リンク先決定における特徴の抽象性を利用した wikification の精度向上

村上 凌悠　　綱川 隆司　　西田 昌史　　西村 雅史

静岡大学大学院総合科学技術研究科 情報学専攻

1. はじめに

Wikification[1]とは Wikipedia へ自動的にリンクを付与する技術のことを言う．Wikification の実現によりリンクの記事の編集コストの削減のほか，曖昧性解消が必要となる機械翻訳などの技術に応用できる．Wikification を行うためには，リンクが張られる単語の曖昧性解消を行ってリンク先となる記事を決めるリンク先決定が必要である．本研究は wikification のリンク先決定の精度向上を目的としている．

リンク先候補となる単語（アンカー）は，抽象的特徴（リンク先候補が持っている性質，上位概念など）とそれ以外の特徴を内包している．リンク先決定においてリンク先候補間で抽象的特徴の類似度が高い場合にリンク先決定の精度が下がるという問題がある．本研究では抽象的特徴がそれ以外の特徴によるリンク先決定に影響を与えないように，抽象的特徴による分類とそれ以外の特徴による分類を分けて処理する方法について検証した．また，Wikipedia と Wikification コーパスの異なる２つのテストセットで，wikification の影響を分析した．

2. 関連研究

Wikification をはじめとする曖昧性解消タスクにおいては，対象のエンティティの周辺の単語や品詞などの情報を素性とした統計的学習の手法が用いられる[2][3][4].

また，曖昧性解消を行う語義の対象が Wikipedia のような辞書など対象の概念の説明のあるドキュメントの場合，そのドキュメントと曖昧性解消を行う単語の周辺の情報との類似度を用いることが可能であり，その手法が提案されている[1][5]．また，オンライン辞書の場合，リンク構造が利用可能である．Wikipedia はオンライン辞書であり言語間リンク，カテゴリなどさまざまなリンク構造を持ち，それらを利用した手法が提案されている[2][6].

3. 提案方法

3.1. リンク先決定の概要

複数の抽象的特徴が類似していればリンク先決定が難しくなると考えられる．例えば，アンカー“野村”のリンク先候補として“野村克也”“野村謙二郎”が存在するが，これらのリンク先候補は“野球監督”といった抽象的特徴で共通している．アンカーの周辺の文脈から得られる情報が，そのアンカーが“野球監督”として使用されたときのものであるなら，どちらのリンク先候補にも共通するため，リンク先の分類においては精度を下げる要因となりうる．逆に抽象的特徴が異なっているリンク先候補間ではリンク先決定が比較的容易であると考えられる．したがって，初めに抽象的特徴が似ているものでクラスタ分類を行い（図 1（①）），リンク先決定時にはどのクラスタに属するのかを決定する（図 1（②））．属するクラスタが決定するとそのクラスタ内のリンク先候補でリンク先決定を行う．クラスタ内のリンク先候補は抽象的特徴が似ているため，抽象的特徴のみを用いると，1 で述べたようにアンカーの周辺の文脈から得られる情報が共通するため，リンク先決定を行うことは難しいと考えられる．したがって，抽象的特徴以外の特徴（以下，非抽象的特徴）を用いてリンク先決定を行う（図 1（③））．次にリンク先決定の各段階の詳細を説明する．

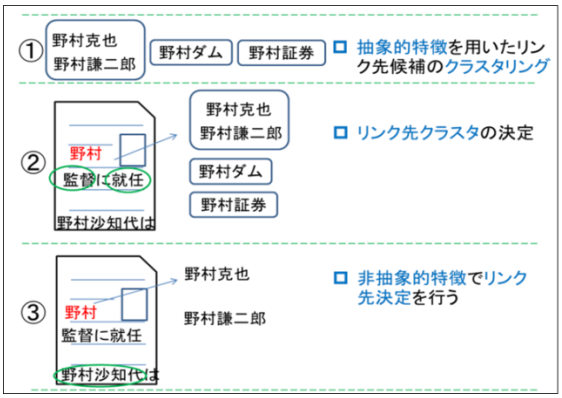

図 1 リンク先決定の概要

3.2. 抽象的特徴を用いたクラスタリング

リンク先候補の抽象的特徴の類似度でクラスタリングを行う．クラスタリングを行うために抽象的特徴をベクトルとして表す．リンク先候補の属する Wikipedia のカテゴリをリンク先候補の抽象的特徴とみなし，各ベクトルの成分とカテゴリを利用して得られる単語を対応させる．抽象的特徴として表すのに相応しいと考えられる単語に対応する抽象的特徴ベクトルの成分の値が大きくなるよう重みづけを行う．次にカテゴリを利用した単語の取得法と抽象的特徴ベクトルの生成法を記述する．

3.2.1. カテゴリを利用した単語の取得法

リンク先候補の属するカテゴリと共通のカテゴリに属する記事を取得し，それらの記事にリンクを張っているアンカーの周辺の単語（アンカーの前後５語，助詞，助動詞，記号は除く）を取得する．図 2 に例を示す．アンカーの“野村”のリンク先候補“野村克也”は“日本の野球選手”といったカテゴリに属する．“野村克也”のほか“大谷翔平”,“山田哲人”などの記事も同様に“日本の野球選手”のカテゴリを持つ．それらの記事にリンクを張っているアンカーの周辺の単語を取得し，そのカテゴリの単語集合とする．

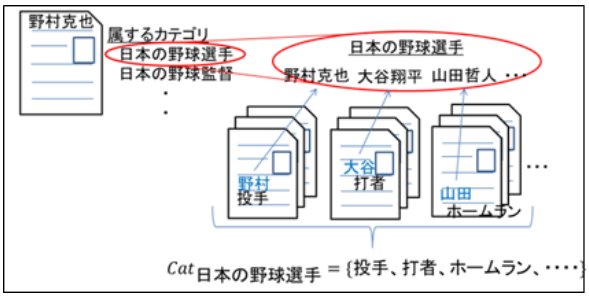

図 2 カテゴリの単語集合の取得

リンク先候補の各カテゴリの単語集合を取得した後，それら

の単語集合の和集合をとり，リンク先の抽象的特徴を表す単語集合とする（図 3）．

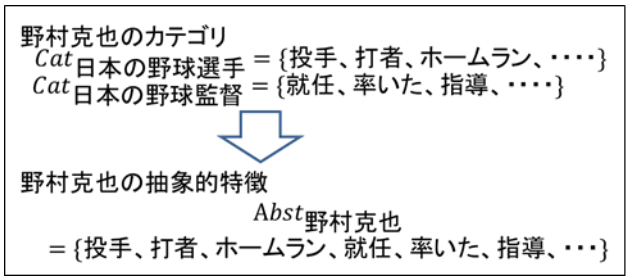

図 3 リンク先候補の単語集合の生成

3.2.2. 抽象的特徴ベクトルの生成

次に得られた各リンク先候補の単語集合を用いて，抽象的特徴ベクトルを構成する．抽象的特徴ベクトルの各成分は各カテゴリから得られたすべての単語と対応付ける．リンク先候補ごとに，得られた単語集合のそれぞれの単語と対応するベクトルの成分の値を求める． 第 n 成分の値は以下のように求める．

$$f_{w_n} = CoProb_{w_n} \times Specificity_{w_n} \quad \ldots\ldots (1)$$

w_nは第 n 成分に対応する単語である．

式(1)の$CoProb_{w_n}$はw_nとw_nを取得する際に用いたアンカーとの共起確率である．同じカテゴリを特徴として持つアンカーそれぞれで共通して得られる単語が，そのカテゴリを特徴づける単語として適切であると考えたため，確率によって重み付けする．w_n が“ホームラン”である場合の例を図 4 に示す．カテゴリ“日本の野球選手”に含まれる記事が 3 記事あり，それぞれの記事にリンクを張っているアンカーが 1 つずつ存在するとする．そのうちアンカーと共起する“ホームラン”は 2 つなので確率は 2/3 となる．

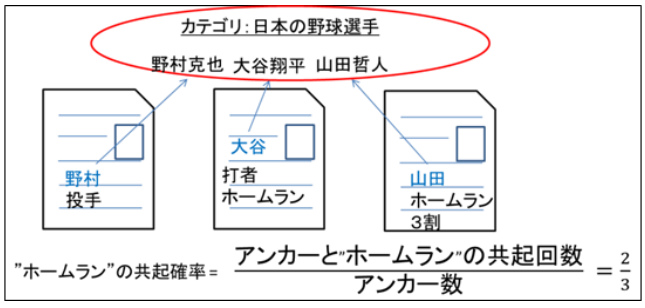

図 4 “ホームラン”とアンカーの共起確率

式(2)の$Specificity_{w_n}$はw_nを取得するのに用いたカテゴリの多様性の低さを表す尺度である．多様性のあるカテゴリ，例えばリンク先候補“野村克也”の属するカテゴリ“存命人物”には，様々な分野の人物の記事が属している．つまり非常に高い抽象度を持ったカテゴリであると言える．抽象度の高いカテゴリから得られた単語を用いてクラスタリングを行うと，一つのクラスタ内に多様なリンク先候補が含まれることになり，非抽象的特徴でのリンク先決定の精度を下げるおそれがある．したがって，抽象度のより高いクラスタから得られた単語の重みを小さくすることを考える．

ここで，Wikipedia の冒頭部に着目する．Wikipedia の冒頭部はその記事の概念を端的に表した表現が使用される．例えば，記事“ジャガー”の冒頭部は“ジャガー（Panthera onca）は，食肉目ネコ科ヒョウ属に分類される食肉類．”となっている．あるカテゴリの記事集合で，この冒頭部の情報に大きなバラツキが見られれば抽象度の高い多様性のあるカテゴリであるとみなせる．このバラツキを計算するために冒頭部に出現する名詞を用いて（文頭の“～は，”の後の文の名詞を用いる），その名詞の出現確率で平均情報量を計算する．平均情報量が大きければバラツキが大きいと言えるので，その逆数をとることで多様性のあるカテゴリから得られた単語の重みを小さくする．したがって$Specificity_{w_n}$は以下のようになる．

$$Specificity_{w_n} = \frac{1}{entropy(Category)} \quad \ldots\ldots (2)$$

$$entropy(Category) = -\sum_{w_c \in W_{category}} P(w_c) log P(w_c) \quad \ldots\ldots (3)$$

ただし，$Category$はw_nを取得する際に用いたカテゴリ，$W_{category}$は $Category$ に属する記事から取得した単語集合，$P(w_c)$は単語w_cの出現確率とする．

なお，単語w_nの値が複数のカテゴリから得られた場合，それぞれのカテゴリから求めた値のうち最も大きい値を用いる．

3.2.3. クラスタリング

上記で得られた抽象的特徴ベクトルを用いてクラスタリングを行う．クラスタリングは凝集型クラスタリングを用いる．リンク先候補間の類似度は抽象的特徴ベクトル間のコサイン類似度とし，クラスタ間の類似度は単連結法で求める．

3.3. クラスタ決定

アンカーの属するクラスタを決定する．3.2.1 と同様に前後の単語を用いてリンク先対象のアンカーをベクトル化する．ベクトルはクラスタリングに用いた抽象的特徴ベクトルを用い，ベクトルの成分は取得した単語と対応する成分を 1，それ以外を 0 とする．得られたベクトルを用いて最も類似しているクラスタを決定する．クラスタリングでは単連結法で最も類似しているクラスタを求めたため，同様にクラスタ決定においても単連結法と同様に，最も近いアンカーが属するクラスタに決定する．また，前後の単語から十分な情報が得られないこともあるので，どのクラスタとの類似度も 0 の場合，クラスタ決定を行わずにすべてのリンク先候補を対象に非抽象的特徴を用いた方法のみでリンク先を決定する．

3.4. 非抽象的特徴でのリンク先決定

3.3 においてアンカーの属するクラスタが決定したので，クラスタに含まれるリンク先候補のうち，どのリンク先候補が相応しいのかを決定する．リンク先決定には袁ら[3]の決定リストによる手法を用いる．

決定リストは，各アンカーに対し，アンカーとリンク先の組を正解データとし，そのアンカーと共起するアンカーとの関連を学習する．学習を行うことで以下のようなルールのリストが得られる．

「IF “BMW” co-occurs with “ジャガー” THEN link “ジャガー” to “ジャガー（自動車)”」

上記のルールは，アンカー“BMW”がアンカー“ジャガー”と共起すれば，アンカー“ジャガー”は記事“ジャガー（自動車)”にリンクすることを意味する．このようなルールがトレーニングデータから求めた確信度（リンク先決定におけるルールの確信度）順に並べられる．アンカーのリンク先決定時には，決定リストの上位のルールから順に該当する共起アンカーが存在するかを調べ，共起アンカーが存在する上位 3 つのルールの多数決で決定する．

クラスタに含まれるリンク先候補は抽象的特徴において類

似しており，抽象的特徴を用いてリンク先決定してしまうと 1 で述べたようにアンカーの周辺の文脈から得られる情報が共通し，リンク先決定が難しくなる可能性があるため，抽象的特徴を決定リストのルールから取り除く．上記で示したルールの場合，もし“BMW”が抽象的特徴であるなら上記のルールをリストから取り除く．取り除く抽象的特徴はその単語と対応する抽象的特徴ベクトルの要素が閾値 t 以上のものとする．

4. 実験

4.1. 実験設定

手法の有効性を検証するため，提案手法と決定リストの場合の比較実験を行う．提案手法においてクラスタリングの有無が与える影響について検証するため，クラスタリングを行わずに非抽象的特徴による手法のみの結果も取得する．また，テストセットによる影響を検証するため，テストセットとして先行研究[7]で用いた 2016 年 2 月 3 日時点の日本語 Wikipedia と Wikification コーパスの 2 つを用いて実験を行う．テストデータに Wikipedia を用いる場合と Wikification コーパスを用いる場合では実験手法が異なる．

Wikipedia を用いる場合 5 分割交差検定により実験を行う．Wikipedia 全体のリンクデータを扱うため分割は記事単位で行う．また，処理に時間がかかるためテスト対象のアンカーを 100 個に絞る．アンカーの選択は十分にデータを確保するため各リンク先候補へのリンクがそれぞれ 10 回以上出現する曖昧なアンカー（リンク先候補が 2 つ以上あるアンカー）をランダムに 100 個選択している．リンク先候補は Wikipedia で実際にリンクが張られているものを対象にしている（曖昧さ回避ページは除く）．

Wikification コーパスは学習に用いるためには十分なデータ量がないため，学習には Wikipedia のデータのみを用いている．Wikification コーパスには対象リンク先として NIL リンクが含まれている．NIL 判定に関しては決定リストでのリンク先決定時にすべてのリンク先候補へのリンクがトレーニングデータに存在しなかった場合 NIL とする．実験対象のアンカーは，リンク先候補が 1 つのものも含め，すべてのアンカーとする．このため，リンク先決定が必ず正解になるものが一定数含まれる．リンク先候補は Wikipedia で実際にリンクが張られているものを対象にしている（曖昧さ回避ページは除く）．

また，Wikipedia のテストと Wikification コーパスのテストで同様に，提案方法での各パラメータは次のように設定した．リンク先候補の凝集型クラスタリングにおいて，クラスタ間の閾値を 0.5 とし，閾値以上のものを同じクラスタとした．また，非抽象的特徴を用いた決定リストのリンク先決定において抽象的特徴を取り除くための閾値 t は 0.1 とした．なお単語を取得する際に行った単語分割は MeCab を使用した．MeCab の辞書データとして 2017 年 7 月 10 日時点の NEologd を使用した．

4.2. 実験結果

エラー! 参照元が見つかりません。にテストデータに Wikipedia を用いた場合，表 2 に Wikification コーパスを用いた場合の結果を示す．

表 1 結果（テストデータ Wikipedia）

手法	抽象的特徴によるクラスタリング	決定リストからの抽象的特徴の削除	正解率
提案手法	○	○	91.24%
提案手法	×	○	93.40%
決定リスト	×	×	91.23%

表 2 結果（テストデータ Wikification コーパス）

手法	抽象的特徴を用いたクラスタリング	決定リストからの抽象的特徴の削除	正解率
提案手法	○	○	67.44%
提案手法	×	○	85.03%
決定リスト	×	×	82.29%

Wikipedia,Wikification コーパスのいずれのテストセットにおいても，クラスタリングを行わずに抽象的特徴を取り除いた場合に最も高い性能が得られ，決定リストの場合と比較して，Wikipedia では 2.17%，Wikification コーパスでは 2.74%改善された．一方で，抽象的特徴を用いたクラスタリングでは，テストセットで Wikification コーパスを用いた場合，決定リストとの場合と比較して 14.85%，テストセットに Wikipedia を用いた場合と比較して，23.80%低下した．

5. 考察

結果より，リンク先決定の対象を Wikipedia とした場合では抽象的特徴が有効であるが，Wikification コーパスでは抽象的特徴が有効でないことがわかる．Wikification コーパスでのミスのうち，クラスタ決定時点の誤りは 82%を占める．Wikipedia の場合だと全体のミスのうちクラスタ決定でのミスは 61%で大きな違いが見られる．大半を占めた Wikification コーパスでのクラスタ決定ミスの要因は以下のようなものが考えられる．

一つ目が正解のリンク先の抽象度が高いアンカーが多く含まれていることが挙げられる．例えば，アンカー“先生”はテストデータに 35 個含まれるが Wikification コーパスでは，リンク先はすべて“先生”になる．しかし，学習データである Wikipedia にはリンク先候補として“先生”以外に，“先生（ドラえもん)”，“手塚治虫”などがある．リンク先の抽象度が高いとそのカテゴリの抽象度も高くなり， 3.2.2 で示した方法を適用すると，抽象度が高いカテゴリから得られた単語の重要度が下がり，比較的抽象度の低いカテゴリを持つ他のリンク先候補（例，“先生（ドラえもん)”）を含むクラスタに決定される可能性が高くなることが考えられる．

二つ目が，Wikipedia カテゴリの具体性の高さが影響していることが考えられる．具体性の高いカテゴリの“三重県の警察署”は“警察署”といった概念に“三重県の”といった具体的な情報が付与されている．したがってカテゴリ“三重

県の警察署”からは三重県と関連のある語句が得られやすくなり，対応する抽象的特徴ベクトルの成分の値は大きくなる．そうなると“三重県津市”のようにアンカー“三重県”の後ろに関連のある語句“津市”がある場合，クラスタ決定時に，カテゴリ“日本の都道府県”に属するリンク先候補“三重県”を含むクラスタよりも“三重県の警察署”に属するリンク先“三重県警察”を含むクラスタに決定されやすくなりミスが生じると考えられる．

Wikipedia をテストデータとした場合は上記のようなケースが少なかったことと，正解リンク先が Wikification コーパスのように 1 つに偏ることがなかったことが，Wikipedia での正解率の方が高かった要因であると考えられる．また，テストデータによる違いが強く影響していることがわかる．次に，用いたデータの問題点について考察する．

エンティティリンキングでは，リンク対象とする語句に固有名詞だけでなく一般名詞も含めるかどうかといった問題があるが，Wikification コーパスには比較的多く含まれていた．Wikification コーパスで多く見られた一般名詞として“先生”，“首相”などの敬称がある．敬称が文中で用いられる場合，例えば“首相”では次の文では初めに出てくる首相はエンティティの対象としては“首相”であるが，次に出てくる首相は“小泉純一郎”を指すことがわかる．

「小泉首相は二十一日，来年一月の通常国会前に国会の常任委員長や副大臣・政務官の人事を行う方針を固め，・・・．

首相は同日夜，首相官邸で記者団に対し，副大臣と政務官の交代について，・・・・．」

Wikification コーパスでは“首相”のリンク先がすべて“首相”であるが，Wikipedia では“首相”のみではなく，具体的な対象“安倍晋三”もリンク先として存在する．エンティティリンキングでは語句が何を示すかよりも語句の意味を推定するものであるので，どちらも“首相”にリンクを張るべきであり，首相がどの首相を示すかは照応解析の問題として扱うべきであると考える．Wikipedia では，編集者によってはその対象が何かを考えたうえでリンクを張っている人もいる．つまり，Wikipedia でのリンク付与基準においては照応解析も必要となる場合がある．

また，エンティティリンキングの問題として換喩をどう扱うかに関しても問題である．次の 1 文は Wikification コーパスから取得した一文であるが，“都”が示すリンク先は“東京都”となっている．しかし厳密にいえばこれは行政機関である“東京都庁”である．

「経費の六割は都の助成金だ．」

ここでエンティティリンキングではどちらに判定すべきかを考えたときに，それはエンティティリンキングが何を目的として行われているのかに依存すると考える．例えば，機械翻訳の場合，“都”を“Tokyo”と“Tokyo Metropolitan Government”どちらに翻訳すればいいかというと，どちらでも正しいがどちらかと言えば英語でも換喩的に用いることができるので“Tokyo”と訳すのが一般的であり，“東京都”に判定すべきである．一方，機械による意味理解ではより厳密に何を意味するのかを示す必要があるため後者の“東京都庁”を指すべきである．つまり，エンティティリンキングは目的に応じて設計されるべきであり，それに応じて手法やコーパスを提案するべきではないかと考える．今回用いたデータは単にリンクを張り可読性を向上させるだけであれば問題はないが，他に応用するのであれば問題があると考える．

6. おわりに

本論文では，リンク先候補の抽象的特徴の類似度で生成されたクラスタで，どのクラスタに属するかを決めた後，そのクラスタ内のリンク先候補の抽象的特徴を取り除いた特徴を用いることでリンク先決定を行う方法を提案した．実験では，抽象的特徴を取り除くことによる有効性が確かめられた一方で，クラスタリングの効果は十分ではなく，また，テストセットによって性能に大きな違いがみられた．抽象度が高いアンカーに適していないこと，Wikipedia のカテゴリに具体的な表現が付与されていることの弊害が提案手法の欠点であることが考えられる．また，Wikipedia が編集者によってリンクを張る基準が異なること，テストで用いたコーパス，提案手法がどういった目的に応用するのかを考えて設計されていないことが問題として考えられる．

謝辞

本研究は JSPS 科研費 JP15K16096 の助成を受けたものです．

参考文献

[1] Mihalcea Rada, Andress Csomai：Wikify! linking documents to encyclopedic knowledge, In Proc. of CIKM 2007, pp. 233-242（2007）．

[2] D. Milne, I. H. Witten：Learning to Link with Wikipedia, In Proceeding of the 17th ACM Conference on Information and Knowledge Management, pp. 509-518（2008）．

[3] 袁楊, 綱川 隆司, 梶 博行：決定リストの機械学習による wikification，言語処理学会第 21 回発表論文集, pp. 688-691（2015）．

[4] 菅原 拓夢，笹野 遼平, 高村 大也, 奥村 学:単語の分散表現を用いた語義曖昧性解消，言語処理学会 第 21 回年次大会 発表論文集, pp. 648-651（2015）.

[5] M. Lesk：Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, In Proceedings of the SIGDOC Conference, pp. 24-26（1986）．

[6] 林 義彦，山内 健二，永田 昌明，田中 貴秋：言語間の情報補完を用いた対訳文の Wikification，2014 年人工知能学 会全国大会，pp. 1A2-2（2014）

[7] 村上 凌悠，綱川 隆司，西田 昌史，西村 雅史：リンク先の抽象的特徴を利用した wikification の精度向上，WINF 2017