 — 固有表現抽出・分類	33
5.2	代名詞の概念クラス分類手法の提案	33
5.2.1	提案手法の概要	33
5.2.2	係り受け特徴量の特徴量選択	35
5.2.3	概念クラス分類の評価実験	36
5.3	共参照解析への適用手法	37
5.4	共参照解析性能の性能評価	38
5.4.1	実験条件	38
5.4.2	評価	38
5.4.3	実験結果の詳細	39
第6章	結論	41
6.1	まとめ	41
6.2	今後の展望	42

目 次

2.1	共参照の例	5
2.2	言及ペアモデル	8
4.1	置換可能性と共参照	17
4.2	提案識別的言語モデルに用いる正例・負例の作成方法	20
4.3	学習に用いるデータ数を変化させたときの識別的言語モデルの表現力評価実験結果	23
5.1	言及の概念クラス	31
5.2	代名詞の概念クラスと共参照解析	31

第1章 序論

1.1 背景

インターネットの発達に伴い膨大な量のテキストデータが存在するようになってきた。自然言語処理 (Natural Language Processing; NLP) は、英語や日本語といった人間が日常使用する自然言語によって記述されたデータを、計算機によって処理し活用するという分野であり、情報過多問題に対する有望な技術として近年大きな注目を集めている。

共参照解析 (coreference resolution) は自然言語処理の重要なタスクのひとつである。文章中の語が現実世界における同じ実体 (entity) を指し示している現象を共参照 (coreference) と呼び、これらの語同士は共参照関係にある、と言う。共参照の関係にある語は言及 (mention) と呼ばれ、二つの言及に着目した場合、手前に出現するものは先行詞、後から出現するものは照応詞と呼ばれる。共参照解析のタスクは、共参照関係にある言及を繋ぐリンクのすべて (チェーン) を明らかにすることである。なお、本稿で対象とするのは英語における名詞句の共参照とする。

共参照解析は自然言語処理において重要な役割を果たす。共参照の関係は、機械翻訳や質問応答システム、自動要約といった自然言語を扱う様々なアプリケーションで利用できる。重要なタスクであるため、文書データに共参照関係を人手で付与して作成されたコーパスも複数存在し [1, 14, 30], 対象の問題に対する提案手法の性能を競う共通タスクも複数行われてきている [13, 21]。

共参照には様々な言語現象が関与しているため解析は容易でなく、現状最高レベルの性能のシステムでも F 値で 60 数% 程度である。毎年新しい手法が提案されている [3, 27, 11] が、依然として性能向上の需要のある問題である。

共参照解析では、意味を考慮することが行われている。意味を考えるための手法の一つとして、辞書が活用されている。言及の語について辞書に記されている情報を用いることができ、WordNet などの類義語辞書情報の利用 [26] は有効性が知られている。しかし、多義性のある語の場合にはどの意味を採用するかを決定する問題を解く必要があり、また代名詞についてはほとんど無情報になってしまうことも問題である。

言及そのものについての情報だけでなく、言及の周辺に存在する情報が解析時に重要と考えられる共参照の例が存在する。たとえば、

- (1) When I saw a bike in the park, it was driving at high speed.

という文を考える。文中の “it” の先行詞候補には “I”, “a bike”, “the park” がある。言及そのものについての情報からは、一般に “I” は人間を指す代名詞で、“it” は人間を指さない代名詞であることがわかる。この不整合から “I” は “it” の先行詞でないと判断できる。しかし、“a bike” と “the

park” はどちらも単数形・生物ではないものを指す言及であり，“it” の持つ性質と矛盾しない。つまり，言及そのものだけの情報からは，“it” と共参照関係にある単語としてどちらの先行詞候補も適格と考えられる。この例で正しく “it” の先行詞が “a bike” であると答えるためには，“it” の文脈である “it was driving” という部分の情報が必要である。drive の主語として “the park” は不適切であり，“a bike” は適切であることから，正しい先行詞 “a bike” を選択できる。ここで用いた，言及の単語や句そのものの情報だけでなく周辺の情報を，文脈情報と呼ぶことにする。辞書から得られる情報は言及を構成する語についての情報だけであり，文脈情報が重要となる例には有効でない。

文脈情報を考慮した共参照解析の試みは複数行われてきている。その一つに，ウェブ情報を用いる手法がある [3]。この手法は，Web のドキュメントの N-gram データである Google N-gram コーパス [5] 内における特定のパターンの出現回数の特徴量とすることで文脈情報を捉えている。この手法の利点は，ウェブ情報という膨大な量の文書データを利用できる点である。また，直接 Web の検索ヒット数を利用する手法も提案されている [23]。この手法では，ヒット数が多いほど表現ほど文脈的に妥当であると考えられる。Google 検索ヒット数の特徴量は，代名詞である照応詞の先行詞を発見するタスクにおける性能向上に正答率 7.09% 寄与している。ラベルなしの生テキストから文脈情報を得るアプローチとして，生テキストからのパターン抽出により文脈情報を獲得する研究がなされている。情報抽出の手法を応用したアプローチ [4] では，照応詞の文脈と先行詞の文脈の組のパターンを用い，同一の実体を持つ妥当な組み合わせの文脈であるかどうかを推定する。パターンは共参照関係のラベル付きコーパスから有用そうなものを獲得している。

本研究では，文脈的な置換可能性という新しい観点を提案する。文脈的に置換可能とは，置換しても文章の文脈的な妥当性が保たれる性質のことで，先行詞と照応詞は同じ実体のことを表しているため文脈的に置換可能と考えられる。置換可能性の推定を構文解析結果などの深い情報を利用できる手法で行うことで，文字列パターンの数え上げといった表層から得る情報に比べてより正確な文脈情報を取得可能となることが期待される。

1.2 本研究の目的

本研究では共参照解析システムの性能向上を大きな目的とする。そのために，三段階の目的を設定する。

- 共参照には様々な言語現象が関与しているため共参照解析は容易でない。現状の共参照解析システムで生じているエラーの解析によって，解決に取り組むべき課題点を明らかにする。
- 共参照における文脈情報を捉えるため，先行詞と照応詞の文脈的な置換可能性を推定する。
- 文脈的非整合という課題点の解決のため，先行詞と照応詞の文脈的な置換可能性を考慮する共参照解析手法を提案し，有効性を確かめる。

1.3 本研究の貢献

本研究の貢献は大きく二点である。

- エラー解析によって文脈的非整合という課題があることを明らかにした。
- 文脈的非整合の解決を目的として、先行詞と照応詞の文脈的な置換可能性という観点を提示し、置換可能性推定方法と推定結果を共参照解析に適用する手法を提案し、実験により有効性を確かめた。

後者については、置換可能性推定方法として 1) 識別的言語モデルによる手法 2) 代名詞の概念クラス分類による手法の二通りを提案し、評価を行った。それぞれについての貢献を以下に記す。

識別的言語モデルによる置換可能性推定を利用するアプローチでは、以下が挙げられる。

- 豊富な種類の特徴量の利用によって言語モデルとしての識別能力が向上することを確認した。
- 共参照解析システムへの適用手法として、言語モデルによる置換可能性推定値を利用する新しい特徴量を提案した。実験により、照応詞が代名詞である言及ペアにおける特徴量の有効性を確認した。

代名詞の概念クラス分類による置換可能性推定を利用するアプローチでは、以下が挙げられる。

- 対象の代名詞の原形という表層情報に加えて、係り受け情報という文脈情報を利用する分類手法を提案し、概念クラス分類実験において分類性能が改善することを示した。
- 共参照解析システムへの適用手法として、先行詞と照応詞の概念クラスの一致不一致を置換可能性であるとして推定した結果を提案した。共参照解析への適用実験を行い、代名詞を照応詞とする組に関する性能が向上し、文書全体における性能も僅かながら向上することを確認した。

1.4 本論文の構成

まず第 2 章で共参照解析の問題設定と代表的な手法を概説し、従来システムのエラー解析の結果から文脈的整合性が大きな課題であることを述べる。続く第 4 章では共参照解析問題に適した擬似負例を学習に用いた識別的言語モデルによる置換可能性推定の手法を示す。推定した置換可能性を既存の共参照解析手法 [11, 28] へ適用し、代表的な言語モデルである N-gram 言語モデルを利用した手法との比較実験とその結果の考察を行う。第 5 章で、代名詞の概念クラス分類による置換可能性推定を行い、前章と同様に共参照解析に適用した実験結果と考察を述べる。最後に第 6 章でまとめと今後の課題を述べる。

第2章 共参照解析

2.1 共参照解析とは

本章では共参照解析タスクの問題設定と代表的な手法を概説し、共参照解析において文脈情報が有用となる例を述べる。

2.1.1 問題設定

文章中の表現には、現実世界に存在する実体 (entity) を指し示しているものがある。たとえば、図 2.1 に示した以下の二つの文

(2) **Barack Obama** was elected in this election. People welcomed **the new president**.

においては、“**Barack Obama**” はある人物を指し示しており、“this election” はある選挙を指し示している。こうした、実体を参照している表現は言及 (mention) と呼ばれる。

例文 (2) において、言及 “**the new president**” と言及 “**Barack Obama**” は同一の実体 (Barack Obama という、大統領である人物) を参照している。このように異なる言及が現実世界における同一の実体を参照する現象は共参照と呼ばれ、これらの表現同士は共参照関係にあると言われる。

共参照関係は、文章の内容を理解するために理解すべきことのひとつである。文章を計算機でまとめる自動要約や、計算機によって他の言語に翻訳する機械翻訳といったアプリケーションで利用される。共参照関係を知ることは、計算機が文章と事物を対応させ、文章の内容を理解するための重要な一歩と言える。

共参照解析とは、文章中の共参照関係を正しく検出することである。共参照解析は、それぞれの実体に着目すると、それを参照する言及をクラスタリングする問題に帰着できる。ただし、言及の参照先はただ一つの実体に決定できるとする。参照先が複数考えられ、一つに決定できない場合があるが、本論文では扱わない。たとえば下の例文では、“It” の参照先は “the restaurant”, “some sushi” もしくは「レストランで寿司を食べた」という出来事が考えられ、一つに定めることはできない。

(3) I went to the restaurant and ate some sushi. It was so good.

共参照関係にある2つの言及に注目した場合、文章中で手前に出現する言及を先行詞 (antecedent) , 後に出現する言及を照応詞 (anaphor) と呼ぶ。

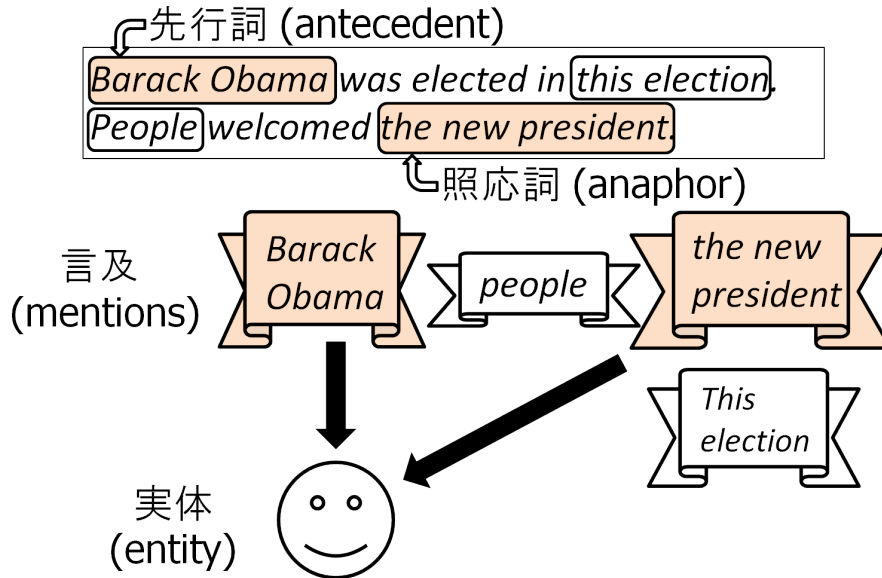


図 2.1: 共参照の例

2.1.2 評価指標

共参照解析システムの評価には、人手で共参照関係の情報が付与された文章群であるコーパスを用いる。コーパスの共参照ラベルとの一致が多いほど性能が良いとするが、共参照解析はクラスタリングの問題と言えるため、どのような評価指標が適切であるかについては様々な議論がある。コーパスに記された人手で付けられた共参照関係を正解ラベルと呼ぶことにする。様々な指標が提案されており、研究の評価では複数の指標について計算することが多い。

以下に本稿で用いるスコア指標を説明する。それぞれの指標について、評価値は再現率・適合率・F 値を計算する。再現率は真の結果のうち推定結果と一致するものの割合、適合率は推定結果のうち真の結果と一致するものの割合、F 値とは、再現率と適合率の調和平均である。

二値分類問題においては、推定結果は 4 通りに分けられる。共参照であるかないかの二値に分類する問題について考えると、正解ラベルと一致する推定結果は、そのラベルが「共参照である」のときは真陽性 (True Positive), 「共参照でない」のときは真陰性 (True Negative) と呼ばれる。同様に、正解ラベルと一致しない推定結果は、推定結果が「共参照である」のときは偽陽性 (False Positive), 「共参照でない」のときは偽陰性 (False Negative) と呼ばれる。以上の内容を表 2.1 にまとめる。

ペアワイズスコア

言及ペアを基準とした評価指標である。すべての言及の、他のすべての言及とのペアを考える。各ペアについて、両方の言及が推定データで参照先の実体 (推定クラスタ) が同一のとき、両方の言及が正解データにおける参照先の実体 (真のクラスタ) が同一ならば真陽性、異なれば誤陽性とし、また両方の言及が推定データで参照先の実体 (推定クラスタ) が異なるとき、両方の言及が正解データ

における参照先の実体（真のクラスタ）が同一ならば偽陰性，異なれば真陰性とする再現率は（真陽性言及ペアの数 / (真陽性 + 偽陰性) 言及ペアの数），適合率は（真陽性言及ペアの数 / (真陽性 + 偽陽性) 言及ペアの数），により計算する。

全てのペアは独立に扱われ，ペア間の関係は表現されない。

MUC スコア [29]

実体クラスタ内のリンクを基準とした評価指標である。再現率の誤り率は，正解の言及間リンクに対して，推定されなかった言及間リンクの割合によって定義される。つまり，推定クラスタを繋ぎあわせて真のクラスタに属する言及すべてを含むようなクラスタを構成するために追加が必要なリンクの本数を分子，真のクラスタを構成するために必要な言及間のリンクの本数を分母とする。これは，（真のクラスタに含まれる言及の個数 - 真のクラスタに含まれる言及を 1 つ以上含む推定クラスタの個数）を分子，（真のクラスタに含まれる言及の個数 - 1）を分母とすることで計算できる。ただし，分母分子それぞれ真のクラスタに関しての和とする。適合率の誤り率は，再現率と同様に正解の言及間リンクに対して，推定されなかった言及間リンクの割合によって定義される。つまり，真のクラスタを繋ぎあわせて推定クラスタに属する言及すべてを含むようなクラスタを構成するために追加が必要なリンクの本数を分子，真のクラスタを構成するために必要な言及間のリンクの本数を分母とする。これは，（真のクラスタに含まれる言及の個数 - 推定クラスタに含まれる言及を 1 つ以上含む真のクラスタの個数）を分子，（真のクラスタに含まれる言及の個数 - 1）を分母とすることで計算できる。ただし，分母分子それぞれ推定クラスタに関しての和とする。

MUC 指標では，余分な言及を含むクラスタに対するペナルティが小さくなりがちであることが指摘されている [16]。

B^3 スコア [2, 28]

実体クラスタを基準とした評価指標。

真のクラスタと推定クラスタの両方に属する言及の割合によって定義される。再現率は，すべての推定クラスタに関して，真のクラスタと推定クラスタの両方に属する言及の数の和を分子に，推定クラスタに属する言及の数の和を分母に取る。適合率は，すべての真のクラスタに関して，（真のクラスタと推定クラスタの両方に属する言及の数の和）を分子に，（推定クラスタに属する言及の数の和）を分母に取る。なお，数はすべてのクラスタについての和とする。

実際の実験では，推定された言及は真の言及と必ずしも一致しない。推定結果の言及もしくは真の言及に対応する真の言及もしくは推定結果の言及が存在しない場合を特別に定義する。[28] ならび

表 2.1: 推定結果の分類

推定結果 \ 正解ラベル	共参照	共参照でない
	共参照	真陽性
共参照でない	偽陰性	真陰性

に本稿では、 $B^3 - all$ として、真の言及に対応する推定言及が存在しない、もしくは推定言及に対応する真の言及が存在しない場合は、再現率は (1/ 対応先のない真の言及の数)、適合率は (1/ 対応先のない推定言及の数) とする。

CEAF スコア (for Constrained Entity-Alignment F-Measure) [12]

推定クラスタと真のクラスタは互い的一对一对応であるという制約の下で、推定クラスタが真のクラスタに適合している度合いを任意のスコア関数について計算する。再現率は (真のクラスタと対応する推定クラスタ間のスコア / 推定クラスタの自身との間のスコア) とし、適合率は (推定クラスタと対応する真のクラスタ間のスコア / 真のクラスタの自身との間のスコア) とする。

スコア関数 $\phi(A, B)$ には、推定クラスタと真のクラスタをそれぞれ A, B とすると、

$$\phi(A, B) = (2 \times |A \cap B|) / (|A| + |B|)$$

を用いる。

CoNLL スコア

ここまで述べてきたように、共参照解析の評価指標は一つに定まっていない。そこで、CoNLL 2011 で開かれた共参照解析の共通タスク [21] (同一のデータを用いて同一のタスクに対するスコアを競うもの) では、MUC スコアと B^3 スコアと CEAF スコアそれぞれの F 値の平均を、システムの順位付けに用いる評価指標とした。

2.1.3 共参照解析の流れ

共参照解析の処理は大きく下記の二段階に分けられる。

- 言及検出 (mention detection) : 言及を認識
- 共参照解析 : 前段の言及候補にある共参照関係を認識

言及検出は、文章中にある言及を認識するという一つのタスクである。構文解析器や固有表現抽出システムの出力を利用し、主に固有表現、名詞句、代名詞を言及候補とする。共参照関係の段階では、前段の処理で得た言及候補にある共参照関係を探していく。そのため言及検出で認識から漏れてしまった言及は解析の対象とならず、関係する共参照も発見されることはない。

共参照解析を含む多くの自然言語処理のタスクでは、他の自然言語処理タスクを解いた結果を利用することが多い。たとえば構文解析・品詞タグ付け・固有表現抽出・分類・センテンス分割といったタスクの推定結果は広範な有効性が知られている。固有表現抽出・分類の結果は多くの共参照解析手法で用いられている [15]。

2.1.3.1 言及ペアモデル

共参照関係を発見するための単純なアプローチは、文章内のすべての言及ペアについて共参照であるかないかを判断していくことである。すなわち、文章内に言及候補が N 個検出されたとき、各

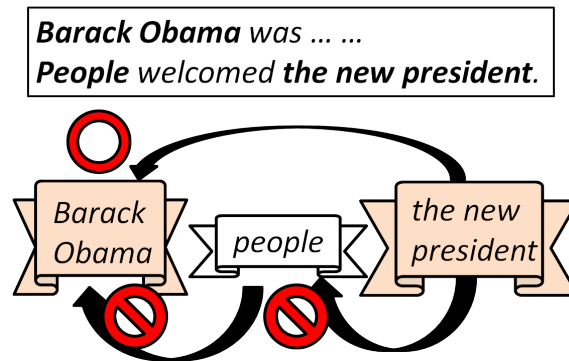


図 2.2: 言及ペアモデル

言及候補について、他のすべての言及候補との関係 $N - 1$ 個について共参照関係であるかどうかを推定してゆく。この解析の方針は言及ペアモデルと呼ばれる。解析の様子を図 2.2 に示す。

2.2 ルールベースの共参照解析システム

2.2.1 ルールベースの共参照解析

人手で作成したルールをそのままデータに適用する手法をルールベースの手法と呼ぶ。たとえば、単数複数異なるものは「必ず」共参照ではないというルール、copular 動詞 (be 動詞など) で繋がれたものは「必ず」共参照であるというルール、といったものであり、条件付けは事前に決定している。

他の自然言語処理タスクと同様、共参照解析においても機械学習を利用した手法が近年盛んである。しかし現在の最高レベルの精度を上げている共参照解析手法にルールベースの手法 [11] が含まれている通り、決して時代遅れの使えないアプローチではない。また、完全にルールにしたがって結果を出すため、一部ブラックボックス的である機械学習による実験と比べて挙動が明らかで結果を解析しやすい。そのため、共参照解析において有用な情報を明らかにするためにルールベースの手法を設計することも行われている [9]。

2.2.2 Stanford Deterministic Coreference Resolution System [11]

Stanford Deterministic Coreference Resolution System [11] という、CoNLL2011 の共通タスク [21] において最も優れたスコアを提出したルールベースのシステムについて説明する。特徴は、優先したいものに対して良い評価を出すルールを優先的に適用し、評価順に次々にルールを適用していくという複数のルールの使い方にある。

手続きとしては、まず解析したい言及について、それよりも手前に存在する言及達を右から左へ、近い順に判定してゆき、共参照である言及を発見したときに終了となる。

2.2.2.1 言及検出

言及検出では、再現率を優先する。まず、名詞句、所有代名詞、固有表現をそれぞれすべて言及候補として取り出し、再現率の高いルールから順に適用し、ルールに該当する言及候補を省いていくという手順を取る。詳しくは以下の通りである。番号の若い順に実行する。

1. 名詞句、所有代名詞、固有表現、のすべてを言及候補として取り出す
2. ある言及 (m_A) を含む言及 (m_B) があり、主要語を共有する言及が存在する場合、 m_A を削除する
3. 数字を表す語を削除する
4. 一部分であること・数量指定を表す修飾語の付いている語を削除する
5. 特定の構文であった場合、実体を参照しない冗語の *it* と判断し *it* を削除する (例: *it is ... that ...*)
6. 国名が形容詞化した語を削除する
7. ストップワードに指定した語を削除する

2.2.2.2 共参照解析

同じ実体を表す言及が一つのクラスタになるようにクラスタリングを行う。解析開始時には各言及1つを含むクラスタが言及の数だけ存在する状態である。言及のペアについて共参照関係にあるかどうかを判断し、ペアを構成する言及のそれぞれを含むクラスタ二つを一つのクラスタに併合することで徐々にクラスタを大きくしていく。

共参照解析では、間違った共参照関係を判断し誤ってクラスタを繋ぐことのないようにすることが重要であるため、言及検出とは逆に、適合率の高い順にルールを適用していく。

用いるルールを以下に示す。番号の若い順に適用し、共参照関係であるかないかの推測結果が決定した場合、そこで終了する。ルール名中に出現する主要語という用語は、その名詞句の中で最も主要な語のことを表す。たとえば、“the correct runway”の主要語は“runway”である。

1. 談話情報

会話文に登場する人物に対するルール。

例：ダブルクォーテーションに囲まれた文中に出現する一人称代名詞 (*I, my, me* など) は発言者と同一である (共参照関係にある)

2. 文字列完全一致

言及の文字列が完全に一致するものを共参照関係であるとするルール。

3. 文字列緩い条件の一致

関係代名詞節や前置詞句を除いたときに、言及の文字列が一致するものを共参照関係であるとするルール。

4. 構文合致

特定の構文に該当する場合についてのルール。例：同格構文にある言及同士は共参照関係である。

5. 厳しい条件の主要語一致

主要語が一致するならば共参照関係であるとするルール。ただし、いくつかの制約条件がある。たとえば、後に出現する方の言及に係る修飾語がそれ以前に出現する言及に係る修飾語の中に存在しない場合は、共参照関係ではないとする。

6. 固有表現の主要語一致

主要語が固有表現であり、一致するならば共参照関係であるとするルール。

7. 緩い条件の主要語一致

5. 厳しい条件の主要語一致 と同様だが、いくつかの制約を除いたルール。

8. 代名詞

代名詞の言及に関するルール。1. 談話情報に該当するもの以外の代名詞は、すべてこのルールによって処理される。以下の制約の全てに抵触しない場合、共参照関係であると推定する。

- 単数形・複数形が矛盾しない
- 性別が矛盾しない
- 人間であるかどうか矛盾しない
- 生物であるかどうか矛盾しない
- 代名詞の言及とその先行詞との間の距離は 3 センテンス以上離れない

性別・人間であるかどうか・生物であるかどうかの情報は、代名詞以外については構文解析器・固有表現分類器・辞書を用いて取得する。代名詞については、各代名詞の性質を記したリストを予め用意し、それに従い取得する。たとえば、“it” は単数形・性別なし・人間でない・生物でない とされる。

2.3 教師あり学習ベースの共参照解析システム

2.3.1 教師あり学習ベースの共参照解析

共参照関係にあるかどうかの判断に教師あり学習を用いる解析手法がある。教師あり学習では、予めラベルの付与されている訓練データから学習を行う。ルールベースの手法とは異なり、事前にルールを設計する必要がなく、容易に特徴量を追加することができる。自然言語処理分野において機械学習の手法は広く活用されており、共参照解析に対しても多数の教師あり学習ベースの手法が提案され続けている [3, 17, 16, 24, 26]。

2.3.2 Reconcile [28]

Reconcile [28] という教師あり学習ベースのシステムについて説明する。このシステムは言及ペアモデルを採用している。各ペアを独立に解析するため、最終的な推定結果である共参照クラスタを得るためにクラスタリング処理が必要となる。そこで、言及検出、共参照解析、クラスタリングの三工程を順番に行っている。以下では共参照解析とクラスタリングの手法について説明する。

2.3.2.1 共参照解析

教師あり学習の手法により推定する。学習で用いる特徴量は、大きく分けて以下の四通りである。

- 表層
言及の表層を利用する特徴量。主要語の一致など。
- 品詞・構文形
同格構文，定冠詞の有無 など。
- 距離
言及間にある単語数・センテンス数 など
- 意味情報
辞書に記されている意味の分類，類語情報 など

このシステムでは複数の学習器が用意されているが，本稿ではシステム内でデフォルトの学習器とされている平均化パーセプトロンを用いる。平均化パーセプトロンとは，パーセプトロンの改良版である。パーセプトロンは二値の識別器であり，入力 \mathbf{x} に対し出力 $y \in \{-1, 1\}$ を返す。

$$y = \text{sgn}(\mathbf{w} \cdot \mathbf{x})$$

ただし， \mathbf{w} は重みベクトルを， sgn は符号関数を表す。

$$\text{sgn}(x) = \begin{cases} 1 : x > 0, \\ 0 : x = 0, \\ -1 : x < 0 \end{cases}$$

学習の手順を以下に示す。

1. まず重みベクトル \mathbf{w} を初期化しておく。
2. 訓練例として入力と正解出力のペア (\mathbf{x}, t) を与えたとき，推定結果 $y = \text{sgn}(\mathbf{w} \cdot \mathbf{x})$ が正解出力 t と一致する場合には，何もせず，不正解であった場合には，重みベクトル \mathbf{w} を以下の更新式に従い更新する。

$$\mathbf{w}_{\text{new}} = \mathbf{w} + \alpha(t - y)\mathbf{x}$$

ただし， α は更新幅を表す定数とする。

3. 異なる訓練例を与え，2. を繰り返す。

学習終了時の重みベクトル w が学習結果である。

平均化パーセプトロンでは、上記手順の 2. の際に更新後の重みの値と重みの更新回数を記録しておく。学習の終了時に 重みの平均値 = 各更新後の重みの合計値 / 重みの更新回数 を計算し、この平均値を学習結果とする。

2.3.2.2 共参照リンクのクラスタリング

このシステムでは複数のクラスタリング手法が用意されているが、本稿ではシステム内でデフォルトの手法とされている single-link 法を用いる。これは推移閉包と呼ばれるものと同じである。まず、各言及ペアに対して推定結果を $[0, 1]$ の範囲の実数値で取得する。次に、この値が閾値以上であるペアを全て共参照リンクとして取得する。最後に、前の手順で取得した共参照リンクをすべて繋げることで、共参照のクラスタが得られる。

第3章 ルールベースの共参照解析システムの エラー解析

本章では、現状の共参照解析の課題点を明らかにするために、従来手法の出力で生じている誤りについて、原因による分類を人手で行ったエラー解析について説明する。

3.1 設定

第2章で説明したルールベースの共参照解析システムである Stanford Deterministic Coreference Resolution System [11] を MUC-6 コーパス [1] のテストデータに適用した結果を分類した。このデータには新聞記事から作成されており、30 文書に 2,068 個の言及が含まれている。全体の性能は、ペアワイズスコアで適合率 90.3%，再現率 53.3%，F 値 67.1%であった。生じた誤りのうち 123 個を分類した。

3.2 エラー分類の結果

分類結果を表 3.1 に示す。各数値（ポイント）は、その分類に該当するエラーの数を表す。ただし各エラーが複数の分類に跨る場合は、エラー一つにつき値の合計が 1 となるように、該当する分類の数の逆数を各分類に割り振っている。

以下では、各分類項目について説明する。

●文脈依存の類義語

特定の文脈においてのみ類義語となるような名詞（句）間の共参照に対する誤り。

(4) “the Clinton camp” – “Clinton transition officials”

例 “the Clinton camp – Clinton transition officials” では、“camp” と “officials” が類語関係になりうるという情報が不足していたために共参照関係が発見されなかった。“camp” と “it officials” は常に類語関係にあるわけではなく、文脈が役所関係の話であるという条件であるために類語となっている。しかしこの類語関係を常に許容してしまうと間違った共参照を判断してしまう。ゆえに単純な辞書の利用だけでなく、文脈の情報を活用することが有効と考えられる。

表 3.1: Stanford dcoref [11] における共参照解析誤りの分類結果。●を付けた項目が本研究で着目する点である。

原因	ポイント
●文脈依存の類義語	14.5
特殊な共参照（数値，時間など）	12.5
原因不明	11.5
修飾語の後出し	10.8
文章外知識	10.3
直接話法	8.0
●代名詞の概念クラス	7.0
他の実体との関係	5.8
分類不能	5.0
関わりの深い語との関係	3.0
その他（前処理でのミス，アノテーションミス）	30.5
合計	119.0

特殊な共参照

数値や時間など，主に物質以外を表す表現に関する共参照例に対する誤り。こうした共参照関係は，参照先の実体の性質が多くの共参照の対象とは異なるため，その性質特有の情報を考慮しなければ判断できない。

原因不明

誤りの生じた原因を特定できなかったもの。

修飾語の後出し

共参照関係にある手前に出現した言及には付いていなかった修飾語が，後から出現した言及には新たに追加されている例に対する誤り。エラー解析の対象のシステムでは，こうした例は制約により共参照関係にはないと判断するルールを含んでいる。システム固有の問題と言える。

文章外知識

copular 動詞，題名特有の表現，常識といった，文章以外の部分の情報が重要であるような例に対する誤り。

直接話法

直接話法に関わる部分の，話者・聞き手といった視点を考慮しなければならない例に対する誤り。

●代名詞の概念クラス

代名詞を照応詞とする言及ペアに対する誤り。先行詞の概念クラスと、代名詞の概念クラスに不整合があるものの、代名詞に対しては概念クラスが付与されていないために誤って共参照関係であると推定されてしまっているもの。

- (5) While **IBM's balance sheet** and core finances have remained stable, **it** has lost much credibility on Wall Street and has been stripped of **its** prized triple-A credit rating.

上の例において、“It” が指し示しうる候補に“IBM”と“IBM's balance sheet”の二通りがある。“It”の概念クラスは、“It”を含む文の内容を考慮すると、書類“balance sheet”よりも組織“IBM”の方がふさわしいと考えられる。このことから共参照関係にあるのは“IBM”であると解析することができる。

- (6) Patrick Purcell, chief executive of **News Corp.'s News America Publishing unit**, said yesterday, “**We** have been asked by various public officials if we would consider taking a look at it, and we are.”

この例では、**We**のクラスが会社（組織）である。従来システムではWeを必ず人間を指し示す代名詞であると事前に決めてしまっているために、正しい先行詞**News Corp.'s News America Publishing unit**を発見できていない。**We**が会社（組織）を表すことを推定できれば、正しい解析結果を得られる可能性が高まる。

他の実体との関係

言及だけでなく、参照先の実体の関係が重要であるような例に対する誤り。たとえば、

- (7) **Mary Kay Corp.**, a privately held cosmetics company. Mr. Bartlett was chief operating officer of **Mary Kay Cosmetics Inc.**, **the company's** operating subsidiary.

では、*the company*の**Mary Kay Corp.**との共参照関係を正しく取得するためには、同様に[組織]を表す言及**Mary Kay Cosmetics Inc.**ではないことを判断しなければならない。この判断は、まず**Mary Kay Cosmetics Inc.**と**the company's operating subsidiary**が同格関係にあり、*A's subsidiary*はAではありえないという実体間の関係性から行うことになる。

分類不能

いずれの項目にも分類できなかったもの。

関わりの深い語との関係 実際と同じものを指しているわけではないために共参照ではないものの、それでもほぼ同じことを指しているような非常に関わりの深い語が存在し、この語と言及との関係が共参照の判断において重要である例に対する誤り。

- (8) And more recently, somebody plucked out *Dr. Sagan as a code-name honoree*. The nerds reportedly jested: Maybe the name would be lucky, propelling the computers to sales of “billions and billions.” They changed the code name, all right. ... Apple won't, either,

though a spokesman — contrary to previous accounts — says the BHA designation that replaced **the Sagan code name** was “randomly chosen.”

この例文では、**the Sagan code name** と the name が共参照関係にある。この共参照に関しては、*Dr. Sagan as a code-name honoree* から *Dr. Sagan* がコードネームの由来であるという、関わりの深い語との関係が重要な情報となっている。

その他

構文解析や品詞タグ付けといった前処理における誤りや、コーパス自体が間違っているアノテーション誤り。

3.3 課題の発見

誤り分類項目を、以下の三項目に大別して考える。ただし、分類不能・原因不明・アノテーションの誤りは除いた。

- 共参照解析の対象外：その他（前処理でのミス、アノテーションミス）
- 個別の対処が有効であろう特殊な例：特殊な共参照，文章外知識，直接話法
- システム固有の問題：修飾語の後出し
- 共参照解析の課題と考えられるもの：● 文脈依存の類義語，● 代名詞の概念クラス，他の実体との関係，関わりの深い語との関係

共参照解析の課題と考えられるもののうち、エラー解析で数が多く発見されたものは● 文脈依存の類義語，● 代名詞の概念クラス（合計 21.5 ポイント（18.1%））である。これらは共通して、文脈情報の不足が原因であると言える。したがって文脈情報を考慮することが、共参照解析の性能向上に有用であると期待される。

個別の対処が有効と考えられる数値の共参照やタイトル内表現といった特殊な例や、共参照解析の手法が対象としている問題の外にあると考えられる構文解析結果の誤りなどの事前処理の精度の低さに由来する例の割合も小さくないが、本研究の提案手法では解決の対象としない。

第4章 識別的言語モデルによる 置換可能性を利用した共参照解析

本章では、言語モデルの手法を用いて共参照関係の妥当性を文の文脈から判別する提案法について説明する。

4.1 言語モデルによる文脈的な置換可能性の推定

4.1.1 文脈的な置換可能性と共参照

ある2つの表現について、置換したときに文脈に関する文の自然さが保たれる性質を、文脈的な置換可能性と呼ぶこととする。

先行詞と照応詞の置換可能性は共参照関係にない言及同士よりも高いことが期待される。共参照関係にある言及は同じ実体を参照しているため、置換前後で文の内容が変化しないためである。反対に、共参照関係にない言及と入れ替えた文は、誤った文となる可能性が高い。図4.1の例では、言及“it”を、共参照関係にある言及“a bike”と入れ替えたときの

(9) a bike was driving at high speed.

は妥当な文となり、共参照関係にない言及“the park”と入れ替えた文

(10) the park was driving at high speed.

は不自然な、文脈的に誤った文となっている。つまり、照応詞“it”と“a bike”の置換可能性は高く、“it”と“the park”は低い。

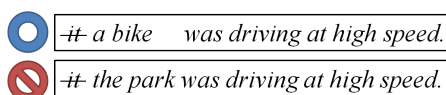


図 4.1: 置換可能性と共参照

4.1.2 言語モデリング手法

4.1.2.1 言語モデリング

自然言語をモデル化することを言語モデリングと言い、モデル自体は言語モデルと呼ばれる。言語モデルによって、任意の文字列に対して自然言語文としての自然さを推定することができる。

事前に人手でアノテーションを付けていない、生の文書データから学習を行う。そのためデータを増やすことへのコストが小さく大量のデータを利用することができる。

4.1.2.2 N-gram 言語モデル

文章中の N 単語からなる単語列は単語 N-gram と呼ばれる。“It is a pen.” というセンテンスに含まれる単語 1-gram は “It”, “is”, “a”, “pen”, “\$” であり、単語 2-gram は “It is”, “is a”, “a pen”, “pen \$” である。ただし \$ は終端記号であり、文の終端を表す。文の終端のピリオドは無視する。以下、N-gram は全て単語 N-gram のこととする。

N-gram 言語モデルは、単語の生起確率は直前の $N-1$ 個の単語列のみに依存すると仮定した確率的言語モデルである。つまり、

$$P(w_n | w_1^{n-1}) = P(w_n | w_{n-N+1}^{n-1}) \quad (4.1)$$

となる。ただし、 w_1^n は n 個の単語からなる単語列 $w_1 w_2 \dots w_n$ を、 $P(w_1^n)$ は w_1^n の生成確率を表す。 $i < 1$ のときの w_i は、予め文頭に $(N-1)$ 個の特別な記号があることによって定義できる。非常に単純なモデルであるが、確率・統計的自然言語処理の分野で最も広範に使われている言語モデルである [31]。

N-gram の確率は最尤推定により計算できる。

$$P(w_n | w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^n)}{C(w_{n-N+1}^{n-1} *)} \quad (4.2)$$

となる。ただし、 $C(w_1^n)$ は単語列 w_1^n の学習データにおける出現回数を、 $*$ は任意の単語を表す。

N-gram 言語モデルは、学習データに出現しない単語に対する確率をゼロとしてしまう。この「ゼロ頻度問題」に対処するために、平滑化と呼ばれる手法を適用することが多い。本研究では Kneser-Ney 平滑化 [10] を用いる。Kneser-Ney 平滑化は、 M 次元の N-gram の確率を求めたいときに、この N-gram が学習データ中に存在しない場合には、 M よりも低次元の N-gram で代用するバックオフ平滑化を改良した手法である。

N-gram 言語モデルにおいては、 n 語からなるセンテンス w_1^n の言語としての自然さは

$$P(w_1^n) = \prod_{i=1}^n P(w_i | w_{i-N+1}^{i-1}) \quad (4.3)$$

によって計算できる。

4.1.2.3 擬似負例を用いた識別的言語モデルの先行研究

擬似負例を利用した識別的言語モデルの手法が提案されている [18]。

識別モデルは負例を学習に用いる文書データには正例である正しい文しか存在しない。間違っている可能性の高い文を生成し、擬似負例として学習に用いている。[18] では文の正誤判定タスクを対象としており、擬似負例には N-gram 言語モデルにより生成された文を採用した。識別モデルとすることで可能となった豊富な特徴量の利用によって、高い識別精度を達成している。

4.1.3 置換可能性と言語モデル

言語モデルによってセンテンスの自然さを推定できる。置換後のセンテンスに対する言語モデルの確率を、置換可能性の推定値とする。

4.2 共参照解析のための識別的言語モデルの提案

先行詞候補となるような言及を用いた擬似負例の作成によって、共参照解析に適した識別的言語モデリングを行う。

4.2.1 擬似負例の作成

識別的言語モデルの学習に用いる擬似負例を作成する。データには生の文書データを用いる。

識別対象のインスタンスは「言及とその文脈の組」とする。それぞれ、正例は文書データ中の文をそのまま用いた「言及とその言及の文脈の組」、擬似負例は「言及と他の言及の文脈の組」とする。たとえば、図 4.2 に示す例，“All the cars” についてのインスタンスの作成を考える。正例は文章をそのまま用いて、

- (All the cars, * were driving slowly there.)

となる。擬似負例は，“All the cars” の先行詞候補である “The park” ・ “cats” についてそれぞれインスタンスを作成する。

- (The park, * were driving slowly there.)
- (cats, * were driving slowly there.)

言及と文脈の組をインスタンスとすることで、言及に関して文脈的に整合しているか不整合であるかを識別する問題になると考えられる。

正例は文書データ中のセンテンスそのままを、擬似負例には言及をその言及の先行詞候補と置換したセンテンスを用いていることに相当する。この擬似負例は、共参照関係にない不正解の先行詞候補を意図したインスタンスである。

手順を以下にまとめる。

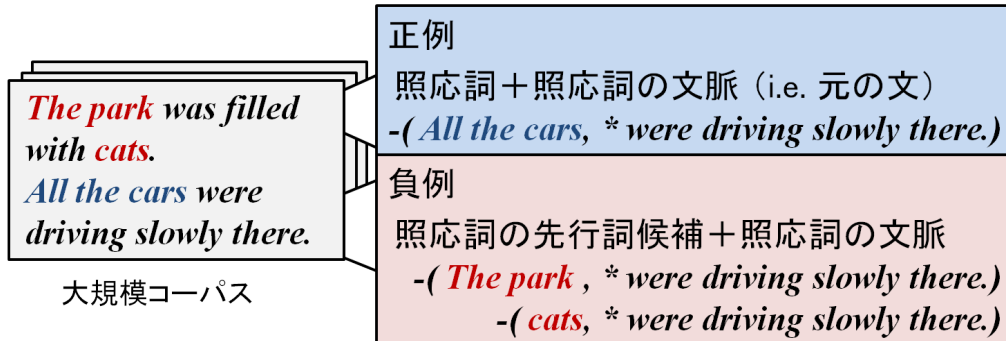


図 4.2: 提案識別的言語モデルに用いる正例・負例の作成方法

1. まず言及検出器により言及候補を取得する。
2. 各言及候補について、正例と擬似負例のインスタンスを作成する。正例には、ただし、置換する先行詞候補は対象の言及に近い順に3つまでとする。

言及検出器は前章で述べた Stanford Core NLP [11] 内のシステムを利用する。

この擬似負例は共参照関係にない先行詞候補との置換結果を意図している。そのため共参照関係にあることが多い以下の場合を除いた。

- 先行詞候補と対象の言及の主要語が同じ語である場合
- 少なくとも一方の言及が代名詞である場合
- 言及同士が重複している場合

主要語とは、名詞句を構成する単語のうちその名詞句を代表するような語である。“the vice president”では“president”が主要語である。言及同士の主要語の一致は共参照解析において識別能力の高い、正の重みを持つ特徴量として知られている [26]。そのため、共参照関係である可能性が高いとして置換対象外とする。代名詞はその性質上、近くに出現する言及を参照していることが多く、共参照関係である可能性が高いため、負例の作成には不向きと考えられる。重複している言及同士は同格関係でない限り共参照にならない、という i-within-i 制約と呼ばれるヒューリスティクスが共参照解析で用いられている [9]。つまり重複のある言及同士には、同格関係にある言及の組み合わせが多く含まれている。同格関係にあるものは共参照関係であるため、負例として不適格である。

実際に作成した擬似負例から無作為に抽出した 59 例について、文の内容を確認した結果を表 4.1 に示す。約 7 割が、意図した負例である文脈的に誤った文となっている。

4.2.2 識別器の学習

識別器には、ロジスティック回帰モデルを用いる。出力として 0 から 1 の実数が得られ、これを確率とみなすことができる。

識別的言語モデルで用いる特徴量は以下の通り。

係り受け特徴量

言及の関わる係り受けについての特徴量。構文解析器 [25] により取得する。

コンテキスト特徴量

言及の周囲に存在する単語・N-gram についての特徴量。各単語・N-gram について、言及からの相対単語距離とその単語の原形の組み合わせを特徴量とする。

使用する単語窓の大きさと N-gram の次元は、それぞれ 3 とする。すなわち、言及の 3 語前から 3 語後に含まれる 1gram, 2gram, 3gram を特徴量として用いた。

4.2.3 識別的言語モデルの表現力測定

有効な特徴量を選択するため、上の手法で作成した擬似負例と正例をテストデータとして識別的言語モデルの性能を測定した。

4.2.3.1 使用データ

Gigaword v4 コーパス [19] を利用した。このコーパスは、実際に出版された新聞をテキストデータ化したものである。出典・内訳を表 4.2 に示す。

4.2.3.2 センテンス分割

Gigaword コーパスのテキストデータにはセンテンスの区切りが付いていないため、センテンス分割を行う必要がある。本研究では、以下の条件を満たす区切りをセンテンスの区切りとみなし分割する。

- 後ろに大文字のあるピリオド：“.”
- クエスチョンマーク：“?”
- 閉じ括弧：“)”

ただし、以下に該当するものはセンテンスの区切りではないとして分割を行わない。

表 4.1: 提案法で作成した擬似負例の内訳

	個数	割合
文脈的に誤った文	40	68%
置換した言及同士は共参照ではないが、文脈的に正しい文	17	29%
置換した言及同士が共参照であったため、正しい文	2	3%
合計	59	100%

4.2 共参照解析のための識別的言語モデルの提案

- 敬称・月名・曜日の略称の末尾のピリオド
- 一部の定型句に含まれるピリオド (a.m., p.m., i.e., e.g.)

4.2.3.3 手法の比較

特徴量を変化させたときの性能の違いを測定する。

- 係り受け特徴量
- コンテキスト特徴量
- 係り受け特徴量+コンテキスト特徴量

また、学習に用いるデータ数を変化させたときの性能変化を測定する。

4.2.3.4 評価指標

テストセットにおける識別能力を測定する。各インスタンスは、正例・擬似負例ともに識別的言語モデルの学習データと同様の方法で作成する。したがってこの評価実験は、正例と擬似負例の識別問題である。テストセットのインスタンス数は400,000とする。

評価値は正例・擬似負例それぞれにおける適合率・識別率・F値、これらのマクロ平均、精度を計算する。マクロ平均は、正例と擬似負例における評価値の平均を取ったものである。適合率・識別率についてそれぞれ平均を計算し、F値はこれらの調和平均を計算する。精度 (Accuracy) は、正答数 / 全インスタンス数である。

表 4.2: Gigaword v4 コーパスデータの出典と内訳 [19]

新聞	略称	時期	センテンス数
Agence France-Presse, English Service	afp_eng	1994/05 – 1997/05, 2001/12 – 2008/12	24,813,831
Associated Press Worldstream, English Service	apw_eng	1994/11 – 1999/11, 2000/01 – 2008/12	49,463,453
Central News Agency of Taiwan, English Service	cna_eng	1997/09 – 2002/03, 2003/08 – 2004/06, 2004/08 – 2008/12	1,215,181
Los Angeles Times/Washington Post Newswire Service	ltw_eng	1994/05 – 1998/06, 2003/08 – 2008/12	10,976,194
New York Times Newswire Service	nyt_eng	1994/07 – 2004/05, 2004/07 – 2008/12	60,993,192
Xinhua News Agency, English Service	xin_eng	1995/01 – 2001/11, 2002/01 – 2009/12	13,734,535

4.2 共参照解析のための識別的言語モデルの提案

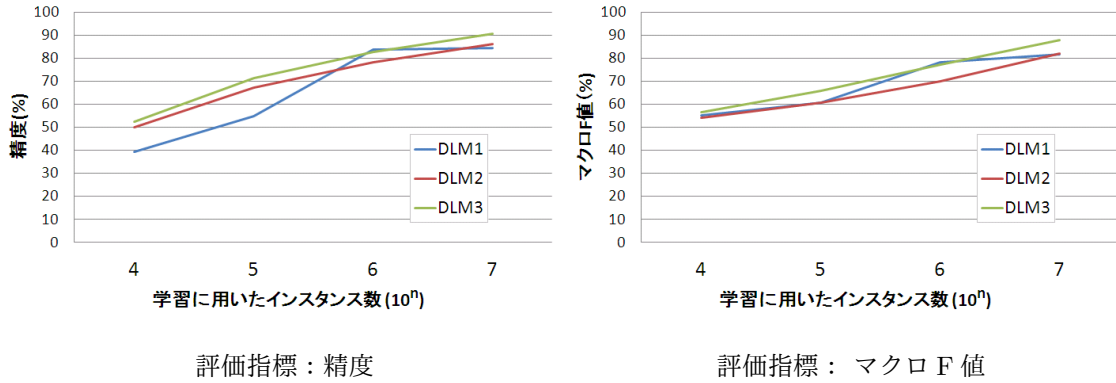


図 4.3: 学習に用いるデータ数を変化させたときの識別的言語モデルの表現力評価実験結果

4.2.3.5 実験結果・考察

特徴量を変化させたときの性能比較を表 4.3 に示す。係り受け特徴量とコンテキスト特徴量を組み合わせて用いたときが最も識別性能が高かった。豊富な種類の特徴量を用いることで文脈情報を捉えることができ、性能が上昇したと考えられる。

学習に用いるデータ数を変化させたときの性能比較グラフを図 4.3 に示す。いずれの特徴量を用いたモデルでも、多くのデータを用いるほど性能が向上する傾向が見られる。言語モデルとして妥当な挙動と言える。

表 4.3: 特徴量を変化させたときの識別的言語モデルの表現力評価実験結果。各モデルの使用特徴量は表 4.4 参照。

名前	正例			擬似負例			マクロ平均			精度
	適合率	識別率	F 値	適合率	識別率	F 値	適合率	識別率	F 値	
DLM1	68.6%	80.2%	73.9%	92.0%	86.1%	89.0%	80.3%	83.2%	81.7%	84.5%
DLM2	89.8%	68.3%	77.6%	89.0%	97.1%	92.9%	89.4%	82.7%	85.9%	89.2%
DLM3	86.6%	77.7%	81.9%	91.9%	95.5%	93.6%	89.3%	86.6%	87.9%	90.6%

表 4.4: 各識別的言語モデルで使用した特徴量

名前	使用特徴量
DLM1	係り受け
DLM2	コンテキスト
DLM3	係り受け+コンテキスト

4.3 共参照解析システムへの置換可能性の適用

第 2 章で述べたルールベースの手法 [11] と教師あり学習ベースの手法 [28] に、識別的言語モデルによる置換可能性推定結果を適用し、性能を確認する。

4.3.1 言語モデルによる置換可能性の推定

照応詞 ana_i について考える。添字はセンテンス内の i 番目の単語であることを示す。つまり、 $w_1 \dots w_{i-1} ana_i w_{i+1} \dots w_n$ というセンテンスを考える。この照応詞を先行詞候補 ant_j に置換したセンテンス $w_1 \dots w_{i-1} ant_j w_{i+1} \dots w_n$ に対する言語モデルの推定値を置換可能性として用いる。

4.3.2 ルールベースのシステムへの適用

従来システムでは、すべてのルールを適用しおえた際に複数の先行詞候補が残った場合、照応詞との構文木上の距離が最も小さい先行詞候補を採用している。

この複数候補から最終的に先行詞を選択する際の基準を、識別的言語モデルによる置換可能性推定結果に変更する。具体的には、以下の二通りの適用方法による変更を行う。

1. 最も置換可能性の高い候補
2. 置換可能性がある閾値以上である先行詞候補の中で、構文木上の距離が最も小さい候補

たとえば、ある照応詞に対して A(0.2), B(0.7), C(0.8) という 3 先行詞候補があるときを考える。括弧内は各候補に対する置換可能性の推定値であり、上記の順番が構文木上の距離の小さい順であるとする。このとき、最も置換可能性の高い候補を選ぶ適用手法 1. においては、置換可能性が 0.8 と最も高い C が先行詞として選ばれる。一方、閾値を用いる適用手法 2. において閾値を 0.3 としたときは、まず構文木上の距離が最も小さい A は置換可能性 (0.2) が閾値未満であるため不適とされる。次に距離が小さい B は置換可能性 (0.7) が閾値以上であるため、先行詞として選ばれる。

4.3.3 教師あり学習ベースのシステムへの適用

置換可能性推定値を用いた特徴量をシステムに追加する。

4.3.3.1 識別的言語モデルによる置換可能性特徴量

ロジスティック回帰モデルの重みベクトルが w であるとき、データ x に対する確率値 p はシグモイド関数 $\sigma(t) = 1/(1 + \exp(-t))$ を用いた以下の式で計算されている。

$$\begin{aligned} p &= \sigma(-wx) & (4.4) \\ &= \frac{1}{1 + \exp(-wx)} \end{aligned}$$

置換可能性特徴量は値の大小が得られればよく、0から1の実数値で表される必要はない。そのため本研究ではデータ x に関する線形関数である wx を置換可能性特徴量とする。ただし、従来システム内で使用する他の特徴量の値と大きく離れないようにするため、15で割った値を用いる。

$$\text{識別的言語モデルによる置換可能性特徴量} = \frac{wx}{15}$$

4.3.3.2 N-gram 言語モデルによる置換可能性特徴量

N-gram 言語モデルによる置換可能性特徴量には、置換後のセンテンスに対する言語モデルの推定値 P の対数を取った $\log P$ を用いる。

N-gram 言語モデルでは、単語の生起確率は手前の単語列のみに依存すると仮定している。しかし、共参照解析で考慮したい文脈情報は言及の後ろ側にも存在する。たとえば言及が主語である場合に係る動詞や、言及に係る前置詞句などは考慮に入れたい。そこで本研究では、後ろ向きの N-gram 言語モデルを学習し、前向き N-gram 言語モデルと後ろ向き N-gram 言語モデルそれぞれによる置換可能性特徴量を、2つの独立した特徴量として用いる。

4.4 共参照解析の性能評価

4.4.1 実験設定

共参照解析の評価実験には、ACE05 [30] コーパスから英語新聞記事のデータを用いた。ACE05 コーパスは、共参照・参照先の各実体のクラスについてアノテーションが付けられたテキストデータである。英語新聞記事は、現実の新聞記事 106 文書を用いたデータが提供されている。データの出典を表 4.5 に示す。訓練セット・テストセットの構成は Rahman ら [22] が公開しているファイルリストに従った。訓練セットとテストセットの比が文書単位で 3:1 になるように分けられている。センテンスの分割は第 4.2.3.2 節と同様の手法で行った。

言語モデルの学習には、Gigaword v4 [19] コーパスの一部のデータを用いた。表 4.6 に示す。ドメインの違いによる性能低下を防ぐため、共参照解析評価実験で用いるデータ（表 4.5）と近い時期の記事を使用した。ただし、出典が全く同じデータは除いている。

識別的言語モデルの学習・推定には機械学習ライブラリ Liblinear [6] を、N-gram 言語モデルには Berkeley LM [20] を用いた。

言語モデルによる置換可能性特徴量は、特に文脈情報の考慮が有効な言及ペアに与えることが望ましいと考えられる。そこで、特徴量を与える対象を変更した二通り

表 4.5: ACE05 [30] コーパス中の新聞記事データの出典と内訳。共参照解析評価実験に用いる。各新聞の略称については表 4.2 参照。

出典の新聞（略称）	時期
afp_eng, apw_eng, nyt_eng, xin_eng	2003/03 – 2003/08

4.4 共参照解析の性能評価

- すべての言及ペア
- 照応詞が代名詞である言及ペア

についてそれぞれ解析性能を測定した。対象外の言及ペアに対しては、置換可能性推定値は常に 0.0 とした。

閾値は、開発セットにおける上位 1/4 のインスタンスに与えられる置換可能性の値を目安に設定し、N-gram 言語モデルに対しては 0.004、識別的言語モデルに対しては 0.09 とした。

4.4.2 評価

4.4.2.1 ルールベースの従来法に適用した場合

結果を表 4.9 に示す。

1. 置換可能性最大の候補を採用の適用手法では、N-gram 言語モデル・識別的言語モデル共に、従来システム以下の性能であった。これは、先行詞と照応詞の距離という共参照解析における重要な情報を失ってしまったことによると考えられる。

2. 置換可能性を閾値として利用の適用手法においても、ほぼ全ての N-gram 言語モデル・識別的言語モデル共に、従来システム以下の性能であった。距離情報と置換可能性の双方を考慮するための適用手法であったが、性能の低下した実験においては、距離情報を完全には保持できなかったことが原因と考えられる。ただし、照応詞が代名詞である言及ペアにのみ、識別的言語モデルによる置換可能性を適用した実験では、ペアワイズ指標で +0.6% の向上が見られた。

4.4.2.2 教師あり学習ベースの従来システムに適用した場合

結果を表 4.10 に示す。

表 4.6: 識別的言語モデルの学習に用いたデータ

新聞略称	時期	センテンス数
afp_eng	2002/03 – 2003/02, 2003/09 – 2004/08	4,833,236
cna_eng	2002/03, 2003/09 – 2004/05, 2004/08	54,237
ltw_eng	2003/08 – 2003/12	572,601

表 4.7: 各言語モデルの学習に用いたデータ

モデル	センテンス数	出典
N-gram 言語モデル	161,196,386	Gigaword v4 コーパス 全て (詳細 : 表 4.2)
提案識別的言語モデル	5,460,074	CNA, 2002 – 2004 新聞記事 (詳細 : 表 4.6)

全ての先行詞・照応詞の組に適用した場合以上に、代名詞が照応詞である組にのみ適用した場合においてより大きな性能向上が確認された。代名詞は表層の情報が少なく文脈情報が重要となるため、言語モデルによって得られる情報が有効に働いたと考えられる。

表 4.8: ルールベースの従来法における共参照解析実験結果 1. 置換可能性最大の候補を採用

全ての言及ペアに適用した場合									
指標	従来システム			N-gram 言語モデルを適用			識別的言語モデルを適用		
	適合率	識別率	F 値	適合率	識別率	F 値	適合率	識別率	F 値
B^3	88.5%	69.9%	78.1%	88.5%	69.9%	78.1%	85.5%	67.3%	75.3%
MUC	84.7%	68.9%	76.0%	84.7%	68.9%	76.0%	80.7%	65.8%	72.5%
ペアワイズ	78.2%	49.7%	60.8%	78.2%	49.7%	60.8%	70.8%	44.9%	55.0%

照応詞が代名詞である言及ペアにのみ適用した場合									
指標	従来システム			N-gram 言語モデルを適用			識別的言語モデルを適用		
	適合率	識別率	F 値	適合率	識別率	F 値	適合率	識別率	F 値
B^3				86.4%	68.3%	76.3%	85.4%	67.3%	75.3%
MUC				81.9%	66.7%	73.5%	80.7%	65.8%	72.5%
ペアワイズ				73.6%	47.3%	57.6%	70.7%	44.9%	54.9%

表 4.9: ルールベースの従来システムにおける共参照解析実験結果 2. 置換可能性を閾値として利用

全ての言及ペアに適用した場合									
指標	従来システム			N-gram 言語モデルを適用			識別的言語モデルを適用		
	適合率	識別率	F 値	適合率	識別率	F 値	適合率	識別率	F 値
B^3	88.5%	69.9%	78.1%	89.4%	65.6%	75.7%	89.2%	66.8%	76.4%
MUC	84.7%	68.9%	76.0%	83.5%	60.9%	70.4%	84.4%	64.1%	72.9%
ペアワイズ	78.2%	49.7%	60.8%	78.5%	45.0%	57.3%	78.0%	45.3%	57.3%

照応詞が代名詞である言及ペアにのみ適用した場合									
指標	従来システム			N-gram 言語モデルを適用			識別的言語モデルを適用		
	適合率	識別率	F 値	適合率	識別率	F 値	適合率	識別率	F 値
B^3				88.5%	69.6%	77.9%	88.7%	69.8%	78.1%
MUC				84.4%	68.4%	75.6%	84.7%	68.6%	75.8%
ペアワイズ				78.3%	49.7%	60.6%	79.2%	50.1%	61.4%

表 4.10: 教師あり学習ベースの従来システムにおける共参照解析実験結果

指標	全ての言及ペアに適用								
	従来システム			+N-gram 言語モデル			+識別的言語モデル		
	適合率	識別率	F 値	適合率	識別率	F 値	適合率	識別率	F 値
B^3	95.0%	58.0%	72.1%	95.0%	58.0%	72.1%	94.6%	58.2%	72.1%
CEAF	25.2%	74.7%	37.2%	25.1%	74.6%	37.1%	28.4%	76.2%	41.1%
MUC	83.1%	46.0%	59.2%	83.1%	46.3%	59.4%	79.3%	44.4%	57.0%
CoNLL	–	–	56.2%	–	–	56.2%	–	–	56.7%

指標	照応詞が代名詞である言及ペアに適用								
				+N-gram 言語モデル		+識別的言語モデル			
	適合率	識別率	F 値	適合率	識別率	F 値	適合率	識別率	F 値
B^3				95.0%	58.0%	72.0%	94.6%	58.3%	72.1%
CEAF				25.1%	74.6%	37.2%	28.5%	76.2%	41.1%
MUC				83.2%	45.9%	59.2%	79.4%	44.6%	57.1%
CoNLL				–	–	56.1%	–	–	56.8%

第5章 代名詞の概念クラス分類による 置換可能性を利用した共参照解析

本章では、代名詞についての概念クラス分類の手法を提案し、分類結果を共参照解析で利用する方法とその実験結果について説明する。

5.1 概念クラスと共参照

5.1.1 概念クラスとは

実体はその性質によってクラス分けをすることができる。たとえば、“the company”の指す実体は[組織]という性質の分類に、“Barack Obama”や“the president”の示すものは[人]という分類に属する、というようなクラス分けが考えられる。言及は実体を参照している。したがって、言及については図 5.1 のように、参照先の実体の分類と同じ分類を与えることでクラス分けが可能となる。この分類を言及の概念クラスと呼ぶことにする。前述の例では、“the company”は[組織]，“Barack Obama”と“the president”は[人]という概念クラスを与えられる。

本研究では、概念クラス・共参照関係のアノテーションが付けられているデータセットとして、前章と同じく ACE05 [30] コーパスを利用する。このコーパスでは表 5.1 に示す 8 つの概念クラスが定義されている。概念クラスと共参照関係のアノテーションは実体に対して付与されている。即ちそれぞれの実体に、概念クラスと、その実体を参照先とする言及が紐付けられている。

表 5.1: ACE05 [30] コーパスにおける概念クラスの一覧

概念クラス名	英語名称	英語略称	例文 (太字: 該当の言及)
施設	Facility	FAC	at the Davao City international airport
政治的地名	Geo-Political Entity	GPE	Miami imposed a curfew because of its riots
場所	Location	LOC	Death toll in Philippines blast could hit 30
組織	Organization	ORG	The Davao Medical Center
人	Person	PER	Many of the victims were shown
車両	Vehicle	VEH	backed by helicopters which fired at least three rockets
兵器	Weapon	WEA	It's a very powerful bomb.

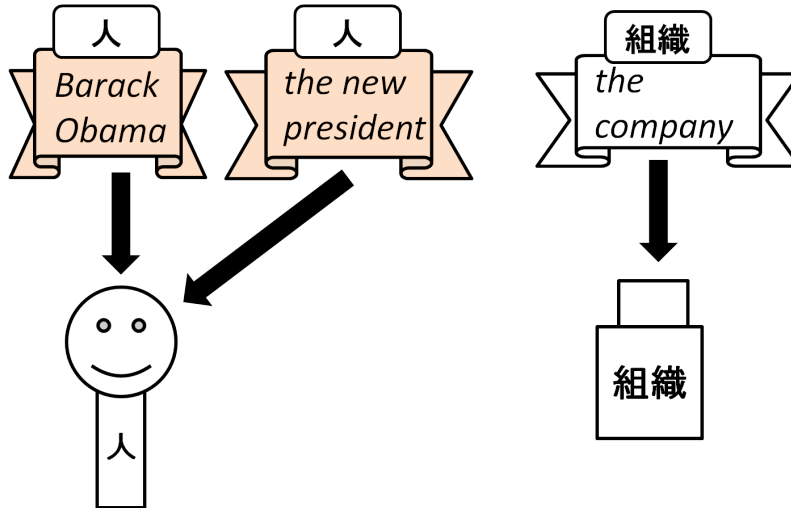


図 5.1: 言及の概念クラス

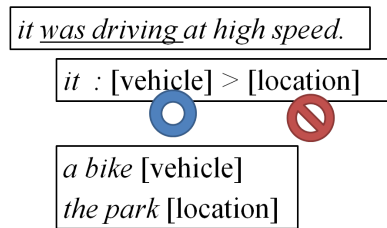


図 5.2: 代名詞の概念クラスと共参照解析

代名詞は参照先を持たない場合がある。例を示す。

(11) **It**'s fine today.

上の例文中の太字で示した“**It**”は参照先を持たない。言及とは実体への参照を持つ表現のことであり、参照の有無の判断は言及検知の問題である。本研究の対象は言及検知が行われた後の共参照解析であるため、概念クラス分けの対象においても、参照先を持たない言及はすでに除かれているとみなす。つまり、以降のクラス分けの対象は参照が存在することが既知である言及に限定する。

5.1.2 共参照解析における概念クラスの利用

5.1.2.1 共参照関係と概念クラス

共参照関係にある言及は、参照先の実体が同一であるため、概念クラスも同一であると考えられる(図 5.2)。

言及の概念クラス推定結果を利用することで、共参照解析の精度が上昇することが知られている。

例を示す。

- (12) Toyota, who joined **Toyota**, based in Toyota city, central Japan, in 1984, spearheaded **the automaker's** Internet and telematics business a technology that puts wireless communications into vehicles. Considered a candidate for future **Toyota** president, Toyoda oversees **its** China business, a critical market for **the automaker**.

例文 (12) では、代名詞 “its” と共参照関係にある言及 “Toyota”, “the automaker” の全てが [ORG] の概念クラスに分類される。

- (13) While **IBM's** *balance sheet* and core finances have remained stable, **it** has been stripped of **its** prized triple-A credit rating.

例文 (13) では、代名詞 “its” と共参照関係にある言及 “it”, “IBM” の全てが [ORG] の概念クラスに分類される。一方、共参照関係にない言及 “balance sheet” は [ORG] には分類されない。

実際に第 2 章で述べたシステムも固有表現の概念クラスの推定結果を使っている。Stanford dcoref [11] では制約として用いられている。まず、各言及には概念クラスが付与される。そして、二つの実体クラスについて共参照関係であるかどうかを判断する際に、互いのクラス内に含まれる言及のうち、付与された概念クラスの合致するものが一組もない場合、制約により共参照関係ではないと判断される。ただし、分類不能ないし推定不能とされた言及は他のどの概念クラスの言及とも合致するとする。教師あり学習ベースの手法 [28] でも一部の特徴量において制約として用いられている。Alias, Animacy, Appositive, ClosestComp の特徴量が Compatible になるのは概念クラスが一致するときのみとしている。

5.1.2.2 共参照解析の先行研究における概念クラス推定

名詞句を対象とした共参照解析において言及は、固有表現、普通名詞、代名詞という三種類に分類することができる。共参照解析で用いられる概念クラスは、名詞の種類によってそれぞれ別の手法で推定されることが多い。

固有表現については、固有表現抽出・分類という研究が共参照解析とは別に広く行われている。文書中に存在する固有表現を検出し、どの概念クラスに分類されるかを推定する問題である。手法は 5.1.3 で後述する。

普通名詞については、辞書情報が利用できる。WordNet [8] には、名詞に対してその上位語が記録されている。上位語を辿ることで、概念クラスに相当する意味の分類を取得することができる。

代名詞に対しては概念クラス分類と言えるような粒度の意味の分類は、先行研究 [11, 28] では行われていない。代名詞では代名詞そのものの表層から得られる情報が少ないために、十分な精度の結果を得ることが難しいためと考えられる。概念クラスではないが、参照先の実体に関する情報の一部は推定され利用されている。詳細を次の節で述べる。

5.1.2.3 共参照解析における代名詞の概念クラス相当の情報の取り扱い

代名詞の概念クラス分類ではないが、代名詞の表層から推定される参照先の実体の性質を利用することは行われている（表 5.2）。

たとえば、単複と animacy という分類が使われている。単複とは、代名詞の単数形・複数形のことを、animacy とはその表層を持つ代名詞が生物・非生物のどちらを参照するかのことを意味する。どちらの情報においても、複数の分類に属する可能性のある言及に対しては無判断としている。実際に先行研究 [11] では、“it”, “its”, “itself” は非生物, “they”, “them”, “their”, “themselves”は無判断, その他の代名詞が生物である, と分類している。

単複や animacy は、概念クラスに比べると分類の粒度が小さい。例文 (13) では、言及 “its” に関しては「単数」「非生物」という分類になるが、先行詞候補 “IBM” と “IBM’s balance sheet” のはどちらも同じ「単数」「非生物」という分類になるため、共参照解析に対して有用とは言えない。この例では、概念クラスのようなより詳細な分類が有用である。

5.1.3 関連研究 — 固有表現抽出・分類

クラス分けは、固有表現に対しては多くの研究があり、固有表現抽出と呼ばれている。機械学習ベースの固有表現抽出器で用いられている特徴量 [15] を表 5.3 に示す。特徴量は、単語に関する情報、辞書から得る情報、文書全体から得る情報に分けられる。単語情報と辞書情報は、固有表現の表層に関する情報である。文書全体情報は、文書中での対象の固有表現の振る舞いの情報である。

代名詞と固有表現は性質が異なる。固有表現では、同じ表現は常に同じ実体を参照することが多い。上記の特徴量でも、表層が多く用いられている。一方代名詞は様々なものを表すことができるため、同じ表現でも決して同じ実体を表すわけではない。表層から得られる情報は固有表現に比べ僅かである。そのため、固有表現分類とは異なる情報が必要であると考えられる。

5.2 代名詞の概念クラス分類手法の提案

5.2.1 提案手法の概要

代名詞の概念クラス分類に対しては、教師あり学習を行う。識別器には第 4 章で述べたように識別的言語モデルの作成でも用いたロジスティック回帰モデルを用いる。

表 5.2: 従来の共参照解析で用いられている代名詞の概念クラス相当の情報

名称	説明	例
animacy 単複	参照先が生物か無生物か 単数形か複数形か	I は生物, it は非生物, they は無判断 he は単数, they は複数

分類の対象とするものは代名詞だが、その代名詞であるということが概念クラスを一意に決めるものを除く。つまり、データ中に出現した代名詞に付与されている概念クラスが複数種類あるものとする。たとえば “it” は [ORG], [LOC], [WEA], [GPE], [PER], [FAC], [ORG], [VEH] のラベルの付いたものが存在するので対象とする。“she” や “her” は [PER] のラベルの付いたものしか存在しないので、対象外とする。この基準に従って決めた分類対象の代名詞は下記の通り。

mine, we, our, us, ours, you, your, he, his, it, its, they, them, their, this, these, that, those, which, who, whose, whom, where, one, ones, ourselves, yourself, itself, themselves, here, there

表 5.3: 固有表現抽出・分類で用いられる特徴量 [15]

性質	名称	説明
単語情報	大文字使用	先頭が大文字かどうか
	句読点	ピリオド、ハイフンなどを含むかどうか
	記号	記号を含むかどうか
	数字	数字を含むかどうか
	文字	アポストロフィ・ギリシャ文字を含むかどうか
	POS	品詞
	長さ	表現を構成する文字数, 単語数
辞書	一般的な辞書	類語辞書など
	略語	略語辞書
	固有名詞辞書	組織名・人名・地名辞書
文書全体	特徴的な表現	組織名・地名によく用いられる単語
	出現回数	大文字・小文字それぞれの出現回数, 共参照解析結果の利用
	局所的な構文	列挙・同格、文書・段落・センテンス内における位置
	メタ情報	括弧・リスト・表・図の内部にあるかどうか
	コーパス内頻度	頻度、共起

表 5.4: 代名詞概念クラス分類で用いる特徴量

性質	名称	説明	例
表層	原形	対象の原形	BASE=it, BASE=they
構文解析結果	係り受け	対象を含む係り受け	(表 5.5 参照)
周辺情報	N-grams	対象の周囲の N-gram	3gram_3=candidate_for_future, 2gram_3=candidate_for
	BoW	対象を含む文を構成するすべての単語	BoW=candidate, BoW=for, BoW=future

まず、固有表現抽出で有効性の知られている特徴量 [15] (表 5.3) から、代名詞にも適用できるものを用いることを考える。しかし、代名詞は表層から得られる情報が少なく、固有表現に対しては有効である手法も、代名詞に対しては意味をなさないものが多い。たとえば代名詞には単語中に記号や数字を含む例はない。そこで提案法では以下に示すような、分類対象の代名詞の周辺情報を利用する。

- Bag-of-Words (BoW)
- 周囲の N-gram
- 係り受け

Bag-of-words とは、対象を含む文を構成する単語の集合のことで、対象の出現する文脈でどのような単語が用いられているのかを表す周辺情報である。周辺の N-gram は、位置情報や単語の連続を考慮するという点で BoW よりも詳細な周辺情報と言える。係り受けの情報は、構文解析器の出力から得られる。構文解析器 Enju [25] を用いた構文解析結果から、対象の代名詞に係る単語や句を用いる。表 5.5 に特徴量として用いる係り受けを示す。係り受け特徴量には係る単語の原形に加えて、品詞と係り受けの種類も用いる。たとえば、例文

- (14) Considered a candidate for future Toyota president, Toyoda oversees its China business, a critical market for the automaker.

では、分類対象となる代名詞は “its” であり、係り受けの特徴量は

[link]-1->noun(business)

となる。それぞれ、[link]: はこれが係り受け特徴量であることを表し、-1-> は分類対象の代名詞 “its” を始点とする係り受けであることを、noun(business) が係り先の品詞と原形を示している。

5.2.2 係り受け特徴量の特徴量選択

係り受け特徴量には種類があるため、どの特徴量を用いるかを定める特徴量選択を行う。本研究では表 5.5 に示した 22 種類から、用いる係り受け特徴量を選択した。特徴量選択の手順は以下の通り。

1. 22 種類すべての特徴量を用いた実験を行い、性能を計算する。
2. 次に、各特徴量を減らした実験を行い、その中で最も高い性能とそのときの特徴量の組み合わせを得る。
3. 減らす前の性能と最も高い性能を比較する。
減らす前の性能の方が高い場合、減らす前の組み合わせを最も性能の良いものとして終了する。
減らした後の性能の方が高い場合、2. に戻る。

表 5.5: 係り受け特徴量

	名称	説明
動詞	2V	対象を目的語とする動詞
	1V	対象を主語とする動詞
	2VS	対象を目的語とする動詞とその動詞の主語
	1VO	対象を主語とする動詞とその動詞の目的語
	VV	対象に係る複数の動詞
	V_V	対象に係る動詞とその動詞に係る動詞
	V_VS	対象に係る動詞とその動詞に係る動詞とその動詞の主語
	V_aux	対象に係る動詞と助動詞
	Vpsv	対象に係る受動態の動詞
	V_prep	対象に係る動詞と前置詞
	V_prepO	対象に係る群動詞と目的語
前置詞	prep	対象に係る前置詞
	prepN	対象に係る前置詞とその前置詞のもう一つの係り先の名詞
	prepV	対象に係る前置詞とその前置詞に係る動詞
	prepVS	対象に係る前置詞とその前置詞に係る動詞とその動詞の主語
等位接続詞	coord	対象に係る等位接続詞とその前置詞のもう一つの係り先の名詞
所有	poss	対象に係る所有を表す語とその語のもう一つの係り先の名詞
	app	対象に係る同格を表す語とその語のもう一つの係り先の名詞
補語	C	対象と補語の関係にある語
副詞	Vadv	対象に係る動詞とその動詞に係る副詞
第四文型	SVOO_C	対象を主語・目的語・第二目的語のいずれかとする第四文型
第五文型	SVOC	対象を主語・目的語・補語のいずれかとする第五文型

5.2.3 概念クラス分類の評価実験

5.2.3.1 実験条件・使用データ

概念クラス分類の問題は複数クラス分類である。今回は一対他分類と呼ばれる手法を用いた。各クラスについて、そのクラスに属するならば1、その他のクラスに属するならば0を出力する分類器を学習する。最も高い値を出力した分類器の表すクラスを推定結果とする。また、バイアス項は1とした。

データは、共参照解析と同様に ACE05 [30] コーパスを用いた。訓練セット・テストセットもそれぞれ同じ文書群とした。評価は、訓練セットにおける 10-fold の交差検定により行った。評価指標には精度 (Accuracy) とマクロ F 値を用いた。定義はそれぞれ以下の通り。

$$\text{精度} = \text{クラス分類正解数} / \text{代名詞総数}$$

$$\text{マクロ F 値} = \sum_i \text{クラス } i \text{ の } F \text{ 値} / \text{クラス数}$$

学習に用いる特徴量を変化させ、性能の変化を確認する。まずベースラインとして、対象の代名詞の原形だけを特徴量とする分類器 A の性能を調べる。これは訓練データ中で、対象と原形が同じである代名詞に最も多く割り当てられるクラスを常に採用することに相当する。そして、周辺情報である BoW, 周辺 N-grams, 係り受け の特徴量を追加し性能を確認した。

5.2.3.2 係り受け特徴量選択結果

前述の手法により選択した、概念クラス分類の精度が最も高くなる係り受け特徴量の組み合わせは、

2V, 1V, 2VS, 1VO, prep, prepVS, coord, poss, C, Vadv, SVOO-C, VV, V-V, V-VS,
V_aux, Vpsv, V_prep, VprepO

の 18 種類を用いたものであった。以降、係り受け特徴量にはこの組み合わせを用いる。

5.2.3.3 代名詞分類実験結果

結果を表 5.6 に示す。各指標における最高値を太字で表した。周辺情報を用いる分類器は、pc4 以外はすべて分類器 pc1 よりも高い性能を示した。表層情報に加えて、対象周辺から得る文脈情報を利用することにより分類性能が改善したと考えられる。

5.3 共参照解析への適用手法

教師あり学習ベースの共参照解析システム [28] に代名詞概念クラス分類結果を利用した新しい特徴量を加える。このシステムでは、言及のペアに対して特徴量を与え、共参照関係の有無を推定する。

表 5.6: 代名詞分類実験結果

分類器名	使用特徴量	精度	マクロ F 値
pc1 (ベースライン)	原形	75.8	22.2
pc2	原形, 係り受け	80.1	41.4
pc3	原形, BoW	80.8	42.2
pc4	原形, N-grams	72.3	30.0
pc5	原形, BoW, 係り受け	80.2	41.5
pc6	原形, BoW, N-grams	80.2	45.1
pc7	原形, BoW, N-grams, 係り受け	80.4	45.6

概念クラスの推定結果が一致するかどうかを表す二値の特徴量を、代名詞概念クラス分類結果による置換可能性特徴量として用いる。

固有表現・普通名詞の概念クラスは、従来システム内で行われている推定をそのまま用いる。つまり、固有表現には Stanford Core NLP [7] 内の固有表現抽出・分類器を用いて、普通名詞には WordNet [8] の情報を用いて概念クラスを推定する。

代名詞概念クラス分類器の学習には訓練セットを用いるが、インスタンスとして代名詞のみを使用すると、訓練例が少ないという問題がある。そこで、概念クラスと周辺の文脈情報の関係は普通名詞においても代名詞と同様であると考え、訓練時には普通名詞もインスタンスとして学習した識別器を利用する実験を行った。なお普通名詞とは、コーパス内で付けられている品詞タグが NAM であるものとする。

5.4 共参照解析性能の性能評価

代名詞概念クラス分類による置換可能性特徴量の有効性を調べるため、特徴量を追加した際に起こる性能の変化を確認する。

5.4.1 実験条件

まず、従来システムと代名詞概念クラス分類による置換可能性特徴量を利用した場合の性能を比較する共参照解析実験を行う。ここで置換可能性の推定に用いる代名詞概念クラス分類器は、代名詞概念クラス分類実験において性能の良かった手法 G とする。また、比較対象として、代名詞概念クラス分類器のベースラインである、対象の原形のみを特徴量として用いる手法 A を利用した場合の性能も評価する。

各代名詞概念クラス分類器については、学習時のインスタンスとして (1) 代名詞のみを用いた分類器 (2) 普通名詞・代名詞の両方を用いた分類器の二通りを比較する実験を行う。

本手法で概念クラス推定を行うのは代名詞に対してであるため、置換可能性特徴量を与える対象は少なくとも片方の言及が代名詞である言及ペアのみとした。

データには、第 4 章の共参照解析実験と同様に ACE05 [30] コーパスを用いた。

5.4.2 評価

結果を表 5.8 に示す。太字は各指標における最大値である。従来システム（性能：表 5.7）に比べて、代名詞概念クラス分類による置換可能性特徴量を追加することで全ての指標の性能が僅かながら向上している。

概念クラス分類器は、周辺情報を用いた pc3 と pc7 のいずれもが、表層だけを用いた pc1 に比べて共参照解析においても高い性能を得た。精度の高い pc3 よりもマクロ F 値の高い pc7 の方が共参照解析適用時に有効であった。

訓練データに代名詞と普通名詞の両方を利用した概念クラス分類器は、いずれの特徴量セットでも代名詞のみを利用したときの共参照性能を下回った。普通名詞における文脈情報は、表層情報の少ない代名詞における文脈情報とは異なる点があり、そのために代名詞の共参照解析に対しては有用な情報とならなかった可能性が考えられる。

5.4.3 実験結果の詳細

代名詞概念クラス分類による置換可能性特徴量を追加したことで、代名詞に関する解析結果が変化した例を示す。

新しく発見される共参照の例

表 5.7: 従来システムの共参照解析性能

指標	従来システム		
	適合率	識別率	F 値
B^3	95.0%	58.0%	72.1%
CEAF	25.2%	74.7%	37.2%
MUC	83.1%	46.0%	59.2%
CoNLL	-	-	56.2%

表 5.8: 代名詞概念クラス分類による置換可能性特徴量を利用した共参照解析実験結果

指標	+代名詞のみから学習した概念クラス分類器								
	pc1			pc3			pc7		
	適合率	識別率	F 値	適合率	識別率	F 値	適合率	識別率	F 値
B^3	95.1 %	58.5 %	72.4 %	94.9 %	58.7 %	72.5 %	95.0 %	58.7 %	72.6%
CEAF	25.4 %	75.0 %	37.5%	25.4 %	74.8 %	37.5%	25.4 %	74.9 %	37.5%
MUC	83.7 %	47.1 %	60.3 %	83.2 %	47.6 %	60.5 %	83.5 %	47.4 %	60.5%
CoNLL			56.7 %			56.8 %			56.9%

指標	+代名詞・普通名詞から学習した概念クラス分類器								
	pc1			pc3			pc7		
	適合率	識別率	F 値	適合率	識別率	F 値	適合率	識別率	F 値
B^3	94.9 %	58.3 %	72.3 %	95.2 %	58.2 %	72.2 %	95.0 %	58.4 %	72.4 %
CEAF	25.4 %	75.0 %	37.5%	25.3 %	74.9 %	37.3 %	25.3 %	74.8 %	37.3 %
MUC	83.2 %	47.0 %	60.0 %	83.9 %	46.5 %	59.8 %	83.5 %	47.0 %	60.1 %
CoNLL			56.6%			56.5 %			56.6 %

- (15) Toyoda, who joined **Toyota**, based in Toyota city, central Japan, in 1984, spearheaded **the automaker's** Internet and telematics business - a technology that puts wireless communications into vehicles.
Considered a candidate for future Toyota president, Toyoda oversees **its** China business, a critical market for **the automaker**.
- (16) **Kurdish fighters** in the north chipped in, when a top leader suggested **they** may agree not to seek control of the northern city of Kirkuk.

例文 15 では、代名詞 **its** に対する先行詞 **the automaker** が新しく発見された。概念クラス分類結果は、**its** は [ORG], **the automaker** は [ORG] であり、一致している。この概念クラスの一致により、共参照関係である可能性が高いと推定された。

例文 16 では、代名詞 **they** に対する先行詞 **Kurdish fighters** が新しく発見された。概念クラス分類結果は、**they** は [PER], **Kurdish fighters** は [PER] であり、一致している。

間違いが正された例

- (17) ISLAMABAD, April 25 (**Xinhua**)
“No one would be allowed to live in Pakistan illegally and use its land for achieving **their** nefarious designs,” Information Minister Sheikh Rashid Ahmed said on Friday.

例文 17 に対する従来システムの解析結果では、代名詞 **their** に対する先行詞として **Xinhua** が推定されていたが、これは誤りである。提案法の解析結果では、共参照関係ではないと正しく判断できていた。概念クラス分類結果は、**Xinhua** は [ORG], **their** は [PER] であり、不一致である。

以上の例において、従来システムでは誤り表層の情報の少ない代名詞に関わる共参照関係を文脈情報を考慮することで正しく解析することができた。

第6章 結論

6.1 まとめ

本論文では共参照解析システムの性能向上を目的として、課題点を明らかにするためにエラー解析を行い、そこで発見された文脈的非整合という課題の解決のために先行詞と照応詞の文脈的な置換可能性という観点を考慮するアプローチを提案した。

エラー解析では、従来システムの出力に生じているエラーについて、原因による分類を人手で行った。分類結果から、文脈的非整合がエラーの大きな原因の一つであり、現状の共参照解析における課題であることを明らかにした。また、今回の提案手法では解決の対象としなかったが、個別の対処が有効と考えられる数値の共参照やタイトル内表現といった特殊な例や、共参照解析の手法が対象としている問題の外にあると考えられる構文解析結果の誤りなどの事前処理の精度の低さに由来する例の割合も小さくなかった。

エラー解析時に見られた文脈的非整合性によるエラーの解消を目的として、先行詞と照応詞の文脈的な置換可能性を考慮するアプローチを提案した。文脈的な置換可能性とは、置換しても文章の文脈的な妥当性が保たれる性質のことである。このアプローチは、先行詞と照応詞は同じ実体のことを表しているため文脈的に置換可能であるという仮定に基づいている。

置換可能性の推定には二通りの手法を検討した。識別的言語モデルによる手法では、照応詞を先行詞候補に置換した後のセンテンスの英語文として自然さを推定し、置換可能性とした。言語モデルの手法を利用することで、大量の生テキストデータから文脈情報を捉えることを期待するものである。モデルの学習は、テキストデータ中に存在する文をそのまま用いた正例、共参照解析の解析結果誤りに相当するように作成した擬似負例から行った。正例と擬似負例の識別実験では、識別モデルであることが可能とする豊富な種類の特徴量の利用によって識別能力が向上することを確認した。共参照解析への適用実験では、まずルールベースのシステムへの適用実験を行った。従来システムにおいて複数候補からの先行詞選択時に照応詞との距離の近さのみを基準としていたところを、新たに置換可能性推定結果も考慮するような基準に変更した。実験の結果、置換可能性の適用以前と解析性能がほぼ変化しなかった。距離情報と置換可能性を完全には保持できなかったことが原因と考えられる。教師あり学習ベースのシステムへの適用実験では、新しい特徴量として置換可能性を追加した。全ての先行詞・照応詞の組に適用した場合以上に、代名詞が照応詞である組にのみ適用した場合においてより大きな性能向上が確認された。代名詞は表層の情報が少なく文脈情報が重要となるため、言語モデルによって得られる情報が有効に働いたと考えられる。

代名詞の概念クラス分類による置換可能性の推定では、先行詞と照応詞の概念クラスの一致不一致を置換可能性とした。表層の持つ情報の少ない代名詞においては、参照先の実体の性質が文脈に

よって表されているという仮定に基づいている。概念クラス分類実験では、対象の代名詞の原形という表層情報に加えて、対象周辺から得る文脈情報を利用することにより分類性能が改善することが示された。共参照解析への適用実験では、代名詞の関わる共参照に対しての性能が向上し、文書全体における性能も僅かながら向上することを確認した。

6.2 今後の展望

識別的言語モデルの学習では、擬似負例の作成方法の変更が考えられる。本研究では照応詞を手前に出現した先行詞候補に置換したものを負例に用いたが、この手法では従来の共参照解析システムで既に解析可能である負例が多く作成されてしまい、新しい情報の獲得には繋がらなかった可能性がある。他の擬似負例の作成方法には、たとえば N-gram 言語モデルにより照応詞を生成することが考えられる。

概念クラス分類の手法では、概念クラスの細分化が考えられる。本研究では従来の多くの固有表現分類に用いられているクラスに基づいた分類を行ったが、より細かいクラス分類結果があればより深い情報を得ることができる。また、概念クラス分類は、実体の意味についての分類であることから、共参照解析の部分問題とも捉えられる。より多くのクラスに分類することは、より共参照解析で解決すべき問題のより広い部分を解くことになり、性能向上に繋がることが期待される。

また、識別的言語モデル・代名詞概念クラス分類器は、いずれも共参照解析以外への適用が可能である。言語モデル・固有表現分類結果が活用されている場面や、文脈情報を重要とするようなタスクへの適用が有効であると期待できる。

研究発表報告

査読なし会議論文

1. 佐藤美沙, 三輪 誠, 鶴岡慶雅, 近山 隆, ルールベースの共参照解析システムにおけるエラーの分析, 言語処理学会第 19 回年次大会, 2013 年 3 月.
2. 佐藤美沙, 三輪誠, 鶴岡慶雅, 近山隆, 照応詞と先行詞の置換可能性を手掛かりとする共参照解析, NLP 若手の会, 2013 年 9 月

謝辞

本研究に取り組むにあたり、多くの方々のお世話になりました。

本研究の実施の機会を与えていただき、その遂行にあたってご指導を頂いた近山隆教授に深謝いたします。主に進捗報告や発表練習などの席で有益なコメントと適切なアドバイスを与えていただきました。

指導教員である鶴岡慶雅准教授に深謝いたします。研究の方向性や、議論、研究の動向、英語表現まで、研究に関わる事項すべてにわたって温かいご指導を頂きました。また、毎日朝早くから規則正しく勤務される先生には、職業研究者のあるべき姿を見る思いでした。

近山研究室 OB のマンチェスター大学の三輪誠氏に深く感謝いたします。日頃から研究の進み具合を気に掛けていただき、テーマの検討から、細かな技術的指導、執筆や発表の作法に至るまで多くのご助言を頂きました。本論文がここに存在するのも氏の親身なご助力に依るものです。

博士課程の浦晃氏をはじめとする研究室の方々からも多大な恩恵を受けました。厚く御礼申し上げます。

最後に、先輩同輩後輩諸氏、学外の方々、友人、インターン時にお世話になった関係者様方、ならびに家族へ、感謝の意を示します。

参考文献

- [1] *MUC6 '95: Proceedings of the 6th Conference on Message Understanding*, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.
- [2] Amit Bagga and Breck Baldwin. Algorithms for Scoring Coreference Chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, number 919, pages 563–566, 1998.
- [3] Mohit Bansal and Dan Klein. Coreference Semantics from Web Features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 389–398, 2012.
- [4] David Bean and Ellen Riloff. Unsupervised Learning of Contextual Role Knowledge for Coreference Resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–304, 2004.
- [5] Thorsten Brants and Alex Franz. *The Google Web 1T 5-gram corpus version 1.1.*, 2006.
- [6] Rong-en Fan, Kai-Wei Chang, Cho-Jui Hsieh, Wang-Rui Wang, and Chih-jen Lin. LIBLINEAR : A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9(2008):1871–1874, 2012.
- [7] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, 2005.
- [8] Miller George, A. WordNet: A Lexical Database for English. *Communications of the Association for Computing Machinery*, 38(11):39–41, 1995.
- [9] Aria Haghighi and Dan Klein. Simple Coreference Resolution with Rich Syntactic and Semantic Features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1152–1161, 2009.
- [10] Reinhard Kneser and Hermann Ney. Improved Backing-off for M-gram Language Modeling. In *Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, 1995.

-
- [11] Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic Coreference Resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916, 2013.
- [12] Xiaoqiang Luo. On Coreference Resolution Performance Metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, 2005.
- [13] Recasens Marta, Marquez Lluís, Sapena Emili, Martí M. Antonia, Taule Mariona, Hoste Veronique, Poesio Massimo, and Versley Yannick, editors. *SemEval-2010 Task 1: Coreference Resolution in Multiple Languages*, 2010.
- [14] Kaufmann Morgan. MUC-7. In *Proceedings of the Seventh Message Understanding Conference*, 1998.
- [15] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January 2007.
- [16] Vincent Ng. Supervised Noun Phrase Coreference Research : The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, 2010.
- [17] Vincent Ng and Claire Cardie. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, number July, pages 104–111, 2002.
- [18] Daisuke Okanohara and Jun’ichi Tsujii. A Discriminative Language Model with Pseudo-Negative Samples. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 73–80, 2007.
- [19] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English Gigaword Fourth Edition. LDC2009T13, 2009.
- [20] Adam Pauls and Dan Klein. Faster and Smaller N -Gram Language Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 258–267, 2011.
- [21] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 Shared Task : Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Shared Task*, pages 1–40, 2012.

-
- [22] Altaf Rahman and Vincent Ng. Supervised Models for Coreference Resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 968–977, 2009.
- [23] Altaf Rahman and Vincent Ng. Resolving Complex Cases of Definite Pronouns : The Winograd Schema Challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, 2012.
- [24] Lev Ratinov and Dan Roth. Learning-based Multi-Sieve Co-reference Resolution with Knowledge. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2012.
- [25] Kenji Sagae, Yusuke Miyao, and Jun’ichi Tsujii. HPSG Parsing with Shallow Dependency Constraints. In *Proceedings of the 44th Meeting of the Association for Computational Linguistics*, 2007.
- [26] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544, 2001.
- [27] Veselin Stoyanov and Jason Eisner. Easy-first Coreference Resolution. In *Proceedings of the 24th International Conference on Computational Linguistics*, volume 2, pages 2519–2534, 2012.
- [28] Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. Conundrums in Noun Phrase Coreference Resolution : Making Sense of the State-of-the-Art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664, 2009.
- [29] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the 6th conference on Message Understanding*, page 45, 1995.
- [30] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. ACE 2005 Multilingual Training Corpus. LDC2006T06, 2006.
- [31] 北 研二, 辻井 潤一. 言語と計算 (4) 確率的言語モデル. 東京大学出版会, 1999.