The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

語義曖昧性解消のための定義文拡張における ネットワーク特徴量を利用したWikipediaの記事評価手法の検討

A Study on Evaluation Method of Wikipedia's Article Using Network Feature Quantity in Definition Statement Extension for Word Sense Disamibiguation

村田 亘 *1　　大沢 英一 *2
Wataru Murata　　Ei-Ichi Osawa

*1公立はこだて未来大学大学院 システム情報科学研究科
Graduate School of Systems Information Science, Future University Hakodate

*2公立はこだて未来大学 システム情報科学部 複雑系知能学科
Department of Complex and Intelligent Systems, School of Systems Information Science, Future University Hakodate

Word Sense Disambiguation is one of the basic tasks of Natural Language Processing. This is the task which distinguishes multi-sense word in a sentence. Wikipedia was regarded as a large scale-network that articles are nodes and links are edges study using the link structure of Wikipedia. Therefore the present study extended define statement of word sense by selecting articles which have strong connectivity with other word sense.

Experiment of Words Sense Disambiguation failed in selecting articles and there are word sense whose correct answer rate was low. Therefore the correct answer rate is considered to rise evaluating usefulness of articles selected quantitatively. The present study focused on network feature quantity and analysed the network structure. It was found from the result of the observation that especially value of Modularity Q affected the correct answer rate.

1. はじめに

近年，インターネットやスマートフォンの普及に伴い，多くの人々がSNSやブログなどを用いて情報を発信することができる．このことから，膨大な量の言語データがインターネット上に存在し，これらを分析することで，情報検索，情報抽出，テキストマイニング，文書要約，翻訳などの応用技術の研究が行われている．そういった研究は，自然言語処理の研究分野に属しており，「言葉がわかる」計算機システムの構築を目指している [新納 16]．

自然言語処理の基本タスクの1つに，語義曖昧性解消 (WSD: Word Sense Disambiguation) がある．これは，文中の多義語の語義を識別するタスクであり，特に翻訳技術の性能向上には必要不可欠である．また，日本語の場合は文中で漢字に変換されていない平仮名の単語を正しい漢字に変換する「かな漢字変換」にも応用できる．

WSDは一般に教師あり機械学習手法を用いた研究が多くなされており，現在の教師あり機械学習の枠組みで処理すれば，90%程度の正解率に達すると言われている [新納 16]．しかし，訓練データの作成コストが高く，対象単語が多くても数百程度に限定されていしまうという問題がある [新納 16]．現実のアプリケーションではすべての単語を対象にする必要があるため．教師あり機械学習手法で解決することは難しい．そこで，すべての単語に語義を付与するWSDはall-words WSDとして，主に教師なし機械学習を用いて研究がされており，現在は60~70%の正解率である [新納 16]．教師なし機械学習の手法には，辞書の定義文から語義の分散表現を求め，評価データの文脈の分散表現との比較を行う手法 [Chen 14] がある．

一方で，知識ベースを利用したWSDの研究もなされており，古くからある知識ベースの手法にLeskアルゴリズム [Lesk 86] がある．これは，辞書の各語義の定義文に含まれる単語と，対象の文の単語との重複が一番多い語義に決定する手法である．しかし，Leskアルゴリズムには，重複する単語が存在しない場合に，語義を決定できないという問題がある．

そこで村田らはWikipediaを大規模ネットワークと捉え，ネットワーク特徴量や共起指標などを利用することで語義と関係の強い記事の選定を行い，定義文の拡張を行った [村田 17]．WSDの実験では，拡張を行うことで正解率が約10%向上した．しかし，村田らの手法では記事が正しく選定されないことで，語義の拡張が正しく行われなかった例がみられた．

もし語義の拡張に有用な記事を定量的に評価できれば，正解率の向上が期待できる．また，WSDだけでなくWWW(World Wide Web) の有用なページの選定などに応用できると考えている．

そこで，本研究では語義の定義文拡張のための有用な記事の識別方法についてネットワーク特徴量を利用して検討する．

連絡先: 村田 亘，公立はこだて未来大学大学院 システム情報科学研究科，北海道函館市中野町 116-2，0138-34-6448，g2117046@fun.ac.jp

2. 前提知識

本章では，本論文の内容で用いる特に重要な技術や知識について，概要を述べる．

2.1 クラスター係数

クラスター係数とは，あるノードの隣接ノードどうしが隣接ノード，つまり三角形の割合を示す指標である．ノード v_i に隣接しているノードの数 (次数) を k_i とすると，v_i のクラスター係数 C_i は以下の式で表される [増田 10]．

$$C_i = \frac{v_i\text{を含む三角形の数}}{k_i(k_i-1)/2} \tag{1}$$

2.2 コミュニティ抽出

図1のように，同じ集団内ではエッジが密で異なる集団間にはエッジがあまりないネットワークはよく見られる [増田 10]．点線内が1つの集団に対応し，これをコミュニティと呼ぶ．

コミュニティ抽出のための指標にはモジュラリティQがよく利用され，Qは値が高いほど良い分割といえる．モジュラリティQは以下の式で表される．

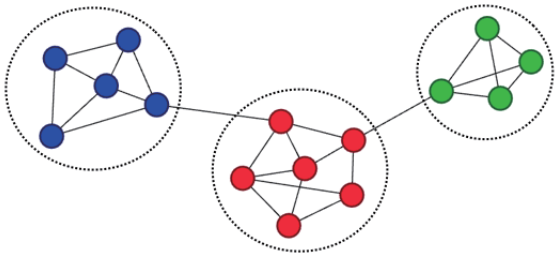

図 1: コミュニティの例

$$Q = \sum_i (e_{ii} - a_i^2) \tag{2}$$

e_{ii} はネットワーク全体におけるコミュニティ内のエッジの割合，a_i はネットワーク全体におけるコミュニティi から他のコミュニティへのエッジの割合である．したがって，コミュニティ内のノード間のエッジが多く，コミュニティ外のノード間のエッジが少ない分割ほど Q は高くなる．

Q を利用したコミュニティ抽出法には，Newman 法 [Newman 04] がある．しかし，Newman 法は大規模ネットワークに適用させた場合に計算量が膨大になるため，計算量を抑えたコミュニティ抽出法である CNM 法 [Clauset 04] が提案されている．

2.3 Jaccard 係数

Jaccard 係数は，2 つの集合間の類似性を表す指標である．X と Y の Jaccard 係数は以下の式で表される．

$$Jaccard(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{3}$$

本研究では，ノード v_i とノード v_j に隣接しているノードをそれぞれの集合とし，$Jaccard(v_i, v_j)$ を算出する．つまり，v_i と v_j が互いに共通のノードと隣接しているほど Jaccard 係数は高くなる．

3. ハイパーリンク構造の Jaccard 係数を利用した定義文拡張による WSD 手法

本章では，[村田 17] の研究の概要を述べる．

3.1 提案手法の概要

提案手法の手順の概要を以下に示す．ここで，本研究でのネットワークは，「Wikipedia の記事をノード，他の記事へのリンクをエッジとする無向ネットワーク」とする．

(i) ソース記事 s に対して被リンク先に距離 2 の無向ネットワークを構築

(ii) Jaccard 係数とクラスター係数，CNM 法を用いた記事の選定

(iii) 選定された記事 (Selected Articles) から，共起ネットワーク (Co-occurrence Network) の構築

(iv) キーワード抽出 (Keyword Extraction)

(v) 語義曖昧性解消 (Word Sense Disambiguation)

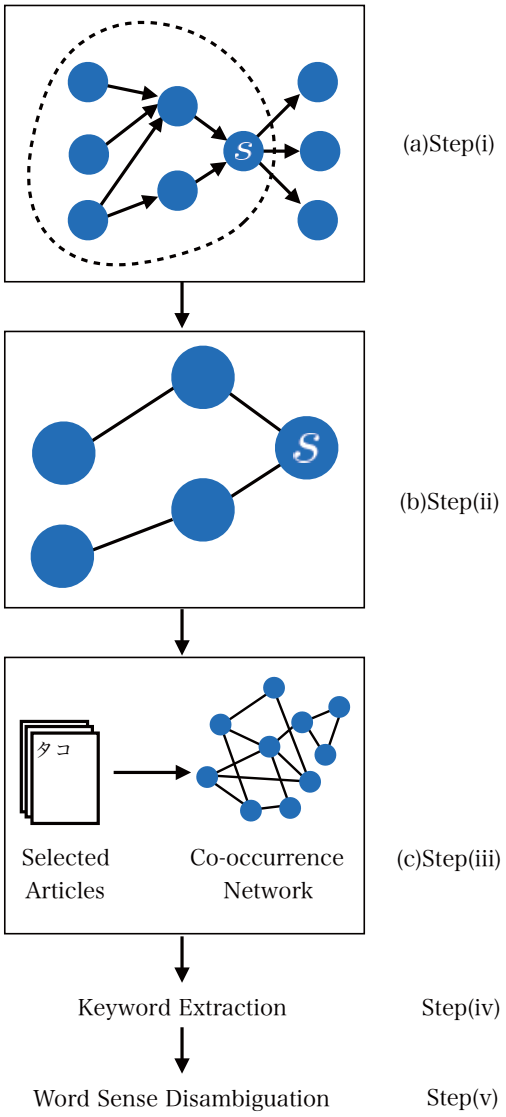

図 2: 提案手法の手順 (出典：[村田 17])

手順 (i)，(ii) は [千田 10]，手順 (iii)，(iv) は [松尾 02] の提案手法を参考にしている．

4. 章以降の説明のため，手順 (i)，(ii) の概要を述べる．

3.1.1 手順 (i)：ネットワークの構築

ソース記事 s(語義) から被リンク先に対して距離 2 の無向ネットワークを構築する．被リンク先とは，ソース記事へリンクを張っている記事のことである．ただし，被リンク先が存在しない場合はソース記事 s のみを用いる．

3.1.2 手順 (ii)：Jaccard 係数とクラスター係数を用いた記事の選定

手順 (i) で構築されたネットワークの全てのエッジに対して Jaccard 係数を算出し，閾値を 0.01 に設定する．次に，Jaccard 係数が閾値未満であればエッジを除去しソース記事 s のクラスター係数を算出する．ただし，エッジの本数が 0 本となったノードは除去する．さらに閾値を 0.01 上げ，エッジの除去，ソース記事 s のクラスター係数の算出を繰り返す．そして，ソース記事 s のクラスター係数が最後に極大をとる Jaccard 係数を最終的な閾値に決定し，ネットワークを再構築する．最後の極

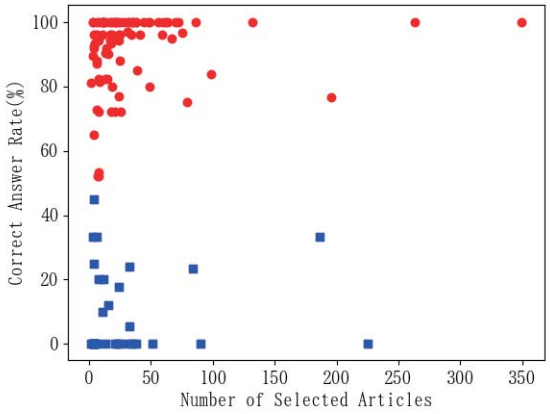

図 3: 選定された記事数と正解率

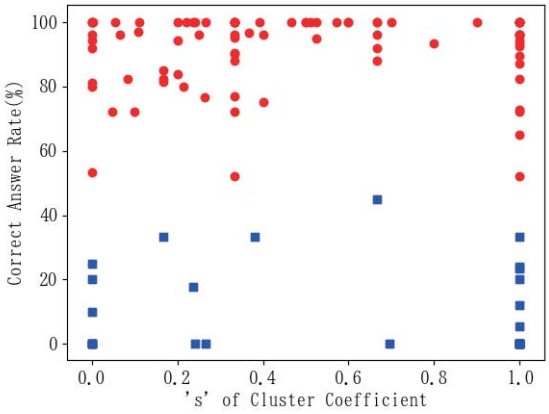

図 4: ソース記事 s のクラスター係数と正解率

大にする理由は，Jaccard 係数が高いため類似性が高い記事が残り，クラスター係数が高いためネットワークが密であるからである．

次に，再構築されたネットワークに対して CNM 法を適用し，ソース記事 s を含むコミュニティに含む記事を選定された記事集合とする．

3.2 実験・評価

実験に用いた評価データは，多義語が 50 個でそれぞれの多義語に対して 50 個ずつ用意した．それぞれの多義語の語義数には偏りがあるため，語義の総数は 119 個であった．つまり，2500 個の評価データを用いて，語義曖昧性解消の実験を行い，正解率を算出した．

$$\text{正解率} = \frac{|\,\text{正解データ} \cap \text{提案手法の結果データ}\,|}{|\,\text{正解データ}\,|} \qquad (4)$$

結果は，拡張を行わない場合の正解率は 62.3%，拡張を行った場合が 73%であり，正解率が約 10% 向上した．しかし，記事の選定が失敗しているために，定義文の拡張が正しく行われず正解率が低い語義もみられた．

そこで，本研究では選定された記事が有用かどうかを識別するために，記事の選定の際のネットワーク構造を解析する．

4. 選定された記事の評価

本章では，ネットワーク特徴量を用いて選定された記事が WSD に有用かどうかを検討する．ここで，正解率が 50%以上の語義の選定された記事集合を「有用」とする．有用な記事集合は 88 個，有用でない記事集合は 31 個あった．

4.1 選定された記事数

図 3 は選定された記事数と正解率のグラフである．赤の丸い点になっているのは正解率が 50%以上，青の四角い点は 50%未満の正解率の語義である．

図 3 から，選定された記事数は 1 ～ 50 個が最も多くなっているが，その範囲においても正解率が 0%の語義から 100%に近いものが多く，記事数を用いた有用な記事の識別は難しい．

4.2 クラスター係数

図 4 は [村田 17] の提案手法の手順 (ii) の閾値を決定し，ネットワークを再構築した際のソース記事 s のクラスター係数と正解率のグラフである．

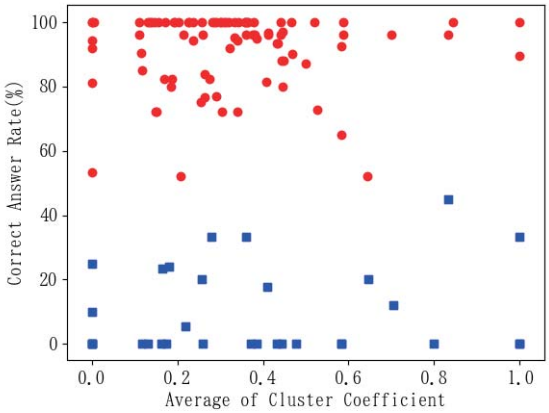

図 5: 平均クラスター係数と正解率

ソース記事 s のクラスター係数が高ければネットワークが密であるので，正解率に影響すると考えたが，有用な記事集合のなかでもクラスター係数が低いものが多かった．また，クラスター係数が高いのにもかかわらず，正解率が低い記事集合もみられた．

図 5 はネットワークを再構築した際の平均クラスター係数と正解率のグラフである．グラフから平均クラスター係数は 0.1 ～ 0.6 の値に集中しており，正解率との関係はあまりない．

4.3 モジュラリティ Q

図 6 はコミュニティ抽出をする際のモジュラリティ Q の値と正解率のグラフである．グラフの分布は図 3 に似ているが，図 6 の方が分散が大きくなっており，$Q = 0.7$ 以上の有用な記事集合は約 9 割存在する．

ここで，Q に閾値を設定する．閾値以上の Q の記事集合を有用な記事集合と識別し，閾値未満の記事集合を有用でない記事集合と識別する．

例として $Q = 0.7$ を閾値とし正しく有用な記事集合かどうかの識別を行う．図 6 の横破線は正解率が 50%，縦破線は $Q = 0.7$ である．この場合，Q が 0.7 以上の有用な記事集合

表 1: モジュラリティQ に閾値を設定した場合の Precision・Recall・F 値

	有用な記事集合 (赤の丸い点)			有用でない記事集合 (青の四角い点)		
Threshold	Precision	Recall	F	Precision	Recall	F
0.1	0.745	**0.965**	**0.841**	0.4	0.064	0.111
0.2	0.75	0.954	0.839	0.428	0.096	0.157
0.3	0.75	0.954	0.839	0.428	0.096	0.157
0.4	0.754	0.943	0.838	0.444	0.129	0.2
0.5	0.754	0.943	0.838	0.444	0.129	0.2
0.6	0.771	0.92	0.839	0.5	0.225	0.311
0.7	0.795	0.886	0.838	**0.523**	0.354	0.423
0.8	**0.844**	0.738	0.787	0.452	0.612	**0.52**
0.9	0.838	0.295	0.436	0.295	**0.838**	0.436

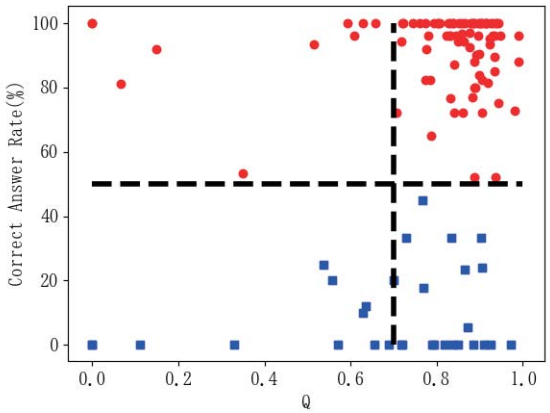

図 6: モジュラリティQ と正解率

(右上) は正しく識別され，0.7 未満 (左上) は誤って識別される．また，0.7 以上の有用でない記事集合 (右下) は誤って識別され，0.7 未満 (左下) は正しく識別される．

表 1 は，Q に閾値を設定し Precision，Recall，F 値を算出したものである．有用な記事集合の F 値は $Q = 0.1$ のとき最大だが，0.1 ～ 0.7 で大きく値は変わらない．これは，有用な記事集合の正解率が $Q = 0.7$ 以上に有用な記事集合が密集しているため Recall の変化があまりないためである．

このため，選定された記事集合の有用性を識別するためには $Q = 0.8$ が最も良い閾値であることがわかる．しかし，そのときの有用でない記事集合の Recall は約 0.6 であることから，4 割が識別できていない．

今後は，他のネットワーク特徴量などを利用する必要がある．

5. おわりに

本論文では．記事の選定が失敗し WSD の正解率が低くなってしまう場合があるという [村田 17] の問題に対して，ネットワーク特徴量を用いることで選定された記事の有用性の識別方法の検討を行った．検討した中ではモジュラリティQ を用いた場合が最も多く識別ができたが，識別に失敗しているものも多かった．

今後は中心性などのネットワーク指標や，[村田 17] の提案手法の手順 (i) から (ii) にかけてネットワークがどのように変化しているのかを解析することで，より良い識別手法が得られると考えている．

また，WWW などのハイパーリンク構造を持つネットワークに対して本手法を適用させ，より良い有用な記事の抽出方法を模索したい．

参考文献

[Chen 14] Chen, X., Liu, Z., and Sun, M.: A Unified Model for Word Sense Representation and Disambiguation, in *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1025–1035 (2014)

[Clauset 04] Clauset, A., Newman, M. E. J., and Moore, C.: Finding community structure in very large networks, *Physical Review E*, Vol. 70, No. 066111 (2004)

[Lesk 86] Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, in *the 5th annual international conference on Systems documentation*, pp. 24–26 (1986)

[Newman 04] Newman, M. E. J.: Fast algorithm for detecting community structure in networks, *Physical Review E*, Vol. 69, No. 066133 (2004)

[松尾 02] 松尾 豊, 大澤 幸生, 石塚 満 : Small World 構造に基づく文書からのキーワード抽出, 情報処理学会誌, Vol. 43, No. 6, pp. 1825–1833 (2002)

[新納 16] 新納 浩幸 : 自然言語処理の現状と展望 語義曖昧性解消, 情報処理学会誌, Vol. 57, No. 1, pp. 20–21 (2016)

[千田 10] 千田 俊輔, 大沢 英一 : リンク構造解析による Wikipedia のナビゲーション情報の抽出, 合同エージェントワークショップ&シンポジウム JAWS2010 (2010)

[増田 10] 増田 直紀, 今野 紀雄 : 複雑ネットワーク 基礎から応用まで, 近代科学社 (2010)

[村田 17] 村田 亘, 大沢 英一 : ハイパーメディアの Jaccard 係数に着目した定義文拡張における語義曖昧性解消, 日本ソフトウェア科学会第 34 回大会講演論文集 (2017)