ルールの自動生成と対話的選択に基づく情報抽出ルール作成支援の提案

河合 剛巨　安藤 真一
NEC メディア情報研究所
E-mail: {t-kawai@bx, s-ando@cw}.jp.nec.com

1．はじめに

膨大な量のテキストデータから必要な情報だけを抽出する情報抽出技術の重要性が高まっている．例えば，地名や組織名等を対象にした固有表現抽出[1]や，製品とその評判などの評判情報抽出[2]，特定の関係にある表現を抽出する関係情報抽出[3]など応用例は多岐に渡る．

情報抽出では一般に，対象文書に対して，辞書や情報抽出ルール（以下抽出ルール）を適用することにより抽出が行われる．抽出ルールの作成は，人手により作成する方式と，機械学習によって構築する方式が挙げられる．前者は抽出ルールの記述に十分な知識とスキルを要し，適切な結果を抽出するよう調整するのも簡単ではないため作成コストがかかる．後者は事前に学習に用いる正解付コーパス等の教師データを大量に必要とするため，コーパスの作成にコストがかかる．また，学習手法によっては得られた抽出ルールに可読性がないために細かい修正が困難という問題もある．

特に，分野の変更や時代の変遷によって新たな抽出要求が発生するケースを考えると，その都度，抽出ルールやコーパスを作成する必要があり，そのコストの大きさは問題である．こうした抽出要求に応えるためにも，少ない労力で簡単に抽出ルール作成が行えることが望ましい．

本稿では，少数の抽出対象の例示から候補となる抽出ルールの自動生成と，抽出ルール間の関係性を用いた抽出結果の対話的選択とに基づく情報抽出ルール作成支援方式について提案する．

以下に，まず関連研究について述べ，次に提案方式を説明し，最後に検証結果を示す．

2．関連研究

前述のように，抽出ルールや正解付コーパスなどを人手により準備することは作成コストが大きく，新たな抽出要求に対応し難い．このような問題に対し，部分的教師付き学習や教師なし学習による方式も試みられている．例えば，固有表現抽出では，少量の抽出ルールや単語リストをシードとして，ブートストラップ法により抽出ルールの学習と抽出を逐次的に行う方式がある[4][5]．宇津呂らの方式[5]では，人手作成した初期単語リストを起点に，ブートストラップにより抽出ルールを決定リストで学習している．これにより，人的コストを抑えて再現率を上げられるが，抽出誤りの影響が伝搬するため，適合精度に問題がある．確信度等により不要な結果や抽出ルールを棄却する方法もあるが，十分ではない．

そこで，本稿では人手による適度な結果の確認を介在させるアプローチのもとに情報抽出ルール作成支援方式を提案する．

3．情報抽出ルール作成支援方式

3.1．提案方式の概要

提案する情報抽出ルール作成支援方式は，まず候補となる複数の抽出ルールを自動生成し，作業者との対話を通じて適切な抽出ルールを絞り込む方式である．本方式では，特に抽出ルール間の関係性を用いて作業者との対話を制御し，その負担を軽減する．

図１に本方式の概要図を示す．

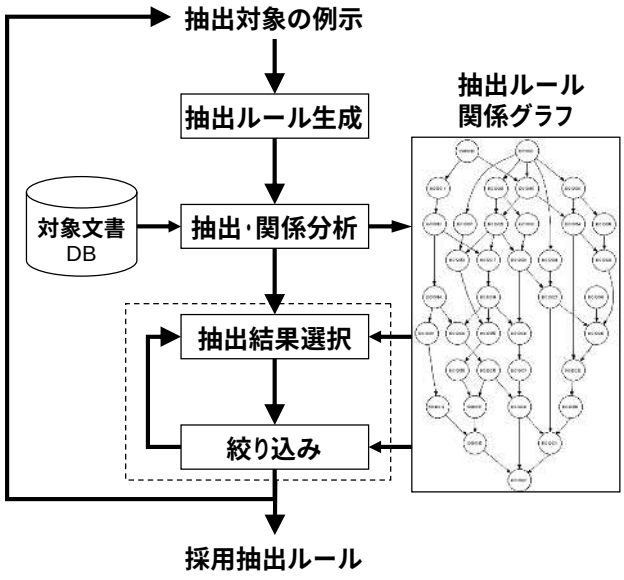

図 1: 提案方式の概要図

提案方式は，大きく以下のステップからなる．

・ステップ１
抽出対象の例示入力から，パタン展開して複数の抽出ルールの候補を自動生成する
・ステップ２
対象文書に抽出ルールを適用した結果から抽出ルール間の関係を分析し，抽出ルール関係グラフを構築する
・ステップ３
抽出ルール関係グラフに基づき，適切な抽出ルールを絞り込む．その際、抽出結果の正否選択を作業者と対話的に行うことで実行する．

上記のステップ１～３を繰り返す．

以下，各ステップについて，詳細を記す．

3.2．抽出ルールの自動生成

ステップ１では，抽出対象の例示データの入力を行い，それを基にして候補となる抽出ルールを複数生成する．

＜抽出対象の例示＞

例示用テキストの表示画面にて，抽出箇所をマークすることで指定して例示データ入力を簡易化する．また，対象属性(対象のクラス)の指定も可能とする．

＜抽出ルールの自動生成＞

次に，例示データから抽出ルールを自動生成する．抽出ルールは，パタン条件部と出力部から構成され，パタン条件部のパタンによって抽出箇所に該当する部分を同定する．

パタンは，以下のように展開して生成される．

まず，抽出箇所を含む例示用テキストの形態素解析結果および構文解析結果より，以下のようなA～Cの基本構造を生成する．

・A：抽出箇所の形態素列
・B：抽出箇所とその周辺まで含めた形態素列
　B1： 抽出箇所＋後置ｎ個の形態素
　B2： 抽出箇所＋前置ｎ個の形態素
　B3： 抽出箇所＋前置，後置ｎ個の形態素
・C：A・B と係り受け関係にある形態素列

これらの基本構造について，さらに各形態素を原型と品詞，意味属性等の素性に展開する．つまり，素性の組み合わせの数だけ，基本構造にそれぞれ素性を付与したものをパタン条件部とする．

また，パタンの組み合わせ爆発を防ぐため，一定の制限を設けて展開数を抑えた．パタンの展開数が多ければ，抽出漏れが減少するが，その分，抽出結果の確認量が増える問題があるためである．

本稿では，形態素列については，文節内に展開を限定した．文節外のパタン展開はCの係り受け関係にある文節を扱った．

図 2 にパタン展開の例を示す．

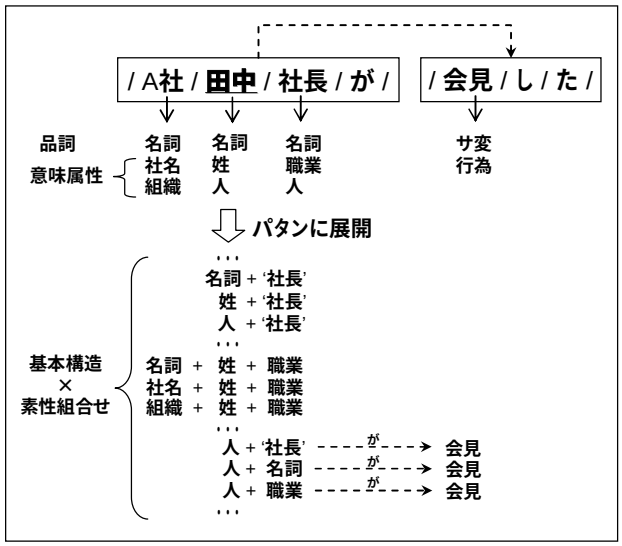

図 2: パタン展開の例

3.3．抽出ルール間の関係分析

ステップ２では，抽出ルール作成用の対象文書に抽出ルールを適用し，その結果から抽出ルール間の関係を分析して，抽出ルール関係グラフを構築する．

抽出ルール関係グラフとは，個々の抽出ルールに対応するノードを抽出結果の包含関係によってエッジで結んだ有向グラフのことである．ここで，ノード集合V と有向エッジ集合E からなる有向グラフを抽出ルール関係グラフ$G=(V,E)$とし，ある抽出ルールa，bによって得られた抽出結果の集合をU_a，U_bとする．抽出ルール関係グラフGは，aとbの各々に対応するノード$v_a, v_b \in V$を持ち，$U_a \subset U_b$の時，$\overrightarrow{v_a v_b} \in E$なる有向エッジを持つ．

Gを構築する際には，$\overrightarrow{v_a v_b} \in E$かつ$\overrightarrow{v_b v_c} \in E$の時，$\overrightarrow{v_a v_c}$は作成しない．また，ダミーのノード$v_{root}$を起点に，他の抽出ルールの抽出結果を内包しない（抽出結果数の少ない）抽出ルールから順次，包含関係によってエッジを結ぶと効率的に構築できる．

図 3 に抽出ルール関係グラフを有向グラフの構造で示す．

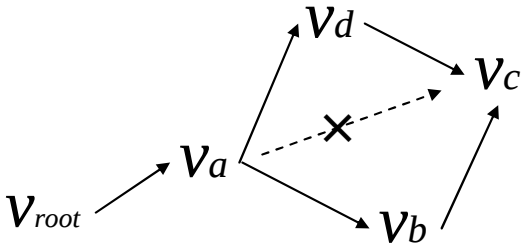

図 3: 抽出ルール関係グラフ

図 4 に，複数の抽出ルールと抽出結果Uの集合の関係を示す[1]．例えば，v_cに対応する抽出ルールによる抽出結果U_cが，v_bに対応する抽出ルールによる抽出結果U_bを包含している関係などを表している．図 4 の抽出結果の各集合U_a, U_b, U_c, U_dは，図 3 の抽出ルール関係グラフのv_a, v_b, v_c, v_dの各ノードにそれぞれ対応している．

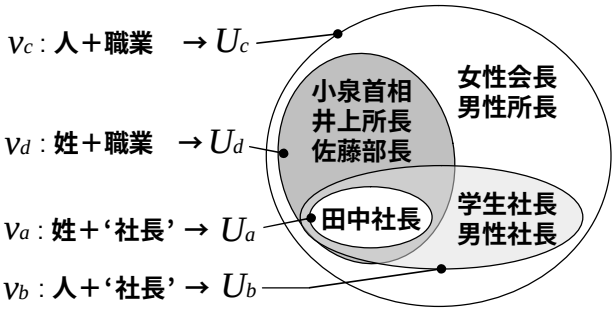

図 4: パタンと抽出結果集合の関係の例

3.4．対話的選択と抽出ルールの絞り込み

ステップ 3 では，ステップ 2 において構築した抽出ルール関係グラフに基づいて，抽出結果の正否選択を対話的に繰り返し，適切な抽出ルールの絞り込みを行う．

具体的には，以下の処理１～４の流れとなる．

1．抽出ルール関係グラフから，エッジ毎に算出した選出スコアに従ってエッジを１つ選出し，そのエッジの両端に位置するノードに対応する抽出結果の差分を作業者に提示し，その抽出結果の正否の選択を受ける．なお，今回は選出スコアとして，各ノードに対応する抽出結果数の比率を用いた．

2．抽出結果の正否の選択結果から，該当する抽出ルールの採否を判定する．

3．２で判定した抽出ルールが採用の場合，抽出ルール関係グラフを基にして，先祖のノードを全てたどり対応する抽出ルールを全て採用にする．２で判定した抽出ルールが非採用の場合，子孫のノードを全てたどり対応する抽出ルールを全て非採用にする．

4．抽出ルールの絞り込みが終了かどうかを判定し，終了するまで１へ戻り繰返す．

この処理の流れにより，作業者は対話的に抽出結果の正否を選択することで，適切な抽出ルールを絞り込むことができる．

次に，抽出結果の対話的選択と抽出ルールの絞り込みの例を，前述の図 3 および図 4 を用いて説明する．ここでは例えば，社長名のみを抽出したいという状況を考える．

まず，図 3 に示す関係の抽出ルール関係グラフからエッジを選出する．選出スコアの大きなエッジ$\overrightarrow{v_a v_d}$を選出し，図 4 から，差分の抽出結果として$U_d - U_a$の抽出結果（ここでは「小泉首相，井上所長，佐藤部長」）を提示する．これらは社長名としては不適であるので作業者は非採用の選択をする．その結果，ノードv_dに対応する抽出ルールは非採用と判定できる．この場合，抽出ルール関係グラフを基に，ノードv_dの子孫ノードをたどり，ノードv_cに対応する抽出ルールも自動的に非採用と判定する．実際に，図 4 の通り，v_cノードに対応するパタンは，v_dのパタンよりもさらに制約が緩く，v_cよりも不要な抽出結果を多く含むため，不適である．

次に，残りの抽出ルール間のエッジ$\overrightarrow{v_a v_b}$が選出される．$\overrightarrow{v_a v_b}$に対応する抽出結果の差分として$U_b - U_a$の抽出結果（ここでは「学生社長，男性社長」）を提示する．これらは，やはり不適であるので非採用と選択され，ノードv_bに対応する抽出ルールを非採用とする．ここでは，未判定の子孫ノードがないので，処理３は不要である．

最後は，ノードv_aに対応する抽出ルールに絞り込まれる．($\overrightarrow{v_{root} v_a}$を選出して)抽出結果を提示すると，適合であるため作業者は正解と選択し，適切な抽出ルールが採用される．

なお，エッジの選出スコアによって，絞り込みの効率が大きく左右されると考えられるが，効果的な選出スコアの算出方法については，さらなる検討が必要である．

4．本方式の妥当性検証

本方式の妥当性を検証するために以下の実験を行った．また，どのような選出スコアが有効かその指針となる知見の獲得を試みた．

4.1．検証実験と結果

実際に，本提案方式に基づいて情報抽出ルール作成支援システムを試作し，簡単な固有表現抽出タスクを設定し検証実験を行った．

本稿では，首相名を抽出するタスクを取り上げ，以下に詳細を述べる．

まず「安倍首相は･･･」という文を例示用のテキストとして用い「安倍」の部分を抽出箇所として指定することで，「人 +‘ 首相 ’」等の 62 の抽出ル

[1] ここでは説明のため，抽出ルールではなく，パタン条件部を簡易表示した．引用符‘ ’は表層を意味し，それ以外は品詞や意味属性など素性中の値を示す．

ールが自動生成された．

続いて，新聞記事 75 文を対象文書として上記の抽出ルールを適用し抽出ルール関係グラフを構築した．構築された抽出ルール関係グラフを図 5 に示す．

図 5 中の各ノードは，左から少数の抽出結果数の抽出ルールとなっており，また右へ向かうほど抽出結果が増大することを意味する．「→」の有向エッジの関係が包含関係を表す．

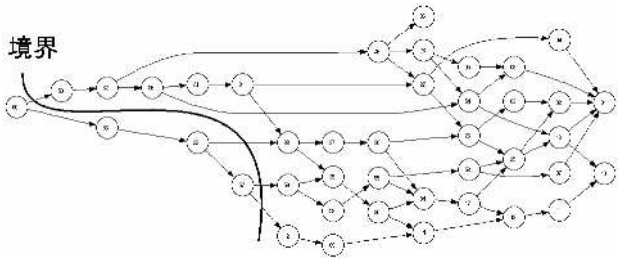

図 5: 抽出ルール関係グラフの一例

なお，全く同一の抽出内容を抽出する抽出ルールが存在したため，これらの抽出ルールは，１つのノードに統合して示している．

自動生成されたルール数が 62 と少数の場合でも，任意の 2 ルール間の組合せは ${}_{62}C_2 = 1891$ 通りあるが，上記の場合，抽出ルール関係グラフ中のエッジの数は 59 である．

最後に，構築された抽出ルール関係グラフに基づき，抽出結果を選択することにより絞り込まれた結果を図 5 の境界で示す．本実験では，境界線の左側４ノードが採用として絞り込まれた．採用された抽出ルールを確認したところ，適切な抽出ルールであった．なお，対話的選択に要した確認回数は 20 回であった．このことは全ての抽出ルール数および抽出ルール関係グラフの総エッジ数よりも少ない回数で絞り込み可能であることを示している．

4.2．考察

検証の結果，作業者は，初回に例示用テキストに対して抽出したい箇所を例示し，提示された抽出結果の正否を対話的に選択するのみで，適切な抽出ルールの作成が簡単に可能であり，提案方式の妥当性が確認できた．

また，抽出ルール間の各抽出結果の差分のみを提示することで，抽出ルールの採否が判定可能であり，確認量を軽減できることが確認できた．

さらに，抽出ルールを記述できる作業者であっても，作成した抽出ルールの結果を確認し修正する試行錯誤の作業が必須であり，本方式はこの作業を支援するためにも役立つと考えられる．

今回，エッジの選出スコアとして，各ノードに対応する抽出結果数の比率を用いた．理由は，パタンの傾向が変わるほど抽出結果数には差異があり，優先して確認した方が効率的であると想定したためである．実際に，パタンの制約の強さと抽出結果数の相関は確認できたが，パタンの制約が緩いほど抽出結果数が増大する影響の方が大きく，抽出ルール関係グラフの終点(図 5 では右側)に近いエッジほど優先される結果となった．また，図 5 の例では，最小で４つのエッジについて抽出結果を確認すれば絞り込み可能であり，より効果的な選出方法が求められる．ただし，タスクにより境界位置は大きく変わるので，始点側から選出すれば良いとも言えない．現在，ノードから出る子孫ノードへのパス数やその他の統計的基準により効率的かつ安定した選出スコアの算出方法について検討中である．

5.まとめ

本稿では，少数の抽出対象の例示から複数の候補となる抽出ルールを自動生成し，各抽出ルールの抽出結果を比較して抽出ルール関係グラフを構築することにより，抽出ルール間の関係性を用いて抽出結果の対話的選択と絞り込みを簡易化する情報抽出ルール作成支援方式を提案した．また，提案方式を簡単な抽出ルール作成タスクに適用し，容易に抽出ルールが作成できることを示した．

今後の課題としては，方式全体としての定量的な評価，およびブートストラップとの組み合わせも含めた更なる再現率向上の仕組み，ならびにより効果的な選出スコアの算出方法の検討等があげられる．

参考文献

[1] 関根聡: 固有表現から専門用語，NLP2004 併設ワークショップ「固有表現と専門用語」発表論文集，pp.1-4，2004．

[2] 立石健二，石黒義英，福島俊一: インターネットからの評判情報検索，人工知能学会誌，Vol.19，No.3，2004．

[3] 薬師寺あかね，宮尾祐介，建石由佳，辻井潤一: 述語項構造パターンを用いた医学・生物学分野情報抽出，言語処理学会第 11 回年次大会発表論文集，C1-7，2005．

[4] M. Collins and Y. Singer: Unsupervised Models for Named Entity Classification，Proc. Joint SIGDAT Conf. on EMNLP/VLC.，pp.100-110，1999．

[5] 宇津呂武仁，颯々野学: ブートストラップによる低人手コスト日本語固有表現抽出，情報処理学会自然言語処理研究会，Vol．2000，No.86，pp.9-16，2000．