自 然 言 語 処 理 154－7
（２００３．３．６）

対称モデルに基づく共参照関係アノテーションスキーマ

川添愛
九州大学大学院*/ 国立情報学研究所†
*〒812-8581 福岡市東区箱崎 6-10-1
†〒101-8430 東京都千代田区一ツ橋 2-1-2
zoeai@nii.ac.jp

ナイジェル・コリアー
国立情報学研究所
〒101-8430 東京都千代田区一ツ橋 2-1-2
collier@nii.ac.jp

要旨

本論文では、現在開発中である共参照関係（同一指示関係）のアノテーションスキーマの概要を述べる。このスキーマは理論的研究により動機づけられた共参照関係の捉え方（本論では「対称モデル」と呼ぶ）に基づいており、なおかつ 1)様々な知的領域のユーザが一貫性のあるアノテーションを行うことを容易にする 2) テキスト間アノテーションを容易にする 3) アノテーションを直接オントロジーに関係付けることができる 等の利点がある。

An Annotation Scheme for Coreference Based on the “Symmetric View”

Ai Kawazoe
Kyushu University*/ National Institute of Informatics†
*〒812-8581 6-10-1 Hakozaki Higashi-ku
Fukuoka, Japan
†〒101-8430 2-1-2 Hitotsubashi Chiyoda-ku
Tokyo, Japan
zoeai@nii.ac.jp

Nigel Collier
National Institute of Informatics
〒101-8430 2-1-2 Hitotsubashi Chiyoda-ku
Tokyo, Japan
collier@nii.ac.jp

Abstract

This paper provides an overview of the annotation scheme for coreference which is now being developed. We will state that our scheme assumes the theoretically-motivated view of coreference which we call the “symmetric view”, and that the scheme enables to link annotations directly to the ontology and allows annotators from various domains to make consistent annotations.

1. 概説

本論文では、近年の理論言語学の成果に基づく共参照関係アノテーションスキーマについて述べる。共参照関係（同一指示関係; coreference）とは、同じ対象を指示する二つ以上の言語表現の間の関係を指す。コーパスにおいて共参照関係の情報が適切な方法によって記述されていることは、情報抽出 (information extraction) 等のタスクにとって重

要なことである。本論文で提案するアノテーションスキーマは、ユーザが言語学者であるか否かにかかわらず、テキスト内の表現間の共参照関係に対して一貫したアノテーションを行えることを目指すものである。

従来、多くの共参照関係アノテーションスキーマが様々な目的のために提案されている。その中の主なものに MUC-7 (Hirschman & Chinchor 1997) 、 MEDSTRACT anaphoric annotation (Castaño et.al. 2002)、MATE (Poesio 2000, Poesio et. al 1999, Davies et. al 1998) 、UCREL anaphoric annotation (Figelstone 1992)等が挙げられる。しかし、従来のスキーマの中で、一般的なアノテーションスキーマとして応用されているものはほとんどない。これは、現行のスキーマのほとんどが共参照関係を、照応表現(anaphoric expression)のその先行詞（antecedent)に対する依存関係、すなわちドキュメント内（intra-document)の関係として記述しているためであると考えられる。

本論文では、共参照関係は基本的に、指示される「もの」と、それを各々独立的に指示している表現の集合との関係であるとする、理論的に動機づけられた見方を採用する。そしてこの見方に基づき、一般的な使用において有用な共参照関係のアノテーションスキーマを提案する。

2. 従来のアノテーションスキーマ

情報抽出等のタスクに広く応用可能な共参照関係アノテーションスキーマは、次のような基準を満たしていることが望ましい。

(1) i. 言語学的知識の有無にかかわらず、様々な知的領域のユーザが一貫した、適切なアノテーションを簡単に行うことができる
ii. 機械学習の教師コーパスとして使用することができる
iii. 特定のテキストの内部だけでなく、異なるテキスト間の共参照関係を記述することができる。

この基準に従って、現行のアノテーションスキーマをいくつか評価してみたい。ここでは、MUC-7、MEDSTRACT、MATE の三つのスキーマについて見ていく。MUC-7 は用語抽出 (named entity task)等、現在様々な用途に使用されているスキーマである。MEDSTRACT は分子生物学、医学の分野のテキストからの情報抽出のために開発されたスキーマである。MATE は、主に対話コーパスのアノテーションのために開発されている。

これらのスキーマはいずれも、「先行詞」の情報を重視し、共参照関係を表現間の非対称的な関係として記述する。MUC-7 や MEDSTRACT ではテキスト内の照応表現 のアノテーションにおいて、その表現の先行詞の ID が REF 素性や Antecedent 素性の値として記述される。MATE では LINK という関係を用いてテキスト外でのアノテーションを行うが、LINK によって記述される共参照関係に入る表現のうち一つを必ず先行詞として指定するという点で MUC-7 や MEDSTRACT と同じ、非対称的なアノテーションスキーマであると言える。

MUC-7とMATEにおいては、共参照関係を記述するところのIDENT関係は対称的(Symmetrical) かつ推移的(Transitive)な関係であって、非対称的(Asymmetrical)ではないと明言されている。よって、どの表現を先行詞として選ぶかは問題でないと述べている。MUC-7の実例においては、ほとんどの場合共参照関係を持つ表現の中で一番近くにあるものが先行詞として指定されている。他方、MEDSTRACTのアノテーションの実例 においては、指示的な表現間の階層関係と、先行詞の選択に関するいくつかの制約（例えば、代名詞 (pronoun) は固有名 (name)の先行詞になれない等）が設けられているようである[1]。

簡単さ・一貫性

しかし、いずれにしても、この非対称的なアノテーションのせいで、現行のスキーマが一般的な使用に応用しにくくなっていること

[1] MEDSTRACT の anaphora resolution corpus は次の Web サイトで参照可能である。
http://www.medstract.org/gold-standards.html

は否定できない。まず、「ユーザが一貫した、適切なアノテーションを簡単に行うことができるか」という観点から見ると、どの表現がどの表現の先行詞であるかを決めるのはユーザにとって決して簡単なことではないという問題がある。MUC-7 や MEDSTRACT のように、先行詞の選択を完全に自由にしてしまうと、例えば代名詞が固有名の先行詞になるというような、直感に反するアノテーションを許してしまう上、ユーザ間で一貫性が保てなくなる可能性が高くなる。他方、MEDSTRACT のように、指示的な表現間の階層関係と先行詞の選択に関するいくつかの制約を設ければ、直感に反するアノテーションを防げるが、ユーザにとって、このような階層関係や制約の記憶を保持するのは容易なことではない。

機械学習

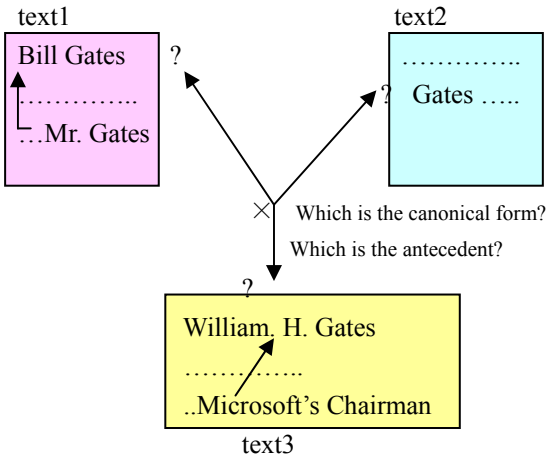
機械学習の観点では、MUC-7 や MATE のように自由に先行詞を選択するスキーマに基づいてアノテーションがなされたコーパスは、そのままでは機械学習の教師コーパスとして使用できない可能性がある。というのは、非対称的な記述の仕方では、同じ関係を記述するのに表面的に何通りもの方法許してしまう。これは、機械が学習対象である言語現象の特徴を適切に捉えることを妨げるからである。

MUC-7 は実際、機械学習を用いた照応解決(anaphora resolution)の研究のいくつかにおいて使用されている (Ng & Cardie 2002 等）が、その中では、まず記述された共参照関係連鎖(coreference chain)から共参照関係にある表現の集合、すなわち同値類（equivalence class）を生成し、それに基づいて更に採用する照応解決アルゴリズムに応じて適切なデータセットを構築するという数段階かの処理がなされている。すなわち、機械学習アプローチにおいて必要とされているのは同値類そのもので、アノテーションの際に記述されている「先行詞」の情報はさほど重要視されていない。もちろん、上に述べた方法でも実質的な問題は起こらないが、可能ならばアノテーションの段階で既に共参照関係にある表現群を生成できるようなスキーマの方が望ましい。

テキスト間アノテーション

更に、情報抽出等のタスクに応用することを考慮すると、アノテーションスキーマが「異なるテキスト間の共参照関係を適切に捉えることができる」ということも望まれるが、非対称モデルを採用するスキーマにとっては実現が難しいことである。非対称的なアノテーションスキーマを用いてテキスト間の共参照関係を捉える際には、ユーザが共参照関係にある表現群の中から最も「慣用的な形式」を決定する必要がある。例えば、“Gates”、“Bill Gates”、“William Gates”、“William H. Gates”のうち、どれが最も適切なラベルであるか判断しなければならない。この点については、ユーザ間での一致が難しいことが予想され、一貫性を欠く可能性がある[2]。

(2)非対称的なテキスト間アノテーション



3. 非対称モデル vs. 対称モデル

では、なぜ従来のスキーマのほとんどにおいて「先行詞」の概念が重視され、非対称的なアノテーションがなされているのだろうか。確かに、照応解決 (anaphora resolution)が、照応表現の先行詞を同定するタスクとしばしば同一視されるという点においては、「先行詞」は重要な概念であるように思われる。しかし、

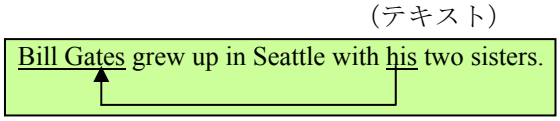
[2] また、MUC-7 に関しては、van Deemter & Kibble (1999)により、IDENT 関係として記述されているものの中に束縛変項照応や関数-値関係など、共参照関係と本質的に異なる関係が存在する等、様々な問題点が指摘されている。

理論的研究においては、必ずしも「先行詞」という概念が共参照関係にとって実体のあるものと見なされている訳ではない。

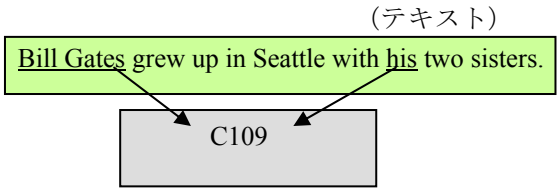
従来、理論言語学においては、共参照関係に対して二つの異なる見方がある。一つは、共参照関係を、二つの表現間の「指示的な依存関係」(の集積) と捉える見方で、本論文では以下これを「非対称モデル」と呼ぶことにする。前節で概説したアノテーションスキーマは、非対称モデル的に共参照関係を記述していると言える。もう一つは、共参照関係を、二つ以上の表現が同じ実世界（または話者の何らかの認知的領域）の対象物をそれぞれ独立に指示する現象とする捉え方である。本論文では以下これを「対称モデル」と呼ぶことにする。対称モデルにおいては、共参照関係は言語表現間の関係ではなく、テキスト内の言語表現とテキスト外の対象物との関係の一つのインスタンスということになる。

以下の(3) (4)にも示されているように、非対称モデルにおいては Bill gates が his に指示的に依存されているのに対し、対称モデルにおいては Bill Gates と his が同じ対象概念 (C109) を指示していることが重要で、Bill Gates が his の指示対象を決定しない。すなわち、後者においては、Bill Gates が his の先行詞かどうかということは重要な問題ではない。

(3) 「非対称モデル」



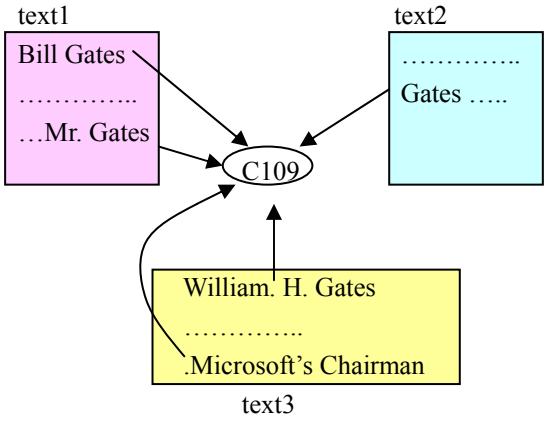
(4) 「対称モデル」



近年の理論的研究（Ueyama 1998, Hoji et.al. 2000 等）においては、束縛変項照応（bound variable anaphora) 等の典型的な依存照応（dependency anaphora）と、共参照関係との間の区別が明らかにされ、その結果として共参照関係を対称モデル的に捉えることの理論的重要性も見直されてきている[3]。

前節で指摘した非対称的なアノテーションスキーマにおける難点を考えると、対称モデルに基づく対称的なアノテーションスキーマを考案し、これに移行することは妥当であるように思われる。対称モデルに基づくスキーマでは、アノテーションの際にユーザが先行詞を指定する必要がない。また、共参照関係にある表現群をアノテーションの段階で直接生成することができる。更に、次の図のように、テキスト間アノテーションの際にも、慣用的な形式を選択したりする必要がない。

(5)
対称的なテキスト間アノテーション(Cf. (2))



ただし、上のようなアノテーションを実現するためには、知識概念体系によるサポートが必要である。次節で提案するアノテーションスキーマは、現在進めている PIA プロジェクト (Portable Information Access Project; Collier et. al 2001, 2002b)において開発中のオントロジー(Ontology) サーバを使用する予定である。対称モデルに基づくアノテーションスキーマを用いて共参照関係を公共のオントロジーに関連づけることにより、機械がテキスト内の

[3] ただし、これらの研究ですべての共参照関係を対称モデル的に捉えているわけではなく、共参照関係の中にも、一部の独自に指示対象を持たない表現（日本語のソ系列の指示詞など）が関わるものは束縛変項照応や E タイプ照応と同じ成立条件に従うことが指摘されている。

要素について理解したり推論したりする際に意味的な一貫性が達成されることが期待できる。

4. スキーマの概要

ここで、本論文で提案する共参照関係アノテーションスキーマの概要を述べる。本論文における基本的なアプローチは、次の二点である。

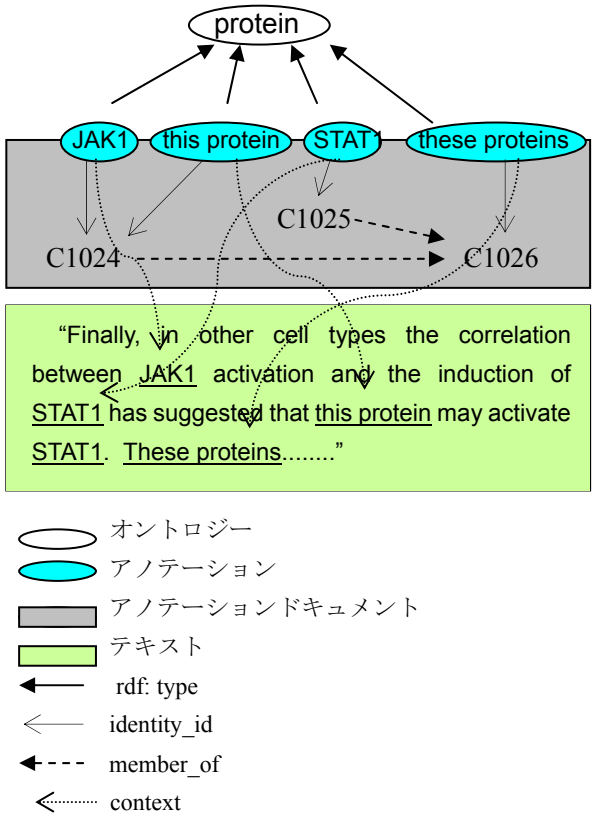
(6) i. 共参照関係にあるすべての表現を、それらがそれぞれ独立に対象概念を指示しているという点で同等と見なす。
ii. アノテーションが元のテキストと独立に現れることを許す。

すなわち、MATE や Collier et. al (2002a)で用いられている「テキスト外アノテーション」を採用し、元のテキストとアノテーションドキュメントを切り離す。アノテーションドキュメントの中では、指示対象となる概念間の関係（例えば集合とそのメンバーの関係など）を記述することができる。

ここで提案するアノテーションスキーマの概要を(7)の図に示す。図中で示されているように、*context* というプロパティが XPointer 値 (De Rose et. al. 2000) を取り、アノテーションを元のテキストと関係づけている。*identity_id* というプロパティによって、各々のアノテーションがアノテーションドキュメント内の概念と関連づけられる。指示対象である概念に対して“C1024”のような ID を与え、それを指示する表現の集合によって指定する。

更に、各々のアノテーションは RDF で定義された関係によって公共のオントロジーに関連づけられる。(7)に示されているように、JAK1 と this protein の間の共参照関係から、これらによって指示されている対象概念の情報を得ることができる。

(7)アノテーションスキーマの概要



PIA プロジェクト(Collier et. al 2001, 2002b)においては、本論文で提案するアノテーションスキーマが、機械学習に有用な教師データの作成に使用される予定である。よって、一貫性のあるアノテーションを実現するため、現段階では音形を持つ名詞句の間の同一指示関係に対してのみアノテーションを行う。すなわち、動詞句、文、ゼロ代名詞の関わる同一指示関係や、束縛変項照応、E タイプ照応、全体－部分の関係等はアノテーションの対象から外している。また、共参照関係にある表現のタイプ(*name, alias, pronoun, definite* and *indefinite* 等）を、*type* というアノテーションクラスを用いて記述する。これは、機械による照応解決 (anaphora resolution)の学習を段階的に行うためである。共参照関係のサブタイプのうち、最も簡単なものから学習を開始し、徐々により難しいものへ移行するという方法をとる。現在、共参照関係のサブタイプの難易度に関する調査を行っており、その結果に即してスキーマを発展させていく予定である。

5.MMAX との比較

これまでの研究の中で、対称的なアノテーションスキーマが全く提案されていないわけではなく、例えば MMAX (Multi-Modal Annotation in XML; Muller & Strube 2001)の照応関係アノテーションスキーマにおいては、一部対称的なアプローチがなされている。このスキーマでは、照応表現に対してアノテーションを行う際に、次の二つの段階を踏む。

(8) i. 照応表現を、それと共参照関係にあるテキスト内のすべての表現と関係づける。（義務的）
ii. 共参照関係にある表現の集合の中から、先行詞を選ぶ。（任意）

アノテーションの最初のステップである(8i)の段階では、対称的なアノテーションがなされている。その次の(8ii)の段階の先行詞の指定も義務的ではない。Muller & Strube (2001)は、先行詞の指定を任意にしたせいで必要な情報が失われるということはないと述べている。

本論文の提案するアノテーションスキーマは、(8ii)のステップを採用しないという点で MMAX と異なる。既に述べたように、共参照関係に限って言えば、先行詞は本質的な概念ではない。また、照応解決を行う際には、どのような照応解決方法を採用するかによって、どう「先行詞」を規定すると最適な結果が得られるかが変わってくることも考えられる。このような理由から、ユーザによるアノテーションの段階で「先行詞」を指定しない方が、一貫した、また後の段階で使いやすいコーパスが得られると考える。

6. 結語

以上、本論文では理論的研究に動機づけられた共参照関係の「対称モデル」に基づくアノテーションスキーマを提案し、その有用性について議論した。

現在、アノテーションのガイドラインと、EMBO Journal の分子生物学の記事に基づくアノテーションセットを構築中である。実際の使用に関わる様々な潜在的な問題については本論では触れなかったが、ガイドラインの中で詳しく述べる予定である。

また、現在開発中の Open Ontology Forge において、ユーザに対するソフトウェア・サポートを提供する予定である。

参考文献

J. Castaño, J. Zhang, J. Pustejovsky (2002). Anaphora Resolution in Biomedical Literature. International Symposium on Reference Resolution, Alicante, Spain. http://medstract.org/papers/coreference.pdf

N. Collier, K. Takeuchi, C. Nobata, J. Fukumoto and N. Ogata. (2002a) Progress on Multi-lingual Named Entity Annotation Guidelines Using RDF(S). In *Third International Conference in Language Resources and Evaluation,* Las Palmas de Gran Canaria, Spain. 29th-31st May 2002. pp. 2074-2081.

N. Collier and K. Takeuchi (2002b) PIA-Core: Semantic Annotation through Example-based Learning. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain, 29th – 31st May, pp. 1611-1614.

N. Collier, K. Takeuchi and K. Tsuji. (2001) The PIA Project: Learning to Semantically Annotate Texts from an Ontology and XML-Instance Data. In *Position paper proceedings of the First Semantic Web Working Symposium (SWWS'2001)*, Stanford University, California, USA, July 30th – August 1st, pp.8-9.

S. Davies, M. Poesio, F. Bruneseaux and L. Romary (1998) Annotating Coreference in Dialogues: Proposal for a Scheme for MATE. First draft. http://www.hcrc.ed.ac.uk/~poesio/MATE/anno_manual.html

S. De Rose, E. Maler, and R. Daniel. (eds). XML Pointer Language (XPointer) Version 1.0,

W3C Candidate Recommendation, 11th September 2001.
http://www.w3.org/TR/sptr, 2000.

S. Fligelstone (1992). Developing a Scheme for Annotating Text to Show Anaphoric Relations. In: G. Leitner (ed.), *New Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter. pp.153-170.

L. Hirschman and N. Chinchor (1997) MUC-7 Coreference Task Definition, Version 3.0. in *Proceedings of MUC-7*.
http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

H. Hoji, S. Kinsui, Y. Takubo and A. Ueyama (2000) Demonstratives, Variables, and Reconstruction Effects. In *Proceedings of the Nanzan GLOW: The Second GLOW Meeting in Asia*, pp. 141-158.

C. Muller and M. Strube (2001) Annotating Anaphoric and Bridging Expressions with MMAX. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue. Aalborg*, Denmark, 1-2 September 2001, pp. 90-95.

V. Ng and C. Cardie (2002) Combining Sample Selection and Error-Driven Pruning for Machine Learning of Coreference Rules. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-02),* pp.55-62, Philadelphia, PA.

M. Poesio (2000) "Coreference" in *MATE Annotation Guidelines*.
http://www.ims.uni-stuttgart.de/projekte/mate/mdag/ /cr/cr_1.html

M. Poesio, F. Bruneseaux, and L. Romary, (1999) The MATE Meta-scheme for Coreference in Dialogues in Multiple Language. In *Proceedings of the ACL Workshop on Standards for Discourse Tagging*. Maryland.

A. Ueyama (1998) *Two Types of Dependency*. Doctoral dissertation, University of Southern California, distributed by GSIL publications, USC, Los Angeles.

K. van Deemter and R. Kibble (1999) What Is Coreference, And What Should Coreference Annotation Be? In *ACL'99 Workshop on Coreference and its applications*, University of Maryland, June 1999.