E-047

新聞記事からの事件情報抽出
Crime Information Extraction from Newspaper Articles

柴田　裕亮†　　山村　毅
Yusuke Shibata　Tsuyoshi Yamamura

1. 序論

我々は電子化して蓄積・配信されている膨大な文書から重要な情報だけを抜き出して整理したいという要求に対して，例えば一覧表のような視覚的に分かりやすい表示形式で情報を提示するということを考えており，その際に必要となる各種情報を抽出することを試みている．

本研究では新聞記事の犯罪事件に関する文書を対象としている．このような文書における重要な情報の一つとして，人物情報がある．我々はまず，誰が何をしたか，つまり誰が犯人で誰が被害者なのかを判別して抽出する方法を提案する．

従来の研究[1]では固有名詞，つまり名前で表現された人物に対してのみ犯人と被害者の判別を行っている．一方，本研究では，犯人と被害者の名前の記述が無く人物を表す名詞で置き換えられている場合にも，犯人と被害者の判別をして抽出する．犯罪事件記事では必ずしも固有名詞が出現するとは限らないので，このような処理は非常に重要であると考える．

2. 新聞記事の分析

記事中の人物は，人名か「男」「少年」「会社員」のように人物を表す名詞や，職業で表現される．これらを犯人，被害者と判断するためには個々の単語だけでなく，前後の文脈も見なければならない．実際に，インターネット上で配信されている犯罪事件記事を'06 の毎日新聞，朝日新聞，読売新聞から計 150 記事収集し，分析したところ以下のような傾向があった．

① 「容疑者」や「被告」という語の直前に犯人の名前が記述される
② 犯人は「逮捕する」の目的語，「逮捕される」の主語として記述される
③ 犯人や被害者が取った，あるいは受けた行為についての記述がある

3. 犯人と被害者の抽出

①は特定のキーワードを検索する事で容易に抽出する事が可能であると考えられる．

②と③では，犯人及び被害者は，記事の係り受け関係を解析し，**助詞と係り先の犯罪を表す動詞**を見ることで抽出する．②は「逮捕する」の対象となった人物（逮捕された人物）を犯人として抽出する．③は，例えば図 1 の「男が女性をはねた」という文では，「男」が「はねる」に係り，助詞「が」を伴っているので犯人と判断できる．「女性」も「はねる」に係っているが，助詞「を」を伴っているのでこれは被害者と判断できる．よって，犯人を主語である「男」，被害者を目的語である「女性」として抽出する．また「女性が男にはねられた」という受身の文では，主語と目的語を反対にして抽出する．

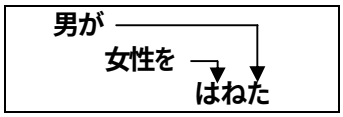

図 1: 係り受け解析の例

4. 動詞の選定

犯罪を表す動詞に係る人物を抽出するためには，その動詞を何らかの方法で得なければならない．

我々はまず，分析のために収集した 150 記事から，犯罪を表す動詞を人手で抜き出し，辞書を作成した．この結果，登録された動詞は全部で 45 語であった．その一部を図 2 に示す．この辞書を元に犯人と被害者の抽出を行ったが，以下のような問題が生じた．

i. 動詞を登録するために大きなコストがかかる
ii. 登録していない動詞に係る人物が抽出できない
iii. 汎用性が無い．他の分野の記事に転用する場合，新たに動詞を登録しなおす必要がある

そこでこの問題を解決するために，*tf-idf* を用いて記事中の重要な動詞を機械的に選定することを考える．文書中に出現するすべての動詞の重要度を測り，重要度の高い動詞に係る人物を抽出する．

$$tf = \text{文書中で単語}w\text{が使われた回数}$$

$$idf = log\frac{\text{単語}w\text{が使われた文書数}}{\text{全文書数}}$$

殴る	殴打する	脅す	脅迫する
奪う	強奪する	盗む	ひったくる
はねる	だます	絞める	着服する
偽る	縛る	倒れる	死亡する

図 2: 人手で選定した動詞の例

5. 評価実験

新聞記事を 200 収集し，あらかじめ人手で正解データを作成しておく．実装したシステムの出力と，この正解データが一致

† 愛知県立大学

した場合に正解とし，評価の尺度として，正解率，再現率，適合率を用いる．なお正解率は，全記事中犯人又は被害者の有無まで含めて正解であるものの割合である．実験 1 として人手で動詞を登録した場合と，実験 2 として *tf-idf* 重み付けを用いた場合の 2 つを行う．なお，*idf* の計算には'95 の毎日新聞記事を用いた（111565 記事）．

部分的に正解したものを含めた場合の各評価尺度の値はそれぞれ表 1，表 2 のようになった．ここで部分的に正解しているものとは，例えば「会社員男性」と「男性」という同一人物についての記述があった場合に後者を抽出したもの，あるいは，複数存在する犯人または被害者の内の一部だけを抽出したもののことである．

また表 3，表 4 は，犯人と被害者の 2 つが共に正解している記事の数とその割合を示す．

	正解率	再現率	適合率
犯人	89.95	90.06	97.02
被害者	87.37	83.44	95.45

表 1: 実験 1 の各評価尺度の値 (%)

	正解率	再現率	適合率
犯人	89.50	88.77	90.21
被害者	76.00	83.77	76.79

表 2: 実験 2 の各評価尺度の値 (%)

	記事数	割合(%)
共に正解	138	69.00
部分的正解を含み共に正解	26	13.00
1 つ以上の不正解を含む	36	18.00

表 3: 実験 1 の正解記事数と割合

	記事数	割合(%)
共に正解	111	55.50
部分的正解を含み共に正解	28	14.00
1 つ以上の不正解を含む	61	30.50

表 4: 実験 2 の正解記事数と割合

6. 結果の分析

6.1. 誤り記事の分析

人手で動詞を選定した場合と *tf-idf* 重み付けを利用した場合の結果を比較すると，犯人については再現率，適合率共に大きな変化は見られない．これは，犯人は「容疑者」や「逮捕する」というような手がかり語から抽出する処理の効果が大きいことと，犯罪事件記事では犯人を主語として書かれている文章が比較的多いので単純に主語を取ってくるとそれが正解である，といった理由が考えられる．

被害者の適合率が大きく下がっている原因は，重みの高い動詞の上位に「検知する」「譲渡する」「販売する」などといったものが現れるためであると考えられる．例えば「男からアルコールを検知した」「男性からバッグを奪った」という 2 つの文があった場合，前者は記事中の犯人であり，後者は被害者であると考えられる．しかし，本手法ではそれぞれ「男」「男性」を被害者として抽出してしまう．これは，「奪う」のような動詞とは違い，「検知する」のような動詞は，被害者が能動態で目的語，あるいは受動態で主語とならないためである．

また両実験共に，詐欺の記事など文書中に虚偽の文が存在する場合の誤りが見られた．これは例えば「男から『息子さんが女性をひいた』という電話があった」という文から被害者として女性を抽出してしまう，といったようなものである．この例の場合では，電話が虚偽の内容であることを理解せねばならず，本手法で対処することは非常に難しいと言える．

6.2. 正解記事の分析

人手で動詞を選定した場合では，犯人と被害者が共に正解しているものは 138 記事あった．部分的な正解を認めると 164 記事あり，この手法によって 82%の記事に対しておおむね正しく人物情報を抽出できたことになる．

tf-idf 重み付けを利用した場合では，6.1 節で述べたように被害者の適合率が低下しているため，共に正解している記事の割合が下がっている．

7. まとめと今後の課題

本研究では，係り受け関係を解析し，動詞に注目した情報抽出と，従来のキーワードからの情報抽出を組み合わせることによって，犯人と被害者を判定して抽出することができた．

人手で動詞を選定する手法では精度の高い抽出が行える半面，いくつかの問題が生じたために *tf-idf* 重み付けを利用した手法を試みた．情報検索の分野などで文書中の重要単語を選ぶ方法としてよく知られ，扱いやすい *tf-idf* を用いることで，部分的に正解したものを含めて 7 割程度の正解を得ることができた．精度を向上させるためには，主語が犯人（目的語が被害者）となりやすい動詞と，目的語が犯人（主語が被害者）となりやすい動詞を機械的に区別する仕組みが必要であると考えられる．1 つの方法としては，動詞の前後にある名詞など他の品詞を利用することが考えられ，どのような情報を与えればこの問題が解決できるのかを検討していきたい．

その他の課題として，本研究の抽出手法では複数の人物が 2 文以上に分かれて登場するような複雑な記事では正しく抽出することはできない．これを解決するためには文間関係や名詞の照応関係を解析する必要がある．

また，本研究の最終的な目標は，文書中の重要な情報を視覚的に分かりやすい形で提示するということであり，抽出項目の拡大や，その他の分野でも適用できるような抽出手法についても検討していきたい．

参 考 文 献

[1] 金山 淳一，北條 孝，田村 直良：“文章の構造解析による新聞記事からの事件情報抽出”情報処理学会研究報告，No.104，2002-NL-152，pp.1-6，2002