言語資源の検索における用途情報の利用

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

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．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．こうした言語資源の用途情報は，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．ここで，言語資源 X の用途情報とは「A のために X を利用する」という表現に言い換え可能な表現 A であると定義する．例えば，WordNet の用途情報として「辞書引き」や「語義曖昧性解消」などが挙げられる．こうした用途情報を検索キーワードとし，用途に適合する言語資源を探し出すことができれば，利用者が目的に適した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．

本論文では，言語資源の効率的利用に向けた新しい知見として，自動抽出した用途情報が言語資源の広範な検索に貢献することを示す．具体的には，自動抽出した用途情報を言語資源メタデータデータベースと組み合わせ，メタデータのみを用いた場合と用途情報とメタデータを併用した場合とで研究トピックのキーワードに合致する言語資源の件数を比較する．

2　用途情報の有用性

2.1　用途情報とは

対象 X の用途情報とは,「A のために X を利用する」という表現に言い換え可能な表現 A であると定義する．以下に WordNet の用途情報の例を示す．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．

- We use WordNet for lexical lookup.
- Finally we assign to each noun its corresponding WordNet code.

ただし，用途情報は具体的な事象を表すものとする．これは “our proposed method” や “this purpose” などの抽象的な情報を対象外とするためである．また，以下の例のように,「X のために X を利用する」と言い換えられる表現 X，つまり，ある言語資源 X そのものの拡張や更新，修正を表す表現も対象外とする．

- We applied an automatic mapping from WordNet 1.6 to WordNet 1.7.1 synset labels.

2.2　用途情報の有用性の検証

本節では，まず，人手で抽出した用途情報を利用することによって，広範に言語資源が検索できるようになることを検証することにより，用途情報そのものの有用性を示す．言語資源の用途を「言語資源の公式ホームページもしくは既存の学術論文で発表されている用途」とし，言語資源のカタログ情報に用途情報を加えたものを用途情報データベースとして用いた．カタログ情報には言語資源メタデータデータベース SHACHI[11] に格納されている約 2100 の言語資源に関するメタデータを用いた．実験では，メタデータのうち，‘type.purpose’ を用いた．‘type.purpose’ には言語資源の利用目的が記載されている．これは，言語資源のホームページから人手により抽出したものである．一方，用途情報としては LREC2004 の論文集から次のようにして人手により抽出したものを用いた．まず，pdftotext ツール[1]を用いて，LREC2004 の論文集からテキストを取り出した．そして，そこから SHACHI に登録されている各言語資源についてそれぞれの言語資源名を含む文を抽出した．ここで抽出されたのは約 2800 文であった．この約 2800 文を分析したところ 247 件の言語資源に関する 440 件の用途情報が含まれていることが分かった．

検証では ACL2008 の研究トピックから抽出した 40 種類のキーワードを検索クエリとして用いた．検索対象はメタデータおよび用途情報とし，40 種類のキーワードと文字列照合した結果，一致した場合に正しく検索されたものと仮定した．検証はメタデータのみを検索対象とした場合とメタデータと用途情報の両方を検索対象とした場合のそれぞれについて行い，検索結果を比較した．実験結果を表 1 に示す．用途情報が加わることにより，9 種類のキーワードにおいて言語資源の数が増加したがわかる．この結果から，抽出対象を他の論文集や Web などへ広げることにより，より幅広い言語資源の検索ができるようになり，言語資源の利用の可能性が広がることが期待できる．

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

表 1: 用途情報の有用性の検証実験.

キーワード	メタデータ	メタデータ+用途情報
Dialogue	8	8
Enbodied Conversational Agents	0	0
Language-enhanced Platforms	0	0
Information Retrieval	52	54
Text Data Mining	0	0
Information Extraction	11	11
Filtering	0	1
Recommendation	0	0
Question Answering	0	2
Topic Classification	0	0
Text Classification	3	3
Sentiment Analysis	0	0
Attribute Analysis	0	0
Genre Analysis	0	0
Language Generation	1	1
Summarization	8	10
Machine Translation	55	56
Language Identification	21	21
Multimodal Processing	0	0
Speech Recognition	211	266
Speech Generation	1	1
Speech Synthesis	32	32
Phonology	0	0
POS Tagging	1	1
Syntax	0	0
Parsing	11	14
Grammar Induction	0	0
Mathematical Linguistics	0	0
Formal Grammar	0	0
Semantics	1	1
Textual Entailment	0	0
Paraphrasing	0	0
Word Sense Disambiguation	0	2
Discourse	10	11
Pragmatics	0	0
Statistical and Machine learning	0	0
Language Modeling	25	25
Lexical Acquisition	0	0
Knowledge Acquisition	0	0
Development of Language Resources	0	0

3 自動抽出した用途情報の有用性の検証

本節では，自動抽出した用途情報が言語資源検索へ与える影響について述べる．検証には言語資源メタデータデータベース SHACHI[11] に格納されている約 2100 の言語資源に関するメタデータに，自動抽出した用途情報を加えたものを用途情報データベースとして用いた．用途情報データベースの作成の流れを図 1 に示す．まず，抽出ルール [2] を用いて論文集から用途情報を抽出する．そして，抽出した用途情報を適切な言語資源に分類し，用途情報データベースに格納する．

3.1 抽出ルールの適用

SHACHI に収録されている言語資源を対象に，LREC2004 と LREC2006 の論文集に抽出ルールを適用した．その結果，91 種類の言語資源に対して，用途情報を含む 728 文が抽出された [3]．ただし，抽出ルールが対象とした言語資源と SHACHI に収録されている言語資源とでは多少異なる点がある．抽出ルールでは，言語やバージョンなどが異なる言語資源が複数存在しても，それらを特に区別することなく，同一の言語資源として扱っている．このような場合，以降では，言語資源クラスと呼ぶ．例えば，English WordNet と Arabic WordNet をともに WordNet として抽出する．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．そのため，抽出した 91 種類の言語資源クラスに対する用途情報が用途情報データベース中の

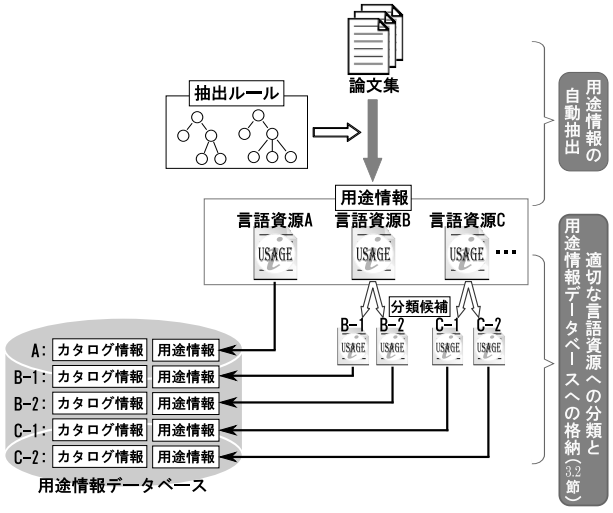

図 1: 用途情報データベース作成の流れ

どの言語資源の用途情報であるかを識別し，適切な言語資源に分類する必要がある．

3.2 適切な言語資源への分類と用途情報データベースへの格納

分類に重要となる情報を調べるため，人手で適切な言語資源に分類し，結果を分析した．91 種類中 57 種類の言語資源クラスに関する 597 文が分類を必要とするものだった．この 597 文に対し，抽出した文，および，その文を含む論文中に出現する言語資源名，分類候補を参照することにより適切な言語資源へ分類した．ただし，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．ここで，分類候補とは例えば，WordNet に対する Arabic WordNet や English WordNet，EuroWordNet のことであり，各言語資源クラスに対して，あらかじめ人手でリスト化している．分類結果を分析したところ，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．

用途情報を含む文を適切な言語資源へ自動分類するため，用途情報を含む 597 文と分類候補とのペアを作成し，各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．597 文のうち，LREC2004 の論文集から抽出した 299 文と分類候補のペア 4189 のデータを学習データとして，LREC2006 の論文集から抽出した 298 文と分類候補のペア 3960 のデータをテストデータとして用いた．素性としては，抽出した文，および，論文中の言語資源名を含む名詞句に共起して出現した言語名やバージョンに関する情報を利用した．例えば，SpeechDat に関する以下の文では，German や Spanish などの言語名や II や FDB-4000 のようなバージョンに関する情報を素性として利用した．

- The automatic speech recognition experiments us-

表 2: 自動抽出した用途情報の有用性の検証実験.

	メタデータ		メタデータ+用途情報	
キーワード	キーワードのみ	キーワード+類義語	キーワードのみ	検索精度 (%)
Dialogue	8	10	8	-
Enbodied Conversational Agents	0	0	0	-
Language-enhanced Platforms	0	0	0	-
Information Retrieval	52	57	53	100 (53/53)
Text Data Mining	0	1	0	-
Information Extraction	11	26	12	100 (12/12)
Recommendation	0	0	0	-
Filtering	0	0	0	-
Question Answering	0	1	3	100 (3/3)
Topic Classification	0	4	0	-
Text Classification	3	4	3	-
Sentiment Analysis	0	0	0	-
Attribution Analysis	0	0	0	-
Genre Analysis	0	0	0	-
Language Generation	1	1	1	-
Summarization	8	8	9	75.0 (9/12)
Machine Translation	55	55	55	-
Language Identification	21	21	21	-
Multimodal Processing	0	2	0	-
Speech Recognition	211	230	274	100 (274/274)
Speech Generation	1	2	1	-
Speech Synthesis	32	37	32	-
Phonology	0	8	0	-
POS Tagging	1	18	36	97.3 (36/37)
Syntax	0	3	0	0 (0/2)
Parsing	11	12	11	84.6 (11/13)
Grammar Induction	0	0	0	-
Mathematical Linguistics	0	0	0	-
Formal Grammar	0	0	0	-
Semantics	1	4	1	33.3 (1/3)
Textual Entailment	0	0	0	-
Paraphrasing	0	0	0	-
Word Sense Disambiguation	0	1	0	-
Discourse	10	10	15	100 (15/15)
Pragmatics	0	5	0	-
Statistical and Machine Learning	0	7	0	-
Language Modeling	25	25	25	-
Lexical Acquisition	0	0	0	-
Knowledge Acquisition	0	0	0	-
Development of Language Resources	0	14	0	-

ing German and Spanish SpeechDat(II) FDB-4000 databases show performance improvement.

また，SVM の学習には TinySVM [2] を用いた．まず，学習したモデルを学習データに適用したところ，92.0%の精度と 61.8%の再現率が得られた．次に，テストデータに適用したところ，精度が 23.8%，再現率が 63.0%となり，分類性能が低くなった．これは，LREC2004 と LREC2006 では，抽出される用途情報や対象となる言語資源が異なっていることが原因であると考えられる．しかし，この問題は，学習データの量を増やすことで解決できると考える．

分類の結果，597 文のうち，250 文は 1 つの言語資源に分類され，114 文は複数の言語資源に対して分類された．残りの 233 文はどの言語資源に対しても分類されなかった．また，57 種類の言語資源クラスに関する用途情報は 212 件の言語資源に分類された．分類を必要としなかった 34 種類を含んだ 246 件の言語資源に対して，用途情報を含む文を用途情報データベースに格納した．

3.3 自動抽出した用途情報の検証実験

まず，用途情報データベースに対して，ACL2008 の研究トピックから抽出した 40 種類のキーワードを用いた実験を行った．本実験では，2.2 節の実験と同様，言語資源の用途を「言語資源の公式ホームページもしくは既存の学術論文で発表されている用途」とする．キーワードを用いて，用途情報データベース中のメタデータ‘type.purpose’ および用途情報と文字列照合した結果，文字列照合に成功した言語資源を正解として検索する．ただし，用途情報には抽出ルールによる抽出ミスが存在するため，抽出に成功しているもののみを正解とした．検索対象としてメタデータのみを用いた場合とメタデータと用途情報の両方を用いた場合についてそれぞれ実験を行い，検索結果を比較した．また，メタデータのみを対象として，40 種類のキーワードに対する類義語を用いた検索も行った．ここで利用した類義語は，対応する類義語辞書が存在しなかったため，‘type.purpose’ に含まれるデータを分析し，人手により生成した．例えば，POS Tagging の類義語には morphological parsing や part of speech tagging などがあり，表層上または意味上の差異を吸収した類義語が設計されている．類義語の設計には主観が含まれるため，検索結果が多少増減する可能性があるものの，メタデータのみを検索対象とした場合に検索される言語資源数の上限値を表すと言える．

実験結果を表 2 に示す．類義語を利用することで 20 種類のキーワードに対して検索された言語資源の件数が増加した．これに対し，用途情報を利用した場合では 10 種類のキーワードに対して，検索される言語資源の件数が増加した．ただし，抽出ミスの影響のため，実際に言語資源数が増加したキーワードは 7 種類であった．一部のキーワードにおいて，検索精度が低下してしまったものの，Speech Recognition や POS Tagging, Discourse においては類義語を利用した場合に勝る結果となっている．このことは用途情報を用いることにより，メタデータのみでは検索できない言語資源が発見で

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

表 3: 被験者実験の結果.

被験者 ID	1	2	3	4	5	6	7
「用途情報なし」で発見した言語資源数	5	2	3	2	5	2	4
「用途情報あり」で発見した言語資源数	5	2	3	4	5	3	2
用途情報が加わることにより発見された言語資源数	2	1	0	1	1	2	2

きるようになることを示している．例えば，Wikipedia はメタデータのみでは検索できないが，以下の用途情報が加わることにより，“Question Answering” で検索できるようになった．これにより，Wikipedia の広範な利用状況を知ることもできる．このように，用途情報は広範に言語資源を検索するのに有用であると言える．

- Results show that Wikipedia is a potentially useful resource for the Question Answering task.

さらに，被験者 7 名に対して，10 分間の制限時間を設け，被験者 ID1〜3 には information retrieval，4 と 5 には summarization，6 と 7 には question answering に関する言語資源を発見するタスクを与え，用途情報の有無により結果を比較した．被験者はまずキーワード検索を行う．検索結果は用途情報またはメタデータのいずれかにキーワードを含む言語資源のリストで構成されている．メタデータには，‘type.purpose’ と ‘description’ を用いた．被験者はリスト中の言語資源のメタデータおよび用途情報を参照することにより，適した言語資源を選択する．実験結果を表 3 に示す．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,「用途情報なし」では検索できなかった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．

4 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）[9] 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 [10] 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 [5] やコーパスから抽出する手法 [6, 12]，Web から抽出する手法 [8] 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 [1, 4]．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を行う．

5 まとめ

本論文では，自動抽出した用途情報が言語資源の検索の効率化に貢献することを示した．自動抽出した用途情報を言語資源メタデータデータベースと組み合わせることにより，用途情報を用いない場合に比べて，目的に合う言語資源を精度良く検索できるようになった．

本論文では論文を対象として用途情報を抽出したが，Web への適用は今後の課題である．なぜなら，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．ただし，Web へ本手法を適用する際には抽出ルールの更なる洗練が必要となるだろう．

参考文献

[1] 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).

[2] 小澤俊介, 遠山仁美, 内元清貴, 松原茂樹: 言語資源の効率的利用のための用途情報抽出, 言語処理学会第 14 回年次大会, pp. 1069-1072 (2008).

[3] 小澤俊介, 遠山仁美, 内元清貴, 松原茂樹: 言語資源の用途情報の抽出と利用, NL-184, pp. 77-82 (2008).

[4] 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).

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

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

[7] Charniak, E.: A Maximum-entropy-inspired Parse, NAACL-2000, pp. 132-139 (2000).

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

[9] Grishman, R. and Sundheim, B.: Message Understanding Conference - 6: A Bried History, COLING-96, pp. 466-471 (1996).

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

[11] SHACHI: http://shachi.org/.

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