社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－NL－184 (12)
2008／3／28

言語資源の用途情報の抽出と利用

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

近年，言語資源を用いた研究が盛んに行われており，言語資源は言語研究に必要不可欠なものとなっている．しかし，Web 検索を用いても，言語資源の内容に関する記述は見つかるが，用途に関する記述を見つけることは容易ではない．そのため，各言語資源の本来の価値が世の中に十分に知られていない場合が多い．本稿では，言語資源の効率的な利用促進を目的に，言語資源の用途情報の抽出手法を提案する．本手法では，構文情報に基づいたルールを学術論文に適用することにより抽出する．抽出した用途情報を言語資源メタデータデータベースに応用し，言語資源の効率的な検索への効果について検証した．

Automatic Acquisition of Usage Information for Language Resources

Kozawa Shunsuke†, Tohyama Hitomi†,
Uchimoto Kiyotaka‡and Matsubara Shigeki†

†Nagoya University ‡National Institute of Information and Communications Technology

Recently, language resources (LRs) are becoming indispensable for linguistic researches. However, existing LRs are not fully utilized because it is not well known that they have a variety of usages. It indicates that the intrinsic value of LRs is not recognized very well. In this research, therefore, we extract a list of usage information for each LR to promote the efficient utilization of LRs. In this paper, we propose a method for extracting a list of usage information from academic articles by using rules based on syntactic information, and show that it helps us to efficiently search LRs by adding the extracted list of usage information to metadata database of LRs.

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．しかし，言語資源が必ずしも十分に活用されているとは言い難い．その理由として，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挙げられる．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．

このような問題に対して，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，用途をクエリとした検索システムが挙げられる．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することができれば，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，検索クエリに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．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，本稿では，言語資源の効率的利用を促進することを目的に，学術論文から言語資源の用途情報の抽出手法を提案する．本手法では，構文情報に基づいたルールを用いる．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のパターンマッチングにより抽出する．また，抽出した用途情報を言語資源メタデータデータベース SHACHI[11, 12] に応用し，用途情報が言語資源の検索に寄与するかについて検証をした．

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．

近年では，Generative Lexicon Theory [5] における概念を用いる目的や機能を表す telic role や概念を生み出す動作を表す agentive role が注目を集めている．例えば，「本」という名詞に関して，「読む」は telic role を表す動詞であり，「書く」は agentive role を表す動詞である．こうした telic role や agentive role を表す名詞と動詞の二項組を WordNet から抽出する手法 [6, 7] やコーパスから抽出する手法 [8, 9]，Web から抽出する手法 [10] が提案されている．しかし，こうした手法で取得できる telic role や agentive role はある名詞の性質に関する情報であり，我々が求める用途情報とは異なる．

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 [2, 3]．しかし，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，動詞にも着目する．また，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する．

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 1 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.

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) 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

抽出ルールの生成には，用途情報を記述する特徴を利用する．そこで，用途情報を記述する特徴を分析するため，用途情報の抽出を行った．

言語資源を WordNet，論文集を LREC2004 に限定して分析し，WordNet に関する用途情報を抽出した．その結果，193 の用途情報を含む文と 214 の用途情報を抽出した．以下に抽出された利用目的，利用方法の例を示す．以下の例では，用途情報に該当する部分を下線で示している．

- 利用目的
  - We use WordNet for lexical lookup.
  - The use of WordNet enables a more systematic and more detailed attachment of such marks.
  - WordNet is a valuable resource for semantic annotation.
  - The assumed baseline is the algorithm that tags the corpus according to the first WordNet sense.
- 利用方法
  - We outline a mechanism for deriving new concepts from WordNet using metonymy.
  - Finally we assign to each noun its corresponding WordNet code.

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

まず，pdftotext ツール [1] を用いて，学術論文からテキストを抽出する．そして，抽出したテキスト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 [4] 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

LREC2004 の論文集から抽出した用途情報を分析したところ，用途情報の記述に，動詞が重要な役割を果たしていることがわかった．そこで，用途情報記述に利用される動詞を抽出し，次の 3 点に着目することにより 6 つのクラスに分類し，分類したクラスごとに抽出ルールを生成する．1 点目は，動詞が一般動詞，be 動詞のいずれであるかである．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．

- **Usage クラス**
  目的語を 1 つとる以下のような一般動詞が用いられ，言語資源名が目的語に含まれる．

  use, utilize, exploit, employ, apply, etc.

  動詞が前置詞を必要としない場合は図 1 の構造，必要とする場合は図 2 の構造となる．

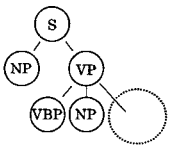

図 1: Usage1

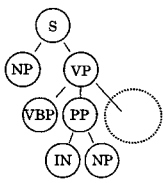

図 2: Usage2

  このとき，点線で囲まれた部分に図 3〜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．ただし，図 3 と図 4 における前置詞 IN は for，in，on，as，towards のいずれかとする．

- **Contribution クラス**
  目的語をとる以下のような一般動詞が用いられ，言語資源名が主語に含まれる．

  contribute, enable, allow, provide, etc.

  図 7 または図 8 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．

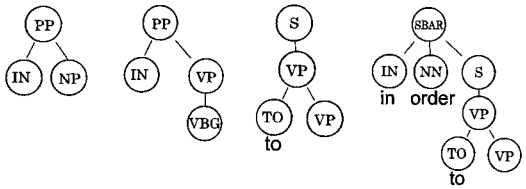

図 3: 用途 1　図 4: 用途 2　図 5: 用途 3　図 6: 用途 4

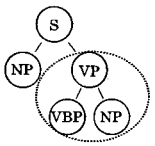

図 7: Contribution1

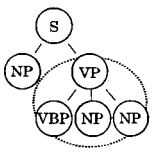

図 8: Contribution2

- **Derivation クラス**
  目的語を 2 つとる以下のような一般動詞が用いられ，言語資源名が前置詞の目的語に含まれる．

  derive, obtain, extract, acquire, etc.

- **Linkage クラス**
  目的語を 2 つとる以下のような一般動詞が用いられ，言語資源名が目的語または前置詞の目的語に含まれる．このクラスは概念辞書である 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．

  assign, match, link, merge, map, etc.

  図 9 または図 10 のような前置詞を含む構造もしくは，and を用いた図 11 のような構造をとるとき，点線で囲まれた動詞句を用途情報として取得する．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．

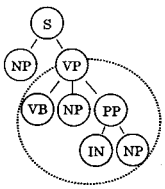

図 9: Linkage1

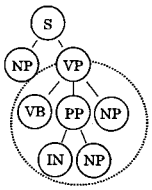

図 10: Linkage2

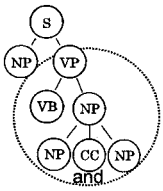

図 11: Linkage3

- **Explanation クラス**
  be 動詞や以下の形容詞が用いられている．

  useful, valuable, available, helpful, etc.

- **Source クラス**
  動詞以外の以下のような表現が用いられている．

  according to, based on, through, etc.

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

上記の6つのクラスのうち，Usage, Contribution, Explanation, Source クラスを利用目的，Derivation, Linkage クラスを利用方法として抽出する．

5 評価

5.1 LREC, WordNet を用いた実験

抽出ルールの生成に利用した LREC2004 の論文集に含まれる 214 の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 を分析し，抽出した 197 の用途情報を正解データとしてオープンテストを行った．

実験結果を表 1 に示す．クローズドテストで 72.9% の再現率と 78.4%の精度，オープンテストで 60.9%の再現率と 72.7%の精度を得ることができた．オープンテストにおいて，再現率・精度の大きな低下が見られなかったことから，抽出ルールの汎用性が示された．

表 1: LREC, WordNet を用いた実験結果.

	LREC2004(closed)		LREC2006(open)	
	再現率 (%)	精度 (%)	再現率 (%)	精度 (%)
利用目的	71.3 (102/143)	80.3 (102/127)	60.9 (70/115)	70.0 (70/100)
利用方法	76.1 (54/71)	75.0 (54/72)	61.0 (50/82)	76.9 (50/65)
合計	72.9 (156/214)	78.4 (156/199)	60.9(120/197)	72.7(120/165)

5.2 ベースラインとの比較

本手法は，構文情報を活用する点に特徴がある．本節では，その利点を明らかにするために，構文情報を用いない方法をベースラインの手法として本手法と比較する．ベースラインの手法では，言語資源名と動詞などのキーワードを含む文を抽出する．この比較では，用途情報を含む文のうち正しく抽出できたものの割合 (再現率) と自動抽出した文のうち用途情報を含んでいたものの割合 (精度) で評価した．本手法については，抽出した用途情報を含む文を出力として評価した．

実験結果を表 2 に示す．構文情報を利用した本手法の有用性を確認した．

表 2: ベースラインとの比較実験結果.

		再現率 (%)	精度 (%)	F 値
LREC2004 (closed)	ベースライン	93.3 (180/193)	28.2 (180/638)	43.3
	本手法	77.7 (150/193)	78.1 (150/192)	77.9
LREC2006 (open)	ベースライン	90.2 (157/174)	27.6 (157/568)	42.3
	本手法	67.8 (118/174)	70.2 (118/168)	69.0

5.3 異なる言語資源を用いた実験

5.3.1 同じ種類の言語資源への適用

WordNet と同種の言語資源である FrameNet に適用し，同種の言語資源に対する用途情報の抽出を試みた．

表 3 に実験結果を示す．言語資源に依存した表現である Linkage クラスを含む利用方法も抽出できていたため，WordNet と同種の言語資源に対してはルールをそのまま適用することで抽出が可能であると考える．

表 3: LREC, FrameNet を用いた実験結果.

	LREC2004(closed)		LREC2006(open)	
	再現率 (%)	精度 (%)	再現率 (%)	精度 (%)
利用目的	75.0 (6/8)	85.7 (6/7)	78.9 (15/19)	78.9 (15/19)
利用方法	50.0 (3/6)	100 (3/3)	69.2 (9/13)	90.0 (9/10)
合計	64.3 (9/14)	100 (9/9)	75.0 (24/32)	82.8 (24/29)

5.3.2 異なる種類の言語資源への適用

WordNet とは異なる種類の言語資源である Penn Treebank への適用を行い，異なる言語資源に対する用途情報の抽出を試みた．まず，LREC2004 の論文集を用いてオープンテストを行ったところ，Linkage クラスを含む利用方法の抽出が失敗したことにより，精度は 75.0%であったものの，再現率が 44.1%と低くなった．そこで，Penn Treebank に特有の表現を抽出するため，LREC2004 の論文集を分析し，以下の動詞を抽出した．

convert, translate, transform, parse, train

上記の動詞をルールに加え，実験を行った．

実験結果を表 4 に示す．Penn Treebank に特有の表現に関して，LREC2006 の論文集を用いたオープンテストにおいても，抽出できていることから，言語資源の種類に依存した表現を抽出することにより，異なる種類の言語資源に対しても適用できると考える．

表 4: LREC, Penn Treebank を用いた実験結果.

	LREC2004(closed)		LREC2006(open)	
	再現率 (%)	精度 (%)	再現率 (%)	精度 (%)
利用目的	83.3 (10/12)	83.3 (10/12)	38.5 (5/13)	55.6 (5/9)
利用方法	65.0 (13/20)	76.5 (13/17)	72.7 (8/11)	72.7 (8/11)
合計	67.6 (23/34)	79.3 (23/29)	54.2(13/24)	65.0(13/20)

表 5: ACL, WordNet を用いた実験結果.

	ACL2004(closed)		ACL2005(open)	
	再現率 (%)	精度 (%)	再現率 (%)	精度 (%)
利用目的	71.7 (91/127)	71.1 (91/128)	66.0 (64/97)	66.0 (64/97)
利用方法	65.8 (52/79)	86.7 (52/60)	40.5 (15/37)	78.9 (15/19)
合計	69.4(143/206)	76.1(143/188)	59.0(79/134)	68.1(79/116)

5.4 異なる論文集への適用

異なる論文集として，ACL への適用を行った．実験結果を表 5 に示す．クローズドテストで 69.4%の再現率と 76.1%の精度，オープンテストで 59.0%の

再現率と68.1%の精度を得ることに成功した．以上より，異なる論文集に対しても適用可能であると考える．

5.5 抽出例

抽出に成功した言語資源WordNetの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
- for query expansion
- to cluster its senses

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，クエリ拡張，クラスタリングなどの具体的な利用目的を抽出することができた．

次に，利用方法の抽出例を以下に示す．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

5.6 考察

異なる言語資源や論文集に適用することにより，抽出ルールの洗練化を図った．抽出ルールの洗練化に関しては，以下の項目が挙げられ，1. が最も変更が大きく，3. が最も変更が少ない．

1. 抽出ルールの追加
2. 抽出ルールの修正
3. 動詞などの特徴的な表現の追加

本研究で行った洗練化では，3. が最も多く行われ，1. はあまり行わなかった．つまり，抽出ルールの大きな変更はあまりなく，細かな修正などが多く行われた．この結果は、本稿で述べた抽出ルールが，様々な言語資源や論文集に対しても汎用的であることを示唆している．

6 用途情報の利用

6.1 SHACHIとの連携

言語資源メタデータデータベースSHACHI[11, 12]との連携を行った．現在，SHACHIには約1800件の言語資源に関するメタデータが格納されている．言語資源の効率的な利用の実現のため，SHACHIに用途情報を格納した．

SHACHIに収録されている言語資源を対象に，LREC2004とLREC2006の論文集に本手法を適用した．その結果，91種類の言語資源に対して，728の用途情報を含む文が抽出された．ただし，本手法が対象とした言語資源とSHACHIに収録されている言語資源では多少異なる点が存在する．本手法では，言語やバージョンなどが異なる言語資源が複数存在する場合，これらを同一の言語資源とみなして用途情報を抽出した．しかし，SHACHIにおいて，これらは異なる言語資源として収録されているため，抽出した用途情報がどの言語資源の用途情報であるかを識別する必要がある．そこで，用途情報を含む文に対して適切な言語資源への分類を行った．

表 6: 用途情報の検索実験.

	用途情報なし			用途情報あり		
キーワード	検索数	正解数	精度 (%)	検索数	正解数	精度 (%)
information retrieval	59	58	98.3	79	72	91.1
summarization	11	6	54.5	18	11	61.1
question answering	7	2	28.6	16	11	68.6

6.2 言語資源への分類

91種類中57種類の言語資源が分類を必要とする言語資源であった．抽出した文とその文を含む論文中に出現する言語資源名，分類候補となる言語資源名を用いて分析，分類を行った．分析の結果，言語資源への分類には言語やバージョン，収録データに関する情報が重要であることがわかった．また，分類対象が特定できない場合，全ての候補に分類することとした．

用途情報を含む597文を，分類候補に対して分類するか否かをSupport Vector Machine (SVM) を用いて判定した．用途情報を含む文と分類候補のペア8137のデータを10分割し，交差検定法を用いて学習・テストを行った．素性として，抽出した文と論文中の言語資源名を含む名詞句に出現した言語やバージョンに関する情報を利用した．SVMの学習にはTinySVM [2] を用いた．分類の結果，精度が88.5%，再現率が61.5%となり，252件の言語資源に分類した．分類結果に基づき，用途情報を含む文をSHACHIに格納した．

6.3 言語資源の検索実験と考察

抽出した用途情報の有用性を確認するため，用途情報が言語資源の検索に寄与するかを検証するための実験を行った．

まず，SHACHIに登録されているメタデータに対して，キーワード検索を行い，キーワードに適した

[2] http://chasen.org/taku/software/TinySVM/

言語資源が検索できるかどうかを用途情報の有無で比較した．結果を表6に示す．用途情報が存在することによって，検索された言語資源の数や有用な言語資源の数が増加していることがわかる．また，用途情報によって検索された言語資源の精度(有用な言語資源の数/検索された言語資源の数)に関しても比較的高い値を示している．以上より，用途情報が言語資源の検索に有用であると考える．

さらに，被験者7名に対して，10分間の制限時間を設け，表6のキーワードに関する言語資源を検索する被験者実験を行った．実験の結果，7名中6名の被験者が「用途情報あり」の検索において，新たな言語資源の発見に成功した．また，アンケートにより，7名中6名が「用途情報あり」の検索の方が言語資源を効率的に検索できると回答した．これらにより，用途情報が言語資源の効率的な検索に寄与すると考える．

分類を行う前の用途情報の抽出精度は57.3%であった．これは言語資源を対象として用途情報を抽出したにも関わらず，組織やプロジェクト，ツールなどの言語資源とは異なるものが対象となる場合が存在したためである．この問題に対しては分類を行う際に，projectやtoolなどの単語を含むものを分類対象から除去することにより，70.4%の精度を達成することができた．十分とは言えないものの，実用可能な範囲と考える．

本稿で抽出した用途情報は，SHACHIに収録されている全言語資源の約20%程度しか網羅できていない．今後，LREC以外の論文集に対して本手法を適用し，網羅率を上げていく必要がある．また，ドメイン毎にクラスタリングなど，抽出した用途情報の整形により，より効率的な検索が可能となるだろう．

7 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．実験の結果，実用的なレベルで言語資源の利用目的，利用方法の抽出ができたと考える．また，言語資源メタデータデータベースSHACHIに本手法により抽出した用途情報を応用することにより，用途情報が言語資源の検索に寄与することを示した．

今後の課題としては，抽出ルールの洗練化，および，Webへの適用が挙げられる．Webからは，論文から抽出される用途情報とは異なる用途情報が得られることが期待される．また，より多くの論文集からの用途情報の抽出や抽出した用途情報の整形，抽出誤りによる検索への影響の調査なども今後の課題である．

参考文献

[1] G. Ralph, S. Beth: 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] E. Charniak: A Maximum-extropy-inspired Parse, NAACL-2000, pp. 132-139 (2000).

[5] J. Pustejovsky: The Generative Lexicon, MIT Press (1995).

[6] M. De Boni, S. Manandhar: Automated Discovery of Telic Relations for WordNet, IWC-02 (2002).

[7] T. Veale: Qualia Extraction and Creative Metaphor in WordNet, IJCAI-03 (2003).

[8] P. Bouillon, V. Claveau, C. Fabre, P. Sebillot: Acquisition of Qualia Elements from Corpora - Evaluation of a Symbolic Learning Method, LREC-2002, pp. 208-215 (2002).

[9] I. Yamada, T. Baldwin: Automatic Discovery of Telic and Agentive Roles from Corpus Data, PACLIC-18, pp. 115-126 (2004).

[10] P. Cimiano, J. Wenderoth: Automatic Acquisition of Ranked Qualia Structures from the Web, ACL-2007, pp. 888-895 (2007).

[11] H. Tohyama, S. Kozawa, K. Uchimoto, S. Matsubara and H. Isahara: SHACHI: A Large Scale Metadata Database of Language Resources, ICGL2008, pp.205-212 (2008).

[12] 遠山仁美, 小澤俊介, 内元清貴, 松原茂樹, 井佐原均: 言語資源メタデータデータベースSHACHIの構築, 言語処理学会第14回年次大会発表論文集 (2008).