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

2007－DBS－142 (2)
2007－ＦＩ－ 87 (2)
2007／5／31

語釈拡張に基づくテキスト項目の同定

大久保幸太† 三浦 孝夫†

† 法政大学 工学研究科 電気工学専攻 〒184–8584 東京都小金井市梶野町 3-7-2

あらまし 本稿では，辞書の統合を目的として，項目どおしを同定する手法を提案する．このオブジェクト同一性を効率よく判定するため，構文解析などの言語特有の知識を用いず，ベクトル空間モデルに基づいて，情報検索アプローチについて議論する．また，実験によって，提案手法の有効性を示す．

Identifying Text Objects Based on Expansion

Kouta OHKUBO† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University 3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan

Abstract In this investigation, we propose a sophisticated approach to identify items in several dictionaries for the purpose of integration. homonymous, synonymous and polysemy words. In this work, we put our attention To identity synonymous words efficiently, we discuss IR approach based on vector space modeling to sentences in item explanations, without any knowledge of grammatical analysis or any other NLP analysis. We show theusefulness by some experimental results.

1. 前 書 き

近年のインターネットの急速な拡大により，簡単に短時間で様々な情報を入手できるようになった．反面，よく知られているように，内容の曖昧さや情報相互の矛盾が多く，信頼性の高い利用を仮定できない．しかし，伝統的な本は深く校正し出版されている．それは，信頼性，一貫性およびよくまとまった構造を保証する．`WikiPedia` では多数の読者による信頼性向上や内容の洗練化を行う，欠点を解決するアプローチのうちの一つである．

本稿では信頼性のある一定の範囲の知識の集合体を **辞書** (dictionary) と呼ぶ．辞書では，知識を表現するために，ある単位で項目化しており，これを **見出し語** (item)，その解釈 (semantics) を **語釈** (explanation) と呼ぶ．

Web サイトは，個々に何かのトピックに関する知識を含む．例えば Shakespeare サイトでは Oscar Wild に関する知識を含まないであろう．利用者は検索エンジンなどを介してあるトピックに関する知識 (見出し語) が複数サイトで存在することを知る．しかし同じ知識内容が異なる表現 (語釈) で表されることが多く，複数サイトを調べても新たな情報を見出せない．内容の重複を取り除き，知識 (辞書) の統合が可能になれば，真に新たな情報のみを効率よく抽出することができる．

辞書統合には数多くの困難がある．まず，重複した項目数が多数存在するという量 (*scalability*) の問題がある．また，**多義項目** (複数の意味を持つ見出し語) が存在する．例えば `bank` という語には "銀行" の他に "堤防" という意味が含まれる．多くの場合，これらは意味別あるいは形態素別に表記上の区別を持ち，1 つの見出し語は複数の意味を混同しない．

これに対し，同義項目 (意味を共有する異なった見出し語) は扱いが難しい．例えば `student` と `pupil` は同じ意味を所持するため，統合にはこれらオブジェクトの **同一性判定** (Object Identification) が必要となる．オブジェクト (見出し項目) の同定とは，ある概念を特定し辞書間の項目統合を可能にする方法である．

本稿では，オブジェクト同一性判定を効果的に行う手法を提案する．同義語を効率的に同一性判定するために，文法解析あるいは NLP 分析の知識を用いず，各見出し語に対して，語釈内容を文書とし，ベクトル空間モデルに基づいた IR アプローチについて議論する．

本論文の構成は次の通りである．2 章では，必要となる

概念の導入と関連した定義を要約する．3 章では，辞書統合の定義とオブジェクト同一性を述べ，同一性判定のアイデアを示す．続く 4 章では，この同一性判定の実験を示し，アイデアの有効性を考察する．関連研究を 5 章で述べ，6 章は結論である．

2. 情報表現と辞書

2.1 ベクトル空間による情報表現

多くの場合，知識は文書で表現され，文書はテキスト情報，即ち語の並びとして構成される．

形態素 (Morphology) とは，これ以上に細かくすると意味を失う最小の文字列を言う．文章を形態素に分解する処理を**形態素解析**と呼ぶ．英語では形態素はそれぞれ，ある種のタグ (品詞) を備えた単語に相当するが，日本語では語の抽出を含む形態素の検出方法も問題である．

テキスト情報を探索するには，出現する各語の (出現頻度等) 特徴を値としてベクトル化する**ベクトル空間モデル**が一般的である [4]．一般にテキスト文書 d は出現する語 $w_1, .., w_n$ のベクトルで表現される:

$d = (v_1, ..., v_n)$

ここで v_i は語 w_i に対応する数値であり一般に出現有無 (Binary Frequency) や出現頻度 (Term Frequency) であることが多い．このとき 2 つの文書 d_1, d_2 の類似度は出現数の分布を用いて定義され，これはベクトルの余弦値によって算出できる．

この方法は，モデル化が単純であり類似度も簡単に算出できることから，広く利用されているが，解が重み付け方法に依存し，次元数が数万にも及ぶ高次元データをそのまま扱うと，効率，計算機容量の確保および即応性への対応が困難になる．このため，テキスト情報の次元を縮小し改善を図る**次元縮小技法**が知られている [4]．次元縮小技法では高次元文書ベクトルを低次元空間に射影し，この部分だけを検索対象とするため，効率よく探索範囲を絞り込むことができる．

2.2 辞書と語釈

本稿の目的は，辞書中のふたつの見出し語が同じ意味を所持するかどうか，それらを同一性判定する (**同定する**という) ことにある．このため，語彙参照ツールとして**ワードネット** (Wordnet) を利用する [8], [15].

WordNet は，約 95,600 語のの名詞, 動詞, 形容詞が約 70, 000 の意味あるいは同義語のセットで組織される一般分野のオンライン語彙参照システムである．それらは一般的知識 (分野に特有でない) を含んでいる．実験では GCIDE および COBUILD 等の一般的な辞書を取り上げるため，そのような一般参照システムを用いる．そのため，本稿では WordNet の特徴および意味関係を詳細に利用する．

WordNet には，いくつかの種類の意味関係が保存されており，**同意語**，**反意語**，**上位語**，**下位語**，**部分語**，**全体語**などが定義されている．

The noun human has 2 senses (first 2 from tagged texts)

1. (7) person, individual, someone, somebody, mortal, **human**, soul -- (a human being; "there was too much for one person to do")
2. (5) homo, man, human being, **human** -- (any living or extinct member of the family Hominidae)

The adj human has 3 senses (first 3 from tagged texts)

1. (47) **human** -- (characteristic of humanity; "human nature")
2. (20) **human** -- (relating to a person; "the experiment was conducted on 6 monkeys and 2 human subjects")
3. (15) **human** -- (having human form or attributes as opposed to those of animals or divine beings; "human beings"; "the human body"; "human kindness"; "human frailty")

図 1 ワードネットによる検索結果 (human)

The noun person has 3 senses (first 2 from tagged texts)

1. (7229) **person**, individual, someone, somebody, mortal, human, soul -- (a human being; "there was too much for one person to do")
2. (11) **person** -- (a person's body (usually including their clothing); "a weapon was hidden on his person")
3. **person** -- (a grammatical category of pronouns and verb forms; "stop talking about yourself in the third person")

図 2 ワードネットによる検索結果 (person)

本稿では，2 つの単語が類似概念を定義しているとき，これを**類義語**という．例えば `lofty` は `high` の類義語である．**多義語**は複数の意味を持つ単語をいう．例えば，`picture` は `movie, figure, photo, illustration` などを意味する．ワードネットはシソーラス辞書とは異なり，辞書を構成する意味の基本単位として *synset*(synonym set) を用いる．synset は，様々な階層関係を提供する同義語の集合である．約 95, 600 単語が含まれている場合，synset は，4 つの品詞 (名詞, 動詞, 形容詞および副詞) によるグループに組織される．すべての synset は，同じ意味を持ついくつかの単語を含んでいる．単語が多数の意味を所持するとき，その単語はいくつかの synset の中に含まれる．

図では，`human` をワードネットで検索した結果を示す．これより頻度や意味，同義語の単語の集合などを得る．またワードネットは品詞毎に結果を出力する．例えば，`human` は名詞と形容詞の両方に使われ，名詞としては `person/individual...` と `homo/man/human-being...` の 2 つの意味を持つ．

通し番号の後ろの () 内の数字は，ワードネットの作成に使われた文書集合で使用された意味の数を表し，意味の**使用頻度**として扱うことができる．例えば，図 1 より `human` は文書集合内で $7 + 5 + ... = 94$ 回出現しており，そのうち 7 回は `person` という意味で使用されていたことを示す．また，使用頻度の低い意味は頻度がつけられて

いない.

ワードネットでは, synset は表記上で明示されることはなく，ワードネット辞書内で矛盾なく管理されている. 例えば，human, person はそれぞれ synset 番号 5303 を持つ. この synset は図 1，図 2 での第 1 番目の意味に対応している. 即ち， human と person は 5303 という同じ意味を持つ. これらの単語が持つ識別番号を表 1 に示す.

	human	person
NOUN	1:5303	1:5303
	2:2130996	2:4465544
ADJECTIVE	1:324678454	
	2:2634237	
	3:2634331	

表 1　単語が持つ意味番号

3. オブジェクト同定と語釈拡張

本章ではオブジェクト同定処理を定義し，さらに精度を向上させるための語釈を拡張する操作を導入する.

与えられた見出し語 d に対して，その語釈文から**索引語** (ここでは名詞，動詞，形容詞のいずれかとする) を抽出してベクトル化する.

$d = (v_1, ..., v_n)$

ただし，単語 w_k の出現の有無に対応して $v_k = 1$ あるいは $v_k = 0$ であり，w_k は**不要語除去**, **ステミング処理**が施された語幹 (標準形) と考える.

例えば book という単語の語釈文は GCIDE(注1) 辞書では a collection of sheets of paper と表され，索引語は collection, sheet, paper となる.

2 つの語釈文の**類似性**は，ベクトル空間モデルと同様に，索引語ベクトルの**余弦尺度**として定義する.

例えば book という単語は COBUILD(注2) 辞書では a number of pieces of paper と表されており，索引語で判断した場合，collection, sheet, paper と number, piece, paper との類似度は 0.33 である.

辞書を詳細に調べると，たとえ類義語であっても形式的な類似性を有さない語釈がある. ここで言う形式的類似性とは出現する単語に共通性がほとんどないこと，即ちベクトル表現した場合の類似度が極めて小さい状況を言う.

実際，ワードネットでは book は physical objects consisting of a number of pages bound together と表され，索引語は physical, object, consist, number, page, bind となって，同じ book に対する語釈でありながら，共通に出現するものが無い.

この問題を効率よく解決するために，**語釈拡張**と呼ぶ操作を導入する. これは，語釈文に表れる索引語を，改めて辞書の見出し語として置き換えることである. ここでのアイデアは，索引語を改めて辞書を介して置き換えることで，意味を考慮した類似度計算を行うことにある. 索引語を (辞書の) 語釈文と置き換え，新たに語釈文から索引語を抽出する. (ここでは特定の品詞によるフィルタリングを行うとする.)

例えば，ワードネットでは，見出し語 student は語釈 a learner who is enrolled in an educational institution に対応する. このとき索引語は learner, enroll, educational, institution となる. 見出し語 learner について辞書を調べ，名詞 learner は次の 2 つの語釈を有することがわかる.

```
learner(1) - someone who learns or takes up knowledge or beliefs
learner(2) - works for an expert to learn a trade
```

ここから索引語を抽出し learn, take, knowledge, belief, work, expert, trade を得るため，これを student の語釈ベクトルと置き換え，learner, enroll, educational, institution, learn, take, knowledge, belief, work, expert, trade を得る. 以下これを全ての索引語について繰り返す.

ふたつの見出し語について同一性を判定するとき，各々の語釈ベクトルの類似度がしきい値以下であるときに限って，語釈拡張を行う.

この操作にはいくつか特徴がある. 明らかに，語釈拡張の結果，語釈文の類似度が増加し，特に類義語どおしの語釈が形式的類似度を有する可能性が生じる. 反面，出現する語だけから同一性を判断するため，精度が低下し，類義語ではない語の類似度を検出する可能性がある. 語釈拡張を繰り返すにつれ，これらの問題が顕在化するであろう.

4. 実　　験

本章では，いくつかの実験を通じて，語釈拡張のオブジェクト同定に対する効果を検証する. このため，ベクトル空間モデルを用いた類似性判定手法と，語釈拡張を介在した判定手法を実験により比較検討する.

4.1 準　　備

本実験では，2 種類の実験を行う. ベクトル空間モデルを用いた従来の類似性判定手法を用いた実験 1 では，語釈文に含まれる全ての単語を抽出し，不要語除去およびステミング処理を行ったあと，類似度計算を行う. そして，類似度が**しきい値** σ 以上である場合, 同定を決定する.

実験 2 では語釈拡張を介在した判定手法を用いる. ここでは，特定の品詞でフィルタリングした索引語だけを対象

(注1) : GNU Collaborative International Dictionary of English
(注2) : Collins COBUILD Advanced Learner's English Dictionary

とし，さらに，類似度がゼロのとき語釈拡張を適用する．本実験では拡張回数に上限を設け，提案する操作がどれほど効果的であるかを検証する．本実験では語釈拡張を 0, 1,2,3,4 回を上限として実験する．

本実験データでは，ワードネット辞書のうち B から始まる副詞 235 項目と，GCIDE 辞書のうち B から始まる副詞 175 項目の語釈文を用いる．ワードネットは本来は同義語辞書として作成されたものである．単語の有する意味ごとに synset (意味番号) が割り振られており，同じ番号を持つ単語は，同じ意味を表す．例えば次に示すように，**student, pupil** は共に synset 番号として 10505881 を有する．本実験では，これを正解と見なし，見出し語の同定は，synset の同一性で判断する．

語釈拡張の効果を評価するため，情報検索分野で尺度として利用されている**再現率** (Recall) および**適合率** (Precision) を考慮する．ここで再現率 R，適合率 P および F 値は次のように定義される．

$$R = \frac{A}{B},\ P = \frac{A}{C},\ F = \frac{2}{\frac{1}{R} + \frac{1}{P}}$$

式中で A, B, C はそれぞれ同定した項目数中の正解数，全体の項目数中の正解数，同定した項目数を表す．

定義から明らかなように，再現率とは同定範囲の大きさを表し，高い値ほど同定できた見出し語が多いことを表している．一方，適合率は同定精度を表し，高い値ほど正しく同定できた見出し語が多いことを表す．再現率と適合率を統一的に判定するために F 値が利用されるが，本実験ではしきい値 σ の設定に利用する．しきい値は予め予備実験で範囲を絞り込んである．

見出し項目は，同定した項目がワードネット上で共に同じ synset 番号を有するとき正解と見なす．また，全体の項目数中の正解数とは，見出し語項目に含まれる同じ synset 番号となる項目数を合計する．例えば見出し語 **student, pupil** だけが synset 番号 10505881 を共有すれば **student** の表す意味の項目数は 2 となる．**pupil** についても同様である．

4.2 結　　果

実験 1 の結果を表 2 に示す．ここでは，しきい値 $\sigma = 0.48, ..., 0.6$ を設け，この値を超えた場合に同一と判定するものとする．しきい値に応じた再現率，適合率および F 値を示す．

次に，最適しきい値 $\sigma = 0.53$ における見出し語 **back** の同定例を示す (F 値が最大となるしきい値 0.53)．ここでは順に synset 番号, 見出し語 A, 見出し語 B, 類似度, A の索引語, B の索引語 を示している．

```
74749, back, back, 0.7071, back rear - rear
74749, back, behind, 0.8164, back rear - back
part rear
74749, back, behind, 0.7071, back rear - back
part rear backward
74543, back, back, 0.8164, past time - time past
ago
74543, back, before, 0.8164, past time - time
past previous
74543, back, behind, 0.6324, past time - backward
time order success past
```

Threshold	0.48	0.5	0.53	0.54	0.55	0.6
Correct	339	339	339	339	339	339
Identified	2352	243	241	240	240	186
Correctly Identified	52	40	40	39	39	32
Recall	0.15	0.115	0.117	0.115	0.115	0.17
Precision	0.02	0.16	0.165	0.162	0.162	0.17
F-value	0.03	0.137	0.137	0.134	0.134	0.12

表 2　実験 1 結果

この結果から，単純な既存手法では，最適しきい値 $\sigma = 0.53$ のときでも再現率 11.7 %，適合率 16.5% 程度の性能しか得られず，しかもこの状況が安定して続いている．この程度の同定範囲および精度しか得られない．

次に実験 2 の結果を示す．語釈拡張を行わず品詞フィルタリングだけを適用した結果を表 3 に示す．実験 1 とほぼ動向は変わらず品詞フィルタリングだけでは主たる効果が得られないことを示している．以下に，語 **back** の同定状況を示す．

back と **before** の類似度は向上するが **back** と **behind** では逆に低下している．

Threshold	0.50	0.55	0.60	0.65
Correct	339	339	339	339
Identified	3060	265	212	203
Correctly Identified	60	43	43	35
Recall	0.17	0.126	0.10	0.095
Precision	0.019	0.162	0.165	0.152
F-value	0.035	0.142	0.127	0.114

表 3　実験 2A 結果

```
74749, back, back, 1.0, rear - rear
74749, back, behind, 0.5773, rear - back part
rear
74543, back, back, 0.9999, past time - past time
74543, back, before, 0.9999, past time - time
past
74543, back, behind, 0.6324, past time - backward
time order success past
```

次に語釈拡張を 1 度まで許した実験 2B の結果を表 4，2 度まで許した実験 2C の結果を表 5 に示す．主たる変化は無い．

Threshold	0.45	0.50	0.55	0.60
Correct	339	339	339	339
Identified	3060	268	265	212
Correctly Identified	60	44	43	35
Recall	0.17	0.129	0.126	0.10
Precision	0.019	0.164	0.162	0.165
F-value	0.035	0.144	0.142	0.120

表 4　実験 2B 結果

Threshold	0.45	0.50	0.55	0.60
Correct	339	339	339	339
Identified	3139	285	274	215
Correctly Identified	64	44	43	35
Recall	0.18	0.129	0.126	0.10
Precision	0.02	0.154	0.156	0.16
F-value	0.036	0.1403	0.139	0.120

表 5　実験 2C 結果

しかし, 語釈拡張を 3 回まで許した実験 2D では, 状況に大きな違いが現れる. 実験結果を表 6 に結果を示す. F 値に変動はないものの最適しきい値が $\sigma = 0.77$ に上昇し, 再現率 9.4%, 適合率 36.0%となっている. しきい値 $\sigma = 0.85$ のときには, 適合率が 76%にも向上している.

Threshold	0.75	0.76	0.77	0.78	0.80	0.85
Correct	339	339	339	339	339	339
Identified	225	147	88	63	40	30
Correctly Identified	34	32	32	29	27	23
Recall	0.10	0.094	0.094	0.085	0.079	0.067
Precision	0.13	0.21	0.36	0.46	0.67	0.76
F-value	0.110	0.13	0.149	0.144	0.142	0.120

表 6　実験 2D 結果

```
74749, back, back, 1.0, rear - rear
74543, back, back, 0.9999, past time - past time
74543, back, before, 0.9999, past time - time
past
```

さらに語釈拡張を 4 回まで許した実験 2E の結果を表 7 に示す. また同定例についても同様で, **back** に関して示す. 最適しきい値 $\sigma = 0.88$ では, 再現率 8.0%, 適合率 50.9%, F 値 0.137 となり, しきい値 σ0.9 で適合率は最大で 76.6%に達している.

Threshold	0.5	0.55	0.85	0.86	0.87
Correct	339	339	339	339	339
Identified	21673	20921	925	408	126
Correctly Identified	200	197	38	34	30
Recall	0.589	0.581	0.112	0.1	0.0884
Precision	0.00922	0.00941	0.0397	0.0833	0.238
F-value	0.0181	0.0185	0.0586	0.091	0.129

Threshold	0.88	0.89	0.9	0.95
Correct	339	339	339	339
Identified	53	33	30	30
Correctly Identified	27	28	23	23
Recall	0.0796	0.0678	0.0678	0.0678
Precision	0.509	0.696	0.766	0.766
F-value	0.137	0.123	0.124	0.124

表 7　実験 2E 結果

```
74749, back, back, 1.0, rear - rear
74543, back, back, 0.9999, past time - past time
74543, back, before, 0.9999, past time - time
past
```

実験 2A では, 最適しきい値約 0.55 のとき再現率 12.6%, 適合率 16.2%と, 実験 1 と比べても変化が見られず, 同等の精度である. この傾向は実験 2B での最適しきい値 0.5, 2C での 0.5 のときも同様であり, 再現率, 適合率にあまり変化が見られない. 最適しきい値自体も変化があまり無い.

しかし, 実験 2D (語釈拡張を 3 回まで許す操作) では, F 値は変わらないものの, 最適しきい値が 0.77 と高くなり, 再現率 9.4%, 適合率 36% にもなり, 他と比べて, 適合率が上昇している. 実験 2E では, 最適しきい値がさらに 0.88 まで高くなり, 適合率 50% にまで上昇するが, 再現率が 7.9% に低下している. また, F 値は 0.137 となり, 実験 2D (0.149) より低い値を得たことから, 上限回数を 3 回まで許した語釈拡張で, 最良の効率が期待できるといえる.

4.3　考察・評価

拡張回数	最大 F 値	しきい値	適合率
0 (2A)	0.142	0.55	16.2
1 (2B)	0.144	0.50	16.4
2 (2C)	0.140	0.50	15.4
3 (2D)	0.149	0.77	36.0
4 (2E)	0.137	0.88	50.9

表 8　実験での F 値

表 8, 図 3 に実験の要約を示す. 本実験結果を通じて, 語釈拡張の上限回数を増加させると, 適合率が向上することが確認できる. 実際, 実験 2A, 2B, 2C ではそれぞれ, 16.2%, 16.4%, 15.4% と変化がないが, 実験 2D では 36%, 2E では 50.9% に達し, 実験 2D と 2E で適合率が上がっている. しかし, 実験 2E で F 値が 0.137 に下がっていることに注意したい. 上限回数 2,3 回から語釈拡張は活発に動作し, F 値向上の傾向にあるが, 上限回数 4 回の実験 2E では期待に反し, 再現率が下がり F 値が実験 2D よりも低下している. このことより, 語釈拡張の上限回数を 3 回まで許した実験 2D が最良の F 値であるといえる.

本実験で, しきい値は同定できる判定基準として作用することに注意したい. 最適しきい値に関しては, F 値が最大となるしきい値, すなわち 0.55 (実験 2A), 0.50 (2B), 0.50 (2C), 0.77 (2D), 0.88 (2E) により決定した. 定義上, しきい値が高いほど適合率が上がり, 例えば実験 2E では, しきい値 0.50 で再現率 58.9%, 適合率 0.9%, 0.95 では再現率 6.8%, 適合率 76.6%と上昇する. しかし, どちらの場合も同定処理には適さない. 最大の F 値となるしきい値で最大の処理能力を得る.

実際, 語 **back** の同定を見る.

```
(1) 74749, back, back, 1.0, rear - rear
(2) 74543, back, back, 0.9999, past time - time
past
(3) 74749, back, behind, 0.5773, rear - back part
rear
```

例 (1) では位置を, また (2) では時間を意味する語と考えられ, 正しく同定できた例となっているが, (3) **behind**

には 74749 の意味を有さず，類義項目ではあるが同義ではない項目が同定されている．

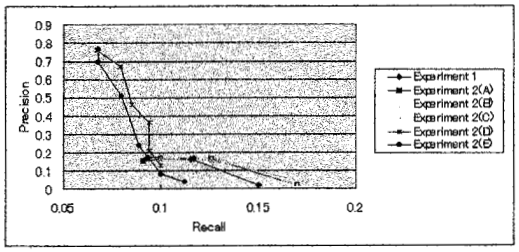

図 3　再現率と適合率 - 概要

5. 関連研究

辞書統合については，これまで自然言語 (NLP)，情報検索 (IR) およびデータマイニング (IE) の視点から，Web ページの統合や遺伝子工学の分野で論じられてきた [1], [2], [6], [12]．オブジェクト同定とは，見出し語を正しくその意図を特定することである．これまでこの問題については 3 つの方向から多くの研究がなされてきた．即ち，**見出し語の認識** (recognition)，**分類** (classification) および**概念の対応付け** (mapping) である．テキスト文書から候補となる用語を抽出する場合や，タンパク質構造を特定する見出し語を判定するなどの分野において論じられてきた．典型的には，**イニシャル文字** (acronym) の解釈を特定する問題などがある．これらに対して，辞書を用いるアプローチ [1], [5]，ルールベースアプローチ [3], [14]，および機械学習アプローチ [2], [13] が提案されている．

類義語・多義語に関する研究も長い歴史を有する．文書内には複数意味を持つ (多義) 単語や複数の単語が同じ意味を持つことが多く，オブジェクト同定処理で考慮することは精度向上に重要である [9], [15]．本稿では，辞書統合に関して語釈の同一性を同定する観点から同義語・多義語を利用する．辞書における見出し項目の一貫性の向上を直接考慮することを考えるわけではない．

本研究では見出し語の対応付けを論じる．この観点からは，用語表現の多様性 (variability) に起因する問題がある [10]．本質的に困難な問題は，**多義性・あいまい性** (ambiguity) による．即ち用語の多義性とは，見出し語に複数の解釈があり，どれを正しく対応付けるかを規定する必要がある．問題領域に依存した用語に限って，発見的にシソーラスを使用する [7] やベクトルモデルによる**パターン学習** [11] が提案されている．しかし，いずれも詳細は発見的であり，統一的議論には至っていない．

6. 結　論

本研究では，辞書統合の背景でのオブジェクト同定の目的で拡張の新方式を提案した．また，その操作が有用で，3 回までの語釈拡張を許すとき，最大の同定精度を確認した．しかし，再現率の低下が見られることから，この改善を図る必要がある．本研究は，辞書統合に向けての第一歩ではある．限られた範囲の中でも様々な対策が考えられ，特定の品詞フィルタリングなどの再検討，分類判定等の機械学習手法，確率的操作の考慮等を検討するべきである．

文　献

[1] Ananiadou, S.: A Methodology for Automatic Term Recognition, proc. *COLING-94*, 1994, pp. 1034-1038

[2] Collier, N., C. Nobata, and J. Tsujii: Automatic Term Identification and Classification in Biological Texts, proc. *Natural Language Pacific Rim Symposium*, 1999, pp. 369-374

[3] Gaizauskas, R., G. Demetriou, and K. Humphreys.: Term Recognition and Classification in Biological Science Journal Articles, Proc *Workshop on Computational Terminology for Medical and Biological Applications*, 2000, pp.37-44

[4] Grossman,D. and Frieder,O.: Information Retrieval - Algorithms and Heuristics, Kluwer Academic Press,1998

[5] Hirschman, L., A.A. Morgan, and A.S. Yeh : Rutabaga by any other name – extracting biological names, *J. of Biomedical Informatics* 35(4), 2002, pp.247-259

[6] Krauthammer, M. and Nenadic, G.: Term Identification in the Biomedical Literature, *Journal of Biomedical Informatics* 37(6), 2004, pp.512-526

[7] Liu, H., S.B. Johnson, and C. Friedman: Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS, *J. Medical Inform Assoc* 9(6), 2002, pp.621-636

[8] Miller, G.A., Beckwith, R. et al.: Introduction to WordNet – An On-Line Lexical Database, *Journal of Lexicography* 3(4), pp.235-244, 1990 (revised 1993, Princeton University)

[9] 那須川 哲哉，河野 浩之，有村 博紀: テキストマイニング基盤技術，人工知能学会誌 16(2)，2001

[10] Nenadic, G., I. Spasic, and S. Ananiadou: Automatic Acronym Acquisition and Term Variation Management within Domain-Specific Texts, proc. *LREC-3*, 2002, pp. 2155-2162

[11] Pustejovsky, J., J. Castano, B. Cochran, M. Kotecki, M. Morrell, and A. Rumshisky: Extraction and Disambiguation of Acronym-Meaning Pairs in Medline, *Medinfo-2001*, 2001

[12] Sebastiani,F.: Machine Learning in Automated Text Categorization, *ACM Computing Surveys* 34(1), 2002, pp.1-47

[13] Shen, D., J. Zhang, G. Zhou, J. Su, and C. Tan: Effective Adaptation of Hidden Markov Model based Named Entity Recognizer for Biomedical Domain, *NLP in Biomedicine in ACL*, 2003, pp. 49-56.

[14] Tejada,S., Knoblock,C.A., and Minton,S.: Learning Object Identification Rules for Information Integration, *Information Systems* 26(8), pp.607-633, 2001

[15] 上嶋 宏，三浦 孝夫，塩谷 勇.: 同義語，多義語の考慮による文書分類の精度向上，電子情報通信学会誌 Vol.J87-D-I No.2, 2004