# 科学技術文献の共参照解析コーパスの整備

岩元文　　能地宏　　進藤裕之　　松本裕治
奈良先端科学技術大学院大学 情報科学研究科
{aya.iwamoto.hr0, noji, shindo, matsu}@is.naist.jp

## 1 はじめに

共参照解析とは，ある文書が与えられたとき，文書内の名詞句で現実の実体を指すもの（メンション）のうち，同じ実体を指すものをクラスタリングすることである．科学技術文献における共参照解析は，同分野における知識抽出等のより応用的な自然言語処理タスクにおいて不可欠である．一方で，CoNLL-2012 shared task [12] など一般分野にアノテーションされたデータは多いものの，科学技術分野に特化した大規模データは今までそれほど多くなかった．今回，科学技術分野に対応した共参照解析器を作るため，Schäfer ら [13] によって作成された ACL anthology に対して共参照情報がアノテーションされたコーパスを用いベースラインとなる解析器を作った．しかし，このデータは MMAX2 という一般的でないデータフォーマットであり，後述するように複数のファイルに情報が分散しており，アノテーションされた情報も少ないため，一般的に利用可能な解析ツールを用いてより一般的で数多くの共参照解析器が作られている CoNLL 2012 shared task と同じフォーマットへ変換した．またその際に付与した自動アノテーションの問題を回避するため，アノテーション単位を変更したデータを作成した．

文献の解析にあたっては，一般的な分野にはない様々な問題がある．まず，多くの統語解析器が一般の分野をもとに作られているため，科学技術の分野では必然的に誤りが多くなることが挙げられる．そのため，共参照解析の前段階の処理となるメンションの候補である名詞句の抽出の精度が低い．

しかし，今回用いるデータは修飾詞や関係代名詞節をすべて含んだ最大の名詞句がメンションとしてアノテーションされているため，初めにルールによってメンションを抽出した時点で正解とされているメンションの範囲に合致しないものが多く，これが後述する MUC, $B^3$, $CEAF_e$ とその平均からなる CoNLL-2012 shared task で用いられた最終的なシステムの評価（CoNLL scorer による出力）にも影響を与える．一方で，その結果を関係抽出など他のタスクで利用する場合は最大の名詞句の範囲は必ずしも必要なく，多くの場合依存構造解析した際の Head word を含む最小の名詞句でよいと考えられる．今回，共参照解析システムの本質的な性能を測るため，言及の抽出の時点でミスが増える最大範囲の名詞句ではなく Head word を含む最小範囲の名詞句（以後 Head NP と記す）へと変換したコーパスの作成・ベースラインシステムでの評価を行った．

## 2 共参照解析について

### 2.1 タスク設定

共参照解析とは，ある文書が入力されたときに，同じ実体を表す語句を同じクラスタにまとめることである．入力は文書であり，出力はクラスタ情報である．この際，対象となるのは名詞句のみには限らない．CoNLL 2012 shared task [12] では，動詞も対象となっている．評価は主に MUC [14]， $B^3$ [1]，$CEAF_e$ [8] の三つとそれらの平均である CoNLL score を指標とするのが近年では主流である．

今回用いるデータでは，科学技術文献が対象となるため，文献内の代名詞や専門用語などの名詞句に加えて，文献の引用も対象とし，それぞれの名詞句のうち同じ実体を表すものをクラスタとしてまとめている．引用に関しては，同じ研究者らの引用を同じクラスタとするだけでなく，彼らから提案された手法・成果なども同じクラスタとしてアノテーションされている．

共参照解析ではある文書が与えられたとき，まずそこから解析の対象となる名詞句などを抽出する．これらが、メンションの候補となる．実際には，必ずしもすべての抽出された名詞が何らかの実体を指しているわけではない．既存の研究ではこの処理をルールで行うものが多い（[5] など）が，ルールで得られるメンションの中にはどのメンションとも同じ実体を指していないもの（singleton mention）ものも多い．この後，集められたメンション（candidate mentions）から同

じ実体を指すと判断されるものを一つのクラスタへまとめていく．

### 2.2 科学技術分野における共参照解析

Bell ら [2] による生物医学分野への共参照解析の応用がある．これは生命科学の反応の流れを把握するための目的で，Lee ら [7] の sieve-based な共参照解析システムをもとにしたものである．また，BioNLP shared task 2011 [11] など，生物化学分野のデータも作成されている．他に，Chaimongkol ら [3] によるデータもある，これは ACM Digital Library から得たいくつかのジャンルの論文の概要に対してアノテーションしたものである．

## 3 データの概要と変換

現在主流となっている教師あり学習で科学技術文献の共参照解析を作るためには，同じ分野で作成された教師データが必要であり，今回 Schäfer ら [13] のものを用いた．この章では，データの概要とその問題点，それを解決するために行ったデータ変換について説明する．

### 3.1 Schäfer らのデータの概要

Schäfer らは，ACL anthology から，2008 年の ACL，2007 年の EMNLP-CONLL，2002 年 COLING の論文を用いた．これらの論文は PDF 形式であり，これを商用 OCR プログラムを用いてテキストに変換されたものに，共参照関係をアノテーションしたものである．データをアノテーションしたのは英語と英文学を専門とする学生である．対象となるのは名詞句や所有格，代名詞や固有表現である．メンションの範囲として，最大の名詞句がアノテーションされている．そのため，すべての修飾詞，関係代名詞節などを含んだ非常に長いメンションが出現する．

### 3.2 共参照解析に用いる際の問題点

#### 3.2.1 データフォーマットの違い

近年の共参照解析システムは，CoNLL-2012 shard task に基づく研究が盛んである．そのため，それに対応した共参照解析器も多数作られている．一方，今回用いるデータは MMAX2 [10] というアノテーションツールに基づいて作られており，テキストとトークナイズ情報，文分割情報と共参照情報がそれぞれ別々のファイルに存在している．既存の手法による実験・比較が行えるようにするためには，これを CoNLL-2012 shard task と同じように全ての情報が一つのファイルにまとまって，フォーマットを変換する必要があった．また，それにあたっては後から品詞や句構造解析等の情報を与えなければならず，そのために一般分野で用いる解析器を用いたが，後述するような新たな問題点も発生した．

#### 3.2.2 論文データの解析誤り

タスク設定でも述べた通り，共参照解析では最初にまずメンションの候補となる名詞句を取り出す必要がある．そのため，この段階で抽出できなかったメンションが多ければ多いほど最終的な CoNLL スコアは低下する．まずこのデータを Stanford CoreNLP [9] を用いて句構造解析したところ，予想よりはるかに誤りが多いことが分かった．誤りが増える原因としては二つ考えられる．一つ目は，CoreNLP が一般的なニュースなどの分野に対して最適化されているため，通常ニュースなどでは出現しないもの（変数や引用など）の解析が難しくなるという点である．二つ目に，文献データの方が CoNLL-2012 shared task で用いられた OntoNotes corpus のような一般的な分野とに比べ一つ一つの文が長くなりがちであることが考えられる．表 1 に平均的な一文あたりの単語数と候補のメンションでどれだけアノテーションされたメンションが被覆できたかを示す．これを見てもわかるように文が長いデータでは被覆率が低くなり，これは句構造解析のエラーの増加によるものと考えられる．

応用に際しては修飾詞を含んだ名詞句の範囲は重要である一方で，最大の範囲は必ずしも必要ではない．これらの問題を解決するため，名詞句全体から Dependency head word を含む最小の名詞句（以後 Head NP と記述する）へとアノテーション範囲を変換する作業を行った．

### 3.3 データフォーマットの共通化

先述の通り，Schäfer ら [13] のデータは複数のファイルに分散している．これを同一ファイルにまとめ，CoNLL-2012 shared task と同じフォーマットへと変換し，比較実験を行えるようにした．このために StanfordCoreNLP を用いて付与した情報は品詞，句構造，固有表現の三つである．これを CoNLL-2012 shared task のフォーマット [1] に従い，解析した結果を付与した．この際，データ中で文境界の情報はあるものの，いくつかの文で正しい文境界とならないものが存在したため，文分割も同時に行った．引用に関しては，引用前後の丸括弧を含むアノテーションと含まないアノテーションが混在したため，後者に統合した．

[1] http://conll.cemantix.org/2012/data.html

表 1: 一文あたりの単語数（平均）と句構造解析で得られた候補となるメンションでどれだけのアノテーションされたメンションが被覆できたかの比率（被覆率）

| | ACL anthology | | CoNLL 2012 shared task data | |
|---|---|---|---|---|
| | train | dev | train | dev |
| 平均文書長 | 25.60 | 26.36 | 17.28 | 16.98 |
| 被覆率 (%) | 78.30 | 79.21 | 93.26 | 92.07 |

### 3.4 Head NP への変換

実際の変更ルールと，どのようなアノテーションがその対象になるのかを説明する．加えて，ルールによる抽出が前後でどう変化したかも記述する．

#### 3.4.1 変更の対象と手順

対象となるのは全ての名詞句，代名詞などと，引用である．名詞句に関しては，先述の通りに Head NP へと変換を行ったが，名詞句の種類（一般のもの，並列句）によってルールを分けた．まず，Head word を抽出するために，アノテーションされたメンションをすべて取り出して Stanford CoreNLP を用いて句構造解析しようとしたところ，すべて文として，つまり (S (…)) として解析されてしまうことがわかった．メンションをすべて名詞句として解析するため，Stanford Parser [6] を Penn Treebank の Wall Street Journal 部分の句構造木から名詞句の部分だけを取り出して名詞句のみからなるコーパスを作り，これ使って学習し，その解析器を使って句構造解析を行った．その後，StanfordDependency [4] に解析された名詞句を与えて依存構造解析を行って Head word を確定し，それを含む Head NP を取り出した．

解析前後の一例を示す．メンションが ‘[the determination algorithm descrived in ([Mohri, 1997])]’ のような場合，全体の Head word は ‘algorithm’ であり，それを含む最小の NP は ‘the determination algorithm’ である．このため，メンションの範囲を変更し，‘the determination algorithm’ を新しいアノテーションとした．

例外として，メンションが並列句 ‘[[violate match], exact match and inside match]’ のような場合アノテーションの変更は行わなかった．

このような場合に Head NP へ変換してしまうと，メンション ‘violate match, exact match and inside match’ の Head NP は並列句の最初の要素 ‘violate match’ と一致し，別の実体を指すメンションと同じ範囲となってしまうので，この場合は範囲を変更しない．

表 2: データの変換の前後での被覆率の変化

| | Maximal NP | | Head NP | |
|---|---|---|---|---|
| 全文書数 | 266 | | 262 | |
| | train | dev | train | dev |
| 被覆率 (%) | 78.30 | 79.21 | 85.37 | 85.89 |

表 3: 実験結果

| | MUC | $B^3$ | $CEAF_e$ | CoNLL F1 |
|---|---|---|---|---|
| Maximal NP | 60.19 | 51.48 | 37.30 | 49.66 |
| Head NP | 62.98 | 54.29 | 40.18 | 52.48 |

### 3.5 変換前後のデータの変化

データ変換にあたって，候補として抽出されたでどれだけのアノテーションされたメンションが被覆できたかの変化を表 2 で示す．Head NP に変換後 4 文書が欠落しているのは，変換過程で見つかった同じメンションが二つの共参照クラスタに属するようなアノテーションがされていたものを除外したためである．被覆率は全ての名詞句，固有表現，代名詞，引用を集めた後に Head word が重複するものをそれぞれ大きい範囲，小さい範囲で統合してメンションを抽出した場合の数値である．今回各データで 9:1:1 になるように学習・開発・テストデータに自動的に分割したため，元データと Head NP 変換後では学習/開発/テストの内容に若干の違いがある．

## 4 既存手法による共参照解析実験

変換されたデータを用い，Berkeley Coreference System（以後 BCS と表記）[5] を使って実験を行った．このシステムでは，Head word の同じメンションは範囲の大きいものへと統合するルールとなっていることに加えて，引用はメンションの対象とならないため，上記のルールを用いて抽出したメンションを外部から与えることで共参照解析を行った．実験結果を表 3 に示す．この表からも分かるように，最大範囲の名詞句を用いることでアノテーションされたメンションの被覆率が低下し，そのことがシステム全体の評価低

下につながっており，不完全な構文解析やそれに基づいた言及抽出とは無縁な Head NP を用いた場合の方がより実際の解析器の性能を表していると考えられる．

## 5 考察

実験の結果，最大名詞句と Head NP でメンションの被覆率が 6%程度異なることが分かった，その差がシステムの性能を 3%弱押し下げていることも分かった．しかし，BCS では最初から最大名詞句を対象とするため，名詞句の修飾詞を恣意的に除いているメンションの抽出方法ではその分情報が欠けており，Head NP を与えた場合若干不利な状況で学習している．今後の方針として，不完全ながらも修飾詞などを素性として含んで学習することが出来れば，精度の改善が期待できる．

## 6 おわりに

今回，科学技術分野における共参照解析のためのコーパスを CoNLL 2012 shared task 形式へ変換し，比較実験を行った．また，変換にあたって句構造解析のエラーが増加する問題に対し，元々アノテーションされていた Maximal NP ではなく Head NP へと変換することによって，エラーによって言及の抽出率が下がっているせいで性能評価の妨げになっていた問題を解決した．

今後，このコーパスを用いてこの分野に特有の問題を考慮した形の共参照解析器に取り組む予定である．

## 謝辞

本研究は JST CREST の助成を受けた．

## 参考文献

[1] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chain. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pp. 563–566, 1998.

[2] Dane Bell, Gustave Hahn-Powell, Marco A. Valenzuela-Escárcega, Gustave Hahn-Powell, and Mihai Surdeanu. An investigation of coreference phenomena in the biomedical domain. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*, 2016.

[3] Panot Chaimongkol, Akiko Aizawa, and Yuka Tateisi. Corpus for coreference resolution on scientific papers. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3187–3190, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1259.

[4] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation(LREC'06)*, pp. 449–454. Association for Computational Linguistics, 2006.

[5] Greg Durrett and Dan Klein. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1971–1982, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[6] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 423–430, Sapporo, Japan, July 2003. Association for Computational Linguistics.

[7] 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*, Vol. 39, No. 4, 2013.

[8] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 25–32, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[9] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60, 2014.

[10] Christoph Mller and Michael Strube. Multi-level annotation of linguistic data with mmax2. In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, 2006.

[11] Ngan Nguyen, Jin-Dong Kim, and Jun'ichi Tsujii. Overview of bionlp 2011 protein coreference shared task. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pp. 74–82, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[12] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pp. 1–40, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[13] Ulrich Schäfer, Christian Spurk, and Jörg Steffen. A fully coreference-annotated corpus of scholarly papers from the ACL anthology. In *Proceedings of COLING 2012: Posters*, pp. 1059–1070, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[14] 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*, pp. 45–52, 1995.