グラフクラスタリングを用いた共参照解析

松田 昇悟　　松本　裕治
奈良先端 科学技術大学院大学 情報学研究科
{matsuda.shogo.mj9, matsu}@is.naist.jp

1　はじめに

近年の共参照解析の手法として，グラフに基づいた手法の研究が多くなされている [1][3][6]．グラフに基づいた手法は，文書内の共参照関係全体を最適化するため，他の手法に比べて比較的高い精度を達成している．しかし，これらの手法は全体を最適化するため手法自体が複雑化していることが問題視されている [13]．そこで Uryupina ら [13] は，より簡易なモデルである言及ペアに基づいたモデルの拡張に取り組んでいる．言及ペアに基づいた手法はグラフに基づいた手法に比べて，高速なプロトタイピングが容易であること，素性エンジニアリングなどの低レベルの振る舞いを理解するために有用であることが挙げられている．しかし，言及ペアに基づいた手法では，基本的に先行詞と照応詞のペアに対して共参照関係にあるかを判定しているため，周囲の言及の情報が利用できないという問題点がある．

これに対して，本稿では，既存の言及ペアに基づいた手法に対して，周辺の言及の情報を考慮するために，従来のコミュニティ抽出で用いられているグラフクラスタリングの手法を共参照解析に取り入れた手法を提案し，その有効性を検証する．結果として提案手法では，既存の言及ペアの手法に比べて 1%ほど精度が向上することを示した．

2　関連研究

2.1　共参照解析

共参照解析は，文書内の現実世界の実体を表現する句などに対して，同じ実体を示す句同士を結びつける処理である．既存の共参照解析の手法では，言及ペアに基づく手法 [9]，言及と実体クラスタを使用する手法 [7]，グラフに基づいた手法 [3] などが提案されている．言及ペアに基づいた手法では，先行詞と照応詞のペアを見て，そのペアが共参照関係にあるかどうかを

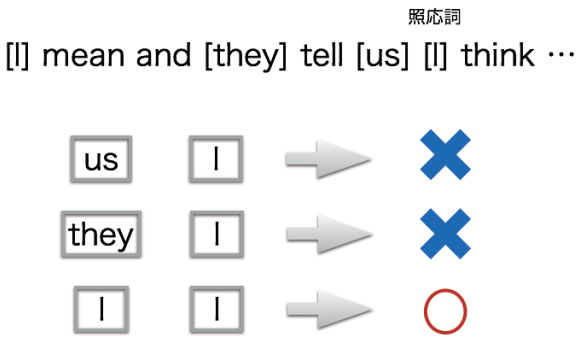

図 1: 言及ペアに基づいた手法の例

判定する手法である．しかし，判定する際に使用できるものが先行詞と照応詞の情報のみであるため，既に構成されている実体クラスタの情報を使用することができない．さらに，どちらか，またはどちらも代名詞であった場合，代名詞自身から得られる情報がほとんど無いため，共参照関係にあるかを決定することが難しいという問題点がある．言及と実体クラスタを使用する手法では，言及と言及のペアではなく，言及と実体クラスタに対して，その言及が実体クラスタに属するかどうかを判定する手法になっている．この手法では，既に決定している実体クラスタの情報が利用できるという利点が挙げられている．しかし，言及と実体クラスタを使用する手法では，逐次的に実体クラスタが構成されるが，その際に誤ったものがクラスタに混入してしまい，判定を誤る原因となっている．グラフに基づいた手法では，文書内全体の共参照関係を一つのグラフであると考え，そのグラフから潜在的な共参照関係を表現する最大全域木を求める手法である．

2.2　コミュニティ抽出

コミュニティ抽出とは SNS のような大規模なグラフから目に見えないコミュニティ構造を得る処理である．コミュニティとはグラフに存在するノードが密と

なっている部分構造である．以下ではコミュニティ抽出の既存の指標と手法について説明する．

2.2.1 モジュラリティ

グラフから抽出したコミュニティ構造の良さを測るための指標として Newman ら [8] が提案したモジュラリティがある．モジュラリティはコミュニティ内のエッジが密，かつコミュニティ間のエッジが疎であるような分割を行う時に高い値を示す．モジュラリティQ は以下の式で表現される．

$$Q = \frac{1}{2M} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2M} \right) \delta \left(C_i, C_j \right)$$

この式で M はグラフにおける全エッジ数を表す，A_{ij} はグラフの隣接行列の ij 要素を表す．k_i, k_j はそれぞれ，ノード i，ノード j の次数であり，δ はクロネッカーのデルタであり，ノード i とノード j の属するコミュニティ Ci, Cj が同一の場合に 1 を返す．

モジュラリティは重みなし無向グラフに対して定義されている．しかし，近年では重み付きグラフにでも適用できる形に拡張したモジュラリティも提案されている [11]．重み付きグラフに適用できる形に拡張されたモジュラリティQ を以下に示す．

$$Q = \frac{1}{2W} \sum_{ij} \left(W_{ij} - \frac{w_i w_j}{2W} \right) \delta \left(C_i, C_j \right)$$

この式では全エッジ数 M が全エッジの重みの総和 W に，ノード i，ノード j の次数 k_i，k_j がノードに隣接する重みの総和になる．

既存のコミュニティ抽出における手法ではモジュラリティ最大化を目的とした手法が主となっている．

2.2.2 Louvain 法

Louvain 法 [2] は局所的な最適化とノードの逐次集約を行うことで高速にモジュラリティを最大化する手法であり，以下のフェーズを繰り返すことでコミュニティを得る．

1. 任意のノードに対して，変化させる前のモジュラリティとそれぞれの隣接するクラスタにマージした際のモジュラリティが最大になるクラスタにノードを統合する．モジュラリティが変化しなくなるまでこの処理を繰り返す
2. モジュラリティが変化しなくなったら，クラスタに属するノード群を一つのノードに集約する．

第一フェーズでは，与えられたネットワークからランダムにノードを選択する．次に選択したノードをクラスタに統合した時のモジュラリティと統合する前のモジュラリティの差分を求める．その差分が最大になる隣接クラスタに選択したノードを統合するという処理である．この処理をモジュラリティの差分が変化しなくなるまで繰り返す．

第二フェーズでは，第一フェーズで求められたクラスタを一つのノードに集約するという処理を行う．この時，クラスタ内エッジは自身から自身へのエッジになり，重みはクラスタ内エッジの重みの総和の二乗になる．クラスタ間エッジの重みは，そのままクラスタ間にあるエッジの総和を用いる．第二フェーズの後は第一フェーズに戻る．この処理をモジュラリティが変化しなくなるか，ノード数が一つになるまで繰り返し，最終的なコミュニティのクラスタを得る．

3 グラフクラスタリングを用いた共参照解析

本稿では，コミュニティ抽出で用いられているグラフクラスタリングを用いた共参照解析の手法を提案する．全体の処理は，共参照関係をグラフで表現し，次に構築されたグラフに対して，Louvain 法を適用するという流れになっている．グラフの構築方法については以下で説明する．

3.1 共参照関係のグラフ

共参照関係をコミュニティ抽出の手法で解くために，文書内全体の共参照関係を無向グラフで考える．この時，各ノードを文書内の各言及，各言及間のエッジを言及間の共参照関係の確率値とする．この確率値は従来の言及ペアで用いられるものと同様に，データから得られた照応詞・先行詞のペアについて共参照関係にあるかどうかのラベルを付与したものを学習データとして使用し，ロジスティック回帰で学習したスコアを使用する．確率値を使用したのは，0・1 のラベルでは表現できない細かい言及間の関係を考慮できると考えたためである．

3.2 余分なエッジの削除

そのまま言及間の関係をエッジとして使用してグラフを構築した場合，後の処理の解析時間が膨大になる

ことや解析の際にノイズになることが考えられる．そのため，使用する必要のないエッジを予め除去する必要がある．

使用する必要がないエッジとして，決して共参照関係にならない言及間のエッジがある．例えば”男性-女性”のような言及間で性別が異なる場合や，”生物-無生物”のような有主性が異なる場合などがある．これらのエッジに関しては，事前にエッジ除去のルールを使用することでエッジを除去する．エッジ除去のルールは Lee ら [4] が構築しているルールベースのシステムの一部と Martschat ら [5] が定義しているルールを使用する．

次に閾値以下のエッジを除去する．モジュラリティを使用したコミュニティ抽出の手法では，独立した言及を検出することが困難であり，値が小さいエッジなのにも関わらずクラスタに取り込まれてしまうことが多くある．そのため，予め閾値以下のエッジを除去する必要がある．ここで独立した言及とは，文書内に存在する言及ではあるものの，文書内の言及以外で自身以外その実体を参照するものがないような言及を指す．

4 実験

共参照解析に周辺の言及を考慮することによる有効性を検証するために，既存の言及に基づいた手法とコミュニティ抽出で使用されている手法を比較した．

4.1 設定

4.1.1 データセット

データセットとして，CoNLL 2012 Shared Task[1][10] で用いられているものを使用した．これは OntoNotes で共参照のアノテーションが行われている部分を使用したものである．データセットは英語・中国語・アラビア語で構成されているが，本稿ではデータセットとして英語のみを使用する．データセットは 7 つのジャンルから構成されており，学習データは 2802 文書，テストデータは 384 文書となっている．

4.1.2 ベースライン

ベースラインとして，既存の言及ペアに基づいた手法として，照応詞に対して各先行詞候補に判定を行ったなかで最も照応詞の近くにあるものを先行詞として決定する手法（closest-first clustering）[12]，各先行詞候補と照応詞で判定を行ったなかで最も高いスコアを持つものを先行詞として決定する手法（best-first clustering）[9]，全ての照応詞とその先行詞候補に対してスコア付けを行った後で，スコアの高い順からクラスタを構成していく手法（easy-first clustering）[13] の 3 つを使用した．さらにグラフに基づいた手法との比較として [3][1] を使用した．素性は [6] の言及ペアに基づいた手法で使用されているものを用いた．

4.2 結果

本節では上記のデータセットに対して行った評価実験の結果を示す．評価指標として，共参照解析で一般に用いられている．MUC，B^3，$CEAF_e$，CoNLL Score を使用する．最終的なシステムの比較は CoNLL Score によって行う．グラフのエッジ除去の閾値は開発セットで調整を行った結果，CoNLL Score が最も高かった 0.4 を閾値として使用した．

実験結果を表 1 に示す．評価は CoNLL 2012 で配布されている公式の評価スクリプトを使用した．学習の際のロジスティック回帰の実装として，scikit-learn[2] を使用した．

結果としてクラスタリングの処理に Louvain 法を用いた手法は他の既存手法に比べて CoNLL Score が 1%程度上回るという結果になった．この結果から周辺の言及を考慮することは精度向上に有効であることが示されている．さらにグラフに基づいた手法と比べて同等の CoNLL Score を達成することができた．

4.3 考察

結果としては，提案法は既存の言及ペアに基づいた手法と比べて，MUC が F 値で 3.18%，グラフに基づいた手法と比べて 1.73%と他の評価値と比べて大きく差がでた．MUC が既存の手法に比べて高い理由として，モジュラリティを指標として使用した場合，実体クラスタ内に多くの言及を保持する傾向があるからだと考えられる．モジュラリティを指標として使用した場合，ノード数の少ないクラスタ．さらに，評価指標として MUC は実体クラスタ内に正解以外の言及を多く含んだ際にペナルティをつけづらいため，結果として多くの言及を含んだクラスタに高い値になっていると思われる．

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

[2] http://scikit-learn.org/stable/

表 1: 実験結果 (%)

手法	MUC			B^3			$CEAF_e$			CoNLL
	P	R	F1	P	R	F1	P	R	F1	
Soon ら [12]	72.35	59.05	65.05	64.28	44.23	52.4	39.05	50.65	44.1	53.85
Uryupina ら [13]	71.41	61.42	66.04	65.04	43.26	51.96	43.91	53.72	48.32	55.54
Ng ら [9]	71.48	67.16	69.25	60.55	51.97	55.93	51.89	51.02	51.45	58.8
提案法	73.67	71.23	**72.43**	61.45	52.95	56.88	56.88	49.31	52.82	60.71
fernandes[3]	75.91	65.83	70.51	65.19	51.55	57.58	57.28	50.82	53.86	60.65
B&K[1]	74.30	67.46	70.72	62.71	54.96	**58.58**	59.40	52.27	**55.61**	**61.63**

次に出力された実体クラスタについてエラー分析を行った結果，代名詞がクラスタ間のハブになることが原因でクラスタが統合されてしまっていることが分かった．代名詞はそれ自身では情報をほとんど持っていないため，どの言及に対してもスコアの差が出にくい．そのため実体が異なるクラスタ対を考えた時に，代名詞はどちらのクラスタ内の言及に対しても同様の確率値を付与してしまい，最終的な結果としてクラスタが統合されしまうと考えられる．この問題は代名詞がほとんど情報を持っていないということが原因であるため，言及と実体クラスタを使用した手法を拡張する形で用いることで精度の向上を行うことができるのではないかと考えている．

5 おわりに

本稿では，既存の言及ペアに基づいた手法に対して，照応詞と先行詞のペアだけでなく，周辺の言及の情報を考慮することができる手法を検証した．言及ペアモデルで周辺の言及との関係を考慮することでより簡易な手法でグラフに基づいたアプローチと同等の精度を達成できることを示した．

今後の課題として，クラスタリングの際のクラスタ間での制約の導入と代名詞に関するエラーについて解決したいと考えている．

参考文献

[1] Anders Björkelund and Jonas Kuhn. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the Association for Computational Linguistics*, 2014.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008, No. 10, p. P10008, 2008.

[3] Eraldo Rezende Fernandes, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú. Latent trees for coreference resolution. *Computational Linguistics*, 2014.

[4] 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, pp. 885–916, 2013.

[5] Sebastian Martschat. Multigraph clustering for unsupervised coreference resolution. 2013.

[6] Sebastian Martschat and Michael Strube. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics*, Vol. 3, pp. 405–418, 2015.

[7] Andrew McCallum and Ben Wellner. Toward conditional models of identity uncertainty with application to proper noun coreference. 2003.

[8] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, Vol. 69, No. 2, p. 026113, 2004.

[9] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 104–111. Association for Computational Linguistics, 2002.

[10] 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. Association for Computational Linguistics, 2012.

[11] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, Vol. 74, No. 1, p. 016110, 2006.

[12] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, Vol. 27, No. 4, pp. 521–544, 2001.

[13] Olga Uryupina and Alessandro Moschitti. A state-of-the-art mention-pair model for coreference resolution. *Lexical and Computational Semantics (* SEM 2015)*, p. 289, 2015.