1K3-3

関係抽出技術の実用性の検討
Study on practical use of relation extraction

谷口 元樹[*1] Motoki Taniguchi
三浦 康秀[*1] Yasuhide Miura
林 光雄[*1] Mitsuo Hayashi
大熊 智子[*1] Tomoko Ohkuma

[*1] 富士ゼロックス(株)研究技術開発本部コミュニケーション技術研究所
Communication Technology Laboratory, Research and Technology Group, Fuji Xerox Company, Ltd.

This paper investigates the feasibility of a relation extraction problem by analyzing the two cases of it. In the first case, we demonstrate that the attributes of criminal person can be extracted by supervised relation extraction. In the second case, we develop a distant supervised relation extraction system where training data is automatically labeled to decrease the production cost of hand-labeled corpus. In addition to a previously proposed distant supervision method, we deploy a simple entity extraction method to obtain entity pairs of a knowledgebase from unlabeled texts. An efficacy of the method is evaluated by the quantity and quality of the obtained entity pairs. We found that the simple method outperforms the previously proposed method in terms of the training data size and the relation classification precision.

1. はじめに

テキストから知識を抽出する手段としてエンティティ間の関係を抽出する関係抽出タスクは ACE(Automatic Content Extraction)、TAC KBP(Text Analysis Conference Knowledge Base Population)などの評価型ワークショップを中心に多くの研究が行われてきた. これらの研究では Web などの膨大で日々作成されるテキスト集合から関係を抽出することで, 大規模な知識ベースを自動的に構築することを目的としている. また, これまでの研究から関係抽出タスクの難易度は高いことが知られている[ボレガラ 2012].

エンティティ間の関係の中でも人物の属性の関係に注目し, 人物情報を抽出する試みが盛んに行われている.人物の属性情報を抽出し, 人物情報データベースを自動的に構築できる. これにより, ある人物に関する情報を知りたい時に人名を検索するだけで, その人の属性情報を全て知ることができる. また, 人物の属性情報を活用することで, テキスト中の人物の名前の曖昧性を解消することができる.

本論文では2つの事例を通して, 関係抽出を利用した人物情報の抽出の実用性を検証する. 事例1では, テキストや関係を限定した人物情報の抽出として, 新聞記事からの犯罪者の属性抽出の精度の評価とエラー分析を行い, 限定的な用途であれば実利用に耐えうる精度を達成できるか検討する. また実用的な精度達成の課題は学習データの作成コストが大きいことであることを示す. 事例2では人的なコストをかけずに学習データを作成するために, 学習データの自動獲得手法に取り組む. 既存の手法である知識源のエンティティとの一致によって獲得される学習データと, 単純な文字列一致によって獲得される学習データの量と質の比較を行う. またこれら学習データを用いた関係分類の精度を評価し, 単純一致によって獲得された学習データを用いた関係分類の精度がより高いことを実証する.

2. 事例1:犯罪者の属性抽出

本章では, 対象となるテキストや関係を限定したタスクに取り組むことで限定的な用途での実用性を検討する.

関係を抽出するテキストとして新聞記事を選択した. 新聞記事は記者がテキストを書いているため, Web などと比較すると標準的な表記で書かれやすく, 形態素解析などの誤りは少ないためである. また日々作成されているため, 大量のテキストを得やすいことも理由にあげられる.

抽出する関係は犯罪者の属性を選択した. これは, 新聞記事には事件の報道記事が多く, また犯罪者であれば名前や年齢などの属性が記述されやすいためである. 具体的な犯罪者の属性は名前, 年齢, 住所, 国籍, 職業, 組織, 罪名, 容疑の 8 つを定義した.

2.1 犯罪者属性コーパス

毎日新聞 10 年分から“逮捕”, “容疑”, “被告”のいずれかの単語を含む 7800 記事をランダムに抽出した. この新聞記事テキストに対して, 下記の 3 種類のアノテーションを行い, 犯罪者属性コーパスを作成した.

1. エンティティ
犯罪者の属性であるかに関わらず, エンティティは名前, 年齢, 住所, 国籍, 職業, 組織, 罪名タグをアノテーションした. 関根の拡張固有表現階層[1]に基づきタグを付与した.
2. 名前とその他の属性の関係
犯罪者の名前とその他の属性をエンティティ間の関係として付与した.
3. 犯罪者フラグ
名前に対して犯罪者である場合には犯罪者フラグをつけた.

図1にアノテーション付与例を,表1には各タグのエンティティ数を示す.

表 1 エンティティ数

タグ	名前	年齢	住所	職業	国籍	組織	罪名	容疑
数	17,609	6,245	8,064	20,522	562	19,052	6,628	8,827

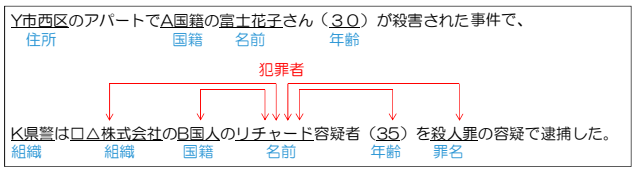

図 1 犯罪者コーパスのアノテーションの付与例

連絡先:谷口元樹, 富士ゼロックス㈱研究技術開発本部, motoki.taniguchi@fujixerox.co.jp

[1] https://sites.google.com/site/extendednamedentityhierarchy/

表 3 精度の評価結果

	ステップ1 エンティティ抽出精度			ステップ2 関係分類精度			システム全体 関係抽出精度		
	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
人名	0.98	0.93	0.94	-	-	-	-	-	-
年齢	0.97	0.98	0.97	0.90	1.00	0.93	0.87	0.87	0.87
住所	0.91	0.89	0.90	0.89	0.88	0.89	0.82	0.81	0.81
職業	0.87	0.83	0.85	0.98	0.98	0.98	0.83	0.82	0.82
国籍	0.74	0.56	0.64	0.94	0.92	0.93	0.80	0.79	0.79
組織	0.83	0.77	0.80	0.86	0.70	0.77	0.68	0.57	0.62
罪名	0.84	0.81	0.82	0.88	0.92	0.90	0.77	0.80	0.78
容疑	0.97	0.99	0.98	0.97	0.93	0.95	0.85	0.81	0.83

2.2 関係抽出手法

犯罪者の属性抽出は機械学習を利用した手法を用いており，2つの計算ステップからなる．ステップ 1 では犯罪者の属性値の候補となるエンティティを抽出する．ステップ 2 では犯罪者の名前とその他の属性を関係であると捉え，同一文中のエンティティのペアに対して関係の有無を分類する．

ステップ1では、固有表現抽出でよく用いられる条件付き確率場を文字単位で用いる．素性は，文字の表層，文字種，形態素の表層，形態素の基本形，形態素の品詞，形態素の活用の種類，形態素の活用形を用いる．

ステップ2の関係分類は属性ごとに 2 値の分類器を作成する．分類器は Support Vector Machine を用いる．素性は下記のものを用いる．

・第1，第 2 エンティティ間の単語列とその品詞列
・第1，第 2 エンティティ間の文字数と形態素数
・名前とその他のエンティティどちらが先に出現しているか
・第1エンティティの左側k個分の単語、品詞 (k=1,2,3)
・第2エンティティの右側k個分の単語、品詞 (k=1,2,3)

ただし，名前とその他の属性のうち文中の出現順が先のものを第1エンティティ，後にあるもの第2エンティティとした．

2.3 負例の作成

ステップ2の関係分類器の学習では負例が必要となるが，犯罪者コーパスでは正例しかアノテーションされていない．そこで，犯罪者の名前と犯罪者の属性ではないエンティティを負例とした．図1の例文における年齢の関係の分類においては，“リチャード”と“35”は正例となり，“リチャード”と“30”は負例となる．表 2 に作成されたエンティティペアの数を示す．

表 2 エンティティペア数

	年齢	住所	職業	国籍	組織	罪名	容疑
正例	3,738	1,202	3,859	120	393	3,015	3,237
負例	2,839	2,827	1,733	85	4,899	924	3,969

2.4 評価

関係抽出精度の評価では計算ステップ毎に評価した．いずれのステップでも5分割交差検定で評価を行った．また，1文中に含まれるエンティティペアが犯罪者の属性であるかどうかで評価した．評価結果を表 3 に示す．

ステップ1のエンティティ抽出精度は国籍に関しては低いものの，その他の属性に関してはおおよそ 0.8 以上を達成している．国籍が低い原因は表1で明らかなように，その他のエンティティと比較して数が少ないためである．

ステップ2単独の関係分類の精度はステップ1のエンティティは全て正しく抽出できているものとして評価を行った．組織以外の関係はおおよそ 0.9 以上と高い精度を達成している．組織エンティティの数は多いが，組織を表すエンティティペアの正例数は少なく，また正例と負例の数に不均衡がある．これらの要因から組織の分類精度が低くなったものと考えられる．

ステップ1と2全体を通した関係抽出の精度は組織以外の関係ではおおよそ 0.8 以上を達成している． 関係抽出の精度は基本的にはステップ1とステップ 2 の精度の積で近似できるため，いずれのステップでも精度の低い傾向にあった組織は全体を通しても精度が低いものと考えられる．

2.5 考察

テキストを新聞記事に，関係を犯罪者の属性に限定することで，組織を除いておおよそ 0.8 以上の抽出精度を達成することができた．組織で関係抽出精度が低い要因は学習データ中の正例の少なさであった．この関係はテキスト中の出現頻度が低いため，大量の正例データを作成するためには，膨大なテキストに対してアノテーションをする必要がある．このため，実用上の課題は学習データ作成のコストが大きいことである．

3. 事例2：学習データの自動獲得

本章では，学習データの作成コストが大きいという課題を解決するために，学習データの自動獲得に取り組む．

近年，関係抽出において Distant Supervision を用いることで学習データの自動獲得する研究が盛んに行われている[Mintz 2009]. Mintz らは既存の知識ベースを利用し，エンティティペアの関係を表す文を自動獲得する手法を提案している．この手法では，知識ベースに登録されているエンティティペアを含む文はそのエンティティペアの関係を表している，というヒューリスティックなルールを用いて，ラベルがついていないテキストに自動的にラベル付けを行う．例えば，“富士ゼロックス”と“1962 年”が“設立年度”という関係であることが知識ベースに登録されていた場合，“富士ゼロックス”と“1962 年”をエンティティとして含む文は全て“設立年度”という関係を表す文であるとラベル付けする．これにより，人的なコストをかけずに大量の学習データを自動的に獲得することができる．

これまで Mintz らの手法を改良した研究が様々行われているが，対象言語は主に英語であり，日本語に適用された例はない．そこで本事例では，Mintz らの手法を日本語に適用し，その実用性を検証する．

これまでの研究では，知識ベース中でエントリ数の多い関係のみを評価の対象にしている．しかし，関係抽出の実応用では，先に抽出する関係が定義される．したがって，本事例では予め関係として人物の属性を定義し，関係を抽出する．

3.1 知識源

Mintz らの手法では Freebase[2]を既存の知識ベースとして利用している．しかし Freebase は英語の知識ベースであり，日本

[2] https://www.freebase.com/

表 4 DBpedia のエンティティペア数と獲得されたエンティティペア数

関係クラス	DBpedia の エンティティペア数	獲得エンティティペア数(一致率)		
		文書内 単純一致	文内 単純一致	文内 エンティティ一致
配偶者	5,052	516(10.2%)	70(1.4%)	38(0.8%)
職業	12,691	2,294(18.1%)	630(5.0%)	357(2.8%)
生誕地	3,049	238(7.8%)	20(0.7%)	11(0.4%)
生年月日	15,795	985(6.2%)	27(0.2%)	2(0.0%)
死没地	2,447	218(8.9%)	21(0.9%)	14(0.6%)
没年月日	3,591	387(10.8%)	64(1.8%)	34(0.9%)
全体	42,625	4,638(10.9%)	832(2.0%)	456(1.1%)

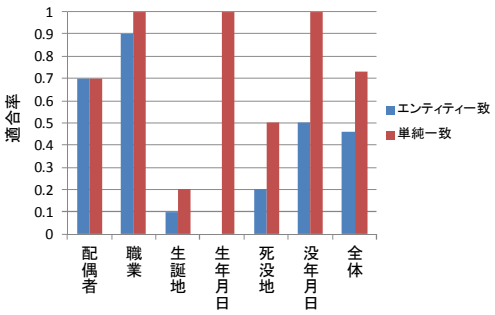

図 2 学習データの適合率

語のものはない. 本事例では代替として Wikipedia の Infobox から半自動的に構築された日本語の知識ベースである DBpedia Japanese[3]の 2014 年 1 月のダンプデータを用いる.Wikipedia の記事中の Infobox の多くは, その記事が属するジャンルのテンプレートを基本に構成されている. 例えば “Infobox 人物”テンプレートでは氏名, ふりがな, 生年月日, 生誕地, 職業などの関係が定義されており、様々な人物に関する記事の Infobox に利用されている. 今回はこの“Infobox 人物”で定義されている関係のうち, DBpedia にエンティティペア数が多い関係である“職業”, “配偶者”, “生誕地”, “生年月日”, “死没地”, “没年月日”の6つを対象とした.

3.2 Distant Supervision による学習データの自動獲得

Mintz らの手法では, まず Wikipedia のテキストからエンティティを抽出する. 同一文中にあるエンティティのペアを知識ベースに参照し, 知識ベースに登録されているエンティティペアと一致した場合はその文を正例, 一致しない場合は負例としてラベル付する. このラベル付けされた文を学習データとして関係分類器の学習を行う.

従来研究とは異なり, 本事例では人物に関する関係クラスのみを対象としているため, 獲得される学習データの量が少ないことが予測される. そこで, Mintz らの手法に加えて, エンティティの抽出は行わずに, DBpedia にあるエンティティペアを単純に文字列として含む文を正例として学習データを獲得する. この手法では, 正しくないペアも正例としてしまう可能性があるが, エンティティ抽出の漏れをなくすことができるため, データ量を増やすことができる.

3.3 学習データの量と質の評価

抽出されたエンティティのみを知識ベースの一致をとる対象とする“エンティティ一致”, エンティかどうかに関わらず全ての文を対象とする“単純一致”の2つの獲得方法による学習データの違いを量と質を評価する.

エンティティの抽出には事例1のステップ1で用いた条件付き確率場による手法を用いた. エンティティ抽出の学習データは拡張固有表現タグ付きコーパス[橋本 2008]のうち新聞記事を利用し, エンティティのタグは関根の拡張固有表現の最下層のタグを用いた. 学習データ構築テキストは Wikipedia10 万記事, 知識源は DBpedia のエンティティペアの 50%を用いた.

学習データの獲得数を表 4 に示す. エンティティ一致と比較して, 単純一致はおおよそ 2 倍を獲得できた. しかし, 一致率で見ると 2%であり, DBpedia の多くのエンティティペアが学習データとしては獲得されていない. これは Wikipedia の記事には DBpedia のエンティティペアが書かれていないためである. 表 3 の“文書内の単純一致”は一致をとる対象を文ではなく, 記事内とした場合である. この一致率を見ると, 10.9%であり, 90%近くの DBpedia のエンティティペアに関する記述が Wikipedia の記事中にはないことが分かる.

自動獲得された学習データの文がエンティティペアの関係を実際に正しく表しているかを人手で評価することで, 学習データの質を評価した. 人物に関する 6 つの関係クラスから各 10 組のエンティティペアとそのエンティティペアが含まれている Wikipedia の文を 1 文ずつランダムに抽出した. ただし, エンティティ一致の“生年月日”の関係は獲得されたデータ数が 2 であるため, サンプル数は2である. 人手での評価結果を正解とした各関係の適合率を図 2 に示す. 1関係あたり 10 サンプルであるため, 統計的に有意であるとは言えないが, いずれの関係においても単純一致の適合率がエンティティ一致を上回っている.

3.4 関係分類

関係分類においては多クラスのロジスティック回帰を用いる. また下記の素性を用いる.

・エンティティペア間の単語列と品詞列
・エンティティの出現順序
・第1エンティティの左側k個分の単語、品詞 (k=1,2,3)
・第2エンティティの右側k個分の単語、品詞 (k=1,2,3)

ただし, エンティティペアのうち文中の出現順が先のものを第1エンティティ, 後にあるもの第2エンティティとしている. また上記の素性をエンティティペアが含まれる文毎に抽出し, エンティティペアが含まれる文全体で統合して一つの素性ベクトルとする.

3.5 学習・評価

DBpedia のデータを二分割し, 一方のデータで学習データの自動獲得に, もう一方の DBpedia のデータを評価データに用いることで自動的な精度評価を行う.

関係抽出タスクとして評価を行う場合には, エンティティ一致による学習データ獲得時と同様の処理を行うことで評価データを作成する必要がある. しかし, 3.3 節で述べたように, 十分な量のデータを作成することが難しいため, 関係分類タスクとして評価する. 関係分類タスクではエンティティペアを入力としてそのエンティティペアを適切な関係に分類する.

関係分類では“関係なし”を表すエンティティペアである負例が必要となる. 提案手法では, DBpedia に登録されているエンティティペアをランダムに2つ選択し, 互いのエンティティを交換することで作成されたエンティティペアが DBpedia に登録されていなければ負例とする. 学習データの負例数は 1000, 評価データの負例数は 500 とした.

学習データ量による分類精度の違いを比較するため, 学習データ構築のテキストは Wikipedia10 万記事と 70 万記事を用

[3] http://ja.dbpedia.org/

いる．また学習時とは別の Wikipedia の 34 万記事を用いて評価データを構築する．

精度比較のためのベースラインは最もインスタンス数が多いラベルである“職業”を全評価データに出力した場合の値とする．

3.6 評価結果

図 3 に関係分類精度の評価結果を示す．分類されたエンティティペアのクラス所属確率の高い順に 1 組ずつ真偽を判定し，正しく人物の関係に分類されたエンティティペアの組数に対する適合率をプロットしている．

学習データの獲得方法による違いを比較すると，エンティティ一致よりも単純一致の精度が高いことがわかる．これは 3.3 節で述べたように，単純一致の学習データの量は 2 倍であり，またデータの質も良いためであると考えられる．ただし，評価データ全体の適合率はいずれもベースラインよりも低いため，分類器としては十分な精度には達していない．

学習データ構築テキスト量の精度への影響をみるために，学習データ構築テキストを 70 万記事に増やすと，正しく分類されたエンティティペア数が多い領域では精度が向上しており，ベースラインの適合率を超えている．表 5 に学習データ構築テキスト 70 万記事における学習データの量と評価データ全体の関係クラス別の分類精度を示す．テキスト量を 7 倍に増やしたことにより，自動獲得される学習データの量が 11 倍に増えている．また学習データ量が多い関係クラスである“職業”，“生年月日”は高い精度を達成できているものの，学習データ量が少ない関係クラスでは精度が低いことがわかる．

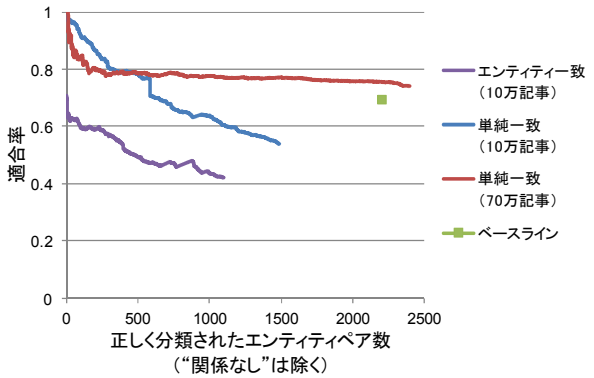

図 3 関係分類精度

表 5 クラス別の分類精度と学習データ量

関係クラス	適合率	再現率	F 値	学習データ量
職業	0.79	0.90	0.84	4,343
配偶者	0.50	0.39	0.44	589
生誕地	0.27	0.15	0.19	255
生年月日	0.78	0.77	0.77	3,029
死没地	0.21	0.06	0.09	209
没年月日	0.63	0.31	0.41	901
全体	0.74	0.74	0.74	9,326
関係なし	0.66	0.62	0.64	1,000

3.7 考察

エンティティ一致による学習データを用いた分類器よりも単純一致による学習データを用いた分類器が高い精度が達成できることがわかった．しかし，単純一致による学習データを用いた分類器においても獲得される学習データ量が少ないクラスの分類精度は実用的なレベルには達していない．この学習データ量が少ない理由は 3.3 節で述べたように DBpedia の多くのエンティティペアに関する記述が Wikipedia 記事にはないことがあげられる．この書かれない要因は(1)学習データ構築テキスト：Wikipedia，(2)知識源：DBpedia の二つに分けられる．

(1) 学習データ構築テキスト：Wikipedia

あるエンティティに関する Wikipedia の記事では，そのエンティティ名は記事の最初の文にしか書かれないことが多く，それ以降には書かれない．つまり，人物の名前は記事の最初の文にしかかかれず，それ以降には人物の名前が省略された形で，属性に関する記述が続くことがある．このため，同一文中のエンティティペアが書かれている文を学習データとして獲得する今回の手法では，十分な量が得られなかったものと考えられる．

(2) 知識源：DBpedia

英語で用いられている Freebase と比較すると日本語の DBpedia に蓄積されたエンティティペアの数は少なく，また誤ったエンティティペアも多い．誤ったエンティティが発生することは DBpedia が Wikipedia の Infobox から半自動的にマッピングされて作られていることに起因する．エラー分析を行ったところ，エンティティの範囲の誤りと種類の誤りが多く見られた．範囲の誤りでは“・ゲイツ”のようにエンティティが途切れているもの，“小説家，評論家，文学者”のように複数のエンティティがまとめられているものがあった．種類の誤りでは，“夏目漱石”と“東京”に対する“生年月日”という関係のように，関係の要素として不適切な種類のエンティティであるものが見られた．

4. おわりに

関係抽出の実用性を検討する目的で 2 つの事例に取り組んだ．

事例1ではテキストや関係を限定することで，用途によっては実利用に耐えうる精度を達成できることを示した．しかし，高い精度を達成するためには多くの学習データが必要であることがわかった．大量の学習データを人手で作成することは非常に人的コストがかかるため，実用上の課題となる．

事例2においては Distant Supervision 手法を用いることで学習データを自動獲得する手法に取り組んだ．既存手法に加えて，単純一致による学習データの獲得を行った．この結果，既存手法よりも単純一致による学習データの方が量と質が優れていることを明らかにした．またこの学習データを用いた関係分類の精度を評価し，精度がより高いことを実証した．これにより，日本語の関係分類タスクにおいても，学習データが大量に獲得できれば実用的な精度は達成できることを示した．

エラー分析を行い，十分な量の学習データを獲得できていない原因が Wikipedia の書かれ方，DBpedia のノイズの多さであることを示した．今後は Wikipedia 以外のテキストを学習データ構築テキストとして用いること，DBpedia のノイズを削減する手法に取り組むことで精度の改善を行いたい．

参考文献

[Mintz 2009] M. Mintz, S. Bills, R. Snow, D. Jurafsky: Distant supervision for relation extraction without labeled data, Proceedings of Association for Computational Linguistics (ACL) '09, ACL, 2009.

[ボレガラ 2012] ボレガラダヌシカ, 谷直紀, 石塚満: 属性値間の関連性を用いた属性抽出の精度向上, 人工知能学会論文誌 27 巻 4 号, 社団法人人工知能学会, 2012 .

[橋本 2008] 橋本泰一, 乾孝司, 村上浩司 : 拡張固有表現タグ付きコーパスの構築, 情報処理学会研究報告 第 2008-NL-188 巻 , 社団法人情報処理学会, 2008.