既存の汎用辞書を利用したオンデマンド概念辞書の構築と質問応答システムへの利用

研究代表者　　鈴木良弥　　山梨大学大学院医学工学総合研究部・准教授

1 はじめに

本研究は，汎用概念辞書に新たな単語を追加することで該当分野に特化した概念辞書を構築することを目的とする．また，概念辞書の評価のために，構築した辞書を質問応答システムに組込み，概念辞書利用による性能向上を評価する．

文書要約，文書分類，質問応答システムなど，多くの自然言語処理アプリケーションにおいて意味情報の利用はシステムの精度向上に有効である．意味情報を利用するためには，概念辞書が必要になる．システムで利用できる概念辞書（類語辞書）は，日本語 WordNet[JWordNet 2009]，EDR 日本語辞書[EDR Dictionary 1996]，分類語彙表[Bunruigoihyo 1964]などはあるが，種類は多くなく，しかも登録語彙は一般語がほとんどである．しかし，現在の自然言語処理アプリケーションは分野を限定したものが多く，利用辞書も分野ごとに特化したものが望まれるが，分野ごとの辞書はほとんど存在しない．

本研究は，種々の自然言語処理アプリケーションで利用するため，各分野で利用可能な概念辞書を該当分野の文書集合を利用して自動生成し，作成した辞書を活用することで様々なアプリケーションの性能向上に寄与することを目的とする．具体的には，まず，新聞記事，特許文書，ホテルレビューから関連語対抽出を行う．関連語対の類似度と汎用辞書内の単語との類似度などを利用して，辞書に未追加の語を概念辞書の階層構造に追加する．構築した辞書を文書分類，質問応答システムの中に組み込むことによってシステムの分類性能，質問応答システムの性能向上を目指す．

2 関連研究

本研究は，目的にあった概念語辞書を既存の辞書を基に作成し，作成した概念辞書を自然言語アプリケーションに利用したときの効果を確認する．特に（1）関連語抽出とシソーラス拡張（2）レビュー文分類について研究を行ったため，それぞれについて関連研究を挙げて簡単に説明する．

2-1 関連語抽出とシソーラス拡張

語と語の類似度を計算し，類似単語を抽出する研究は Hindle[Hindle 1990]，Lin[Lin 1997]，Hagiwara[Hagiwara 2006]などがある．いずれも類似単語は統語的に類似した役割を持つという仮説を基にしており，語間の依存関係を基づいた手法を提案している．また，萩原ら[萩原 2005]は潜在意味モデルである PLSI を用いて名詞間の類似関係を求めている．彼らの研究では一般語を対象に類似単語を抽出しており，専門用語については言及していない．本研究は彼らの研究を基に日本語文書を対象に，限定された分野の専門的な単語間の類似度を計算する．

シソーラス構築・拡張についても多くの研究がある．Tokunaga[Tokunaga 1997]と Uramoto[Uramoto 1996a]は既存の辞書の中で新しい単語を分類することによってシソーラスを拡張する手法を提案している．しかし，対象は一般語だけであり，専門用語に関しては言及されていない．本研究では現在の自然言語処理アプリケーションでの利用を考え，目的にあった分野に限定した場合の概念辞書を作成する．

2-2 レビュー文分類

現在，インターネット通信販売サイトやホテル予約サイトには利用者が書いた沢山のレビューが蓄えられており，利用者は商品購入やホテル予約の判断材料に利用している．しかし，蓄えられているレビューが多いため，全てのレビューを読むことは不可能であり，レビューを読みやすくする工夫が求められている．我々はホテルレビューの評価項目リストの抽出と評価表現の抽出を行った．そのアプローチは評価分析とテキスト分割に分類される．評価分析は自然言語処理の研究の中でも難しいタスクである．それは幅広く研究されており，多くの手法[Beineke 2004]，[Yi 2005]，[Hu 2004]が提案されている．例えば Wei らは製品レビュー中の製品の項目とそれらの関連する評価をラベル付けするために HL-SOT 手法[Wei 2010]を提案している．テキスト分割も多くの手法が提案されている．Hearst は文書中の同じ単語の繰り返しに関する語彙的な結束性を基にした TextTiling[Hearst 1997]によるテキストセグメンテーション手

法を提案している. Utiyama と Isahara は分野に依存しない文書分割のための統計的手法[Utiyama 2001]を提案している. Hirao らは語彙的結束性と単語の重要度の利用[Hirao 2000]を試みている. 彼らは2つの異なった手法を文書分割に利用している. すなわち単語の共起に関する語彙的結束性と, 文書中の各文の重要度の変化である.

3 関連語抽出

ホテル予約サイトのレビューのデータを用いて関連語抽出を説明する. ホテルレビューは様々な宿泊客によって投稿されている. そのため, 同じ内容を違う表現で記述する場合がある. 例えば「部屋」,「お部屋」と「ルーム」は同じ「部屋」の意味として使われている. その上,「客室」と「部屋」は本来違う意味であるが, ホテルレビュー内では同じ意味として使われる. 表 1は楽天トラベルのホテルレビュー内で「部屋」を意味する頻出単語とその頻度を示している.

表 1　抽出された類似単語(部屋)

単語	頻度
部屋	171,796
お部屋	38,547
ルーム	17,203
客室	4,446

レビュー文の分類を精度良く行うためには関連語(類義語, 上位語/下位語, 同属語)を抽出する必要がある. そこで我々は Lin の方法[Lin 1998]を用いてホテルレビューから類似単語を抽出した. まず係り受け関係を用いた類似単語ペア抽出手法について説明する. 2つの単語の係り受け関係は意味的な類似単語の抽出に使われる. Lin は係り受けの三つ組[Lin 1998]を提案している. 係り受け関係の三つ組は入力文中の2つの単語 w, w' とその2つの文法的な関係 r からなる. $\|w,r,w'\|$ は係り受けの三つ組 (w,r,w') の頻度を表す. $\|w,r,*\|$ はコーパス内で関係 (w,r) のトータルの頻度を表す. ここで "*" はワイルドカードを表している.

我々は文法的な関係 r として3種類の日本語の格助詞集合(A,B,C)を用いた. 集合 A は2つの格助詞「が」と「を」からなる. これらはそれぞれ主語と目的語に対応する. 集合 B は集合 A を含む6つの格助詞からなる. 集合 C は集合 B を含む17個の(格)助詞からなる. 類似単語ペア候補として2つ以上の集合によって抽出される単語ペアを選択する. 単語「w」と「w'」との関係 r に関する類似度を計算するために我々は式(1)を用いた.

$$I(w,r,w') = \log \frac{\|w,r,w'\| \times \|*,r,*\|}{\|w,r,*\| \times \|*,r,w'\|} \qquad (1)$$

$T(w)$ を $\log \frac{\|w,r,w'\| \times \|*,r,*\|}{\|w,r,*\| \times \|*,r,w'\|}$ が正の数になるペア (r,w') の集合とすると, 2つの単語 w_1 と w_2 の類似度 $Sim(w_1,w_2)$ は式(2)で定義される.

$$Sim(w_1,w_2) = \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)} (I(w_1,r,w) + I(w_2,r,w))}{\sum_{(r,w)\in T(w_1)} I(w_1,r,w) + \sum_{(r,w)\in T(w_2)} I(w_2,r,w)} \qquad (2)$$

表 2 は抽出された類似単語ペアを表す. 表 2 にはいくつかの表記のゆれがある. 一般に「朝刊」と「新聞」,「朝食券」と「チケット」は違う意味であるが, 2つのペアはホテルレビュー内ではほぼ同じ意味で使われる.

表 2 助詞セットA, B, Cを用いて抽出された類似単語ペア

No.	単語1	単語2
1	好感	大変好感
2	道順	行き方
3	お腹	おなか
4	埃	ほこり
5	ネット	インターネット
6	改修	リニューアル
7	排水口	排水溝
8	口コミ	クチコミ
9	朝刊	新聞
10	朝食券	チケット

4 実験

4-1 特許文書の文書分類

図 1に特許文書分類システムの概要を示す．提案システムは以下の３つのフェイズからなっている．すなわち（１）シソーラス拡張，（２）文書分類のための語の重み付け，（３）文書分類である．各フェイズについて簡単に説明する．

（１）シソーラス拡張

シソーラス拡張は 3 節の関連語抽出を利用する．最終的に新しい単語の対応する意味素性候補は既存のシソーラス[JST Thesaurus1999]の階層的意味素性を利用して決定される．

（２）文書分類のための語の重み付け

文書分類フェイズでは我々は各文書をいくつかの関連テーマに分類する．特許文書には申請者によって付けられたタグが複数存在する．例えば「発明の名称」,「産業上の利用分野」,「発明の目的」,「課題を解決するための手段」などである．各文書には平均 56 の申請者タグがある．ほとんどのタグは多くの文書で使われているが,中には表記上の揺れのあるタグも使われている.そこで我々はタグを６つの意味タグに分類した．

６個の意味タグのラベルとそのタグに分類された申請者タグを表 3に示す.多くのテーマは意味タグの「目的」と関連がある．従って，我々は「目的」が最も重要な意味タグと考え，「目的」内の単語に対して拡張シソーラスを用いて語彙拡張を行った．文書分類のために Bag-of-Words 法と単語の分布を利用した．多くの学習データを使用しているとはいえ,テストデータに現れる単語が学習データに現れない場合もある.従って，拡張シソーラスを使って語彙拡張を行う必要がある．語彙拡張は効果的ではあるが，全ての単語に対して語彙拡張を行なってしまうとノイズにより分類結果が悪くなってしまう．そこで我々は「目的」タグの文書だけに拡張シソーラスを用いた語彙拡張を行った．

テーマ名は文書のテーマごとの分類に有用であるため，テーマ名中の名詞を用いた．我々は下に示す手法を用いた．

１．テーマ名に現れる名詞が学習データのテーマの名詞集合に含まれていない場合，名詞集合にその名詞を追加する．

２．テーマ名に現れる名詞が学習データのテーマの名詞集合に含まれている場合，それらのシソーラス内の上位語とそれらの類似単語に重みをつける．

表 3　意味タグのラベルと分類された申請者タグ

意味タグ	申請者タグの例	タグ内の平均名詞数
技術分野	産業上の利用分野	80.5
目的	発明の名称，発明の目的	134.1
手法	課題を解決するための手段	71.2
特許申請の範囲	特許申請の範囲	151.2
説明	発明の詳細な説明	166.4
例	具体的な例	72.5

（3）文書分類

我々はNaive Bayes分類器を用いて文書分類を行った．関連テーマとして選択されたテーマ $\hat{theme}$ は式(3)で表される．

$$\hat{theme} = \arg\max_{themes} P(theme)\prod_i P(w_i \mid theme) \quad (3)$$

ここで w_i は文書内の i 番目の単語を表している． w_i が拡張されたシソーラス内の単語である場合，シソーラス内の w_i の近隣単語も w_i' として文書分類に利用される．

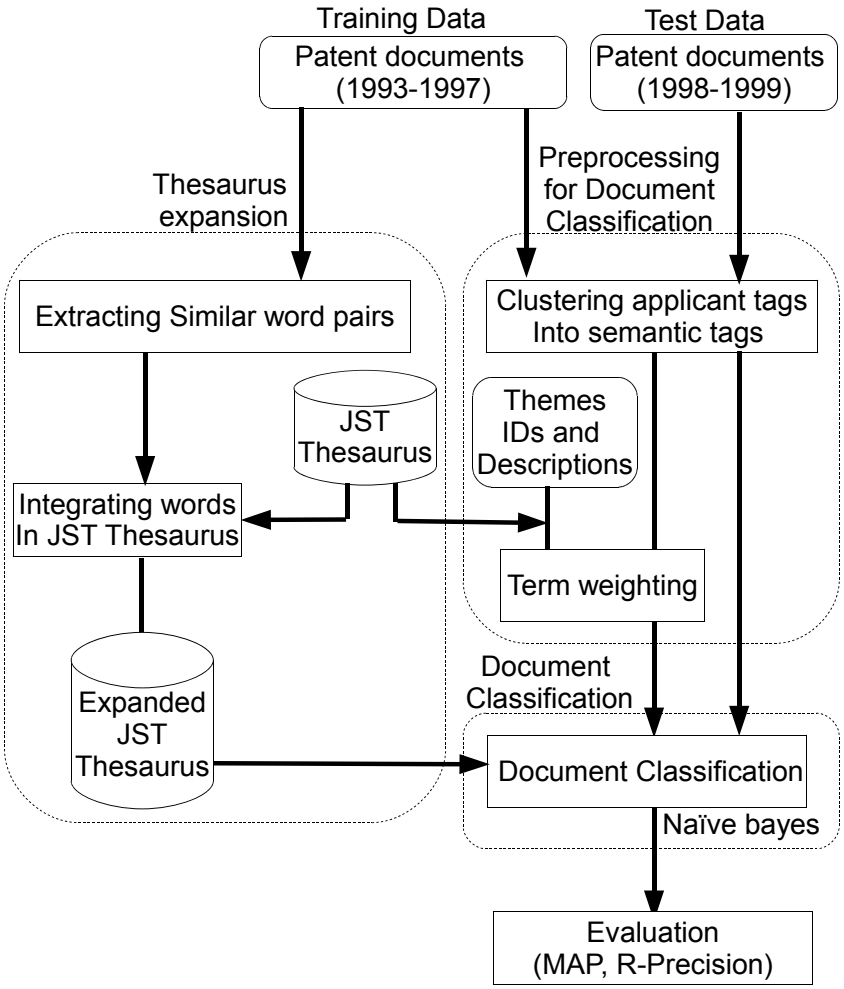

図 1　特許文書分類システムの概要

我々は文書分類のために拡張したシソーラスが有効であるかを調べるために実験を行った．使用した辞書は拡張した特許文書分類のための辞書である．実験では各文書は約2,900のテーマに分類される．実験のために日本語特許文書と提案手法により拡張された専門用語シソーラスを使用する．我々はNTCIR５の特許検索タスクによって提供された日本語特許請求文書（1993-1999）を用いた．学習データは1993年から1997年の文書を用いた．テストデータは1998年と1999年を用いた．学習データは1,707,194文書．テストデータは2,008文書である．学習データもテストデータも複数のテーマに分類される．テーマの数は2,903である．各文書の平均テーマ数は約2.26である．まず，我々は類似単語ペアを抽出し，既存のシソーラスであるJSTシソーラスに追加した．JSTシソーラスは43,314の見出し語を持つ．各見出し語は平均して６個の関連する単語を持つ．関連単語は３種類に分類される．すなわちNT（下位語），BT（上位語），RT（関連語）である．表 4にJSTシソーラス内の関連単語の数を示す．

表 4　JST シソーラス内の関連単語の数

関連単語	単語数
NT（下位語）	102,645
BT（上位語）	122,606
RT（関連語）	26,958

図 2は JST シソーラスの見出し語（フーリエ変換）とその関連語である.

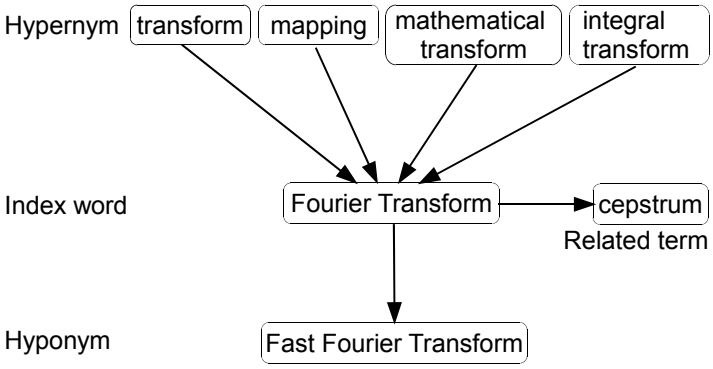

図 2　JST シソーラスの構造（一部）

シソーラスを利用して日本語特許文書からの関連文書検索を行った. また NTCIR５での結果と提案手法を用いた結果とを比較した.結果の評価のため MAP と R-Precision を利用した.MAP は式(4)によって定義される.

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\Pr ecision(R_{jk}) \quad (4)$$

ここでQはテスト文書の集合，m_jは文書jの関連文書の数，R_{jk}は文書jのk番目の関連文書を表す.

表 5に分類実験の結果を示す．表 5中の BOLA1, JSPAT2, WGLAB9 と FXDM3 は NTCIR5 の特許検索タスクの参加システム名である．BOLA1 は k-NN 法と特許文書の構造を利用している．JSPAT2 は Naive Bayes を利用している．WGLAB9 は k-NN 法を，FXDM3 はベクトル空間法を利用している．我々は日本語特許文書を用いて専門用語シソーラスを拡張した．拡張したシソーラスが文書分類に有用であるかを調べるため，シソーラスを使用した場合と使用しない場合との比較を行った．その結果，拡張したシソーラスの利用が文書分類タスクに対して有用であることが分かった[Suzuki 2011a]．また，提案手法と NTCIR5 の特許分類タスクの他の手法との比較を行った．提案手法は非常にシンプルではあるが，他の手法と十分比較可能であることがわかった[Suzuki 2011b].

表 5　分類実験の結果（MAP と R-Precision）

手法	MAP	R-Precision
cosine	0.45	0.41
cosine+JST Thesaurus	0.47	0.43
cosine+expanded Thesaurus	0.47	0.43
Naive Bayes	0.63	0.53
Naive Bayes + JST Thesaurus	0.66	0.56
Naive Bayes + expanded Thesaurus	0.67	0.56
k-NN(BOLA1)	0.69	0.59
Naive Bayes (JSPAT2)	0.66	0.56
k-NN (WGLAB9)	0.62	0.53
VSM (FXDM3)	0.49	0.39

4-2 ホテルレビュー文の分類

(1) システムの概要

図 3は提案するホテルレビュー文分類システムの概要を表している.

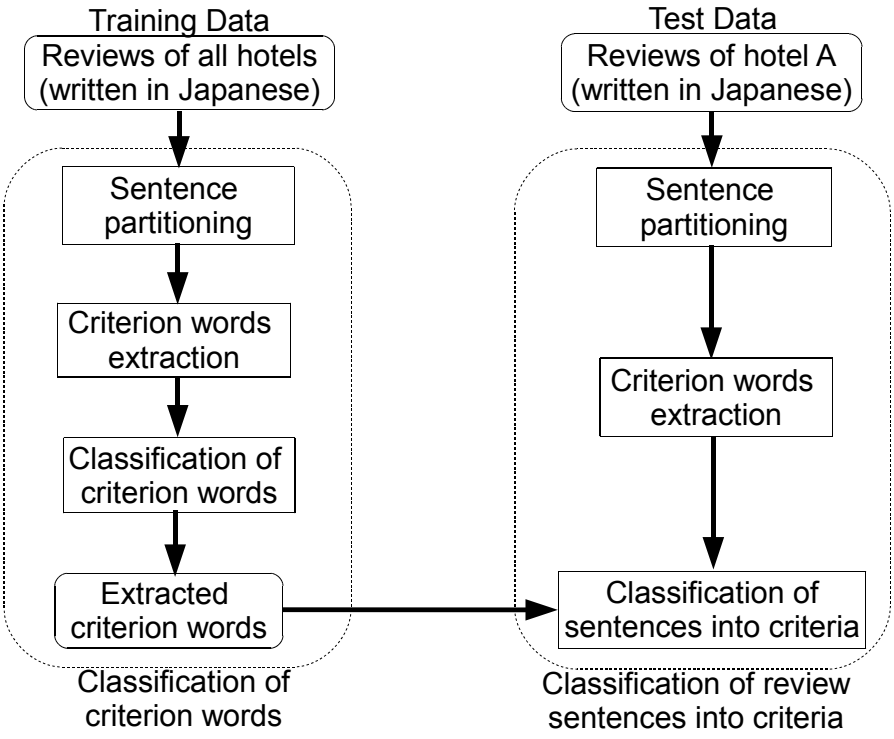

図 3　ホテルレビュー文の分類システムの概要

システムは２つの部分からなる．評価項目単語の分類とレビュー文の評価項目への分類である．ホテルレビュー文はシステムによって評価項目毎に分類される.

(2) 文の分割

レビュー文章には複文が頻繁に使用される.その上,複文には２つ以上の評価項目が含まれることが多い.例えば「朝食のバイキングも美味しいですし，部屋も広いし，一番気に入っているのが部屋に置いてあるシャンプーとリンスの香りがとてもいいと思います.」という文は複数の評価項目に関連するため，評価項目毎に分割することは必要である．福島らはテレビニュースのテキスト要約のために文の分割の方法[Fukushima 1999]を提案している．彼らは文分割のためにルールベースの手法を用いている.

本研究では各複文は複文マーカーと CaboCha（日本語係り受け解析器）[CaboCha 2002]によって評価項目に分割される．表 6は我々の用いた複文マーカーを表す.

表 6　文分割のための複文マーカー

No.	マーカー	頻度
1	ですし	5,597
2	だし	4,888
3	あるし	1,119
4	ったし	2,830
5	ますし	2,408

(3) 評価項目単語の抽出

まず，評価項目単語をレビュー内でレビュー投稿者が言及する単語と定義する．評価項目単語はレビュー内で頻繁に「後置詞「は」＋形容詞」と共起する．レビュー内の評価項目単語を抽出するために，まず「単

語 A＋は＋形容詞」のパターンを抽出する．次に単語 A を抽出し，最後に 3 節で説明する手法と日本語 WordNet[JWordNet]の単語 A の上位単語･下位単語を用いて単語 A の類似単語として抽出される単語を抽出する．表 7はパターン「単語 A＋は＋形容詞」内に頻繁に現れる形容詞を表している．表 7中の形容詞の多くは評価に関する形容詞であることがわかる．

表 7　名詞＋は＋形容詞のパターンで頻出する形容詞

No.	形容詞	頻度	No.	形容詞	頻度
1	良い	142,719	6	美味しい	33,318
2	ない	73,186	7	安い	28,463
3	よい	67,643	8	おいしい	27,310
4	広い	55,524	9	多い	23,122
5	近い	52,423	10	狭い	20,345

表 8は抽出された評価項目単語とそれらの頻度を表している．表中のこれらの単語はホテルの評価項目に関連している．

表 8　評価項目の候補単語(top 10)

No.	単語	頻度	No.	単語	頻度
1	部屋	56,888	6	サービス	11,270
2	朝食	25,068	7	浴室	9,864
3	食事	17,107	8	騒音	8,695
4	サポート	16,677	9	料理	8,252
5	立地	14,866	10	温泉	7,774

我々はレビュー文を日本語 WordNet の語彙情報と単語の類似度を使って分類する．**表 8**内に示される結果から１２個の評価項目を選び出す．まず，各文を１２の評価項目と「その他」の項目に分類する．次に 2 種類のナイーブベイズ（MNB と CNB [JRennie2003]）を用いて各文を分類する．ナイーブベイズ分類器はテキスト分類によく利用される．なぜならナイーブベイズは処理スピードが速く，プログラムが簡単で学習データが少なくても比較的効果的であるからである．ナイーブベイズ分類器を用いて高精度で分類するためには，各クラスに対して多くの学習データが必要である．しかしこのタスクでは多くの学習データを集めることが難しいクラスが存在する．そこで我々は CNB(Compliment Naive Bayes)を用いる．CNB は学習データとしてそれぞれのクラスの補集合を用いるので各クラスに対してより多くのデータを利用することができる．学習データを拡張するために我々は MNB と CNB で同じ評価項目に分類された文を用いた．表 9 は MNB と CNB を使った分類結果を表している．

表 9　MNB と CNB を用いた分類結果

手法	Precision	Recall	F-score
MNB	0.72	0.63	0.67
CNB	0.75	0.64	0.69
MNB と CNB	0.81	0.61	0.70

表 9からわかるように MNB と CNB で同じ評価項目に分類された時，ほとんどの場合分類された評価項目は正しい． そこで我々は MNB と CNB で同じ評価項目に分類された文を追加の学習データとする．MNB 分類器は式(5)で得られる．

$$MNB(d) = \arg\max_{c}\{\log \hat{p}(\theta_c) + \sum_i f_i \log \frac{N_{ci} + \alpha_i}{N_c + \alpha}\} \quad (5)$$

ここで $\hat{p}(\theta_c)$ は事前確率である． f_i はレビュー d 内の単語 i の出現頻度． N_{ci} は単語 i がクラス C の学習

データ中に現れる回数を表す．N_cはクラスC中の学習データに現れる単語の数を表す．α_iとαのためにそれぞれ1と語彙の数を用いた．CNB分類器は式(6)で定義される．

$$CNB(d) = \arg\max_c \{\log \hat{p}(\vec{\theta}_c) + \sum_i f_i \log \frac{N_{\tilde{c}i} + \alpha_i}{N_{\tilde{c}} + \alpha}\} \quad (6)$$

ここで$N_{\tilde{c}i}$はC以外のクラスに属する文書に現れる単語iの出現回数，$N_{\tilde{c}}$はC以外のクラスでの出現頻度である．α_iと αはスムージングパラメータである．ベクトル$\vec{\theta}_c$ は$\vec{\theta}_c = \{\theta_{c1}, \theta_{c2}, \cdots, \theta_{cn}\}$とする．

レビュー文分類実験のために楽天株式会社から許諾を頂き，楽天トラベルのレビューを使用した．表 10で楽天トラベルのレビューデータを示す．

表 10　楽天トラベルのレビューデータ

データの量	250MB
レビューの数	350,000
ホテルの数	15,437
1レビューあたりの単語数	375
1ホテルあたりのレビュー数	23

レビュー文を12の評価項目と「その他」の13項目に分類する実験を行った．表 11は実験で使用した12の評価項目である．

表 11　12種類の評価項目とその評価項目単語

No.	評価項目	評価項目単語	No.	評価項目	評価項目単語
1	立地	立地，アクセス	7	風呂	浴室，バスタブ
2	設備	スイミング・プール	8	アメニティ	髭剃り，歯ブラシ
3	サービス	サポート，サービス	9	ネットワーク	Wi-Fi，ブロードバンド
4	食事	朝食，食事	10	飲料	ビール，コーラ
5	客室	部屋，ノイズ	11	ベッド	ベッド，枕
6	ロビー	ロビー，ラウンジ	12	駐車場	駐車スペース，自動車

日本語類語辞書として日本語 WordNet 1.1[JWordNet]を用いた．ホテルレビュー内の類似単語抽出のためLinの手法[Lin98]を用いた．レビュー分類用実験データとして5つのビジネスホテルのレビューを用いた．ホテル毎のレビューの数は51.2である．表 12はテキスト分割の結果を表す．表 12から分かるように，CNBを用いた結果はMNBを用いた結果よりも精度が向上した．

表 12　クラスタリング結果

手法	Precision	Recall	F-score
MNB	0.74	0.65	0.69
CNB	0.76	0.67	0.71

(4) まとめ

2種類のナイーブベイズ分類器（MNBとCNB）を用いてレビュー文の分類を行った．CNBを用いた結果はMNBを用いた結果よりも精度が良かった．その理由の一つはクラス間のデータのバランスが良いことである．CNBで用いた学習データの数はMNBよりも多くなった．なぜなら各評価項目クラスに関連する限られた単語からなるデータを用いたからである．従って各評価項目クラスの学習データの数はそれぞれ違う．それに対してCNBによる学習データは各クラスの補集合からなるため,学習データ内の単語の数はMNBよりも大きくなる．また学習データは各クラスに対してバランスが良くなる

ホテルレビュー文の分類を行うためにホテルレビューから作成した概念辞書を利用し，分類精度の評価を行った[Takubo2011, Suzuki2012a, Suzuki2012b]．具体的には２種類の Naive Bayes 分類器（Multinomial

Naive Bayes と Compliment Naive Bayes）を利用して，楽天トラベルのホテルレビューを１２個の分類項目（立地，設備，サービス，食事，客室など）に分類した．２段階のレビュー文分類手法とホテルレビューから事前に作成しておいた概念辞書の利用により,F-score による評価で 0.71 という高い結果を得たことから，ホテルレビュー文の項目毎の分類タスクに対して，自動作成した概念辞書の利用は有効であると言える．

5 質問応答システム

作成した辞書を利用した時の効果を調べるため，新聞記事を情報源とした質問応答システムを構築した．質問応答システムは（１）質問解析，（２）情報検索，（３）情報抽出，（４）回答選択の４つの部分から構成される．各部について簡単に説明する．

(1) 質問解析：入力された質問文が何に関する質問か，何を回答すればよいかを解析するために，質問文を形態素解析し，疑問詞を抽出する．

(2) 情報検索： 質問文に含まれる単語が多く含まれる新聞記事を抽出する．質問文に含まれる単語は非常に少ないため，解答が含まれる記事であっても質問文に現れる単語すべては含まない記事がある．そこで前もって作成しておいた概念辞書を利用して質問文に含まれる単語の類義語を含めて，それらの単語が多く含まれる新聞記事を抽出する．

(3) 情報抽出：抽出した新聞記事から質問解析で得られた疑問詞に対応する単語を抽出する．単語抽出時に固有表現情報や辞書情報を利用している．

(4) 回答選択：(3)で抽出した回答候補から出現頻度を考慮して回答を出力する．

現在，疑問詞を含む質問文に対しては回答を提示することができるようになっている．また，質問応答システムは研究室の Web ページで公開準備を進めている．また，現在質問応答システムによる概念辞書利用の効果を調査中であり，結果を国際会議で報告するための準備を進めている．

6 まとめ

本研究により概念辞書作成とその利用について一定の成果を得ることが出来たが，辞書作成時の正確性，概念辞書の利用技術など，改良すべき課題も明らかになった．今後，概念辞書作成手法の改良を行うとともに，様々な分野での文書分類，文書要約システムなどによる概念辞書利用の効果を検証していくつもりである．

【参考文献】

[JWordNet 2009] Francis Bond and Hitoshi Isahara and Kiyotaka Uchimoto and Takayuki Kuribayashi and Kyoko Kanzaki, “Enhancing the Japanese WordNet”, The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP, 2009.

[EDR Dictionary 1996], "Japan Electronic Dictionary Research Institute, Ltd. (EDR)", “EDR ELECTORONIC DICTIONARY VERSION 2.0 TECHNICAL GUIDE”, *National Institute of Information and Communication Technology*, 1996.

[Bunruigoihyo 1964], National Language Research Institute, "Bunruigoihyo", *Shuei publisher* (In Japanese)", 1964.

[Hindle 1990] D.Hindle, “Noun Classification from Predicate-Argument Structures”, *Proceedings of 28th Annual Meeting of the Association for Computational Linguistics*, pp.268-275, 1990.

[Lin 1997] Dekang Lin, “Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity”, Proceedings of ACL/EACL-97,pp.64-71, 1997.

[Hagiwara2006] Hagiwara, M. and Ogawa, Y. and Toyama, K.,”Selection of Effective Contextual Information for Automatic Synonym Acquisition”, *In Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp.353-360, 2006.

[萩原 2005] 萩原正人, 小川泰弘, 外山勝彦, “シソーラス自動構築における PLSI の利用”, 情報処理学会研究報告, 自然言語処理研究会, 2005-NL-166, pp.71-78, 2005.

[Tokunaga 1997] T.Tokunaga,"Extending a thesaurus by classifying words", *In Proceedings of the ACL-EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*, pp.16-21, 1997.
[Uramoto1996a] N.Uramoto, "Corpus-based Thesaurus -- Positioning Words in Existing Thesaurus Using Statistical Information from a Corpus", *Journal of Information Processing Society of Japan*, VOLUME 37, NUMBER 12, pp.2182-2189, 1996.
[Uramoto1996b], N.Uramoto, "Positioning unknown words in a thesaurus by using information extracted from a corpus", *In proceedings of COLING'96*, pp.956-961, 1996
[Beineke 2004] P. Beineke and T. Hastie and S. Vaithyanathan, "The sentimental factor：Improving review classification via human-provided information", *the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004.
[Yi 2005] J. Yi and W. Niblack,"Sentiment mining in webfountain", *Proceedings of the 21st International Conference on Data Engineering*", 2005.
[Hu 2004] M. Hu and B. Liu, "Mining opinion features in customer reviews", *Proceedings of Nineteenth National Conference on Artificial Intelligence*, 2004.
[Wei 2010] Wei Wei and Jon Atle Gulla, "Sentiment Learning on Product Reviews via Sentiment Ontology Tree", *Annual Meeting of the Association for Computational Linguistics*, pp.404-413, 2010
[Hearst 1997] Hearst, M. A., "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages", *Association for Computational Linguistics*, pp.111-112, 1997.
[Utiyama 2001] Masao Utiyama and Hitoshi Isahara, "A statistical model for domain-independent text segmentation", *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pp.499-506, 2001.
[Hirao 2000] Tsutomu Hirao and Akira Kitauchi and Tsuyoshi Kitani, "Text Segmentation Based on Lexical Cohesion and Word Importance",*Information Processing Society of Japan*, Volume41, Number SIG3(TOD6), pp.24-36, 2000.
[Lin 1998] Dekang Lin, Automatic Retrieval and Clustering of Similar Words,
Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Proceedings of the Conference, pp.768-774, 1998.
[JST Thesaurus1999] JST, "JST Thesaurus 1999",http://jois.jst.go.jp/JOIS/html/thesaurus¥_index.htm", 1999.
[Fukushima1999] Takahiro Fukushima and Terumasa Ehara and Katuhiko Shirai, "Partitioning long sentences for text summarization", *Journal of Natural Language Processing (in Japanese)*,
Volume6,Number6, pp.131-147, 1999.
[Suzuki 2011a] Yoshimi Suzuki and Fumiyo Fukumoto, "Multi-Labeled Patent Document Classification using Technical Term Thesaurus", *Proc. of 3rd International Conference on Knowledge Engineering and Ontology Development*, pp.425-428, 2011.
[Suzuki 2011b] Yoshimi Suzuki "Patent Document Classification using expanded Technical Term Thesaurus", The 5th Language and Technology Conference (LTC'11),, pp.151-155, 2011.
[CaboCha 2002] Kudo, T. and Matsumoto, Y.}, "Japanese Dependency Analysis using Cascaded Chunking", *CoNLL 2002:Proceedings of the 6th Conference on Natural Language Learning 2002*, pp.63-69, 2002.
[JRennie 2003] Jason D. M. Rennie and Lawrence Shih and Jaime Teevan and David R. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers", *Twentieth International Conference on Machine Learning*, pp.616-623, 2003.
[Takubo2011] 自動抽出した利用者の視点によるレビュー要約:田窪直人, 鈴木良弥, 言語処理学会第 18 会年次大会(NLP2012), pp.295-298, (広島), 2011.
[Suzuki2012a] Yoshimi Suzuki and Fumiyo Fukumoto, "Segmentation of Review Texts by Using Thesaurus and Corpus-based Word Similarity", Proc. of 3rd International Conference on Knowledge Engineering and Ontology Development, pp.381-384, 2012.

[Suzuki2012b] Yoshimi Suzuki, “Classifying Hotel Reviews into Criteria for Review Summarization”, *Proc. Of 2nd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2012)*, pp.65-72, 2012.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
Multi-Labeled Patent Document Classification using Technical Term Thesaurus	Proc. of 3rd International Conference on Knowledge Engineering and Ontology Development (KEOD2011), pp.425-428	2011.10
Identifying Event and Subject of Continuous News Streams for Multi-Document Summarization	The 5th Language and Technology Conference (LTC’11), pp.197-201,	2011.11
Patent Document Classification using expanded Technical Term Thesaurus	The 5th Language and Technology Conference (LTC’11), pp.151-155,	2011.11
Segmentation of Review Texts by Using Thesaurus and Corpus-based Word Similarity	Proc. of 3rd International Conference on Knowledge Engineering and Ontology Development (KEOD2012), pp.381-384	2012.10
Classifying Hotel Reviews into Criteria for Review Summarization	Proc. Of 2nd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2012), pp.65-72	2012.12