K-N-match 手法に基づく辞書項目の同定

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, 項目同士を同定する手法を提案する. 辞書統合では, 類義語・多義語といった項目内容の重複を扱う必要があり, 項目内容の類似性判定が最も困難である. このオブジェクト同一性を効率よく判定するため、ベクトル空間モデルに基づき, K-N-match 手法を用いたオブジェクト同定法を提案する. オブジェクトの類似性の尺度として一般的なユークリッド距離などは, 2 つのオブジェクトの部分的な類似性を無視してしまう問題があるが, K-N-match 手法を用いることで, N 次元の部分的類似性を掴んだオブジェクト同定を行うことができる.

キーワード 辞書, オブジェクト同定, K-N-match, 語釈拡張, N-gram

Identification of dictionary item based on k-n-match technique

Kota OKUBO† 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. In the integration process, we should solve several issues caused by homonymous, synonymous and polysemy words. In this work, we put our attention on the synonym problem. To identity synonymous words efficiently, we propose object identification approach using k-n-mach technique based on vector space modeling. The general Euclid distance as the similarity measure of object has the problem ignoring the partial similarity of two objects. We propose the object identification approach that can get partial similarity of the N dimension by using K-N-match technique.

Key words Dictionary, Object Identification, K-N-match, Expansion, N-gram

1. 前 書 き

現在, インターネットが普及し，簡単に短時間で様々な情報を入手できるようになったが, その多くの情報は曖昧さや情報相互の矛盾が多く高い信頼性を保証しない. 一方，出版されている書籍は詳細まで校正されており, 信頼性や一貫性が整っていると期待できる.

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，それぞれ何かのトピックに関する知識を含んでおり, 利用者は検索エンジン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ことができる. しかし, 同じ知識の内容が異なる表現 (語釈) で表されることが多く，複数のサイトを調べても新たな情報を見出すことができない. 内容の重複を取り除き, 知識 (辞書) の統合が可能になれば，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 [16]. 項目の重複, 多義項目, 同義項目が存在し, 特に同義項目は扱いが難しい.

オブジェクト同定に関しての例を示す. `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. 同義語の同一性を効率よく判定するために, 文法解析や自然言語分析などの知識を用いず, 各見出し語の語釈内容を文書とし, ベクトル空間モデルに基づいた K-N-match 手法を用いてオブジェクト同定を行う.

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. 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．

形態素とは，意味を持つ最小の文字列を言う．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．

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, ベクトルの余弦値によって算出できる．

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, 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 WordNet を利用する [9], [17].

WordNet は, 名詞, 動詞, 形容詞からなる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．

本稿では，2 つの語が類似概念を定義しているとき，これを**類義語**という．類義語の例を示す． `lofty` は `high` の類義語，`get there` は `arrive at` の類義語である．また，**多義語**は複数の意味を持つ語をいう．多義語の例を示す．`picture` は `movie, figure, photo, illustration`, `act as` は `function, pretend` などを意味する．

WordNet 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, 同義語の集合で, 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．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．

また, synset には 1 つの意味番号が付いており．その synset 内の語は同じ意味番号を持つ同義語となる．

これらの語が持つ識別番号を表 1 に示す.

2.3 K-N-match 手法

K-N-match 手法とは, オブジェクトの類似検索に用いられる手法である [19]. 一般的なオブジェクト類似検索は最近傍探索などでモデル化され, その距離の尺度として, ユークリッド距離などが用いられている．2 点のオブジェクト $\mathtt{a} = (a_1, a_2, a_3, ..., a_n)$, $\mathtt{b} = (b_1, b_2, b_3, ..., b_n)$ に対してのユークリッド距離は,

	human	person	cough up	spit up
NOUN	1:5303	1:5303		
	2:2130996	2:4465544		
ADJECTIVE	1:324678454			
	2:2634237			
	3:2634331			
VERB			1:2179927	1:2179927
			2:62764	2:62764

表 1　語が持つ意味番号

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

となる．しかし, ユークリッド距離を用いる場合には欠点がある．オブジェクトの特徴のセットを上式のように各次元の差の総和で距離計算を行うため, 部分的な類似性が無視されてしまう．また, 距離が大きな相違を持つ次元によって影響してしまうという問題がある．

K-N-match 手法は N 次元を用いた類似検索手法で, N は $N < n$ の範囲で動的に決定される．類似している N 次元のみに注目することで, 大きな相違点の影響を抑えることができる．

表 2 を用い, この手法の直感的な例を示す．クエリーオブジェクトを Q=(1,1,1,1) とする場合, ユークリッド距離を用いると項目 B が返答されてしまう．これは項目 A, C が 3 つの次元で類似しているにもかかわらず, 1 つの次元で大きな相違を持つためである．

一方, 手法を用い 3-match-query($N = 3$) とした場合, 3 つの次元で一致した値を持つ項目 A が返答される．適切な N を決定することができれば, 部分的な類似性を掴んだオブジェクト同定が可能になる．

項目 ID	d_1	d_2	d_3	d_4
A	1	1	1	100
B	20	20	20	20
C	1.2	1.1	1.3	100

表 2　N-match

しかし, 項目 A のように完全な一致を持つとは限らない．そこで, 閾値 σ を設け柔軟化を行う．q_i を次元 i のクエリー値, p_i を次元 i のデータ値とし, $|q_i - p_i|$ が σ 内なら q_i と p_i は一致すると見なす.

例として, $\sigma = 0.3$ とした場合, 表 2 では 3-match-query の補足の答えとして項目 C が返答される．

本稿での K-N-match 手法の定義は, 全オブジェクト中, 最小の $|q_i - p_i|$ を N 次元備えているオブジェクトを上位 K 位まで見つけ出すことである．

3. K-N-match 手法を用いたオブジェクト同定と語釈拡張

本章ではベクトル空間モデルに基づく K-N-match 手法を用いたオブジェクト同定処理を定義し，さらに, 語釈拡張処理 [16] を導入する．

K-N-match 手法を用いるため, 与えられた見出し語 d に対して，その語釈文から**索引語**を抽出してベクトル化し, 索引語・文書行列を作成する．索引語抽出には N グラム処理を用い, 名詞，動詞，形容詞のいずれかを含む 2 グラム (N=1, 2) を索引語とする．

本稿で実験する K-N-match 手法を用いたオブジェクト同定のアプローチを示す．

N グラム N=1, 2 としたとき, 1 グラムは単語, 2 グラムは共起語に相当する．このアプローチでは単語, 共起語ともに考慮して同定を行う．見出し語 A の語釈から, 名詞, 動詞, 形容詞を含む 1 グラム, 2 グラムを索引語として抽出し, ベクトル化する．

$A = (v_1, v_2, v_3, ..., v_n, (v_1v_2), (v_3v_4), ..., (v_nv_{n+1}))$

索引語の重みは索引語頻度を用い, 不要語処理, ステミングを行う．2 グラム "w_i w_{i+1}" の不要語処理は, w_i と w_{i+1} が両方不要語の場合, 不要語 2 グラムとして除去する．また, 2 グラムのステミングは w_i, w_{i+1} それぞれを処理する．

索引語抽出に関する例を示す．見出し語 `book` の語釈は, `a collection of sheets of paper` に対応し, この語釈から, 索引語 `collection, sheet, paper, a collection, collection of, of sheet, sheet of, of paper` を得る．

次に, 索引語を増やす手段として, 語釈拡張 [16] を適用する．以下に語釈拡張の例を示す．

WordNet の見出し語 `coin bank` では, 語釈 `a container for keeping money at home` に対応し, 索引語は `container, keeping, money, home, a container, container for, for keeping, keeping money, money at, at home` となる．2 グラム `at home` を GCIDE 辞書で調べると, 語釈 `At one's own house` を得る．この語釈から索引語を抽出すると, `house, own house` が得られ, これを `coin bank` の語釈ベクトルと置き換え, 新たな索引語 `container, keeping, money, home, a container, container for, for keeping, keeping money, money at, at home, house, own house` を得る．以下, 1 グラム 2 グラムすべての索引語について同様に繰り返す．

そして, 得られた索引語を用い索引語文書行列を作成する．ここで K-N-match 手法を適用するが, 索引語頻度を用いているため, ほとんどの次元で $|q_i - p_i|$ が 1 または 0 になってしまう可能性がある．$|q_i - p_i|$ の最小を見つけることを容易にするため, まず得られた索引語・文書行列の特異値分解 [5] を行い, 次元縮小した近似行列を用い K-N-match 手法を適用する．

K-N-match 手法を用いたオブジェクト同定の簡単な例を表 3 を用いて示す．表 3 は各次元の項目の ID と $|q_i - p_i|$ のセットを意味する．また, 各セットの $|q_i - p_i|$ は閾値 σ 以下とする．

d_1	d_2	d_3
B, 0.1	B, 0.2	A, 0.3
A, 0.3	C, 0.3	B, 0.4
C, 0.5	A, 1.2	C, 0.5

表 3　K-N-match

本稿の K-N-match 手法を用いたアプローチは, 各次元のクエリー値との最小差 ($|q_i - p_i|$) を N 次元備えているオブジェクトを上位 K 位まで見つけ出すことである．1-1-match の場合, 項目 B が最小差を持つため, 項目 B が同定される．1-2-match の場合は, 1 次元目と 2 次元目で最小差を持つ項目 B(d_1:0.1, d_2:0.2) が同定される．2-2-match の場合, 同様に項目 B と次に小さい差を 2 つ持つ項目 A(d_3:0.3, d_1:0.3) が同定される．

4. 実　　験

本章では，いくつかの実験を通じて，K-N-match 手法を用いたオブジェクト同定に対する効果を検証する．このため，閾値 σ, K, N の値を変化させオブジェクト同定を行い, 最適な値を見つけ出す．また, K-N-match 手法を用いず, 余弦値を類似尺度として行った実験 [18] をベースライン実験とし, 比較する．

4.1 準　　備

本実験では，2 種類の実験を行う．実験 1 では，語釈文に含まれる全ての 1 グラムと 2 グラムを索引語とし，不要語除去およびステミング処理を行ったあと，次元縮小した近似行列を用い, K を 1 から 2, N を 1 から 6 の範囲の中で K-N-match 手法を適用した同定を行う．

実験 2 ではさらに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 (名詞, 形容詞, 動詞) でフィルタリングした 1 グラムと 2 グラムの索引語だけを対象とし，同様に K-N-match 手法を適用した同定を行う．また, 拡張回数に上限を設け，実験 2A として, 0 回拡張, K を 1 から 2, N を 1 から 6 の範囲の中で, 実験 2B として, 1 回拡張, K を 1 から 2, N を 1 から 50 の範囲の中で同定を行う．実験 2B では, 拡張により次元数が増えるため, N の範囲を増やしている．

本実験のデータでは，WordNet 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．WordNet 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．

オブジェクト同定の正誤判定に関して例を示す．`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．

手法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．

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

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．

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す．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，本実験では閾値 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 WordNet 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意味番号となる項目数を合計する.

全体の項目数中の正解数に関しての例を示す. 見出し語 `student, pupil` だけが 意味番号 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. また, 2 グラムの場合も, 見出し語 `cough up, spit up` だけが意味番号 2179927 を共有すれば `cough up` を表す意味の項目数は 2 となり, `spit up` も 2 となる.

4.2 結　　果

実験 1 において, 閾値 $\sigma = 0.3, 0.4, 0.5$ を設け, 閾値に応じた再現率, 適合率および F 値をそれぞれ表 4, 表 5, 表 6 に示す. また, 索引語・文書行列とその近似行列の誤差は 34% とする.

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	121	42	19	13	2	1
正解数	22	26	14	11	2	1
全体の正解数	339	339	339	339	339	339
再現率	0.0648	0.0766	0.0412	0.0324	0.00589	0.00294
適合率	0.181	0.619	0.736	0.846	1	1
F 値	0.0956	0.136	0.0782	0.0625	0.0117	0.00588
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	200	20	2	2	0	0
正解数	16	9	1	1	0	0
全体の正解数	339	339	339	339	339	339
再現率	0.0471	0.0265	0.00294	0.00294	0	0
適合率	0.08	0.45	0.5	0.5	NaN	NaN
F値	0.0593	0.0501	0.00586	0.00586	NaN	NaN

表 4　実験 1 結果 (閾値 0.3, 近似行列の誤差 34%)

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	123	52	30	20	3	2
正解数	22	31	17	14	3	2
全体の正解数	339	339	339	339	339	339
再現率	0.0648	0.0914	0.0501	0.0412	0.00884	0.00589
適合率	0.178	0.596	0.566	0.7	1	1
F 値	0.0952	0.158	0.0921	0.0779	0.0175	0.0117
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	200	20	2	2	0	0
正解数	16	9	1	1	0	0
全体の正解数	339	339	339	339	339	339
再現率	0.0471	0.0265	0.00294	0.00294	0	0
適合率	0.08	0.45	0.5	0.5	NaN	NaN
F値	0.0593	0.0501	0.00586	0.00586	NaN	NaN

表 5　実験 1 結果 (閾値 0.4, 近似行列の誤差 34%)

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	125	58	36	24	4	3
正解数	22	31	17	14	3	2
全体の正解数	339	339	339	339	339	339
再現率	0.0648	0.0914	0.0501	0.0412	0.00884	0.00589
適合率	0.176	0.534	0.472	0.583	0.75	0.666
F 値	0.0948	0.156	0.0906	0.0771	0.0174	0.0116
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	206	40	22	18	0	0
正解数	17	9	1	1	0	0
全体の正解数	339	339	339	339	339	339
再現率	0.0501	0.0265	0.00294	0.00294	0	0
適合率	0.0825	0.225	0.0454	0.0555	NaN	NaN
F値	0.0623	0.0474	0.00554	0.00560	NaN	NaN

表 6　実験 1 結果 (閾値 0.5, 近似行列の誤差 34%)

$\sigma = 0.3$ を設けた結果において, K=1 としたときでは, 1-2-match のとき, 再現率 7.6%, 適合率 61%, F 値が 0.136 で, 最大の効率となっている. また, K=2 としたときでは, 2-1-match のとき, 再現率 4.7%, 適合率 8%, F 値が 0.0593 で最大効率である. $\sigma = 0.4$ を設けた結果において, K=1 としたときでは, 1-2-match のとき, 再現率 9.1%, 適合率 59%, F 値が 0.158 で, 最大効率を得ている. また, K=2 としたときでは, 2-1-match のとき再現率 4.7%, 適合率 8%, F 値が 0.0596 で, 最大となっている. $\sigma = 0.5$ を設けた結果において, K=1 としたときでは, 1-2-match のとき, 再現率 9.1%, 適合率 53%, F 値が 0.156 で, 最大の効率を得ている. また, K=2 としたときでは, 2-1-match のとき, 再現率 5%, 適合率 8.2%, F 値が 0.0623 で, 最大となっている. 閾値を変化させた中で, $\sigma = 0.4$, 1-2-match のとき最も高い F 値 0.158 が得られた.

また, 近似行列の誤差を 1%と少なくし, $\sigma = 0.3, 0.4, 0.5$ を設けた結果をそれぞれ表 7, 表 8, 表 9 に示す.

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	133	55	28	13	1	1
正解数	28	38	19	12	1	1
全体の正解数	339	339	339	339	339	339
再現率	0.0825	0.112	0.0560	0.0353	0.00294	0.00294
適合率	0.210	0.690	0.678	0.923	1	1
F 値	0.118	0.192	0.103	0.0681	0.00588	0.00588
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	204	26	4	0	0	0
正解数	18	12	2	0	0	0
全体の正解数	339	339	339	339	339	339
再現率	0.0530	0.0353	0.00589	0	0	0
適合率	0.0882	0.461	0.5	NaN	NaN	NaN
F値	0.0662	0.0657	0.0116	NaN	NaN	NaN

表 7　実験 1 結果 (閾値 0.3. 近似行列の誤差 1%)

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	133	57	30	15	1	1
正解数	28	40	21	14	1	1
全体の正解数	339	339	339	339	339	339
再現率	0.0825	0.117	0.0619	0.0412	0.00294	0.00294
適合率	0.210	0.701	0.7	0.933	1	1
F 値	0.118	0.202	0.113	0.0790	0.00588	0.00588
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	204	26	4	0	0	0
正解数	18	12	2	0	0	0
全体の正解数	339	339	339	339	339	339
再現率	0.0530	0.0353	0.00589	0	0	0
適合率	0.0882	0.461	0.5	NaN	NaN	NaN
F値	0.0662	0.0657	0.0116	NaN	NaN	NaN

表 8　実験 1 結果 (閾値 0.4. 近似行列の誤差 1%)

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	131	57	29	16	3	3
正解数	27	39	20	14	2	2
全体の正解数	339	339	339	339	339	339
再現率	0.0796	0.115	0.0589	0.0412	0.00589	0.00589
適合率	0.206	0.684	0.689	0.875	0.666	0.666
F 値	0.114	0.196	0.108	0.0788	0.0116	0.0116
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	204	28	4	0	0	0
正解数	19	12	2	0	0	0
全体の正解数	339	339	339	339	339	339
再現率	0.0560	0.0353	0.00589	0	0	0
適合率	0.0931	0.428	0.5	NaN	NaN	NaN
F値	0.0699	0.0653	0.0116	NaN	NaN	NaN

表 9　実験 1 結果 (閾値 0.5. 近似行列の誤差 1%)

$\sigma = 0.3$ を設けた結果において, K=1 としたときでは, 1-2-match のとき, 再現率 11.2%, 適合率 69%, F 値が 0.192 で, 最大の効率となっている. また, K=2 としたときでは, 2-1-match のとき, 再現率 5.3%, 適合率 8.8%, F 値が 0.0662 で最大効率である. $\sigma = 0.4$ を設けた結果において, K=1 としたときでは, 1-2-match のとき, 再現率 11.7%, 適合率 70%, F 値が 0.202 で, 最大の効率を得ている. また, K=2 としたときでは, 2-1-match のとき, 再現率 5.3%, 適合率 8.8%, F 値が 0.0662 で最大となっている. $\sigma = 0.5$ を設けた結果において, K=1 としたときでは, 1-2-match のとき再現 11.5%, 適合率 68%, F 値が 0.196 で, 最大の効率である. また, K=2 としたときでは, 2-1-match のと

き, 再現率 5%, 適合率 8.2%, F 値が 0.0623 で最大となっている. 閾値を変化させた中で, $\sigma = 0.4$, 1-2-match のとき最も高い F 値 0.202 が得られた. また, 近似行列の誤差を 34%としたときの最良の結果 ($\sigma = 0.4$, 1-2-match, F 値 0.158) と比較すると, 誤差が小さいほうが, 同定効率は良い.

次に, ベースライン実験での結果を表 10 に示す.

Threshold	0.23	0.24	0.25	0.26	0.27	0.28
Correct	339	339	339	339	339	339
Identified	2010	2003	243	243	220	220
Correctly Identified	57	55	50	50	45	45
Recall	0.168	0.162	0.147	0.147	0.132	0.132
Precision	0.0283	0.0274	0.205	0.205	0.204	0.204
F-value	0.04853	0.04696	0.1718	0.1718	0.161	0.161

表 10　ベースライン実験 結果

ベースライン実験の結果では, 類似度の閾値 0.25 のとき, 再現率 14%, 適合率 20%, F 値 0.171 で最大の同定効率を得る. 今回の K-N-match 手法を用いた実験と比較すると, 最大効率が得られる近似行列の誤差 1%, $\sigma = 0.4$, 1-2-match のとき, 再現率 11.7%, 適合率 70%, F 値が 0.202 と, 今回の K-N-match 手法を用いたほうが同定効率が良いと言える.

次に実験 2A 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, 閾値 $\sigma = 0.3, 0.4, 0.5$ を設け, 近似行列の誤差を 31%とした結果をそれぞれ表 11, 表 12, 表 13 に示す.

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	173	59	31	16	1	1
正解数	14	32	18	14	1	1
全体の正解数	339	339	339	339	339	339
再現率	0.0412	0.0943	0.0530	0.0412	0.00294	0.00294
適合率	0.0809	0.542	0.580	0.875	1	1
F値	0.0546	0.160	0.0972	0.0788	0.00588	0.00588
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	290	30	2	2	0	0
正解数	14	9	1	1	0	0
全体の正解数	339	339	339	339	339	339
再現率	0.0412	0.0265	0.00294	0.00294	0	0
適合率	0.0482	0.3	0.5	0.5	NaN	NaN
F値	0.0445	0.0487	0.00586	0.00586	NaN	NaN

表 11　実験 2A 結果 (閾値 0.3, 近似行列の誤差 31%)

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	173	62	34	20	3	2
正解数	14	33	19	16	3	2
全体の正解数	339	339	339	339	339	339
再現率	0.0412	0.0973	0.0560	0.0471	0.00884	0.00589
適合率	0.0809	0.532	0.558	0.8	1	1
F値	0.0546	0.164	0.101	0.0891	0.0175	0.0117
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	290	30	2	2	0	0
正解数	14	9	1	1	0	0
全体の正解数	339	339	339	339	339	339
再現率	0.0412	0.0265	0.00294	0.00294	0	0
適合率	0.0482	0.3	0.5	0.5	NaN	NaN
F値	0.0445	0.0487	0.00586	0.00586	NaN	NaN

表 12　実験 2A 結果 (閾値 0.4, 近似行列の誤差 31%)

近似行列の誤差 31%の場合, $\sigma = 0.3$ を設けた結果において, K=1 としたときでは, 1-2-match のとき, 再現率 9.4%, 適合率 54%, F 値が 0.160 で, 最大の効率となっている. また, K=2 としたときでは, 2-2-match のとき, 再現率 2.6%, 適合率 30%, F 値が 0.0487 で, 最大効率となっている. $\sigma = 0.4$ を設けた結果において, K=1 としたときでは, 1-2-match のとき, 再現率 9.7%, 適合率 53%, F 値が 0.164 で, 最大の効率を得ている. また, K=2 としたときでは, 2-2-match のとき再現率 2.6%, 適合率 30%, F 値が 0.0487 で, 最大となっている. $\sigma = 0.5$ を設けた結果において, K=1 としたときでは, 1-2-match のとき, 再現率 9.7%, 適合率 51%, F 値が 0.163 で, 最大の効率となっている. また, K=2 としたときでは, 2-2-match のとき, 再現率 2.6%, 適合率 25%, F 値が 0.048 で, 最大となっている. 閾値を変化させた中で, $\sigma = 0.4$, 1-2-match のとき最も高い F 値 0.164 が得られた.

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	174	64	36	22	4	3
正解数	14	33	19	16	3	2
全体の正解数	339	339	339	339	339	339
再現率	0.0412	0.0973	0.0560	0.0471	0.00884	0.00589
適合率	0.0804	0.515	0.527	0.727	0.75	0.666
F値	0.0545	0.163	0.101	0.0886	0.0174	0.0116
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	294	36	8	8	0	0
正解数	14	9	1	1	0	0
全体の正解数	339	339	339	339	339	339
再現率	0.0412	0.0265	0.00294	0.00294	0	0
適合率	0.0476	0.25	0.125	0.125	NaN	NaN
F値	0.0442	0.048	0.00576	0.00576	NaN	NaN

表 13　実験 2A 結果 (閾値 0.5, 近似行列の誤差 31%)

また, 近似行列の誤差を 4%と少なくし, $\sigma = 0.1, 0.4, 0.5$ を設けた結果をそれぞれ表 14, 表 15, 表 16 に示す.

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	180	64	36	15	1	1
正解数	25	40	22	14	1	1
全体の正解数	339	339	339	339	339	339
再現率	0.0737	0.117	0.0648	0.0412	0.00294	0.00294
適合率	0.138	0.625	0.611	0.933	1	1
F値	0.0963	0.198	0.117	0.0790	0.00588	0.00588
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	302	38	4	0	0	0
正解数	28	15	2	0	0	0
全体の正解数	339	339	339	339	339	339
再現率	0.0825	0.0442	0.00589	0	0	0
適合率	0.0927	0.394	0.5	NaN	NaN	NaN
F値	0.0873	0.0795	0.0116	NaN	NaN	NaN

表 14　実験 2A 結果 (閾値 0.1, 近似行列の誤差 4%)

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	181	67	39	18	1	1
正解数	25	43	25	17	1	1
全体の正解数	339	339	339	339	339	339
再現率	0.0737	0.1268	0.0737	0.0501	0.00294	0.00294
適合率	0.138	0.6417	0.641	0.944	1	1
F値	0.0961	0.2118	0.132	0.0952	0.00588	0.00588
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	304	38	4	0	0	0
正解数	28	15	2	0	0	0
全体の正解数	339	339	339	339	339	339
再現率	0.0825	0.0442	0.00589	0	0	0
適合率	0.0921	0.394	0.5	NaN	NaN	NaN
F値	0.0870	0.0795	0.0116	NaN	NaN	NaN

表 15　実験 2A 結果 (閾値 0.4, 近似行列の誤差 4%)

K-N-match	1-1	1-2	1-3	1-4	1-5	1-6
同定数	182	68	39	18	1	1
正解数	25	43	25	17	1	1
全体の正解数	339	339	339	339	339	339
再現率	0.0737	0.1268	0.0737	0.0501	0.00294	0.00294
適合率	0.137	0.6323	0.641	0.944	1	1
F値	0.0959	0.2113	0.132	0.0952	0.00588	0.00588
K-N-match	2-1	2-2	2-3	2-4	2-5	2-6
同定数	304	38	4	0	0	0
正解数	28	15	2	0	0	0
全体の正解数	339	339	339	339	339 339	
再現率	0.0825	0.0442	0.00589	0	0	0
適合率	0.0921	0.394	0.5	NaN	NaN	NaN
F値	0.0870	0.0795	0.0116	NaN	NaN	NaN

表 16　実験 2A 結果 (閾値 0.5, 近似行列の誤差 4%)

近似行列の誤差を 4%と少なくした場合, $\sigma = 0.1$ を設けた結果において, K=1 としたときでは, 1-2-match のとき, 再現率 11.7%, 適合率 62%, F 値が 0.198 で, 最大の効率となっている. また, K=2 としたときでは, 2-1-match のとき, 再現率 8.2%, 適合率 9.2%, F 値が 0.0873 で, 最大となっている. $\sigma = 0.4$ を設けた結果において, K=1 としたときでは, 1-2-match のとき, 再現率 12.68%, 適合率 64.1%, F 値が 0.2118 で, 最大の効率となっている. また, K=2 としたときでは, 2-1-match のとき再現率 8.2%, 適合率 9.2%, F 値が 0.0873 で, 最大となっている. $\sigma = 0.5$ を設けた結果において, K=1 としたときでは, 1-2-match のとき, 再現率 12.68%, 適合率 63.2%, F 値が 0.2113 で, 最大の効率となっている. また, K=2 としたときでは, 2-1-match のとき, 再現率 8.2%, 適合率 9.2%, F 値が 0.0873 で, 最大となっている. 閾値を変化させた中で, $\sigma = 0.4$, 1-2-match のとき最も高い F 値 0.2118 が得られた. また, 近似行列の誤差を 31%としたときの最良の結果 ($\sigma = 0.4$, 1-2-match, F 値 0.164) と比較すると, 誤差が小さいほうが, 同定効率は良い. なお, $\sigma = 0.2, 0.3$ の結果の最大 F 値は $\sigma = 0.4$ と変わらない.

次に, ベースライン実験での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 17 に示す.

Threshold	0.37	0.38	0.39	0.4	0.45	0.5
Correct	339	339	339	339	339	339
Identified	50	50	46	44	41	29
Correctly Identified	37	37	37	35	33	24
Recall	0.109	0.109	0.109	0.103	0.0973	0.0707
Precision	0.74	0.74	0.804	0.795	0.804	0.827
F-value	0.1902	0.1902	0.1922	0.1827	0.1736	0.1304

表 17　ベースライン実験 結果

ベースライン実験では類似度の閾値が 0.39 のとき再現率 10%, 適合率 80%, F 値 0.192 で最大の同定効率を得る. 今回の実験と比較すると, 最大効率が得られる近似行列の誤差 4%, $\sigma = 0.4$, 1-2-match のとき, 再現率 12.68%, 適合率 64.1%, F 値が 0.2118 と, 再現率がわずかに上昇し, 今回の K-N-match 手法を用いたほうが同定効率が良いと言える.

次に閾値 $\sigma = 0.2, 0.4, 0.5$ を設け, 近似行列の誤差を 33%とし, 1 回の拡張操作を含む実験 2B の結果をそれぞれ表 18 表 19 表 20 に示す.

K-N-match	1-19	1-20	1-21	1-22	1-23	1-24
同定数	134	134	132	132	87	70
正解数	21	21	23	23	16	14
全体の正解数	339	339	339	339	339	339
再現率	0.0619	0.0619	0.0678	0.0678	0.0471	0.0412
適合率	0.156	0.156	0.174	0.174	0.183	0.2
F値	0.0887	0.0887	0.0976	0.0976	0.0751	0.0684
K-N-match	2-19	2-20	2-21	2-22	2-23	2-24
同定数	212	210	208	208	112	76
正解数	22	22	22	25	15	10
全体の正解数	339	339	339	339	339	339
再現率	0.0648	0.0648	0.0648	0.0737	0.0442	0.0294
適合率	0.103	0.104	0.105	0.1201	0.133	0.131
F値	0.0798	0.0801	0.0804	0.0914	0.0665	0.0481

表 18　実験 2B 結果 (閾値 0.2, 近似行列の誤差 33%)

K-N-match	1-19	1-20	1-21	1-22	1-23	1-24
同定数	138	138	137	137	94	78
正解数	22	22	24	24	17	15
全体の正解数	339	339	339	339	339	339
再現率	0.0648	0.0648	0.0707	0.0707	0.0501	0.0442
適合率	0.159	0.159	0.175	0.175	0.180	0.192
F値	0.0922	0.0922	0.1008	0.1008	0.0785	0.0719
K-N-match	2-19	2-20	2-21	2-22	2-23	2-24
同定数	216	214	214	214	130	94
正解数	24	24	24	27	17	14
全体の正解数	339	339	339	339	339	339
再現率	0.0707	0.0707	0.0707	0.0796	0.0501	0.0412
適合率	0.111	0.112	0.112	0.126	0.130	0.148
F値	0.0864	0.0867	0.0867	0.0976	0.0724	0.0646

表 19　実験 2B 結果 (閾値 0.4, 近似行列の誤差 33%)

K-N-match	1-19	1-20	1-21	1-22	1-23	1-24
同定数	140	140	139	139	96	80
正解数	22	22	24	24	17	15
全体の正解数	339	339	339	339	339	339
再現率	0.0648	0.0648	0.0707	0.0707	0.0501	0.0442
適合率	0.157	0.157	0.172	0.172	0.177	0.187
F値	0.0918	0.0918	0.1004	0.1004	0.0781	0.0715
K-N-match	2-19	2-20	2-21	2-22	2-23	2-24
同定数	224	220	220	220	138	102
正解数	25	25	25	28	19	15
全体の正解数	339	339	339	339	339	339
再現率	0.0737	0.0737	0.0737	0.0825	0.05604	0.0442
適合率	0.111	0.113	0.113	0.127	0.137	0.147
F値	0.0888	0.0894	0.0894	0.1001	0.0796	0.0680

表 20　実験 2B 結果 (閾値 0.5, 近似行列の誤差 33%)

近似行列の誤差 33%の場合, $\sigma = 0.2$ を設けた結果において, K=1 としたときでは, 1-21-match のとき, 再現率 6.7%, 適合率 17.4%, F 値が 0.0976 で, 最大の効率を得ている. また, K=2 としたときでは, 2-22-match のとき, 再現率 7.3%, 適合率 12.0%, F 値が 0.0914 で, 最大となっている. $\sigma = 0.4$ を設けた結果において, K=1 としたときでは, 1-21-match のとき, 再現率 7.07%, 適合率 17.5%, F 値が 0.1008 で, 最大の効率を得ている. また, K=2 としたときでは, 2-22-match のとき再現率 7.9%, 適合率 12.6%, F 値が 0.0976 で, 最大効率となっている. $\sigma = 0.5$ を設けた結果において, K=1 としたときでは, 1-21-match のとき, 再現率 7.07%, 適合率 17.2%, F 値が 0.1004 で, 最大の効率となっている. また, K=2 としたときでは, 2-22-match のとき, 再現率 8.25%, 適合率 12.7%, F 値が 0.1001 で, 最大となっている. 閾値を変化させた中で, $\sigma = 0.4$, 1-21-match のとき最も高い F 値 0.1008 が得られた. なお, $\sigma = 0.3$ としたときの結果の最大 F 値は $\sigma = 0.4$ と, 変わらない.

また, 近似行列の誤差を 1%と少なくし, $\sigma = 0.02, 0.4, 1.0$ を設けた結果をそれぞれ表 21, 表 22, 表 23 に示す.

K-N-match	1-27	1-28	1-29	1-30	1-31	1-32
同定数	86	83	82	81	80	79
正解数	26	26	27	26	26	25
全体の正解数	339	339	339	339	339	339
再現率	0.0766	0.0766	0.0796	0.0766	0.0766	0.0737
適合率	0.3023	0.313	0.329	0.320	0.325	0.316
F値	0.122	0.123	0.128	0.123	0.124	0.119
K-N-match	2-22	2-23	2-24	2-25	2-26	2-27
同定数	212	174	138	102	86	76
正解数	22	23	22	18	17	15
全体の正解数	339	339	339	339	339	339
再現率	0.0648	0.0678	0.0648	0.0530	0.0501	0.0442
適合率	0.103	0.132	0.159	0.176	0.197	0.197
F値	0.0798	0.0896	0.0922	0.0816	0.08	0.0722

表 21　実験 2B 結果 (閾値 0.02, 近似行列の誤差 1%)

K-N-match	1-27	1-28	1-29	1-30	1-31	1-32
同定数	88	85	84	83	82	81
正解数	28	28	29	28	28	27
全体の正解数	339	339	339	339	339	339
再現率	0.0825	0.0825	0.085	0.0825	0.0825	0.0796
適合率	0.318	0.329	0.345	0.337	0.341	0.333
F値	0.131	0.132	0.137	0.132	0.133	0.128
K-N-match	2-22	2-23	2-24	2-25	2-26	2-27
同定数	212	176	140	104	86	76
正解数	22	24	23	19	17	15
全体の正解数	339	339	339	339	339	339
再現率	0.0648	0.0707	0.0678	0.0560	0.0501	0.0442
適合率	0.103	0.136	0.164	0.182	0.197	0.197
F値	0.0798	0.0932	0.0960	0.0857	0.08	0.0722

表 22　実験 2B 結果 (閾値 0.4, 近似行列の誤差 1%)

K-N-match	1-27	1-28	1-29	1-30	1-31	1-32
同定数	197	196	193	191	188	186
正解数	29	29	29	28	28	27
全体の正解数	339	339	339	339	339	339
再現率	0.0855	0.0855	0.0855	0.0825	0.0825	0.0796
適合率	0.147	0.1479	0.150	0.146	0.148	0.145
F値	0.108	0.108	0.109	0.105	0.106	0.102
K-N-match	2-26	2-27	2-28	2-29	2-30	2-31
同定数	396	392	390	384	378	368
正解数	33	34	35	34	33	31
全体の正解数	339	339	339	339	339	339
再現率	0.0973	0.1002	0.103	0.1002	0.0973	0.0914
適合率	0.0833	0.0867	0.0897	0.0885	0.0873	0.0842
F値	0.0897	0.0930	0.0960	0.0940	0.0920	0.0876

表 23　実験 2B 結果 (閾値 1.0, 近似行列の誤差 1%)

近似行列の誤差 1%の場合, $\sigma = 0.02$ を設けた結果において, K=1 としたときでは, 1-29-match のとき, 再現率 7.9%, 適合率 32%, F 値が 0.128 で, 最大の効率を得る. また, K=2 としたときでは, 2-24-match のとき, 再現率 6.4%, 適合率 15%, F 値が 0.0922 で, 最大となっている. $\sigma = 0.4$ を設けた結果において, K=1 としたときでは, 1-29-match のとき, 再現率 8.5%, 適合率 34%, F 値が 0.137 で, 最大の効率となっている. また, K=2 としたときでは, 2-24-match のとき再現率 6.7%, 適合率 16%, F 値が 0.0960 で, 最大の効率を得ている. $\sigma = 1.0$ を設けた結果において, K=1 としたときでは, 1-29-match のとき, 再現率 8.5%, 適合率 15%, F 値が 0.109 で, 最大の効率となっている. また, K=2 としたときでは, 2-28-match のとき, 再現率 10.3%, 適合率 8.9%, F 値が 0.0960 で, 最大となっている. 閾値を変化させた中で, $\sigma = 0.4$, 1-29-match のとき最も高い F 値 0.137 が得られた. また, 近似行列の誤差を 33%としたときの最良の結果 ($\sigma = 0.4$, 1-21-match, F 値 0.1008) と比較すると, 誤差が小さいほうが, 同定効率は良い.

なお, $\sigma = 0.03$ から $\sigma = 0.3$, $\sigma = 0.5$ から $\sigma = 0.9$ までの結果の最大 F 値は $\sigma = 0.4$ と, 変わらない.

次に, ベースライン実験での 1 回拡張を適用したの結果を表 24 に示す.

Threshold	0.37	0.38	0.39	0.4	0.45	0.5
Correct	339	339	339	339	339	339
Identified	50	50	46	44	41	29
Correctly Identified	37	37	37	35	33	24
Recall	0.109	0.109	0.109	0.103	0.0973	0.0707
Precision	0.74	0.74	0.804	0.795	0.804	0.827
F-value	0.1902	0.1902	0.1922	0.1827	0.1736	0.1304

表 24　ベースライン実験 結果

ベースライン実験では類似度の閾値が 0.39 のとき再現率 10%, 適合率 80%, F 値 0.192 で最大の同定効率を得る. 今回の実験と比較すると, 最大効率が得られる近似行列の誤差 1%, $\sigma = 0.4$,1-29-match, のとき, 再現率 8.5%, 適合率 34%, F 値が 0.137 と, ベースライン実験のほうが適合率が高く, 同定効率も良い.

ここで, 各実験の最大 F 値を表 25 に示す. また, ベースライン実験での最大 F 値を表 26 に示す.

k-n-match	閾値	近似行列の誤差	最大 F 値	適合率	再現率
1-2-match (実験 1)	0.4	34%	0.158	0.596	0.0914
1-2-match (実験 1)	0.4	1%	0.202	0.701	0.117
1-2-match (実験 2A)	0.4	31%	0.164	0.532	0.0973
1-2-match (実験 2A)	0.4	4%	0.2118	0.641	0.126
1-21-match (実験 2B)	0.4	33%	0.1008	0.175	0.0707
1-29-match (実験 2B)	0.4	1%	0.137	0.345	0.0855

表 25　各実験での最大 F 値 (K-N-match)

拡張回数	しきい値	最大 F 値	適合率	再現率
0 (実験 1)	0.25	0.171	0.205	0.147
0 (実験 2A)	0.39	0.192	0.804	0.109
1 (実験 2B)	0.39	0.192	0.804	0.109

表 26　各実験での最大 F 値 (ベースライン実験)

表 25 より, 索引語文書行列とその近似行列の誤差に注目すると , 誤差を小さくしたほうが F 値が高く, 同定効率が良いことが言える. また, F 値を比較すると, K-N-match 手法を用い, 近似行列の誤差を 1%としたときの, $\sigma = 0.4$, 1-2-match で, 最大 F 値 0.2118(適合率 0.641, 再現率 0.126) を得た. ベースライン実験, 表 26 での最大 F 値と比較するとベースライン実験の F 値の最大は 0.192(適合率 0.804, 再現率 0.109) で, 今回の K-N-match 手法を用いた結果のほうが再現率が若干上昇しているため同定効率が良い.

4.3　考察・評価

本実験を通じて, 今回行った K-N-match 手法を用いたオブジェクト同定のほうが, 同定の効率が良い. 0 回拡張 (実験 2A), 近似行列の誤差 4%, $\sigma = 0.4$, K=1,N=2 のとき F 値 0.2118, 適合率 0.641, 再現率 0.126 で最良であった. ベースライン結果の最良は 0 回拡張 (実験 2A), コサイン類似度の閾値 0.39 のとき F 値 0.192, 適合率 0.804, 再現率 0.109 で, 今回の実験では再現率が上昇している. これは同定の正解数が増加したためで, 0 回拡張 (フィルタリングのみ), 近似行列の誤差 4%, $\sigma = 0.4$, 1-2-match のときの正解数は 43 で, ベースライン結果の 0 回拡張, 余弦値の閾値 0.39 のときは 37 である.

また, 今回の実験で, 1 回拡張を行うと同定効率が減少してしまう. これは間違えた同定が多いためだと考えられる. 実際に意味番号 400050 を持つ見出し語 `basely` の同定を見てみると, 最大の効率が得られた実験 2A の近似行列の誤差 4%, $\sigma = 0.4$, 1-2-match のときでは, 2 つの次元で最小の $|q_i - p_i|$(5.06E-6, 8.29E-6) を持つ `basely` が同定され, 正解しているのに対し,1 回拡張 (実験 2B) で最大 F 値が得られた近似行列の誤差 1%, 1-29-match のときでは, 5 つの次元で最小の $|q_i - p_i|$(4.46E-10, 6.65E-10, 1.29E-9, 2.11E-9, 2.14E-9) を持つ `bestially` が同定されるが, `bestially` は意味番号 400050 を共有しないため, この同定は間違いである.

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], [7], [13]．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, この問題については**見出し語の認識** (recognition)，**分類** (classification) および**概念の対応付け** (mapping) の 3 つの方向から多くの研究がなされてきた．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, 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], [6]，ルールベースアプローチ [3], [15]，および機械学習アプローチ [2], [14] 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 [10], [17].

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 [11]．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 [8] やベクトルモデルによる**パターン学習** [12] 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, 辞書統合の背景として, K-N-match 手法を用いたオブジェクト同定を提案した．また, 0 回拡張時の $\sigma = 0.4$, 1-2-match のとき高い精度が得られ, この手法が有効であることを確認した．しかし, 依然再現率が低く改善を図る必要がある．対策として, 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] 北研二, 他: 情報検索アルゴリズム, 共立出版, 2002

[6] 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

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

[8] 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

[9] 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)

[10] 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

[11] 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

[12] 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

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

[14] 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.

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

[16] 大久保 幸太, 三浦 孝夫: 語釈拡張に基づくテキスト項目の同定, 情報処理学会研究報告 Vol.2007, No.54(20070531) pp. 7-12

[17] 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

[18] Okubo,K., Miura,T.: Object Identification using N-gram Based on Expansion, First International Conference on Advanced Intelligence (CAI-08)

[19] Tung, A.K.H., Zhang, R., Koudas, N. and Ooi, B.C.: Similarity search: a matching based approach, 32nd VLDB, pp.631 - 642, 2006