# 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

佐藤 美沙　　三輪 誠　　鶴岡 慶雅　　近山 隆
東京大学 工学系研究科
{m-sato, miwa, tsuruoka, chikayama}@logos.t.u-tokyo.ac.jp

## 1　はじめに

共参照解析の精度は現在のところ70~90%といまだ応用に耐えうる値には至っていない．更なる精度向上のために，まずは解決されるべき問題を知ることが必要である．そこで本稿では，CoNLL2011の共参照Shared Task [3] で最高精度を記録したシステム [2] の出力から，誤りの分類・分析を行った．なお，本稿では英語の共参照解析を対象としている．

## 2　対象とする共参照解析システム

CoNLL2011の共参照Shared Task [3] で最も優れたスコアを提出した共参照解析システムであるStanford Core NLP [2]（以下，対象システムと呼ぶ）の出力のエラーを分析した．

対象システムは，共参照を判断する根拠となる現象をそれぞれルールの形に落とし，その複数のルールを順番に適用していく，複数のルールに基づいたシステムである．ルールの適用は，訓練用データで測ったPrecisionの高い順に行う．

共参照は，複数の句が同一のものを指し示すことであり，このときそれぞれの句を言及 (mention)，指し示されるものを実体 (entity) と呼ぶ．対象システムでは，共参照関係にある言及で構成されるクラスタを求めることになる．初期状態では言及ひとつだけを含む，言及と同じ数のクラスタを生成し，それらのクラスタを徐々に連結していく．各ルールの適用の結果起こる動作は，妥当な共参照関係と考えられる関係が発見された場合，それらの言及の属するクラスタ同士を結びつける，または，絶対に共参照ではないと考えられる関係が発見された場合，それらの言及の属するクラスタ同士の連結を禁止する，である．そのため，早い段階での共参照関係の誤検出すなわちfalse positiveの誤りは，そこで間違ったクラスタが形成されてしまうため，クラスタ単位で判断する対象システムにおいては後の段階に大きな悪影響を及ぼす．そこで対象システムでは複数のルールをPrecisionの高い順に適用している．このPrecisionは事前に各ルールを訓練用データに対して適用することで測ったものである．ルールのリストを表1に示す．

表1: Stanford Core NLPの共参照解析で用いられるルール．番号の若い順に実行される．

| | |
|---|---|
| 1. | Mention Detection |
| 2. | 談話情報の利用 |
| 3. | 文字列一致（厳） |
| 4. | 文字列一致（緩） |
| 5. | 構文合致（同格・copular verb） |
| 6. | 主要部の単語の一致（厳） |
| 7. | 固有名詞の主要部の一致 |
| 8. | 略語 |
| 9. | 主要部の単語の一致（緩） |
| 10. | 語彙情報の利用 |
| 11. | 代名詞関係 |

## 3　誤り分析

### 3.1　対象文書

英語における共参照解析で広く用いられているコーパスであるMUC6 [1] のテストデータ[1]の一部，5ドキュメントを利用した．言及の数は780であった．

### 3.2　誤りの分類

前節の文書を前章のシステムで処理した結果生じたすべての誤り123個を人手で分類した．分類結果を表2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

[1]本当はエラー分析には訓練データを用いるべきである．

表 2: 誤りの分類

| 原因 | ポイント |
|---|---|
| 前処理でのミス | 18.5 |
| コンテキスト依存の類語 | 14.5 |
| 特殊な参照 | 12.5 |
| アノテーションの誤り | 12.0 |
| 原因不明 | 11.5 |
| 修飾語の後出し | 10.8 |
| 外部知識 | 10.3 |
| 談話 | 8.0 |
| 代名詞の概念クラス | 7.0 |
| 他のエンティティとの関係 | 5.8 |
| 判断根拠不明 | 5.0 |
| 関わりの深い語との関係 | 3.0 |
| 合計 | 121.0 |

る場合は，1 エラーにつき値の合計が 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，各分類について例を挙げながら定義を述べる．括弧内の数値はその分類のポイントを表す．例文中の太字は参照元の言及を，下線を引いた文字は正しい参照先を示し，それ以外の着目点をイタリックで表す．

**前処理でのミス (18.5)** 共参照解析を行う以前の，前処理操作での誤りが共参照解析に悪影響を及ぼすことがある．

**・同格構文 (11.0)** 同格構文の取得に失敗すると，同格関係に依る共参照関係は取得できない．

> “Newspapers are something he **Mr. Murdoch** loves.

**・構文解析 (4.5)** 構文解析での誤りによって共参照解析の誤りが生じる．下の例では，With から始まる前置詞句は ***Mr. Akers*** までと解析すべきところで *a lame duck* までであると誤っているために，ex-chairman John Akers との共参照関係の取得に失敗している．

> Its longtime chief financial officer, Frank Metz, took early retirement in the January executive shake-up that also pushed aside ex-chairman John Akers. ... With ***Mr. Akers*** *a lame duck*, vice chairman and acting CFO Paul Rizzo ran the search.

**・言及の切れ目 (3.0)** 以下の例では BURNS FRY Ltd. (Toronto) が言及であるべきところで *firm* までが名詞句であると誤って判断されている．

> *BURNS FRY Ltd. (Toronto)* — *Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm.*

**コンテキスト依存の類語 (14.5)** 常に類語というわけではないが文脈によっては類語であるものが存在し，この知識が欠けているために判断できない共参照がある．以下の例では，Makerと IBMと **giant** と **company** がいずれも会社組織を指しうる類語であるという知識が必要である．

> Computer Maker’s Move Signals Strategy of Cuts. ... His appointment is a strong sign that IBM’s new chairman, Louis V. Gerstner Jr., plans a similar strategy at **the wounded computer giant**. ... While it may look outside **the company**, it has several inside candidates.

**特殊な参照 (12.5)** 共参照関係は，同じ実体を指し示す言及と言及の間の関係である．しかし一部の共参照関係には，指し示す物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ものが存在する．以下で詳細な分類を示す．

**・数字表現 (6.0)** 数字はある量を示す言及であるが，後にその数字を指し示す言及があるときにその量に変化のないことを知っておく必要がある．

> Early this year Mr. York led Chrysler through one of the largest stock sales ever for a U.S. industrial company, raising $1.78 billion. Chrysler is using most of **the proceeds** to reduce its $4.4 billion unfunded pension liability.

**・時間表現 (3.5)** 時間の流れを考慮しなければ判断できない．以下の例では， **two years later** は *1986* 年基準であるため 1988年を指している，という判断が必要である．

> They also said that the Post may not survive long enough for Mr. Murdoch to get

the necessary approval to buy the paper, which he owned from 1976 to 1988. ... After Mr. Murdoch bought the Post, he acquired local television station WNYW-TV in *1986*, and he was forced to sell the paper **two years later**.

**・話題表現 (3.0)** 話題を示す言及とその話題を指し示す言及の共参照がある．

IBM and Mr. York wouldn't discuss his compensation package, which could easily reach into seven figures. **The subject** is sensitive at a time when IBM is running big losses and laying off thousands of employees.

**アノテーションの誤り (12.0)** アノテーションが誤っていると考えられるものもある．たとえば以下の **car** は他の car と共参照とされているが，これは特定の車を指しているのではなく単なる物の種類としての車であると考えられるため，共参照とするべきではない．

In response, Chrysler cut in half the time and money invested in new **car** lines ...

**原因不明 (11.5)** 誤りの原因が不明であったもの．

**修飾語の後出し (10.8)** 共参照関係にある言及では，後から出現する言及には手前で出た言及に付いていない修飾語は付いていないことが多い．対象システムではそのルールを完全に適用しているので，後から修飾語の増える以下の様な例では失敗している．

Carl Sagan, the noted astronomer, is a funny man. ... This traveled at light speed to **the 59-year-old astronomer**. **He** sued Apple last week in U.S. District Court in Los Angeles.

**外部知識 (10.3)** テキスト内に含まれていない情報が必要となるもの．

**・Copular verb(4.0)** 対象システムでは扱っていない copular verb により共参照が示される．

Mr. Wright*resigned as* **president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co.**, to succeed Mark Kassirer, 48, who left Burns Fry last month.

**・テキスト外情報 (3.3)** テキストだけでない情報が必要である．各ドキュメントにはそれぞれタイトルが付属されており，タイトル内の言及にも共参照が付いている．対象システムではタイトルのテキストも本文と同様に処理しているが，タイトルは本文よりも手前に出てくるにもかかわらず本文内の言及よりも修飾語が少ない，タイトル故に冠詞が抜けている，といった特徴があるため，タイトルの処理にはそれ専用の対処が求められる．

IBM Appoints Chrysler's York As Finance Chief
International Business Machines Corp. continued its executive makeover by hiring **Jerome B. York, an architect of the turnaround at Chrysler Corp.**, to become chief financial officer.

**・固有名詞の別表記 (3.0)** 固有名詞の別表記の知識が必要となる場合．以下の例では，Apple Computer Inc. が **Apple** と表記されうるという情報が必要である．略語（例： International Business Machines Corp. – IBM）もこの項に含んでいる．

Nerds over at Apple Computer Inc. were putting together some new computer models. Customarily, **Apple** gives these machines secret code names during development, and real names when they are ready to be sold to the public.

**談話 (8.0)** 談話情報の取得に失敗している．

He added, "Newspapers are something he Mr. Murdoch loves. **I** don't think he has ever lost interest in them."

**代名詞の概念クラス (7.0)** 代名詞の概念クラスの情報が共参照に有用である．以下の例において，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．

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.

次の例では，**We** のクラスが会社（組織）である．

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."

**他のエンティティとの関係 (5.8)**　エンティティとエンティティの関係から推測される共参照関係がある．下の例では，それぞれ Court，**court** に行ったと記述されている *He*，*he* は同一人物であることから Court，**court** も同じ場所であることが推測される．

Cosmic Collision: Apple and Carl Sagan Are Butting Heads — *He* Rockets Into Court, Irate Over a Code Name ... For example: He does not consider himself a Butt-Head Astronomer, and those who have intimated that *he* is have been hauled into **court**.

**判断根拠不明 (5.0)**　共参照関係であると人間が判断している根拠を著者が明確に認識できなかったもの．以下の例では，**he** が *A native of Memphis, Tenn.* を指すことは native が人を表すことがわかれば判断できるが，手前の Mr. Yorkと同一人物であることの根拠が不明であった．

After graduating from West Point, Mr. York got a masters in structural engineering from the Massachusetts Institute of Technology and an M.B.A. at the University of Michigan. *A native of Memphis, Tenn.,* **he** *said* **he** *spends* **his** *spare time working as a "gentleman farmer" and hunting deer, elk, and antelope.*

**関わりの深い語との関係 (3.0)**　実際に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強い語が存在し，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が存在する．以下の例においては，**the Sagan code name** が the name と同一であるというのは，*Dr. Sagan as a code-ame honoree* から *Dr.Sagan* がコードネームの由来であることを利用する，すなわち **the name** と *Dr. Sagan* と *code-name* をすべて共参照関係のクラスタの中に含めて考えると良いと考えられる．

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."

# 4　おわりに

本稿では，共参照解析における誤りの分析を行った．分析の結果得られた，共参照解析の精度向上のために処理が必要と考えられる誤りは，コンテキスト依存の類語，修飾語の後出し，代名詞の概念クラス であった．

今後はこういった誤りの原因の解決に取り組みたい．

# 参考文献

[1] Nancy Chinchor and Beth Sundheim. Message Understanding Conference (MUC) 6. In *Linguistic Data Consortium*, 2003.

[2] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the CoNLL-2011 Shared Task*, pages 28–34, 2011.

[3] Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the CoNLL-2011*, 2011.