Linked Open Dataのグラフ構造特徴量を用いたエンティティ同定

柳瀬 利彦　神田 直之　今一 修　岩山 真
株式会社 日立製作所 中央研究所
{toshihiko.yanase.gm, naoyuki.kanda.kn,
osamu.imaichi.xc, makoto.iwayama.nw }@hitachi.com

1　はじめに

ニュース記事やソーシャルメディアなどの文書中に登場するエンティティ(人物や組織，製品など) を，一意な名称に統制されたエンティティのデータベースと紐付けるタスクは，文書の内容と現実世界とを橋渡しするために重要なタスクである．このタスクは，エンティティの曖昧性解消問題 (Entity Disambiguation) やエンティティの同定問題と呼ばれている．

本研究では，図1に示すように，文書としてはTwitterのようなソーシャルメディアでみられる短文を対象とし，また，エンティティのデータベースとしては，Linked Open Data (LOD) として公開されているグラフ構造を持つデータベースを対象とする．

短文をエンティティ同定の対象にした場合には次の二つの問題が生じる．一つ目は，短文では多くの場合エンティティの正式名称が登場せず，略称や名称の一部のみ表記されることである．二つ目は，共起語などの文脈情報が十分にとれないことである．これらの問題を解決するためには外部情報が必要であり，本研究では同定先であるLODの情報を用いる．

LODには，DBPedia[1] やFreeBase[2], BabelNet[3], YAGO[4] などのResource Description Framework(RDF)[3] で記述されたものが存在する．これらのLODでは，エンティティ間の関係性が含まれている．エンティティをノード，関係性をエッジとしてみるとグラフ構造となり，この情報を同定の際の外部情報として利用する．また，DBPediaやYAGO，BabelNetは，Wikipediaから情報を抽出しており，Wikipediaの持つ知識の網羅的な集積という性質を受けついでいるため，同定先として適している．

[1]http://dbpedia.org/
[2]http://www.freebase.com
[3]http://www.w3.org/RDF/

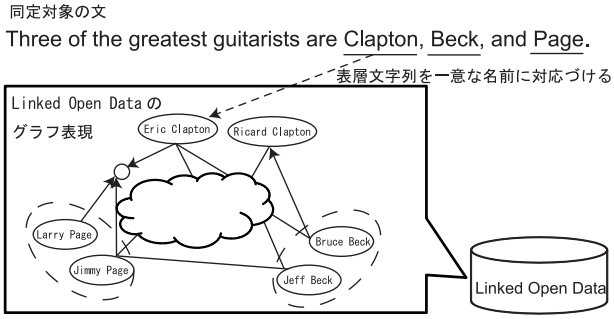

図 1: 短文中のエンティティ同定のタスク設定

一方，LODはそれぞれ別の団体によって管理されているため，構造がLODごとに異なり，また時間の経過とともに編集されていく．そのため，同定には，LODが共通して持つ特徴を利用することが望ましい．

本研究では，RDFで記述されるLODに共通する構造に注目し，文書におけるtf, idfに類似した概念のスコアであるグラフ構造特徴量を定める．そしてグラフ構造特徴量を用いてエンティティを同定する方法を提案する．提案手法を小規模なデータセットに適用し，予備的な評価を行った．

2　関連研究

エンティティ同定の手法としては，エンティティが登場する文書と，対象文書との類似性を用いる方法が提案されている [5]．この方法では，エンティティのデータベースに，グラフ構造を仮定していないため適用範囲が広いという利点がある．一方で，短文の場合には十分な精度で文書間の類似度比較が行えないという問題が起こる可能性がある．

また，LODを利用した手法として，部分的なグラフ構造を利用する方法が提案されている [2]．この手法で利用されるグラフは，LOD全体ではなく，Wikipediaページの参照関係グラフである．このような部分グラ

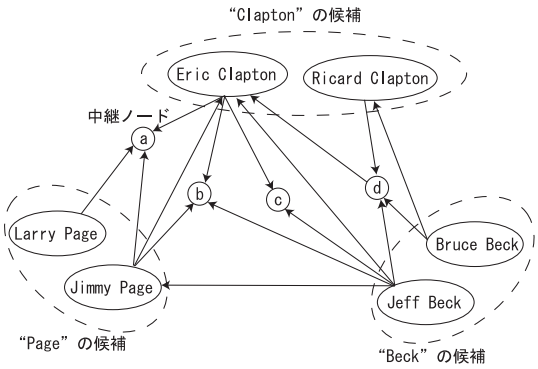

図 2: エンティティ同定の候補と中継ノードが作るグラフの例

フを作るには，LOD にどのような情報が含まれているのか精査する必要がある．

3 提案手法

提案する同定手法においても，LOD の大規模なグラフから，エンティティと関係する部分グラフを抽出し，スコアを計算するが，その際にノードのラベルと接続関係という LOD に共通して見られる特徴を用いる．

LOD のグラフ構造は，RDF の主語，述語，目的語の三つ組から作られる．主語と目的語がノードとなり，主語から目的語に向かって有向エッジが作られる．以下では，まず，部分グラフとスコア計算の詳細を説明し，次に全体の手順を説明する．

3.1 グラフ構造特徴量

図 2 を例として，部分グラフについて説明する．図 2 の部分グラフは Clapton, Page, Beck を表層文字列とする 3 種のエンティティを決めるためのものである．グラフは，エンティティの候補 (Larry Page, Jimmy Page など)，候補と候補が共通して持つプロパティ (中継ノード a, b, c, d)，それらを結ぶエッジから構成される．

この部分グラフから候補の組 (例えば，Eric Clapton, Jeff Beck, Larry Page) を取り出し，グラフ構造特徴量と呼ぶスコアを計算する．全ての候補の組にグラフ構造特徴量を計算した後に，最大値をとる組を同定の解とする．

候補の組 (Eric Clapton, Jeff Beck, Larry Page) に対するグラフ構造特徴量の計算では，候補の組からエンティティのペア (Eric Clapton と Jeff Beck, Jeff Beck と Larray Page, Larry Page と Eric Clapton) を作り，それぞれのペアについて定義されるスコアを計算し，総和をとる．

エンティティ i の候補集合の要素を c_i，エンティティ j の候補集合の要素 c_j とすると，c_i, c_j ペアに対するスコアを 4 種類定めた．1 種類目は，c_i, c_j を結ぶエッジの数 $s_{direct}(c_i, c_j)$ である．従来研究 [2] でも用いられた指標であるが，本手法では Wikipedia 本文のような文書の付加情報は用いないため，重みはすべて 1 としている．2 種類目は，c_i, c_j ペア間を結ぶ中継ノードの数 $s_{common}(c_i, c_j)$ である．ここで，$s_{direct}(c_i, c_j)$ と $s_{common}(c_i, c_j)$ は，共通する中継ノードが多いほど，エンティティ間の関係性が高いという考えに基づいている．これは，文書の場合の tf に類似した指標であると考える．

3,4 種類目の指標では，まず c_i と c_j の中継ノード v に重みを付ける．これら重みを $igef(v)$ と $igcf(v)$ と呼び，ひとつのエンティティを一文書として考えた場合の idf に相当する．これは「多くのエンティティに共通して接続される中継ノードは弁別能力が低い」という考え方に基づいている．2 つの指標があるのは，母集団をエンティティ全体と考えた場合と，同一エンティティの候補内で考えた場合に分けて考えたためである．

母集団がエンティティ全体 E の場合を式 (1) に定める．

$$igef(v) = \log(|E|/gef(v)) \tag{1}$$

ここで $|E|$ はエンティティの個数であり，$gef(v)$ は v がいくつのエンティティと接続しているかを表す．

母集団が同一エンティティの候補集団 C であった場合を式 (2) に定める．

$$igcf(v) = \log(|C|/gcf(v)) \tag{2}$$

ここで $|C|$ はあるエンティティの候補数であり，$gcf(v)$ はあるエンティティの候補がいくつ v に接続しているかを表す．

c_i, c_j ペアのスコアとしては，すべての v に対する重みの和を取り，$s_{igef}(c_i, c_j) = \Sigma_v igef(v)$, $s_{igcf}(c_i, c_j) = \Sigma_v igcf(v)$ と定めた．

3.2 手順

提案するエンティティ同定の処理手順を図 3 に示す．システムは，固有表現抽出などの手法で抽出された，文中に登場するエンティティの組を入力とし，各エンティティに対応する Uniform Resource Identifier(URI) の組を出力とする．最終的な同定には，グラフ構造特徴量を用いているが，その計算には候補の組合せ数に比例した計算時間がかかる．そのため，前処理により候補を絞り込む．

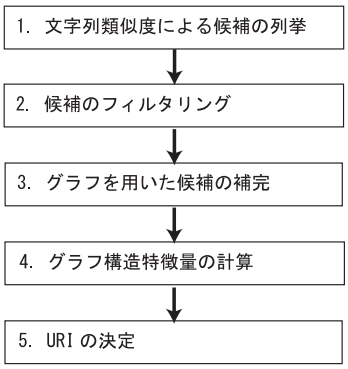

図 3: エンティティ同定の処理手順

システムは，候補の列挙，候補のフィルタリング，候補の補完，グラフ構造スコアの計算，URI の決定の 5 つのコンポーネントから成っている．それぞれの機能を説明する．

候補の列挙では，文中の表層文字列と，LOD のラベル文字列との類似度に基づいて URI の候補を列挙する．LOD のラベル文字列は，事前に全文検索エンジンに登録してあり，表層文字列をクエリとして，ラベルを検索する．

次に，候補のフィルタリングでは，候補の重要性で対象を絞り込む．具体的には，エンティティごとに，候補を次数でソートし，上位 n_t 件を抽出する．次数でソートする理由は，次数が多いノードは多くのエンティティに影響力があり重要であると仮定したためである．

3 番目に，前段で取り漏らされた候補を補う．各候補のプロパティを調べたときに，他のエンティティの候補となっているものは，グラフ構造特徴量の計算対象として抽出した．各エンティティに対して追加する候補の上限を n_r 件とした．

4 番目に，前節で述べた方法で，LOD から部分グラフを作成し，グラフ構造特徴量 $s_{direct}, s_{common}, s_{igef}, s_{igcf}$ を計算する．

最後に，グラフ構造特徴量をもとに，エンティティの URI を決定する．本研究では，予備実験をもとに 2 段階のソートを用いた．はじめに最も強いつながりを表すと考えられる s_{direct} でソートし，次に $s_{common}, s_{igef}, s_{igc}$ のいずれかで 2 次ソートし，エンティティの URI を求めた．

4 評価実験

4.1 実験設定

評価用データとして，KORE50 データセット [1] と，YAGO2S の 2012-12-01 Wikipedia から作成されたバージョンを利用した．KORE50 は，50 個の独立した文からなる．KORE50 データセットの文の例を下記に示す．

“Three of the greatest guitarists started their career in a single band: Clapton, Beck, and Page.”

この文中の Clapton, Beck, Page が IOB2 形式でタグ付けされ，エンティティであることが示されている．各タグには表層文字列と，正解の YAGO2 の URI が付加されている．ただし，人物や組織の区別は無い．本実験では，エンティティタグとその表層文字列が既知であるとする．

表 1 にデータセットの概要を示す．エンティティの総数は 144 であり，そのうち手元の YAGO2S データセットに URI が含まれていた 46 文，132 件を名寄せの対象とした．また，エンティティの異なり数は，手元の YAGO2S 中に存在するもののみでカウントした．

表 1: 実験に用いた文書

項目	-
文	46
エンティティ総数	144
URI 付きエンティティ	132
1 文あたりのエンティティ	2.64
エンティティの異なり数	116

計算機環境は，CPU が 4 コア 2.93GHz，メモリが 8GB，HDD が 1TB の一般的な PC を用いた．OS は 64bit 版の Ubuntu 12.04LTS である．実装としては，文字列類似度による候補の列挙に，全文検索エンジン Solr 4.0 を，グラフ構造特徴量の計算に WebGraph [4] を用いた．

実験は，ベースライン手法として，文字列類似度が最大のものを答えとした場合，次数が最大のものを答えとした場合の 2 種類を評価とした．また，提案手法として，ベースラインに加えて，グラフ構造特徴量を組み込んだ手法を評価した．

4.2 結果

評価結果を表 2 に示す．n_t, n_r の値は，予備実験からそれぞれ 5 と 3 とした．精度は，正解したエンティティの数を全エンティティ数で割った値とした．また，平均順位は，エンティティごとに正解候補の順位を求め，平均を計算した．なお，図 3 の手順 3 の段階で正解エンティティが候補に含まれなかった場合には，平均順位の計算に含めなかった．完全一致は，文中の全てのエンティティ正解した文数を，全文数で割った値とした．

[4] http://webgraph.di.unimi.it/

表 2: KORE50 データセットでの精度評価結果

項番	指標	精度 (%)	平均順位	完全一致 (%)
1	文字列類似度	15.9	208	2.0
2	1 + 次数最大	34.8	40.5	3.0
3	$2 + s_{direct}$	62.9	1.58	16.7
4	$3 + s_{common}$	63.6	1.58	15.9
5	$3 + s_{igef}$	64.4	1.55	16.7
6	$3 + s_{igcf}$	62.9	1.58	16.7

5 考察

5.1 各特徴の有効性

文字列類似度による同定では，1 エンティティあたり平均 1375 個と，多数の候補が抽出された．候補の上位を見ると，類似度スコアが同じであり，区別できていないことが確認された．これは，例えば，人名の場合，姓か名の部分一致が大量に起こるためである．

次に，候補の次数を考慮することで，精度は約 20 ポイント改善した．しかしながら，s_{direct} のようにグラフ構造スコアを考慮する場合と比べて精度は約半分程度と低い．これは，Steve Jobs と Steve Ballmer のように，同名の有名人が多く存在するためであると考えられる．

最後に，グラフ構造スコアを考慮した場合では，約 62〜63%の精度で同定することができた．手法間の差は約 1 ポイントであり，エンティティに換算して 1, 2 個の差であるため，手法の比較にはより大規模なデータでの実験が必要である．一方で，正解エンティティの平均順位は約 1.6 であり，多くの場合で 1 位か 2 位の間に入っていることがわかる．これは，自動的に決定するには不十分な精度であっても，人間の作業者の支援という使い方では，十分な性能である．一方で，グラフ構造スコアを計算事前の候補絞り込みで 132 件中 22 件 (16.7%) の漏れが発生している．漏れの少なさと計算時間は，候補の限定のパラメータである n_t, n_r に依存し，データごとに調整が必要である．

5.2 実行時間

グラフ構造スコアの計算は，エンティティの候補の組み合わせを評価するため，エンティティ数の階乗のオーダーで急速に大きくなる．しかしながら，本実験の場合は，グラフスコア計算部の実行時間は全文合計で約 450 秒，1 文あたりの 9 秒程度であり，現実的な時間内に計算できた．これは，1 文が短く，多くの文でエンティティ数が 3 以下と小さかったこと，また，候補を事前に絞り込んだことが要因である．このように，短文など候補数が絞り込める場合には，多くの組合せを計算するアルゴリズムでも，現実的な時間での実行が可能である．

6 おわりに

短文中に登場する人物や組織などのエンティティを LOD の URI に対して同定するため，グラフ構造特徴量を用いる同定手法を提案した．実験では，各エンティティが高い曖昧性を持つ文書データセット KORE50 と，LOD の YAGO2S を用いて有効性を確認した

本研究では，文書データセットのサイズが 46 文と小さいため，有効性を統計的に評価するためには十分ではない．今後は，より大規模なデータを用いて検証を行う．

また，本研究では，同定に用いる特徴としてグラフ構造のみを用いた。今後は，短文の場合にも利用可能なコンテクスト情報を取り入れることで，さらに精度を改善したい。

提案した同定手法が言語に依存する部分は，文書中の表記と LOD のラベルとの類似度を計算する部分のみである．そのため，本同定手法は，英語以外の言語に適用することは容易であり，今後は，日本語を含めた多言語での評価を行いたい．

参考文献

[1] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. Kore: Keyphrase overlap relatedness for entity disambiguation. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 545–554, 2012.

[2] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 782–792, 2011.

[3] R. Navigli and S. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[4] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago - a core of semantic knowledge. *16th international World Wide Web conference*, 2007.

[5] 小林のぞみ, 松尾義博, 菊井玄一郎. 共起語を手がかりとした固有表現とデータベースレコードの対応付け. 言語処理学会第 15 回年次大会, pp. 821–824, 2009.