学術論文データを用いた言語資源の用途情報抽出

小澤　俊介 †　　遠山　仁美 †　　内元　清貴 ††　　松原　茂樹 †

† 名古屋大学　　†† 情報通信研究機構

Automatic Acquisition of Expressions Representing Purposes and Methods of Using A Language Resource from Academic Articles

Shunsuke Kozawa†　　Hitomi Toyama†　　Kiyotaka Uchimoto††　　Shigeki Matsubara†

†Nagoya University
††National Institute of Information and Communications Technology

1　はじめに

近年，言語学や音声言語処理，自然言語処理の研究分野では，言語現象を実例に基づいて客観的に分析することの重要性が認識され，コーパスや辞書などの言語資源を用いた研究が盛んに行われてきた．中でも，コーパスを用いた確率・統計的手法は応用範囲が広く，これまで音声認識や情報抽出，機械翻訳などに適用され，顕著な成果をあげている．言語資源は，このような間接的な利用に留まらず，辞書編纂や言語教育などに直接利用されることも多い．こうした点から，その重要性は広く認識され，これまで研究基盤として数多くの言語資源が作られており，今後もますます研究に不可欠なものになると考えられる．しかし，残念ながら，各言語資源は十分に活用されていない場合が多い．実際には様々な使い方があるにも関わらず，その情報が利用者に知られていないことが多いからである．このような言語資源の用途に関する記述は，Web や論文中に存在すると考えられるが，Web 検索を用いても容易には見つからない．

ここで，言語資源の用途に関する情報が整理されれば，各言語資源の持つ本来の価値が十分に活かされる可能性が高くなると考えられる．この用途情報の活用例としては，用途をクエリとした検索システムがあげられる．言語資源の用途を検索キーワードとして入力し，用途に適合する言語資源名や言語資源に関する情報を出力として返すことができれば，利用者が言語資源を発見する手助けとなり，言語資源の効率的利用に繋がるだろう．逆に，検索により適合する言語資源がない場合は，ニーズがあるにも関わらず適した言語資源が存在していないということが分かり，今後の言語資源開発に大きく役立つと考えられる．さらに，用途情報の中には，開発者が想定していなかったようなものが含まれることもあり，そこから新たな知見が生まれることも考えられる．このように，用途情報は様々な可能性を秘めた必要かつ有用な情報であると期待される．

そこで，本研究では，言語資源の効率的な利用を促進することを目的として，言語資源の用途情報の抽出を行う．本稿では，構文情報に基づいたルールを用いて学術論文から言語資源の用途情報を抽出する方法を提案する．ルールは用途情報の記述を含む文の構文的特徴に着目して生成し，言語資源の用途情報は，生成した抽出ルールとパターンマッチングを行うことにより抽出する．

2　関連研究

テキストからの情報抽出は米国における MUC (Message Understanding Conference) [1] を起源として，現在も盛んに行われている．MUC では，新聞記事の中から人事異動に関する情報の抽出を行っており，他にも様々な情報の抽出を試みる研究が行われている．

本稿で我々が抽出する用途情報に着目した研究もいくつか行われている．乾ら [2] は，接続詞「ため」を利用して，2 つの出来事間の因果関係を抽出し，出来事間の関係によって 4 種類に分類する手法を提案した．この 4 種類のうち，means 関係が用途情報にあたり，ある行為とその行為を行うための手段の対になっており，コトとコトの関係を抽出している．我々が抽出するのは，モノとコトの関係であり，さらに，我々の提案手法では，接続詞「ため」のような表現だけでなく，動詞にも着目して抽出を行っている．鳥澤 [3] は，名詞，助詞，動詞の 3 項組の共起頻度を用いて，用途表現および準備/用途対の獲得手法を提案した．鳥澤は，対象 X の用途表現を「X を利用することの通常の方法」と定義して，対象 X の一般的な用途表現の獲得を行っている．しかし，我々は，一般的でない用途情報からも新たな知見が得られる可能性があると考え，用途情報の一般性の有無に関わらず抽出を行う．

3　用途情報の分析

3.1　用途情報とは

論文や Wikipedia などのテキストでは，言語資源に関する様々な記述がある．例えば，言語資源のひとつである WordNet に関しては，次の通りである．

(1) We use WordNet for lexical lookup.

(2) We extract lexical relations from WordNet.

(3) WordNet contains semantic relationships.

(4) The most updated version of WordNet is WordNet 2.0.

(5) Resolution of pronouns can be used on top of the WordNet approach.

我々の調査によると，このように言語資源名を含む文は，言語資源に関する内容として大きく次の 4 種類の記述を含むことが分かった．

1. 利用目的
2. 利用方法
3. 言語資源情報
4. その他

ここで，1. は例文 (1) の下線部のように言語資源を利用する目的に関する記述である．2. は，例文 (2) の下線部のようにある目的を達成する為の手段としてどのように言語資源を利用するかに関する記述であり，3. は，例文 (3), (4) の下線部のように言語資源に関する情報の記述である．これらの 3 種類に分類できなかった文は例文 (5) のように，対象の言語資源ではなく主として他の言語資源についての記述を含むことが多い．本研究では，これらのうち，利用目的と利用方法に関する記述を用途情報とする．

3.2 用途情報の抽出

本稿では，用途情報を抽出する対象として論文を用いた．用途情報は Web ページにも存在すると考えられるが，言語資源の用途情報に関しては論文のほうが多く存在すると考えたため，論文を対象とした．

3.1 節でも例を挙げたように，用途情報は様々な形で表現される．そこで，本稿では，論文における用途情報の記述方法の特徴を分析することにより，用途情報の記述パターンを抽出し，このパターンを用いて用途情報の抽出を行う．

言語資源を WordNet，論文集を LREC2004 に限定して分析し，WordNet に関する用途情報の抽出を行った．まず，言語資源名を含む文の抽出を行ったところ，言語資源名である WordNet を含む文が 712 抽出された．次に，抽出した文の分析を行った結果，211 の用途情報を含む文を抽出した．このうち，16 文が 1 文に複数の用途情報を含んでおり，211 文中から 227 の用途情報を抽出した．

3.3 用途情報の記述パターン

抽出された用途情報を含む文を対象に用途情報の記述パターンを分析し，記述パターンを利用目的，利用方法に分類した．

利用目的では，さらに，動詞などの特徴的な表現に着目して，記述パターンを利用・適用，説明，由来，提供の 4 種類に分類した．以下に記述パターンにおける特徴的な表現と利用目的の記述例を示す．以下の記述例では，利用目的に該当する部分を下線で示している．

- 利用・適用
  言語資源の利用，適用に関する記述と共に，利用目的が記述されている．利用，適用の記述には以下の動詞が用いられている．

  use，utilize，exploit，employ，leverage
  capitalize，apply，adapt，adopt，etc.

  記述例
  - We use WordNet for lexical lookup.
  - We use WordNet to understand the links between different parts of the document.

- 説明
  言語資源の説明と共に，利用目的が記述されている．be 動詞または以下の形容詞が用いられている．

  useful，valuable，available，helpful
  favorable，efficient，effective，etc.

  記述例
  - Wordnet is a valuable resource for semantic annotation.

- 由来
  according to，based on，by means of を用いて，言語資源の利用目的が記述されている．

  記述例
  - The assumed baseline is the algorithm that tags the corpus according to the first Wordnet sense.

- 提供
  言語資源が何かをもたらすという表現とともに，言語資源の利用目的が記述されている．提供を表す動詞を以下に挙げる．

  enable，enrich，improve，provide，enhance
  help，give，contribute，allow，bring，etc.

  記述例
  - The use of wordnet enables a more systematic and more detailed attachment of such marks.
  - However, Wordnet due to its structure can provide multiple layers of semantic information, which, when utilized in semantic annotation, can improve the annotations produced.

利用方法では，利用目的と同様に，記述パターンを抽出・取得，照合・統合の 2 種類に分類した．以下に記述パターンにおける特徴的な表現と利用方法を含む記述例を示す．以下の記述例では，利用方法に該当する部分を下線で示している．

- 抽出・取得
  言語資源から情報等を抽出するという利用方法が記述されている．抽出・取得の表現として以下の動詞が用いられている．

  derive，obtain，extract，acquire，mine，etc.

  記述例
  - We outline a mechanism for deriving new concepts from WordNet using metonymy.
  - In what concerns the attribute list generation, syntactic attributes can be obtained from WordNet.

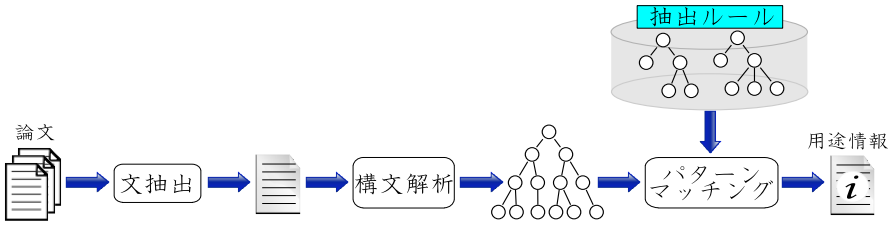


図 1: 処理の流れ

- 照合・統合
  2 つのものを照合，統合するというような利用方法が記述されている．これは，概念辞書である WordNet を対象にしたために出現した表現であり，言語資源に依存した表現だと考えられる．言語資源に依存した記述に用いられる動詞としては以下のものが挙げられる．

  assign，match，link，associate，map
  connect，compare，merge，integrate，etc.

  記述例

  - Finally we assign to each noun its corresponding WordNet code.
  - Each collocation is mapped to the WordNet sense inventory in a semiautomatic manner and transformed into a relatedness edge.

また，上記の 6 種の記述パターンに分類できなかったものとして，用途情報の記述が複雑であるもの，代名詞などが用いられ照応関係を考慮する必要がある文も存在した．

4 構文情報に基づくルールを用いた用途情報の抽出手法

4.1 処理構成

図 1 に用途情報抽出の流れを示す．まず，pdftotext ツール [4] を用いて，学術論文からテキスト情報を抽出する．そして，抽出したテキスト情報から言語資源名を含む文の抽出を行う．次に，言語資源名を含む文に対して，Charniak Parser [5] を用いて構文解析を行う．最後に，構文解析の結果と抽出ルールのパターンマッチングを行うことにより，言語資源の用途情報を抽出する．

4.2 抽出ルール

用途情報の抽出ルールを生成する．3.3 節で分類した用途情報の記述パターンの特徴より，用途情報を記述するのには，動詞が重要な役割を果たしているとみなせる．そこで，本稿では，主として動詞に関して次の 3 点に着目することにより動詞の分類を行い，分類した動詞ごとに抽出ルールの生成を行う．1 点目は，動詞が一般動詞であるか，be 動詞であるかである．2 点目は，動詞が目的語をいくつとるかであり，3 点目は言語資源名の位置である．

- 目的語を 1 つとる一般動詞および言語資源名が目的語に含まれる場合
  動詞が前置詞を必要としない場合は図 2 の構造，必要とする場合は図 3 の構造となる．

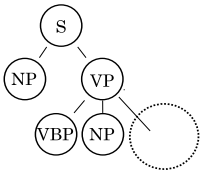


図 2: 利用・適用 1

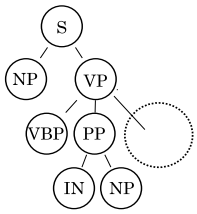
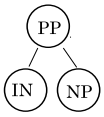
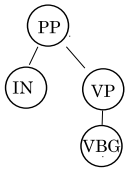
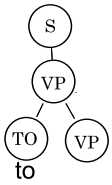


図 3: 利用・適用 2

このとき，動詞が利用・適用を表す動詞であり，点線で囲まれた部分に図 4〜図 7 の構造が含まれていれば，点線部分を用途情報として抽出する．ただし，図 4 と図 5 における前置詞 IN は for，in，on，as，towards のいずれかとする．







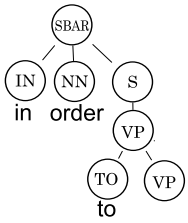


図 4: 用途 1　図 5: 用途 2　図 6: 用途 3　図 7: 用途 4

- 目的語をとる一般動詞および言語資源名が主語に含まれる場合
  動詞が提供を表す動詞であり，図 8 または図 9 の構造をとるとき，点線で囲まれた動詞句を用途情報として抽出する．

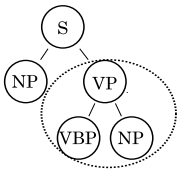


図 8: 提供 1

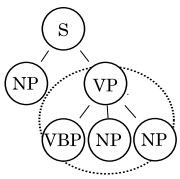


図 9: 提供 2

- 目的語を 2 つとる一般動詞および言語資源名が前置詞の目的語に含まれる場合
  図 10 の構造をとり，動詞が抽出・取得を表す動詞，前置詞が from のとき，点線で囲まれた動詞＋名詞句を用途情報として抽出する．

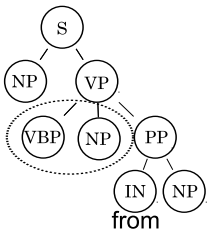


図 10: 抽出・取得

- 目的語を 2 つとる一般動詞および言語資源名が目的語または前置詞の目的語に含まれる場合
  図 11 または図 12 のような前置詞を含む構造をとり，動詞が照合・統合を表す動詞であるとき，点線で囲まれた動詞句を用途情報として取得する．また，前置詞を用いずに and が用いられる図 13 のような構造をとる場合も同様に点線部分の動詞句を抽出する．このように，目的語を 2 つとる動詞の場合，構文情報を用いないと，抽出部分の決定が困難である．

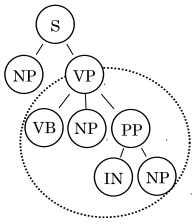


図 11: 照合・統合 1

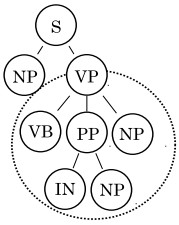


図 12: 照合・統合 2

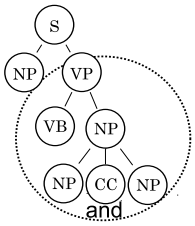


図 13: 照合・統合 3

- be 動詞
  - 形容詞を含まない場合
    言語資源名が目的語中に含まれ，図 14 の構造をとるとき，点線部分に図 4〜図 6 の構造が含まれている場合は点線部分を用途情報として抽出する．また，言語資源名が主語中に含まれ，図 15 の構造をとる場合も同様に，点線部分に図 4〜図 6 の構造が含まれていれば，その部分を用途情報として抽出する．

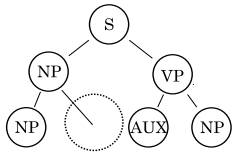


図 14: 説明 1

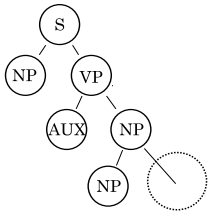


図 15: 説明 2

  - 形容詞を含む場合
    図 16 の構造をとり，特定の形容詞が用いられているとき，点線部分に図 4〜図 6 の構造が含まれていれば，点線部分を用途情報として抽出する．

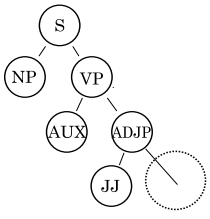


図 16: 説明 3

- 動詞以外
  動詞以外の特定の表現を用いて用途情報の記述がなされている場合を以下に示す．

  - according to, by means of
    図 17 や図 18 の構造を含んでいる場合，点線で囲まれた動詞句を用途情報として抽出する．

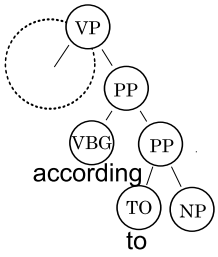


図 17: 由来 1

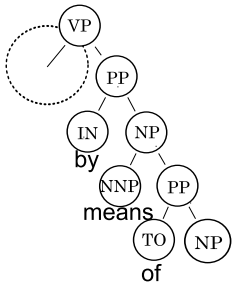


図 18: 由来 2

  - based on
    図 19 の構造を含んでいる場合，点線で囲まれた名詞句を用途情報として抽出する．

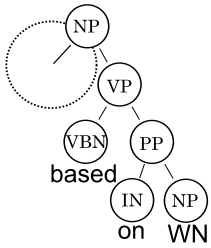


図 19: 由来 3

上記の一般動詞に着目した抽出ルールは能動態および動詞が現在形のときの抽出ルールである．しかし，一般動詞を用いた用途情報の記述には，様々な記述方法が存在する．そのため，以下では，“ 目的語を 1 つとる一般動詞および言語資源名が目的語に含まれる場合 ”を例に上記以外の抽出ルールを説明する．

動詞が現在分詞もしくは過去分詞で，名詞句に係っている図 20 や図 21 の構造をとる場合，点線で囲まれた名詞句部分を用途情報として抽出する．このように，動詞が名詞句に係っているかの判断は，構文情報を必要とするため，構文情報を用いないと抽出が困難だと考えられる．

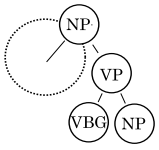


図 20: 現在分詞など

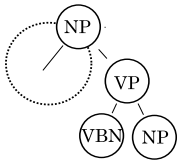


図 21: 過去分詞など

動詞が現在分詞であり，図 22 のように，by+現在分詞となっている場合，点線部分の文を用途情報として抽出する．

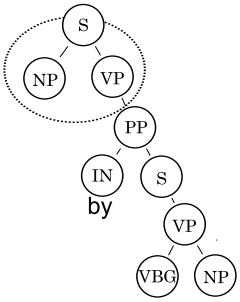


図 22: by+現在分詞

また，動詞が名詞化され，“ the exploitation of ”のように図 23 の構造となる場合，点線部分に図 4〜図 6 の構造が含まれていれば，点線部分を用途情報として抽出する．

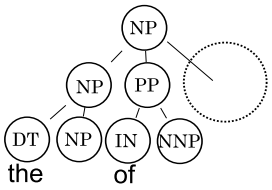


図 23: 名詞化

5 評価

5.1 評価実験

4 章で生成した抽出ルールの評価実験を行った．3 章で分析し，抽出した LREC2004 の論文集に含まれる 227 の WordNet に関する用途情報を正解データとして，クローズドテストを行った．また，新たに，LREC2006 の論文集を分析したところ，LREC2006 には WordNet を含む文が 667 文，そのうち，WordNet に関する用途情報を含む文が 181 文存在した．また，181 文のうち，6 文が複数の用途情報を含んでおり，181 文から 187 の用途情報を確認した．そして，これを正解データとして，オープンテストを行った．

5.2 実験結果

クローズドテストの結果を表 1，オープンテストの結果を表 2 に示す．実験の結果，クローズドテストで 80.2%，オープンテストでも 62.6%の再現率を得ることに成功した．また，利用目的と利用方法では，利用目的のほうが高い再現率を得ることができた．これは，利用方法において，解析ミスによる失敗が多く見られたことと，照合・統合の記述に様々なバリエーションがあり，それに対応する抽出ルールの生成が行えなかったことが原因である．適合率に関しては，クローズドテストで 91%，オープンテストで 83.6%とともに高い値を得ることができた．その他に関しては抽出ルールを生成していないため，再現率，適合率ともに 0 となっている．

抽出に成功した用途情報の例を示す．まず，利用目的の抽出例を以下に示す．

- for NLP
- for word sense disambiguation
- to cluster its senses
- for query expansion
- for finding the type of an entity

利用目的では，1 つ目の例の“ 自然言語処理のため ”といった漠然とした目的のものも存在したが，2 つ目から 4 つ目のように，曖昧性解消，クラスタリング，クエリ拡張などの実用的な利用目的を抽出することができた．また，5 つ目の例のように，抽出部分のみからでは利用目的が分かりづらいものも存在した．しかし，5 つ目の例の主語を見てみると，主語は“ Question Answering systems ”であったため，利用目的が質問応答における質問タイプの同定であることがわかった．このように，より明確な利用目的を得るには，主語を取得する必要もあるだろう．

表 1: クローズドテストの結果

		抽出数	出現数	再現率	適合率
利用目的	利用・適用	63	76	0.829	0.900
	由来	15	17	0.882	1
	説明	8	9	0.889	1
	提供	45	50	0.900	0.918
利用方法	抽出・取得	16	19	0.842	0.941
	照合・統合	35	49	0.714	0.854
	その他	0	7	0	0
利用目的	———	131	152	0.862	0.923
利用方法	———	51	68	0.750	0.879
全体	———	182	227	0.802	0.910

表 2: オープンテストの結果

		抽出数	出現数	再現率	適合率
利用目的	利用・適用	39	56	0.696	0.867
	由来	12	17	0.706	0.706
	説明	1	4	0.250	0.500
	提供	19	26	0.731	0.760
利用方法	抽出・取得	11	19	0.589	0.917
	照合・統合	35	53	0.660	0.897
	その他	0	12	0	0
利用目的	———	71	103	0.689	0.798
利用方法	———	46	72	0.639	0.902
全体	———	117	187	0.626	0.836

次に，利用方法の抽出例を以下に示す．利用方法では，曖昧性解消やクラスタリングなどを目的とした利用方法を抽出することができた．

- extract a lexical expression
- assign WordNet senses to cluster labels

抽出に失敗したものについて，表 3 に原因とその数を示し，以下で失敗例を挙げる．

表 3: 抽出失敗の原因と失敗数

失敗の原因	クローズド	オープン
解析ミス	20	19
関係代名詞	3	8
熟語	6	1
その他	9	30
全体	38	58

解析ミスに関しては，抽出・取得と照合・統合でクローズド，オープン共に解析ミスの約半数を占めていた．これは，抽出・取得，照合・統合の構造に前置詞が含まれているためである．前置詞や and，or などは構造解析の際，係り先の判断が難しいと考えられるため，抽出・取得と照合・統合で多くの解析ミスが見られた．

関係代名詞が含まれていたために抽出ルールによって抽出されなかった例を，以下に示す．

A new slim subset of the WordNet ontology that is then used for the classification process.

この文では関係代名詞が含まれていなければ，抽出可能であったが，関係代名詞が含まれていたために抽出することができなかった．そのため，関係代名詞を考慮した抽出ルールの生成が必要となるだろう．

熟語に関しては，抽出ルールが複雑かつ各熟語に対して 1 つずつルールを生成する必要があるため，今回の抽出ルール生成では，考慮に入れなかった．そのため，以下のような熟語が利用されている文の抽出が行えなかった．

Enriching the synsets with information from SMD makes the usage of a wordnet in an information retrieval task more efficient.

熟語 1 つ 1 つに対して抽出ルールを生成するのは困難である．しかし，用いられる頻度が高い熟語に着目し，ルールを生成することにより抽出失敗数を減少させることができるだろう．

その他に関しては，抽出ルールとして生成していなかったものや考慮していなかった動詞，表現が存在していた．これらのうち汎用性の高いものについては，今後，WordNet 以外の言語資源も分析対象としルールを追加することによって対応できるようになると考えている．

5.3 考察

実験の結果，失敗例で挙げたように，関係代名詞や熟語，未生成の抽出ルールなど，まだ考慮すべき点はあるが，オープンテストにおいてもまずまずの結果が得られたと考えられる．今回，用途情報を抽出した LREC は言語資源に関する会議であるため，WordNet から派生した言語資源を開発するという表現が多く用いられていた．そのため，照合・統合に関する表現が他の会議の論文集に比べて多く，これらの表現は LREC ならではの表現とも考えられる．したがって，異なる会議の論文集に本手法を適用した場合，再現率が下がることが考えられる．照合・統合に関する用途情報の記述については，言語資源の種類に依存するため，異なる言語資源の用途情報を抽出する際には再度検討する必要があるだろう．しかし，照合・統合の抽出ルール以外はある程度汎用性を持ったルールであるため，再現率は大きく下がることはないと考える．

次に，LREC2004 と LREC2006 に含まれる用途情報の違いを調査したところ，LREC2006 中には，LREC2004 中に含まれない以下のような用途情報が約 3 割存在していることが分かった．

- to create training sets
- combined geographical databases with WordNet

これは WordNet という言語資源に限っても，次々と新しい使い方が生み出されていることを示している．同じ会議の論文集において，このような違いが得られたということは，他の会議の論文集も対象にすればさらに多くの新しい用途情報が抽出できると期待される．

本稿では，言語資源名を含む文を対象として用途情報を抽出した．したがって，用途情報に関する記述を含んでいても言語資源名を含まない文の場合，対象外となってしまう．しかし，このような文では，代名詞や言語資源を表す記述を用いて言語資源の用途情報が記述されることが多いため，次のように対処することにより，用途情報を抽出することが可能になると考えている．まず，代名詞を用いた文に対しては，照応関係を考慮することで抽出が可能となると考える．また，言語資源を表す記述を用いた文に関しては，次のようにして抽出できるようになると考える．まず，記述パターンの“ 説明 ”に関するルールを用いて言語資源を表す記述を抽出する．次に，既に抽出した言語資源名とその説明の情報をもとに言語資源名を特定する．そして，特定した言語資源名が抽出対象の言語資源名と一致する場合に，本稿で述べた抽出ルールを適用する．

6 まとめ

本稿では，言語資源の効率的利用の促進を目的として，構文的特徴に着目したルールを用いてパターンマッチングを行うことにより，学術論文から言語資源の用途情報を抽出する手法を提案した．実験を行った結果，クローズドテストで 80.2%の再現率と 91.0%の適合率，オープンテストで 62.6%の再現率と 83.6%の適合率を得ることができた．これは，言語資源の利用目的，利用方法の抽出については実用的なレベルであると考える．

今後の課題としては，今回抽出に失敗した用途情報を抽出するため，関係代名詞などを考慮した抽出ルールを生成することが考えられる．また，異なる論文集に対する抽出ルールの生成や，各パターンにおける動詞の網羅性の向上が挙げられる．さらに，本稿で生成した抽出ルールのうち，照合・統合に関するルールは WordNet など具体的な言語資源に依存したルールであるため，異なる言語資源に対しては，用途情報がどのように記述されているかを調査する必要がある．本稿では論文を対象として用途情報の抽出を行ったが，Web への適用も今後の課題である．Web は，論文から抽出される用途情報とは異なる用途情報を含んでいる可能性があるためである．Web 上と論文上とでは用途情報の記述の特徴に多少差異があると考えられるため，Web へ本手法を適用する際には更なる抽出ルールの洗練が必要となってくるだろう．また，抽出ルールが洗練され，多くの用途情報を抽出することが可能になれば，抽出してきた用途情報を正解データとして機械学習を適用することも考えられる．

参考文献

[1] Ralph Grishman, Beth Sundheim: Message Understanding Conference - 6: A Brief History, COLING-96, pp. 466-471 (1996).

[2] 乾孝司, 乾健太郎, 松本裕治：接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得, 情報処理学会論文誌, Vol. 45, No. 3, pp. 919-933 (2004).

[3] 鳥澤健太郎：対象の用途と準備を表す表現の自動獲得, 自然言語処理, Vol. 13, No. 2, pp. 125-144 (2006).

[4] http://www.foolabs.com/xpdf/ .

[5] Eugene Charniak: A Maximum-extropy-inspired Parse, NAACL-2000, pp. 132-139 (2000).