2C3-1

共参照関係を利用したMarkov Logicによる医学生物学文書中のイベント抽出

Coreference Based Event Extraction with Markov Logic on Biomedical Text

吉川克正[*1] Katsumasa Yoshikawa　平尾努[*2] Tsutomu Hirao　リーデル セバスチャン[*3] Riedel Sebastian　浅原正幸[*1] Masayuki Asahara　松本裕治[*1] Yuji Matsumoto

[*1]奈良先端科学技術大学院大学情報科学研究科
Nara Institute of Science and Technology, Graduate School of Information Science

[*2]NTTコミュニケーション科学基礎研究所
NTT Communication Science Labratory

[*3]マサチューセッツ大学アマースト校 知能情報検索センター
University of Massatusetts Amherst, Center for Intelligent Information Retrieval

This paper presents an approach to extract arguments of events that cross sentence boundaries using coreference relations. In order to use coreference relations effectively, we introduce a non-deterministic model, formulated in Markov Logic. This allows us to improve intra-sentence attachments based on cross-sentence attachments, and vice versa. We show the effectiveness of this approach on biomedical event corpus. Our primary contributions include (1) the presentation of a coreference based approach for cross-sentence attachments and (2) the demonstration that the higher accuracy of coreference resolution leads to the more improvements of event extraction.

1. はじめに

これまでの医学生物学文書における事象-項関係同定を扱った研究は，事象と項が同一の文内に出現する場合だけをその対象としてきた [Björne 09, Buyko 09]．しかし，GENIA Event Corpus [Kim 08] や BioNLP 2009 Shared Task データ [Kim 09] など，近年多く利用されている医学生物学文書のコーパスでは，ある事象に対する項を同定する際，文内だけではなく文外の言及も項候補として扱うべき場合がある．即ち，互いに関係のある事象と項が文境界を越えている場合が存在するのである．図 1 は事象 “inducible”に対する項が，共参照関係を媒介として文境界を越えている事例である．この例のように共参照関係にある項を同定できることは，単純にその解析精度が向上するという以上の意味がある．共参照関係にある項はその文書内での主題性の高さを示しているため，書き手が伝えたいと考えている新たな情報を多く含む傾向にある．逆に，読み手にとってはそのような情報こそ文書理解のために不可欠な最重要なものである．つまり，共参照関係にある項を同定することは文書理解のために価値の高い事象-項関係を積極的に同定することになる．

しかしながら，一般に文境界を越えるような項の同定は，文内に項候補を限定した場合と比較して，項の探索空間が爆発的に広くなり，精度・計算量の両面において困難な問題となる．本研究ではこの問題に対し，次のような二つの戦略で臨んでいる．まず一つ目は共参照関係を利用することによる項候補の限定である．即ち，前処理で共参照解析を行った上で，文外の項を同定する際には，文内の言及と共参照関係にあるものだけを項候補とする．これによって，共参照解析と項同定を同時に行う手法とは対照的に，文外の項であっても文毎に解析することが可能になる．二つ目は，Markov Logic [Richardson 06] を利用した非決定的なモデルの構築である．Markov Logic により，

連絡先: 吉川克正 奈良先端科学技術大学院大学情報科学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
E-mail: katsumasa-y@is.naist.jp

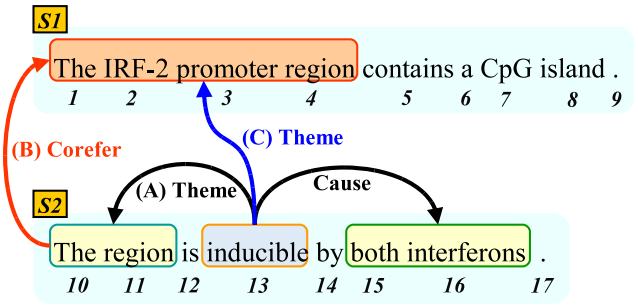

図 1: 文境界を越えた事象-項同定

共参照関係にある項を候補として複数の事象-項関係を同時に扱うことで，文内の項同定と文外の項同定の解析性能を相互に向上させられるようにする．本研究ではこの Markov Logic による非決定的なモデルと，文内の事象-項を同定した上で，その文内の項と共参照関係にある項を（決定的に）全て同じ事象と関係があるものとみなす決定的パイプライン手法との比較を行う．

本研究の提案手法の有効性を示すため，GENIA Event Corpus で実験を行い，決定的パイプラインと比較して，文境界を越える事象-項同定においては 8.9(%)，共参照関係に関連した文内の事象-項同定においては 5.0(%) の F 値の向上を確認した．

2. 関連研究と問題点

近年行われている事象-項関係同定の研究のほとんどは，文内の事象-項関係（文内リンク）のみを同定する対象としており，文境界を越えた事象-項関係（文間リンク）の同定を実現できていない．例えば，BioNLP’09 [*1]において，Riedel らは

[*1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/

Markov Logic を利用した手法を提案している [Riedel 09]．彼らのシステムは大域的な素性を有効に利用して，同じ文内にあるトークン全てを考慮して最適な事象，項，意味役割を同時に推定できた．しかしながら，Markov Logic を用いた手法は大域的な素性を柔軟に利用できる代わりに，計算コストが高く，文内から文書全体への拡張は困難であったため，文間リンクを捉えるには至っていない．

Björne らは文間リンクの推定に共参照関係が有効であること示唆しているが [Björne 09]，共参照解析器の学習に十分なデータが BioNLP'09 Shared Task では与えられなかったことから，共参照関係を利用して文間リンクを捉えるシステムは実現できなかったと述べている．さらに彼らは，共参照関係を利用せず，直接に文間リンクを推定する手法を試みたが，十分な性能を達成できなかったと補足している．

従って，共参照関係を利用した手法を実現する上で，本研究では事象-項関係だけでなく，十分な数の共参照関係がアノテーションされた GENIA Event Corpus (GEC) [Kim 08] を利用することとする．これによって共参照解析器の学習も可能になり，文間リンクの同定に必要な共参照関係を十分に得ることができる．

3. 共参照利用の Markov Logic モデル

文間リンクの同定を行うため，本研究で提案する手法は以下の二点を主軸とする．

1. 共参照関係を利用して項候補を限定する
2. Markov Logic を利用して文内リンクと文間リンクの同定を同時に行う非決定的なモデルを構築する

項候補を限定することにより，コンパクトかつ効率的なモデルを構築することができ，計算コストを大幅に削減できる．また非決定的なモデルを構築することで，再帰的に文内・文間のリンクを同定することが可能になり，全体の解析性能が期待できる．

ここで，本研究で定義した Markov Logic Network(MLN) について説明する．まず，本研究で推定したい対象に基づき，次の表 1 のような三つの推定述語を定義する．

表 1: 推定述語（hidden predicate）

event(i)	トークン i は事象である
eventType(i, t)	トークン i タイプ t の事象である
role(i, j, r)	トークン i はトークン j を項に持ち，その意味役割は r である

本研究の手法は Riedel らの手法 [Riedel 09] を元にして，表 1 で表現される事象，項，意味役割の三つを同時に推定するモデルとなっている．この三つに加えて，共参照関係を表現する述語 corefer(i, j) を定義する．この述語は，トークン i がトークン j と共参照関係にある（二つのトークンが同じエンティティのクラスタに属する）ことを表現している．

本研究で扱う事象や項の表現は，単一のトークンで構成されるとは限らず，複数のトークンから構成された句である場合もある．従って，ここで句を捉えるために，"範囲"を考える必要があることを補足しておく．本研究では複数のトークンからなる句を係り受け木の部分木と見なすことにより，この範囲の問題を解決している．即ち，その部分木の根にあたるトークンにのみ着目して，そのトークンを "アンカー"とするのである．例えば，図 2 に表した句，"The IRF-2 promoter region"の

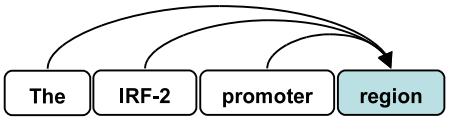

図 2: 係り受け部分木に対するアンカー

場合，係り受け部分木の根になっているのは "region"なので，この "region"をアンカーとする．このようにアンカーを利用することで，複数のトークンから構成される表現を捉える時に必要とされるであろう多くの組み合わせを省略して扱うことができる．

3.1 共参照関係を利用する論理式

前述の通り，図 1 は文間リンクの事例を示している．文 *S2* では，"inducible"が事象を表す [*2]．文 *S2* 内の項候補のみに着目して事象-項関係を同定した場合，"inducible"の項となるのは，"The region"と "both interferons"であり，その意味役割はそれぞれ Theme と Cause である．しかしながらこの事例における "The region"は真の Theme ではない．なぜなら，"The region"はその前の文 *S1* 内にある "The IRF-2 promoter region" と共参照関係にあるからである．このように，この事例における "inducible"の Theme は "The IRF-2 promoter region"であるべきで，事実この句の方が "The region"よりも多くの情報を含んでいる．一方，"The region"は真の項を指示するだけの照応詞にすぎないと言える．

では，この図 1 の事例における事象-項及び共参照関係を，先に定義した MLN の述語を利用して表現することを考える．まず，文 *S2* の文内リンクは role(13, 11, Theme) 及び role(13, 15, Cause) のように表現することができる [*3]．次に，共参照関係を corefer(11, 4) で表現する．最後に，この事例に存在する文間リンクは role(13, 4, Theme) で表現することができる．

図 1 の例において，共参照関係を利用するために本研究で提案する三つの論理式について説明する．まず一つ目は "素性複写"を行う論理式である．一般に "The region"のような照応詞となる項は基本的な情報が不足しているため，その同定も困難である．そこで本研究では照応詞の項に対し，それと照応する先行詞の素性を複写することにより不足している基本素性を補うものとする．図 1 の例に従えば，

$$\text{wordChild}(4, \text{"IRF-2"}) \wedge \text{corefer}(11, 4) \Rightarrow \text{role}(13, 11, \text{Theme})$$

は，照応詞 "The region"に対して，先行詞側から単語の表層形に関する素性 "IRF-2"を複写することを表現している [*4]．一般式は次の形式となる．

$$F(k, f) \wedge \text{corefer}(j, k) \Rightarrow \text{role}(i, j, r) \quad (1)$$

ここで F はトークン k に対する係り受け部分木の持つ基本素性（表層形，品詞，固有表現タグ）を表す述語である．この式 (1) によって，文境界を越えた項（先行詞）の素性を文内の項（照応詞）に複写することができ，文内リンクの同定に先行詞側の素性も併用できるようになる．

---

*2 ここでは直接関係しないが，この事象のタイプは *Positive_regulation* である

*3 複数トークンからなる項は，アンカーとなるトークンによって表現する

*4 wordChild(i, w) はトークン i に対する（係り受け部分木上の）子が，単語の表層形 w を持つことを示している

二つ目の論理式は文間リンクを同定するための推移律である．同じく図 1 の例で考えると，次のような論理式を書くことができる．

$$\text{role}(13, 11, \text{Theme}) \wedge \text{corefer}(11, 4) \Rightarrow \text{role}(13, 4, \text{Theme})$$

この論理式が示すのは，「事象 “inducible”が “The region”を Theme に持ち且つ “The region”が “The IRF-2 promoter region” と共参照関係にあるならば，“The IRF-2 promoter region”も “inducible”の Theme になる」という推移関係である．この論理式を一般化すると次のようになる．

$$\text{role}(i, j, r) \wedge \text{corefer}(j, k) \Rightarrow \text{role}(i, k, r) \quad (2)$$

本研究で提案する MLN はこの式 (2) を組み込んでいる．

この論理式 (2) を利用することによる利点は，文内リンクと共参照関係を同定するだけで，文間リンクも同定することができる点である．即ち文間リンクの項候補は文内の単語と共参照関係にあるものに限定され，結果として項候補の探索空間は殆ど拡がらない．

しかし逆に言えば，式 (2) による性能向上は，文内リンク $\text{role}(i, j, r)$ と $\text{corefer}(j, k)$ の性能に依存している．そこで，共参照関係に関わる文内リンクの性能を向上させるため，次の式 (3) を追加する．

$$\text{corefer}(j, k) \Rightarrow \exists r.\text{role}(i, j, r) \quad (3)$$

式 (3) が表すのは，トークン j が別のトークン k と共参照関係にあるならば, トークン j を項とする事象-項関係が少なくとも一つ存在するということである. この式 (3) が成立する理由は二つある．一つは本研究で構築する共参照解析器が常に事象と関係した項候補を抽出するからである [*5]．もう一つは文書中で繰り返し出現する表現は談話構造上の主題性が高いことを示しているため，何らかの事象の項となり易いと考えられるからである．本研究では分野を問わず，このような主題性の高い表現こそ文書理解のために価値が高く，項として積極的に同定すべき対象であると考えている．

3.2 医学生物学文書における共参照解析

本研究で構築する共参照解析器はペアワイズ共参照モデルに基づいている [Soon 01]．即ち単純にあらゆる名詞句対に対して，その対が共参照関係にあるか否かを二値分類するモデルである．本研究では WordNet などの外部リソースを利用せず，表層形や品詞などの基本的な素性のみで解析を行う．単純な手法にも関わらず，GEC において五分割交差検定で評価した結果，F 値で 59.1 となっている．

4. 実験と評価

4.1 実験設定

まず本研究で利用したデータとツールについて説明する．学習及び評価に利用したデータは GENIA Event Corpus (GEC) [Kim 08] である. GEC は事象-項関係及びそれに関連した共参照関係に対して詳細な情報が付与されている．GEC の事象-項アノテーションは BioNLP'09 のデータよりも詳細であり，多くの共参照関係が網羅されている．文間リンクが事象-項関係全体に占める割合は 4.6% と決して多くはないが，その約 80% は共参照関係を利用して文内リンクと関連づけられており，一貫性の高いデータとなっている．このように，GEC は事象-項関係，共参照関係の両方に関して，本研究の手法を実験・評価するに適した条件を備えている．

素性を生成するため次のようなツールを本研究では利用している．品詞及び固有表現タギングのために GENIA Tagger [*6]，構文解析のため Charniak-Johnson reranking parser [*7]を利用した上で，pennconverter [*8]によって依存構造木へと変換した．共参照解析器の学習と解析には SVM-light [*9]を利用．事象-項関係の学習と解析には Markov thebeast [*10]を利用している．これは自然言語処理のために調整された Markov Logic エンジンである．

本研究の実験は以下のようなステップで行う．まず，全ての文書に対して共参照解析を行い，共参照関係にあるトークン対のテーブルを生成する．次にそのテーブルを用いて共参照関係を利用した MLN モデルを学習した上で，そのモデルにより，評価データの事象-項関係について，最適な解を推定する．

4.2 実験結果と考察

表 2 には本研究の提案手法を評価するために用意した六つのシステムの実験結果を示している．一行目の (a) は共参照関係を全く利用せずに文内リンクのみを同定したスコアである．以降，残り五つのシステム (b)-(f) は全て式 (1) を利用して共参照関係にある項候補の基本素性を複写している．(b) は文間リンクを決定的パイプラインモデルによって同定したシステムのスコアである．この決定的なモデルは，文内リンクを (a) と同じ MLN のモデルによって同定した上で，後処理によって，文内リンクの項と共参照関係にある項を全て決定的に文内リンクと同じ事象の項として同定する．尚，この時の共参照関係にある項とは，文内・文外の両方が考えられるが，その場合も同じ事象の項として同定する．本研究ではこの決定的システムをベースラインと考える．

三行目 (c) と四行目 (d) には，それぞれ式 (2) と式 (3) を利用した MLN の非決定的システムのスコアを示している．また，3.1 節で提案した論理式を全て含める本研究のフルシステムのスコアを示したのが五行目の (e) である．最終行の (f) は (e) のフルシステムにおいて，共参照関係のゴールドアノテーションを利用した時のスコアであり，このスコアが本研究で提案した共参照関係を利用する手法による性能の上限であると言うことができる．

各システムは全て 5 分割交差検定によって評価した．表 2 では次のような事象-項関係について評価している．文間リンク（Cross-sentence），共参照関係と関連した文内リンク（Corefer intra-sentence） [*11]，事象-項関係全体（All links）の 3 種類である．尚，この 3 種類の事象-項関係について精度（P），再現率（R），そして F 値（F）を示している．

では，表 2 における結果について考察する．まず F 値で見ると，(b)-(f) の共参照関係を利用したシステムは全て共参照関係を利用しないシステム (a) のスコアを上回っていることが分かる．次に非決定的システム (c)(d)(e) と決定的システム (b) とを比較することにより，非決定的システムの有効性を示す．(c) のスコアから式 (2) が確実に “Cross-sentence”の性能を向上させていることを確認できる．一方，(d) のスコアからは式 (3) が “Corefer intra-sentence”の性能改善に貢献していることが分かる．フルシステム (e) はベースラインである決定的システム (b) と比較して，文間リンクにおいて 8.9(%)，共参照に関係した

*5 GEC にある共参照アノテーションは全て事象と直接関係したものに限定されている

*6 http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

*7 http://www.cs.brown.edu/~dmcc/biomedical.html

*8 http://nlp.cs.lth.se/software/treebank_converter/

*9 http://svmlight.joachims.org/

*10 http://code.google.com/p/thebeast/

*11 文内リンクのうち共参照関係と関連しているのは 5.8(%) である

表 2: 実験結果

	Cross-sentence			Corefer intra-sentence			All links		
	P	R	F	P	R	F	P	R	F
(a) Without Coreference	-	-	-	86.6	34.9	49.8	72.2	40.3	51.7
(b) Deterministic with Formula (1)	83.8	18.7	30.4	77.5	38.6	51.5	72.9	41.5	52.9
(c) Non-determinstc with Formulae (1)(2)	80.5	23.7	36.7	78.0	38.6	51.7	73.3	41.4	53.0
(d) Non-deterministic with Formulae (1)(3)	-	-	-	72.1	44.0	54.6	72.5	40.5	52.0
(e) Non-deterministic with Formulae (1)(2)(3)	87.8	25.3	39.3	74.0	45.7	56.5	73.0	42.7	53.8
(f) System (e) with Gold Coreference	86.7	58.2	69.7	68.8	64.7	66.7	72.6	46.5	56.7

文内リンクにおいて 5.0(%) の F 値向上が見られた．同じく文内＋文間リンク（All links) においても決定的システムより高いスコアを得られた．この性能改善は統計的に有意である ^{*12}．さらに，(f) のスコアはフルシステム (e) と比較して特に再現率において大きな改善がなされることが読み取れる．このように，共参照解析の性能向上がそのまま文間リンク及び共参照に関係した文内リンクの性能改善につながることが分かる．しかしながら，その文内・文間共にその F 値は 70(%) を越えることができず，これが提案手法の潜在的な性能限界となっている．その主な理由として考えられるのは，前述の通り，文間リンクの約 20% が明示的な照応詞を持って共参照関係にない，ゼロ照応関係になっており，このような場合を本研究のモデルでは表現できていないためである．

5. おわりに

本稿では文境界を越えた事象-項関係同定に関する新たな手法を提案した．本研究で提案した手法は共参照関係を Markov Logic によって利用するものである．共参照関係にある項は談話構造上の主題性の高さから，一般に文書理解のために価値が高く，積極的に同定すべき対象であると本研究では捉えている．本研究のシステムは文間リンクと共参照に関係した文内リンクの双方を非決定的なモデルによって同定，その性能を改善した．さらには，共参照解析の性能を向上させることで，事象-項関係同定のさらなる性能改善が可能になることも確認できた．

しかしながら，提案手法による文間リンクの性能向上は文内リンクと共参照関係の解析性能に依存しているため，その向上も限定的なものとなっている．この問題を克服するため，我々が次に着目するのは文書全体での最適化である．具体的には文書全体を対象として，事象-項-共参照関係を同時推定するモデルを考えている．文書中のトークン全てを同時に考慮して最適化できれば事象-項関係，項-項関係をつないだ照応鎖，そして事象-事象関係までも視野に入れることができ，効果が高いことは容易に想像できる．この着想は Narrative Schema [Chambers 09] とも関連している．しかし問題となるのは計算コストであり，時間・空間ともに計算量を削減する工夫が必須となる．このため，今後は近似手法を含め，効率的な学習・推論について調査を進める予定である．

参考文献

[Björne 09] Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., and Salakoski, T.: Extracting complex biological events with rich graph-based feature sets, in *BioNLP '09: Proceedings of the Workshop on BioNLP*, pp. 10–18, Morristown, NJ, USA (2009), Association for Computational Linguistics

[Buyko 09] Buyko, E., Faessler, E., Wermter, J., and Hahn, U.: Event extraction from trimmed dependency graphs, in *BioNLP '09: Proceedings of the Workshop on BioNLP*, pp. 19–27, Morristown, NJ, USA (2009), Association for Computational Linguistics

[Chambers 09] Chambers, N. and Jurafsky, D.: Unsupervised Learning of Narrative Schemas and their Participants, 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*, pp. 602–610, Suntec, Singapore (2009), Association for Computational Linguistics

[Kim 08] Kim, J.-D., Ohta, T., and Tsujii, J.: Corpus annotation for mining biomedical events from literature, *BMC Bioinformatics*, Vol. 9, No. 1, pp. 10+ (2008)

[Kim 09] Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., and Tsujii, J.: Overview of BioNLP'09 shared task on event extraction, in *BioNLP '09: Proceedings of the Workshop on BioNLP*, pp. 1–9, Morristown, NJ, USA (2009), Association for Computational Linguistics

[Richardson 06] Richardson, M. and Domingos, P.: Markov logic networks, *Machine Learning*, Vol. 62, No. 1-2, pp. 107–136 (2006)

[Riedel 09] Riedel, S., Chun, H.-W., Takagi, T., and Tsujii, J.: A Markov logic approach to bio-molecular event extraction, in *BioNLP '09: Proceedings of the Workshop on BioNLP*, pp. 41–49, Morristown, NJ, USA (2009), Association for Computational Linguistics

[Soon 01] Soon, W. M., Ng, H. T., and Lim, D. C. Y.: A machine learning approach to coreference resolution of noun phrases, *Comput. Linguist.*, Vol. 27, No. 4, pp. 521–544 (2001)

$*12$ $\rho < 0.01$, McNemar's test 2-tailed