NAIST-IS-DD0161012

博士論文

多言語情報アクセスのための
言語資源の構築と利用に関する研究

木村 文則

2007年2月15日

奈良先端科学技術大学院大学
情報科学研究科 情報システム学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
博士 (工学) 授与の要件として提出した博士論文である。

木村 文則

審査委員：

植村 俊亮 教授　（主指導教員）
関 浩之 教授　（指導教員）
宮崎 純 助教授　（委員）

多言語情報アクセスのための言語資源の構築と利用に関する研究*

木村 文則

内容梗概

本論文は，ある言語で書かれた文書群を別の言語による問合せで検索することを可能とする言語横断型情報検索（Cross-Language Information Retrieval: CLIR）あるいは多言語情報アクセスにおいて，ドメインに依存しない言語資源を構築する技術及びその活用について報告する．

世界的なインターネットの発展とともに，利用者は世界中で激増しており，その国籍も、Web 文書の記述に用いられる言語も多様化している．利用者の検索要求によっては，利用者の母国語以外の言語で記述された情報の方が豊富である場合も考えられ，これらを検索したいというニーズは少なくない．外国語文書を電子的に入手することは容易になったが，多くの利用者は母国語以外の言語に精通していない．本研究では言語を越えた言語横断型情報検索 (多言語情報アクセス) に焦点をあて，研究を行った．

従来の Web 検索エンジンは，問合せと同一言語の文書群が検索対象であるため，外国語文書に対する検索は効率的とは言い難い．従来の単言語検索システムにおいてこのような要求を満たすには，利用者自身が辞書などを用いて問合せを翻訳する必要がある．この作業は利用者に負担を強いるだけではなく，不慣れなあるいは全く読み書きができない言語に翻訳する場合は，適切な訳語の選択を誤る可能性が高い．

* 奈良先端科学技術大学院大学 情報科学研究科 情報システム学専攻 博士論文, NAIST-IS-DD0161012, 2007 年 2 月 15 日.

このような要求から，ある言語で書かれた文書群を別の言語による問合せで検索することを可能とする言語横断型情報検索に関する研究が近年盛んである．そこでは、問合せの翻訳や訳語の曖昧性解消などにコーパスを利用する手法などが提案されて，検索精度の向上において一定の成果が得られている．しかしコーパスを利用した手法では，学習に用いるコーパスのドメインに対する依存が大きいため，それ以外のドメインに対しては検索精度が低くなる可能性がある．Web 文書の言語横断検索では文書内容の分野は広範囲に渡っているため，ドメイン依存の問題を改善しなければならない．

本研究は，ドメイン依存の問題を解決するための言語資源の構築および利用を中心テーマとしている．Web ディレクトリを利用してある言語のオントロジーを別の言語に翻訳することにより、二つの言語を対象としたオントロジーを構築する手法を提案した．また，Web ディレクトリを言語資源として用いることを提案し，これを利用した言語横断情報検索システムの構築を行った．さらに、これらのシステムの実証実験を行い，Web ディレクトリを言語資源として利用することが多言語アクセスシステムに有効であることを確認した．

キーワード

多言語情報アクセス，言語横断，情報検索, Web ディレクトリ, 曖昧性解消, 言語資源

Studies on Construction and Utilization of Linguistic Resource for Multi-Lingual Information Access*

Fuminori Kimura

Abstract

This paper reports construction and utilization of domain independent linguistic resource in Cross-Language Information Retrieval or Multi-Lingual Information Access.

The number of the Internet user increases in all over the world, and nationality of the users diversifies. Languages to describe Web documents also diversify. With the worldwide popularity of the Internet, more and more languages are being used for Web documents, and it is now much easier to access documents written in foreign languages. There might also be cases, depending on the user's needs, where valuable information is written in a language other than the user's native language. However, many users are not familiar with foreign languages. Then the support system, such as Multi-Lingual information access, to overcome a language barrier is important in order to reply to these demands. In order to realize multi-lingual information access, it demands retrieval across languages and translation of Web documents. In this study, we focus on retrieval.

Existing Web search engines only support the retrieval of documents that are written in the same language as the query, so there is no efficient way for monolingual users to retrieve documents written in non-native languages. To retrieve

* Doctoral Dissertation, Department of Information Systems, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0161012, February 15, 2007.

across languages in a typical monolingual retrieval system, users have to manually translate queries themselves using a dictionary, etc. This method is not only difficult for the user, it might also result in the query being translated incorrectly, especially when the user is unfamiliar with the language.

To meet these needs, there has been intensive research in recent years on Cross-Language Information Retrieval (CLIR), a technique for retrieving documents written in one language using a query written in another language. A variety of methods, including the use of corpus statistics to translate terms and the disambiguation of translated terms, have been investigated and some useful results have been obtained. In this approach, a source language query is first translated into the target language using a bilingual dictionary, and the translated query is then disambiguated. Our method falls into this category.

However, corpus-based disambiguation methods are significantly affected by the domain of the training corpus, so they may be much less effective for retrieval in other domains. Lin et al. conducted comparative experiments between three monolingual corpora that had different domains and sizes, and concluded that a large-scale, domain-consistent corpus is needed to obtain useful co-occurrence data. In this paper, we studied construction and utilization of linguistic resources in order to resolve domain dependency problem. we proposed the method of constructing the bilingual ontology by translating monolingual ontology using Web directory. Besides, we proposed to utilize web directory as linguistic resources, and constructed Cross-Language Information Retrieval System utilized this linguistic resources. In this paper, we verify that using Web directory as linguistic resource is effective in multi-lingual information access.

Keywords:

Multi-Lingual Information Access, Cross-Language, Information Retrieval, Web Directory, Disambiguation, Linguistic Resource

目 次

1. はじめに **1**

1.1 研究の背景 1

1.2 研究の概要 7

1.3 論文の構成 8

2. 関連研究 **9**

3. 曖昧性を解消した辞書の構築 **15**

3.1 オントロジ 17

3.2 Web ディレクトリを利用した訳語の曖昧性解消 20

3.3 2 言語オントロジの構築実験 24

3.4 2 言語オントロジ構築における考察 26

4. Web ディレクトリの言語横断情報検索への利用 **29**

4.1 提案する手法 30

4.1.1 前処理 31

4.1.2 検索処理 34

4.1.3 問合せの適合カテゴリの選択の実験 39

4.1.4 カテゴリの統合 43

4.2 評価実験 44

4.3 問合せの翻訳 58

4.3.1 検索実験の結果 59

5. 問合せ翻訳手法の改良 **65**

5.1 適合カテゴリの選択 66

5.1.1 実験（適合カテゴリの選択） 66

5.2 複数の訳語の選択 67

5.2.1 実験（複数の訳語の選択） 69

5.2.2 考察 71

6. Web ディレクトリのカテゴリ構造を利用した訳語の曖昧性解消 75
6.1 カテゴリの細分化 77
6.1.1 実験結果 78
6.1.2 2 階層まで利用した検索実験 84

7. 現状と今後の課題 91
7.1 本研究により得られた知見 91
7.2 今後の課題 92

謝辞 95

参考文献 97

研究業績 105

図 目 次

1.1 インターネットの利用率の推移 2
1.2 ユーザの使用言語の割合 (2002 年) 3
1.3 ユーザの使用言語の割合 (2004 年) 3
1.4 ユーザの使用言語の割合 (2006 年) 4
1.5 Web 文書の記述に用いられる言語の分布 (2002 年) 5
2.1 言語横断情報検索: 問合せを翻訳する方式 11
3.1 名詞 “tree” の階層構造 17
3.2 WordNet.1.7.1 による単語 “bank” の概念の抽出結果 19
3.3 WordNet.OWL(抜粋) 21
3.4 2 言語オントロジ構築手法の概要 23
3.5 2 言語オントロジ構築における曖昧性解消の流れ 24
4.1 提案システムの概要 32
4.2 前処理における処理の流れ 33
4.3 検索の流れ 37
4.4 問合せの訳語の決定 39
4.5 カテゴリの統合 44
4.6 NTCIR3 言語横断検索タスク 日本語検索課題 (抜粋) 46
4.7 ストップワード一覧. 48
4.8 再現率・適合率曲線 (a) 56
4.9 再現率・適合率曲線 (b) 57
4.10 再現率・適合率曲線 (c) 57
4.11 カテゴリの特徴語数と得られた訳語獲得率との関係 58
4.12 検索結果の適合率・再現率グラフ 60
4.13 固有名詞を人手で翻訳した場合の適合率・再現率グラフ 62
5.1 適合カテゴリを複数選択した場合の検索結果の適合率・再現率グラフ 67
5.2 適合カテゴリを複数選択した場合の検索結果の適合率・再現率グラフ 70
5.3 適合カテゴリを複数選択した場合の検索結果の適合率・再現率グラフ 71
5.4 適合カテゴリを複数選択した場合の検索結果の適合率・再現率グラフ 72

表 目 次

1.1 ユーザの使用言語の割合 (2006 年) 4
3.1 retrieve に対する訳語 15
3.2 英語版および日本語版 Yahoo! カテゴリにおける最上位の 13 カテゴリ 25
4.1 英語版 Yahoo! カテゴリにおけるカテゴリ “Government” の特徴語（上位 10 語） 35
4.2 内積による適合度を用いた場合の適合カテゴリ (上位 10 件). . . . 41
4.3 提案手法による適合度を用いた場合の適合カテゴリ (上位 10 件). . 42
4.4 NTCIR3 日本語問合せ（TITLE）–前半. 49
4.5 NTCIR3 日本語問合せ（TITLE）–後半. 50
4.6 NTCIR3 日本語問合せ（DESC）問 1–13. 51
4.7 NTCIR3 日本語問合せ（DESC）問 14–26. 52
4.8 NTCIR3 日本語問合せ（DESC）問 27–39. 53
4.9 NTCIR3 日本語問合せ（DESC）問 40–50. 54
4.10 検索結果の平均適合率 61
4.11 固有名詞を人手により翻訳した場合の検索結果の平均適合率 . . . 63
5.1 訳語数限定手法の 11 点平均適合率. 69
5.2 固定閾値手法の 11 点平均適合率. 70
5.3 変動閾値手法の 11 点平均適合率. 72
6.1 階層ごとのカテゴリ数 79
6.2 1 階層までカテゴリを統合した場合の問合せとカテゴリの適合度 . 80
6.3 2 階層までカテゴリを統合した場合の問合せとカテゴリの適合度 . 81
6.4 3 階層までカテゴリを統合した場合の問合せとカテゴリの適合度 . 82
6.5 問合せと適合カテゴリの適合率の平均 83
6.6 2 階層および 3 階層における選択された適合カテゴリ 85
6.7 選択された適合カテゴリの順位の平均 86
6.8 各問合せごとの 11 点平均適合率 87
6.9 各問合せごとの訳語候補 (その 1) 88

6.10 各問合せごとの訳語候補 (その 2) 89

1. はじめに

1.1 研究の背景

1980年代まではインターネットは主に学術利用されていたが，1990年代になると商用利用されるようにもなった．またWWW閲覧ソフトが開発されたことにより，インターネットの利用が容易になった．このようにインターネットに関する環境が整備されていくことにより，1990年代中ごろからインターネットが民間にも普及し始めた．図1.1は，インターネットの利用率の推移である[64]．先進国では1995年ごろからインターネットの利用率が大きく上昇し始め，2004年には50%を超えるにいたった．それでもなお上昇傾向は衰えてはおらず，まだまだインターネットの利用率は増加するものと思われる．先進国の利用率に対して，全世界での利用率は2004年でも十数%とあまり高くはない．これは，発展途上国でのインターネットの普及があまり進んでいないことが原因であると思われる．発展途上国でのインターネットの普及率は2004年でも10%未満であるものの，2000年ごろから上昇し始めており，今後はこれらの国でもインターネットが普及していく可能性は高いと思われる．

1990年代後半以降インターネットが急速に普及したことは，インターネットの利用者を爆発的に増加させ，それは同時に膨大な情報がWeb上に発信されることとなった．例えばGoogle[21]では80億以上のページが登録されており，その数は年々増加の一途をたどっている．また，インターネットの量的な普及だけでなく，国際的な普及も急速に進んでおり，利用者の使用する言語も多様化している．図1.2および図1.3は，2002年と2004年のユーザの使用言語の割合の調査結果[52]である．世界のインターネットユーザのうち英語を母国語とするユーザの割合は，2002年の時点ですでに40%を切っており，中国語，日本語，スペイン語などの英語以外を母国語とするユーザの割合が増えている．また，別の調査結果[23]によると，表1.1に示されているように現在では英語を母国語とするユーザの割合は29.8%であり，わずか4年の間に10%ほど低下している．

これらのことが示すように，英語以外を母国語とするユーザが増加する傾向は年々大きくなっている．利用者の使用する言語も多様化と同様に，Web上に発信

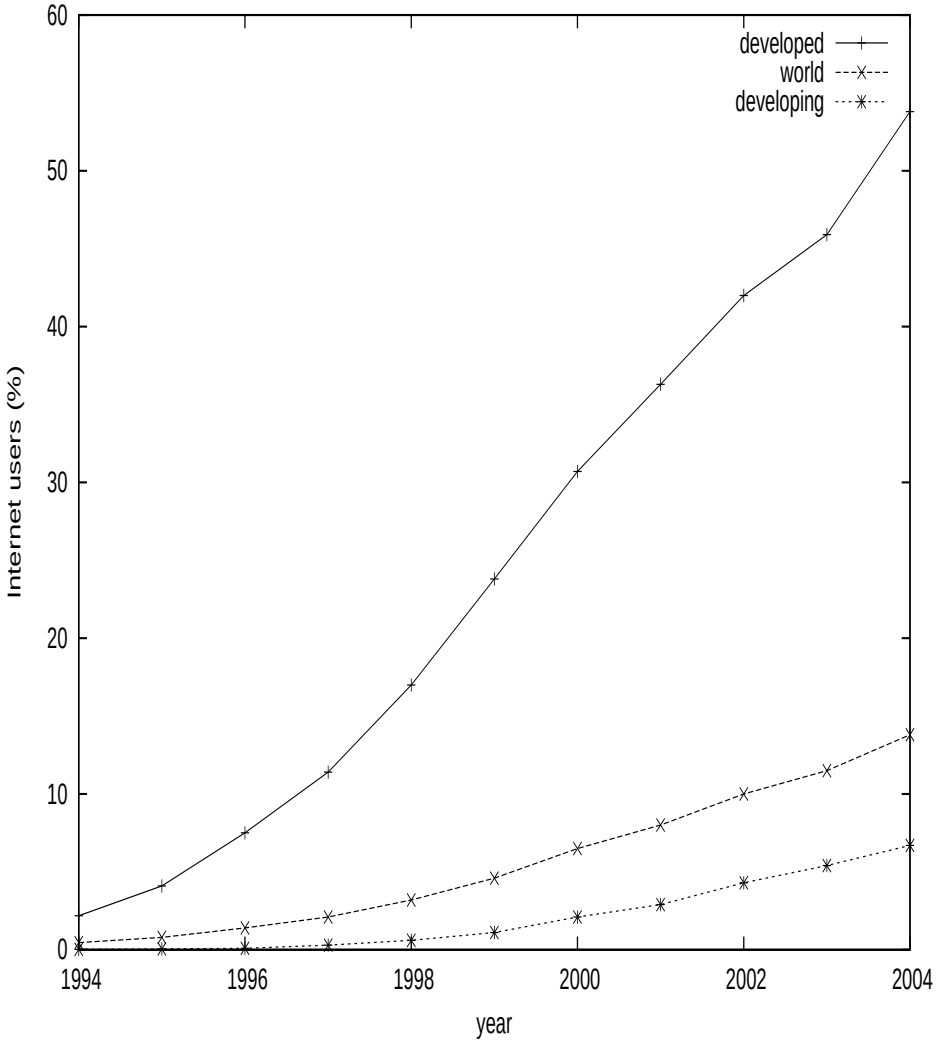

図 1.1 インターネットの利用率の推移

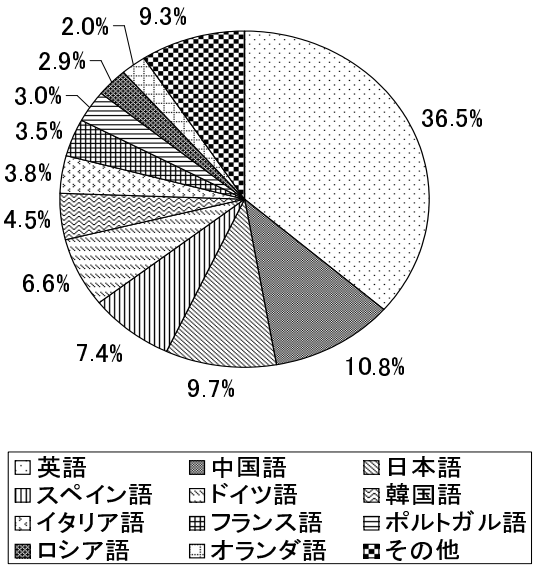

図 1.2 ユーザの使用言語の割合 (2002 年)

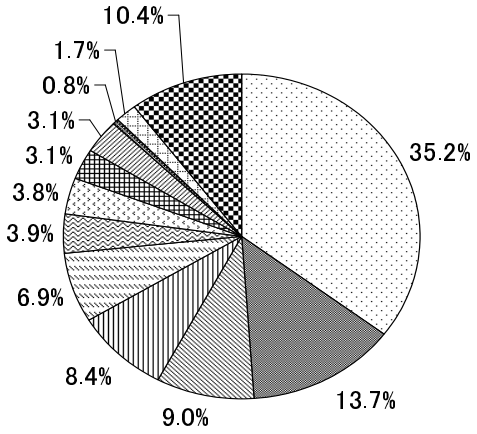

図 1.3 ユーザの使用言語の割合 (2004 年)

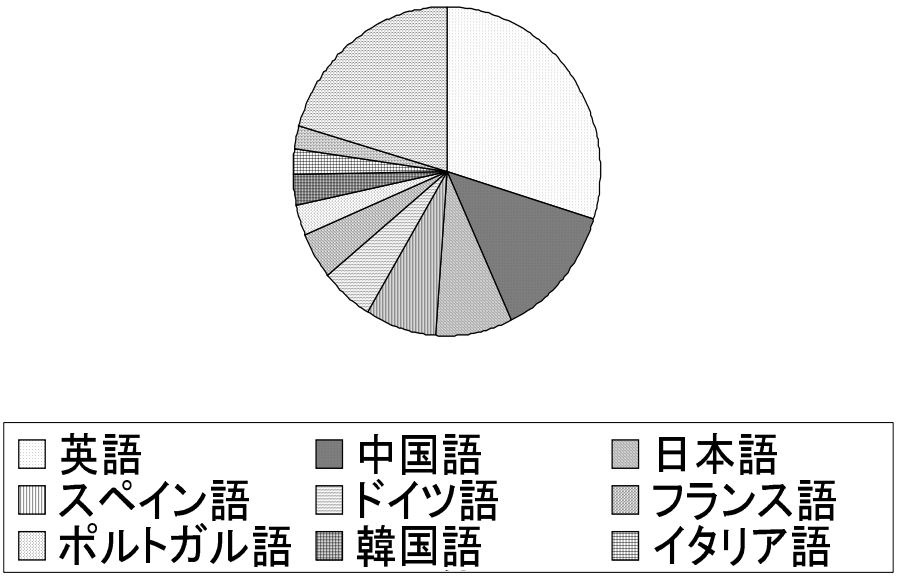

図 1.4 ユーザの使用言語の割合 (2006 年)

表 1.1 ユーザの使用言語の割合 (2006 年)

言語	利用者数 (百万人)	割合 (%)
英語	322	29.8
中国語	144	13.3
日本語	86	8.0
スペイン語	81	7.5
ドイツ語	58	5.4
フランス語	49	4.5
ポルトガル語	34	3.2
韓国語	32	3.0
イタリア語	28	2.6
ロシア語	23	2.1
その他	222	20.6
合計	1079	100.0

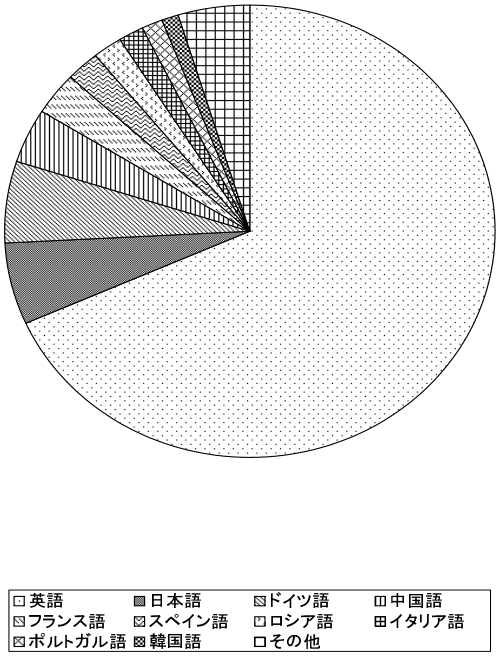

図 1.5 Web 文書の記述に用いられる言語の分布 (2002 年)

される情報の記述にも様々な言語が用いられるようになった．図 1.5 は，2002 年に発表された Web 文書の記述に用いられる言語の分布の調査結果 [52] である．1 位が英語 (68.4%)，2 位が日本語 (5.9%)，3 位がドイツ語 (5.8%) の順となっている．英語の割合が圧倒的に大きいが，2000 年には約 77%を占めていた英語の割合が 2002 年には 68.4%となっていることから分かるように英語の割合は年々低下している．利用者の使用する言語が多様化していることも考慮すると，今後も英語以外の言語の割合がいっそう増大していくと考えられる．

上記の調査結果は，ユーザがアクセスできる情報がより多くなり，またその恩恵にあずかり得るユーザ数も増大していることを示している．しかしながら，全てのユーザが容易にこのような恩恵にあずかることができるわけではない．なぜ

なら，多くのユーザは母国語以外の言語に精通していないからである．Web 上に発信される情報の記述にも様々な言語が用いられるようになったことは利用者の母国語で記述された情報が増加することとなったが，同時にそうでない情報がそれ以上に増大することとなった．このようなユーザにとって，母国語以外の言語で記述された情報を理解することはもちろんのこと，その中からそのユーザが必要とする情報を探し出すことさえ困難である．多国語に精通しているわけではない一般のユーザがこういった情報を有効に活用するためには，いわゆる “言語の壁” を乗り越える必要がある．しかし，全てのユーザが多国語に精通することなど実現不可能である．利用者の側が言語の壁を乗り越えるのが困難であるならば，Web 上で情報を発信する側が各国語版の Web ページを作成するなど，他国語を母国語とするユーザを想定した多言語化が行われれば，言語の壁の問題は解消すると思われる．しかしながら，多言語化に要する労力は多大である．また，情報を発信する側が各国語に精通していることが要求される．さらに利用される言語が増加する傾向にあることも考慮すると，情報の多言語化を実現することは困難であると思われる．

こうした言語の壁を乗り越え，母国語以外の言語にほとんど精通していないユーザであっても他国語で記述された情報を活用することができるようにしようと取り組んでいるのが，“多言語情報アクセス” の研究である [24, 25]．多言語情報アクセスシステムは主に「言語横断情報検索」と「機械翻訳」からなっている．言語横断情報検索により，Web 上に発信されている様々な言語で記述された情報から利用者が必要としている情報を探し出す．機械翻訳は，検索された情報を利用者が読むことができるように翻訳する．

多言語情報アクセスシステムでは，利用者は Web 文書を検索するための問合せを任意のただ一つの言語を入力するだけでよい．システムは利用者から入力された問合せを各言語に翻訳するなどの処理をしたうえで，各言語で記述された文書群に対して検索し，その結果を利用者に返す．このように，利用者は他国語に精通していなくても様々な言語で記述された Web 上にある情報にアクセスすることが可能となる．さらに多言語情報アクセスシステムでは，検索された Web 文書を利用者が精通している言語に翻訳することで，他国語に精通していない利用

者でも検索された様々な言語で記述された情報を読むことが出来るようになることをも目指している．

一方，ある分野の専門家などのように，自分の専門分野における専門用語の訳語は知っていると思われる利用者にも，多言語情報アクセスシステムは有用である．近年は技術の進歩が著しいため，分野によっては専門家でも最先端の知識を追い続けることが困難なこともある．その場合，新たに定義された用語に対する訳語を知らない可能性がある．このようなとき，多言語情報アクセスシステムが有用となる．また，新規に専門以外の分野について学び始める場合にはその分野の用語の訳語を知らない可能性は高くなり，多言語情報アクセスシステムの利用価値も高くなる．

多言語情報アクセスシステムを実現するには，各国語の Web 文書を正しく解析して検索することと，検索した Web 文書を翻訳することの二つのことが要求される．本研究ではこの二つのうちの各国語の Web 文書を正しく解析して検索することについて焦点をあて，研究を行った．

1.2 研究の概要

本研究では，多言語情報アクセスシステムの文書検索における，言語資源の構築と活用方法について研究を行った．多言語情報アクセスシステムのための言語資源の構築方法として，事前に曖昧性解消がなされている辞書の構築を行った．また，多言語情報アクセスシステムのための言語資源の利用方法として，Web ディレクトリを活用することを提案した．これにより，従来のコーパスによる訳語の曖昧性解消手法における問合せとコーパスの対象分野の不一致の問題を解消した．さらに，Web ディレクトリを利用する場合の問合せ翻訳方法について研究を行った．本手法による言語横断情報検索システムを構築し，検索実験により検証を行った結果，Web ディレクトリを言語資源として利用することが多言語情報アクセスシステムに有効であることが確認できた．

1.3 論文の構成

本論文は以下のような構成である．まず，第 2 章では多言語情報アクセスシステムの文書検索に必要となる言語横断情報検索についての本研究との関連研究について述べる．第 3 章では，多言語情報アクセスシステムのための言語資源の構築方法として，2 言語オントロジについて述べる．第 4 章では，多言語情報アクセスシステムのための言語資源の利用方法として，Web ディレクトリの言語資源としての活用方法について述べる．提案手法による言語横断情報検索システムの構築を行い，テストコレクションを用いた評価実験の結果などを示すことにより，提案手法の有効性について検証する．第 5 章では，Web ディレクトリを利用した場合の問合せ翻訳手法について述べる．第 6 章では，Web ディレクトリの階層構造を利用して，より深い階層のカテゴリまで利用することについて述べる．これにより，問合せが対象とする分野をより特定できるようになり，検索精度の向上を図ることができる．最後に第 7 章では，本研究を通して得られた知見，今後の研究の方向性について述べる．

2. 関連研究

多言語情報アクセスシステムを実現するには，言語の壁を越えた検索処理が必要である．これは，“言語横断情報検索 (Cross-Language Information Retrieval: CLIR)” と呼ばれている．言語横断情報検索は，ACM SIGIR などの国際会議，Cross Language Evaluation Forum(CLEF)[9] や NTCIR[48] のなどの情報検索評価会において主要なテーマの一つとなっている．

1990 年代以降インターネットが急速に普及し，膨大な情報がインターネット上で発信されるようになった．また，インターネットの量的な普及だけでなく，国際的な普及も急速に進んでおり，インターネットを利用しているユーザの使用言語も，1990 年代に過半数を超えていた英語を母国語とするユーザの割合が，現在では 30% を切っている状態であり，欧州やアジア各国の言語の台頭が目立っている．しかしながら，全てのユーザが容易に情報にアクセスし，多くの恩恵を受ける訳ではない．なぜなら，多くのユーザは母国語以外の言語に精通していないからである．このようなユーザにとって，母国語以外の言語で記述された情報を理解することを始め，その中からそのユーザが必要とする情報を探し出すことも困難である．

多国語に精通しているわけではない一般のユーザがこういったインターネット上の情報を有効に活用するためには，いわゆる「言語の壁」を乗り越える必要がある．つまり，インターネット上で情報を発信する側が各国語版の Web ページを作成するなど，他国語を母国語とするユーザを想定した多言語化が必要となるが，多言語化に必要となる労力は膨大であり，また同時に情報を発信する側が各国語に精通していることが要求されるため，情報の多言語化は容易に実現できるものではない．

このような要求から，ある言語で書かれた文書群を別の言語による問合せで検索することを可能とする言語横断情報検索に関する研究 [22, 16] が 1990 年代後半から活発に行われるようになった．

言語横断情報検索では，問合せと検索対象文書との間の言語的な相違を吸収し，単言語検索へと帰着させることが主要な課題である．言語横断情報検索に用いられる手法は大きく分けて，検索対象の文書群を翻訳する方式，言語に依存しない

中間言語を用いる方式，問合せを翻訳する方式の三つがある．

検索対象の文書群を翻訳する方式は，事前に全ての検索対象の文書群を問合せと同じ言語に翻訳しておくことにより，問合せを翻訳することなく単言語検索に帰着させて検索を行う．この方式では既存の機械翻訳システムを用いることができ，文脈を考慮できることにより訳語の曖昧性も低くなることから，一般に問合せを翻訳する方式より高い検索精度が得られるとされている [70]．しかしながら，大規模な文書群をすべてあらかじめ翻訳しておくことは現実的ではなく，対応言語の拡張も困難であるため，Web のように多言語が混在し，かつ大規模で更新が頻繁な文書群の検索には不向きである．この問題を改善するために，文書群を翻訳する方式と問合せを翻訳する方式を組合わせた手法 [32] が提案されている．

言語に依存しない中間言語を用いる方式では，シソーラスの意味クラス [20] や Latent Semantic Index[12, 13, 53, 41, 69, 38] などを用いる．この方式では，言語の違いを意識することなく処理することが可能である．しかし，学習に用いるコーパスの規模が大きくなると計算量が膨大となるため，大規模な文書群に対しては実現は困難である．また，Latent Semantic Index を用いる手法では大規模な並列コーパスが必要とされる．並列コーパスとは，ある文書とそれを翻訳した文書が一対になっているのもを収集しコーパスとしたものである．並列コーパスの入手そのものが容易ではなく，それを大量に用意することは困難を極める．

問合せを翻訳する方式は，文字通りユーザによって入力される検索キーワードを翻訳する方法である [27, 11, 14, 30]．問合せを翻訳する方式においては，特に Web 検索エンジンの一般的な利用者が投入する問合せは平均 2 単語程度と短く，単語の羅列である場合が多いため [31]，訳語の曖昧性の解消が問題になる．しかしながら，この方式は，翻訳された問合せを既存の単言語検索エンジンでそのまま用いることができるという利点がある．この方式では，まず対訳辞書を用いて問合せを翻訳し，これに対して訳語の曖昧性を解消する．他の手法と比べて実現が比較的容易であることから，問合せを翻訳する方式が主流になっている．本研究で用いる手法もこの範疇である．

訳語の曖昧性解消の手法として，コーパスを用いる手法などがある [59]．コーパスから共起頻度などの統計情報を取得し，これらを利用して曖昧性解消を行う．

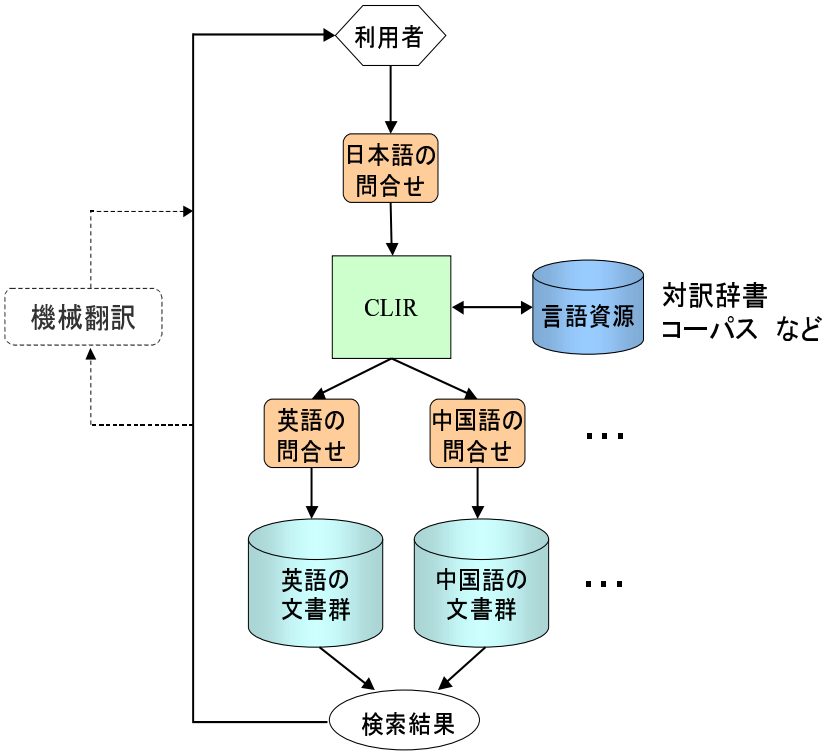

図 2.1 言語横断情報検索: 問合せを翻訳する方式

Ballesteros ら [1] や Gao ら [18] は，共起頻度を相互情報量に基づいて算出し，これを用いて訳語の曖昧性解消を行っている．また，前田らの相互情報量に基づく大域的な共起情報を用いる手法 [43] や，藤井らの Bigram に基づく局所的な共起情報を用いる手法 [17] などがこれまでに提案されている．

しかし，並列コーパスなどの言語資源の入手は一般に容易ではないことは問題点の一つである．この問題を解決する手法として，Web 文書をコーパスとして用いる手法 [42, 54] が提案されている．

また，訳語曖昧性解消にコーパスを用いる手法では，検索要求とコーパス間のドメインの相違による検索精度への影響が指摘されている．Hull[28] および奥村ら [51] は，並列コーパスや類似コーパスを用いる手法において，検索要求とコーパス間のドメインの相違が検索精度に悪影響を及ぼす可能性があることを指摘し

ている．また Lin ら [40] は，単言語コーパスとしてドメインや規模の異なる三つのコーパスを用いて比較実験を行った結果，有用な共起情報を得るには大規模でドメインの一致したコーパスが必要であると結論付けている．

また，対訳辞書やコーパスなどの言語資源が十分にそろっていない言語間においては，仲介言語を介する手法 [33, 56] などが提案されている．

本研究で対象とする Web 検索では，多様な分野の検索要求に対応することが要求される．しかし，そのそれぞれのドメインについて，対応するコーパスをあらかじめ用意することは現実的ではない．そこで本論文では，多様な分野の検索要求に対応するコーパスを利用することなく訳語の曖昧性解消を行うための言語資源の構築および利用について研究を行った．

検索モデルの一つにクラスタモデルがある．一般に，検索対象となる文書集合の中には記述内容の類似した文書が含まれることが多い．そこで，あらかじめ文書集合を類似度に応じてグループ分けしておき，検索時には個々の文書と問合せとの適合度を計算するのではなく，これらのグループと問合せの適合度を計算することによって，計算量を削減することが可能になる．このような検索手法はクラスタ検索と呼ばれる [67]．クラスタモデルではあらかじめ文書集合をグループ化，すなわちクラスタリングする必要がある．クラスタリングの手法には,「階層型クラスタリング」と「非階層型クラスタリング」の二種類に大きく分類される．階層型クラスタリングでは，クラスタは木構造を構成しており，この構造からクラスタ間の関係を知ることができる．Cutting らの“ Scatter ／ gather[10] は，文献クラスタリングを応用して，利用者の要求に合わせて文献集合を絞り込んでいくシステムである．また，Kohonen ら [37] は，文献集合の構造が自己組織化マップによって可視化され，利用者に提示することを提案している．非階層型クラスタリング [4, 49] では，文書をある特定の数のクラスタに分類する．このとき，クラスタ間の関係については考慮しない．それゆえ，階層型クラスタリングに比べてクラスタリング処理が容易であるという利点がある．

クラスタモデルは検索の効率性を改善するだけでなく，検索の有効性も改善できる可能性を持っている．クラスタモデルは,「文書が類似していれば，同じ検索質問に対する適合性も同様に類似している」というクラスタ仮説を前提にして提

案されたモデルである [65]．一般に，利用者が入力する問合せ語の数は 2, 3 語程度とあまり多くないため，問合せ語の表現の違いによって文書中の索引語と問合せ語が一致しない可能性も生じる．そこでクラスタ仮説では，あらかじめ類似した文書をグループ化しておくことにより，このような不一致の可能性を減少させることを狙っている．このように，クラスタモデルは検索の効率性と有効性を改善する可能性を持っているが，これまで実際に使われることは少なかった．その主な理由は，クラスタリングのための計算量が大きいため，大規模な文書集合に対してクラスタリングができなかったからである．最近では，ハードウェアの進歩やアルゴリズムの改善により，ある程度の規模の文書集合であれば現実的な時間でクラスタリングを行うことができるようになってきている [29]．

本研究では，言語横断情報検索の問合せ翻訳における訳語の曖昧性解消に Web ディレクトリを言語資源として用いることを提案している．Web ディレクトリのカテゴリは階層構造を構成していることから，提案手法はクラスタモデルによる検索の利点を享受できる可能性がある．また，Web ディレクトリでは Web 文書は事前にカテゴリに分類されているため，クラスタモデルによる検索で問題となったクラスタリングの処理時間については考慮する必要が無い．それゆえ，Web 文書のように大量に文書が存在するような文書群に対して有効な検索システムを構築することが可能である．

3. 曖昧性を解消した辞書の構築

言語横断情報検索を実現するためには，何らかの方法により言語の壁を越えることが必要となる．それを可能とするために課題となるのが，単語の概念を言語間でどのように結びつけるか，ということである．それができれば，翻訳したい単語を，それと同じ概念として結びつけられた，目的の言語の単語に置き換えることにより言語の壁を乗り越えることができる．一般に，単語の概念を結びつける役割を担っているのが主に対訳辞書である．

2 章で述べたように，言語横断情報検索を問合せを翻訳する方式で行う場合，対訳辞書を用いて翻訳する際に生じる訳語の曖昧性が問題となる．一般に，対訳辞書にはある一つの単語に対する訳語は複数掲載されている．表 3.1 は，ジーニアス英和辞典に掲載されている “retrieve” という英単語に対する日本語の訳語の一覧である．そのため，問い合わせを翻訳するためには，対訳辞書に掲載されている訳語候補のうちから一つ，あるいはいくつかを選択する必要が生じる．このとき，もし翻訳しようとしている単語に対する訳語の候補が全て同じ語義を表しているのであれば，その訳語候補の一つを用いる，あるいは逆に全ての訳語候補を用いることで，言語横断情報検索を行うことができる．しかし一般的には，表 3.1 を見ても明らかなように，一つの単語に対するいくつかの訳語候補の間で微妙に，時には明らかに語義が異なっている．これを「訳語の曖昧性」という．

表 3.1 retrieve に対する訳語

取り戻す	回収する	を回復する
挽回する	救出する	更正させる
埋め合わせる	償う	訂正する
検索する	捜して持ってくる	うまく返す
思い出す	拾い上げる	手に取る
巻き上げる	戻す	

訳語の曖昧性が残ったまま，訳語候補を全て問合せの訳語として用いると，明らかに問合せ語として不要な訳語を含むこととなり，検索結果にも悪影響を与えることとなる．そこで，翻訳前の元の問合せ語が表している語義を推定し，それ

以外の語義を表している訳語候補を排除することにより，訳語の曖昧性解消を行う必要がある．2 章で触れたように，訳語の曖昧性解消に関する研究が行われ，現在ではコーパスを用いて曖昧性解消を行う手法が主流である．コーパスから共起頻度などの統計情報を取得し，これらを利用して曖昧性解消を行う．

コーパスを用いて曖昧性解消を行う手法では，Hull[28] および奥村ら [51] により，検索要求とコーパス間のドメインの相違による検索精度への影響が指摘されている．よって，この手法が十分に機能するためには，問合せが要求すると予想される様々なドメインを網羅できるように，数種類のコーパスを組合わせて用いる必要がある．しかし，コーパスを入手するコストが高く，数種類用意することは容易ではない．また，本研究で対象としている Web 文書の検索では，検索対象となるドメインは広範囲にわたるため，全てのドメインを網羅すること自体が困難である．

そこでまず取り組んだのは，事前に曖昧性解消されている対訳辞書を構築することである．対訳辞書自体が事前に曖昧性解消をされていれば，検索時に曖昧性解消を行う必要は無く，上記のようなコーパスの対象範囲についても考慮する必要は無くなる．このような辞書が構築できれば，それを利用するだけで訳語の曖昧性解消を行うことが出来る．

事前に曖昧性解消されている対訳辞書を構築するためには，通常の対訳辞書のように，ある単語の訳語が全て列挙されているのではなく，異なる語義を表す訳語はそれぞれ別の見出しでまとめられている必要がある．なぜならば，ある一つの見出しを選択した段階で一つの語義に絞り込めていなければ，辞書を利用しただけで訳語の曖昧性解消が行えていることにはならないからである．対訳辞書のように，言語間の対応が単語単位で結びつけられているだけでは不十分であり，より詳細なレベル，すなわち単語の概念単位で対応が結びつけられている必要がある．異なる語義を表す訳語はそれぞれ別の見出しでまとめられている言語資源として，オントロジがある．オントロジは一つの言語における語彙間の関連をまとめた辞書であるため，そのままでは問合せの翻訳に用いることはできない．しかし，言語を越えて単語の概念の関係の表現を，オントロジが表現することができれば言語の壁を乗り越えることが可能となる．そこで本章において，オントロ

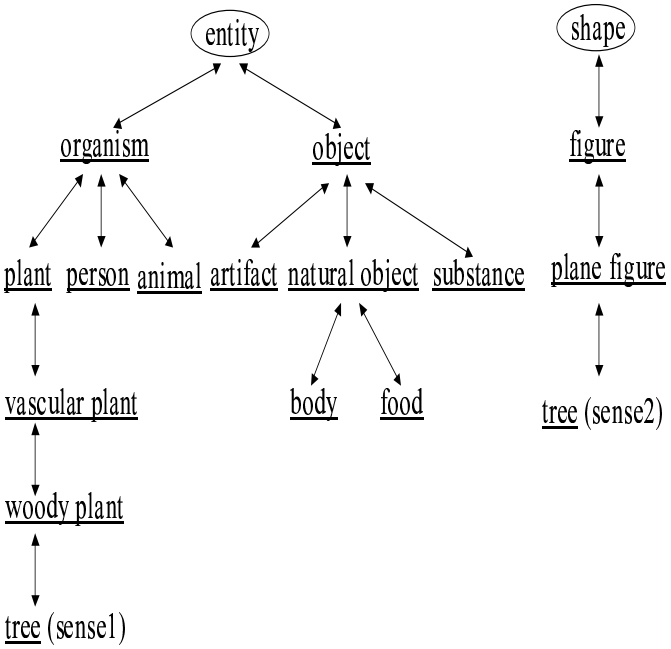

図 3.1 名詞 “tree” の階層構造

ジを事前に翻訳することにより二ヶ国語に対応できるオントロジ（これを「2 言語オントロジ」と呼ぶ）を構築し，この 2 言語オントロジを用いて問合せの翻訳を行うことで，訳語の曖昧性解消を行うことを試みた．

3.1 オントロジ

オントロジはもともと「存在論」を意味する哲学用語であったが，人工知能などの研究においては,「対象とする世界に存在するものごとを体系的に分類し，その関係を記述するもの」として取り込まれた [45]．オントロジには，語と語の関係を 「上位/下位関係」,「部分/全体関係」,「類似関係」,「反意関係」などのさまざまな関係で分類，整理した知識である [62]．図 3.1 は，名詞 “tree” の階層構造を表現している．

同義関係を中心に語と語の関係を分類・整理した辞書のことを,「シソーラス」

とも言う．英語ではRogetのシソーラス [7]，日本語では国立国語研究所の分類語彙表 [47]，角川類語新辞典 [50] などが代表的である．これらのシソーラスはいずれも木構造で表現されている．また，これらのシソーラスは人間の利用者が使用することを想定しており，いずれも書籍として出版されている．

Princeton 大学において開発されたシソーラスである WordNet[44] も，上記のシソーラスと同様に，同義語の集合によって語の意味を表現しようとしている．WordNet が上記のシソーラスと異なっているのは，人間の利用者が使用することを想定しているが最初から電子化されていることである．これは，利用にあたってコンピュータの支援を前提にWordNet が構築されているためである．図 3.2 は，“bank” という単語が持っている名詞の概念を WordNet.1.7.1 により抽出した結果である．

同位語を HTML 文書から自動的に発見する手法も提案されている [60]．HTML の構造に着目し，同レベルに列挙されているような語を同位語である可能性がある語として抽出する．これらの中でも，相互情報量や共起頻度が高いものが同位語である可能性が高いことを利用して，同位語の取得を行っている．また彼らは，同位語の場合と同様にHTML の構造に着目することにより，上位語や下位語を取得する研究も行っている [61]．Herst[26] は，“such as” のような上位語と下位語が現れるようなパターンを発見することで，上位語と下位語を取得する手法を提案した．Sanderson ら [58] は，概念階層の抽出を行う手法を提案した．これは，二つの単語の出現の仕方により，それらの間の包含関係を発見する手法である．また，Glover ら [19] は，ある単語に対して親の概念を表す語，自分自身を指す語，子の概念を表す語を取得する手法を提案した．

WordNet のようにコンピュータの支援が前提ということは，逆にシソーラスやオントロジをコンピュータが活用できることを意味している．オントロジを利用することにより，機械が言葉の意味を理解することが可能となり，文章の表層的な解析だけではできない高度な処理を行うことも可能となる．コンピュータがオントロジを利用することができるようになったことにより，オントロジの工学的研究が行われるようになった [6, 5, 34]．自然言語処理や人工知能の分野などにおいて，オントロジを活用する研究が行われている．実際に，ある特定の分野に

```
The noun bank has 10 senses (first 9 from tagged texts)

1. (883) depository financial institution, bank, banking concern, bank
ing company -- (a financial institution that accepts deposits and chan
nels the money into lending activities; "he cashed a check at the bank
"; "that bank holds the mortgage on my home")
2. (99) bank -- (sloping land (especially the slope beside a body of w
ater); "they pulled the canoe up on the bank"; "he sat on the bank of
the river and watched the currents")
3. (76) bank -- (a supply or stock held in reserve for future use (esp
ecially in emergencies))
4. (54) bank, bank building -- (a building in which commercial banking
 is transacted; "the bank is on the corner of Nassau and Witherspoon")
5. (7) bank -- (an arrangement of similar objects in a row or in tiers
; "he operated a bank of switches")
6. (6) savings bank, coin bank, money box, bank -- (a container (usual
ly with a slot in the top) for keeping money at home; "the coin bank w
as empty")
7. (3) bank -- (a long ridge or pile; "a huge bank of earth")
8. (1) bank -- (the funds held by a gambling house or the dealer in so
me gambling games; "he tried to break the bank at Monte Carlo")
9. (1) bank, cant, camber -- (a slope in the turn of a road or track;
the outside is higher than the inside in order to reduce the effects o
f centrifugal force)
10. bank -- (a flight maneuver; aircraft tips laterally about its long
itudinal axis (especially in turning); "the plane went into a steep ba
nk")
```

図 3.2 WordNet.1.7.1 による単語 “bank” の概念の抽出結果

対するオントロジを構築し，設計支援に利用することなども行われている [63].

電子化されたオントロジの活用した例として，近年では Web 文書の内容を機械が理解できるようにする試みがなされている．その試みとは，Tim Berners-Lee が提唱したセマンティック Web[3] である．セマンティック Web においてもオントロジは，機械が言葉の意味を理解するための知識を与える役割を期待されている．セマンティック Web におけるオントロジの記述言語として，W3C(World Wide Web Consortium) により，OWL(Web Ontology Language) が開発されている [2]. OWL ではある一つの概念はクラスにより記述されるが，他のクラスとの関係が記述できるように設計されており，これを推論ルールとして利用することが可能である．

OWL の形式により記述されたオントロジのうちで，公開されていて利用可能なものに，WordNet.OWL[35] がある．WordNet.OWL は，英語のシソーラスである WordNet1.7.1[68] をもとに OWL により記述することで作られた英語のオントロジである．図 3.3 に，それぞれの概念が WordNet.OWL に記述されている様子を示す．現在のところ，公開されている日本語のオントロジはないが，神崎は日本語の単語を対訳辞書を用いて翻訳し WordNet に対応付けることにより，利用可能な日本語のオントロジを構築する試みを行っている [71]．しかし，この手法では対訳辞書により得られる訳語を全て対応付けているため，概念の異なる単語が同じ概念として扱われてしまう可能性がある．そこで本章では，この問題を解決したオントロジの翻訳手法を提案し，2 言語オントロジの構築を目指す．こうして構築された 2 言語オントロジを言語横断情報検索に利用することにより，検索時に訳語の曖昧性解消を行う必要の無いシステムを構築することを試みる．

3.2 Web ディレクトリを利用した訳語の曖昧性解消

神崎の提案した手法では，単語を対訳辞書により翻訳しその訳語を元の単語と対応付けることにより，オントロジを構築している．対訳辞書を用いる場合，一般に訳語候補は複数存在する．すなわち，“訳語の曖昧性” のことである．これらの表す概念が全て類似している場合は良いが，多義語などのようにそれぞれの訳語候補の表す概念が類似しないこともある．例えば英語の「bank」という単語の

```
<rdf:Description rdf:about="&wn;111467633">
  <wn:wordForm rdf:resource="&wn;love" />
</rdf:Description>
<rdf:Description rdf:about="&wn;201251337">
  <wn:wordForm rdf:resource="&wn;wrap" />
  <wn:wordForm rdf:resource="&wn;envelop" />
  <wn:wordForm rdf:resource="&wn;enwrap" />
  <wn:wordForm rdf:resource="&wn;enfold" />
  <wn:wordForm rdf:resource="&wn;enclose" />
</rdf:Description>
<rdf:Description rdf:about="&wn;400234973">
  <wn:wordForm rdf:resource="&wn;seriatim" />
</rdf:Description>
<rdf:Description rdf:about="&wn;105412780">
  <wn:wordForm rdf:resource="&wn;American_Standard_Version" />
  <wn:wordForm rdf:resource="&wn;American_Revised_Version" />
</rdf:Description>
<rdf:Description rdf:about="&wn;302780094">
  <wn:wordForm rdf:resource="&wn;cervical" />
</rdf:Description>
<rdf:Description rdf:about="&wn;101858287">
  <wn:wordForm rdf:resource="&wn;horse_botfly" />
  <wn:wordForm rdf:resource="&wn;Gasterophilus_intestinalis" />
</rdf:Description>
```

図 3.3 WordNet.OWL(抜粋)

訳語候補として「銀行」,「貯蔵所」,「岸」,「堆積」,「土手」,「堤」,「漕ぎ手」など，類似しているとはいえない単語が対訳辞書には登録されている．神崎の手法では，訳語候補を全て元の単語に対応付けている．この場合，表現する概念が異なっているにもかかわらず，それぞれが類似しているとみなしてしまう可能性がある．

この問題を解決するには，言語横断情報検索の場合と同様に，訳語の曖昧性解消を行うことが必要である．言語横断情報検索においては，コーパスを利用することで訳語の曖昧性解消を行うことが多いことは，2 章において述べた．しかし，本論文が対象とする言語横断情報検索の技術を利用した Web 文書検索のようなアプリケーションを想定した場合，さまざまな分野のコーパスを複数個用意することは現実的ではない．そのため本論文では，Yahoo! カテゴリのような複数の言語で作成された Web ディレクトリに登録されている文書群をコーパスとして用いる．Web ディレクトリには多種多様な分野の Web 文書が登録されているため，Web ディレクトリをコーパスとして用いることは，現存するほとんどの分野に対応したコーパスを利用することに等しいと言える．また，Web ディレクトリによって問合せが対象としている分野を限定することができるので，問合せが対象とする分野と一致したコーパスを利用でき，訳語の曖昧性解消を効果的に行うことができる．

図 3.4 は，本手法の概要を示している．本手法は，英語のオントロジに登録されている単語を対訳辞書を用いて日本語に翻訳することにより，英語版と同じ構造の日本語のオントロジを作成する．ただし単語の翻訳は，後述する Web ディレクトリを利用した曖昧性解消の手法を用いて行う．また，そのときに翻訳元となった概念の番号を対応付けておくことで，英語と日本語を対象とした 2 言語オントロジを構築する．

Web ディレクトリを利用した曖昧性解消を行うには，例えば Yahoo! カテゴリのように複数の言語版がある Web ディレクトリを利用する．本研究では英語から日本語への翻訳なので，Web ディレクトリも英語版と日本語版を用いる．まず，前処理として次の 2 つの処理をおこなう．

- カテゴリごとに属する Web 文書から特徴語を抽出する．

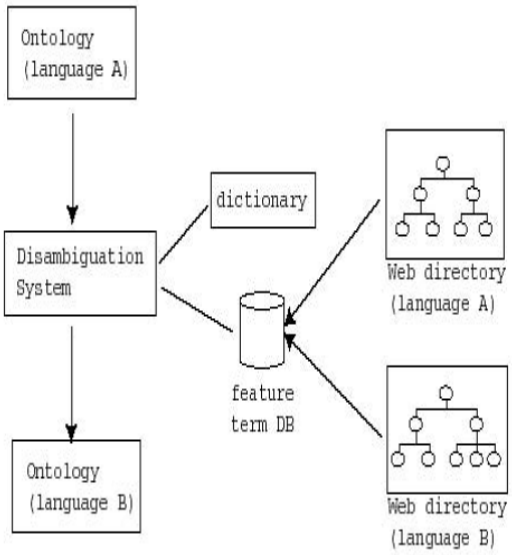

図 3.4 2 言語オントロジ構築手法の概要

- それぞれのカテゴリにおいて，言語間でのカテゴリの対応付けを行う．

これらの前処理が完了すると，訳語の曖昧性解消を行うことができる．
2 言語オントロジ構築における曖昧性解消の流れを図 3.5 に示す．

1. 翻訳したい単語が英語の Web ディレクトリのどのカテゴリに適合するか推定する．

2. 推定された英語のカテゴリに対応付けられている日本語のカテゴリを選択する．

3. 翻訳したい単語の訳語候補を，対訳辞書をから全て抽出する．

4. 抽出されたそれぞれの訳語候補が，選択した日本語のカテゴリの特徴語として存在するか調べ，存在していたものを訳語として用いる．

1. において，一つの単語だけから適合するカテゴリを推定することは困難であるため，複数の単語を用いる必要がある．WordNet.OWL ではある一つの概念に

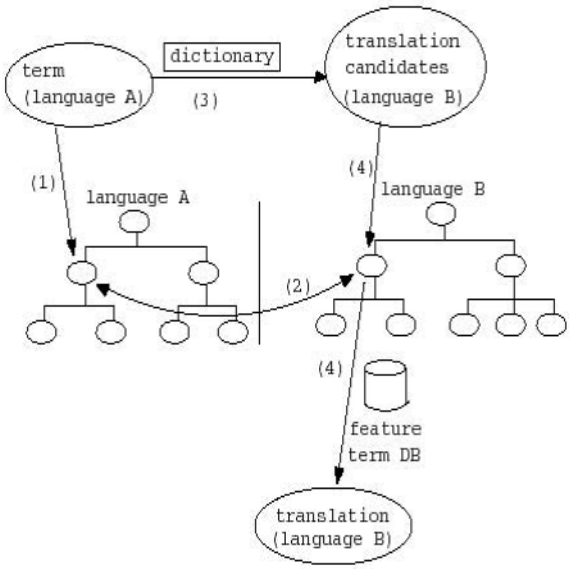

図 3.5 2 言語オントロジ構築における曖昧性解消の流れ

複数の単語が登録されていることも多い．そこで，一つの概念に登録されている全ての単語を用いることにより，適合するカテゴリの推定を行う．

Web ディレクトリの同じカテゴリには類似した分野を主題とした Web 文書が属しているため，訳語候補のうちで同じカテゴリに属しているものは，類似した概念を表している可能性が高いと考えられる．よって，上記手法により類似した概念を表す単語を抽出することが可能である．

3.3 2 言語オントロジの構築実験

提案手法の有効性を検証するため，ある単語の訳語の中から類似する概念の訳語を抽出する予備実験を行った．今回の実験では，英語の “bank” という単語の訳語の中から “銀行” という日本語の単語の概念に類似するものを抽出した．

今回の実験において，Web ディレクトリは Yahoo! カテゴリの英語版と日本語版を用いた．尚，下位の階層のカテゴリを上位のカテゴリに統合することにより，

表 3.2 英語版および日本語版 Yahoo! カテゴリにおける最上位の 13 カテゴリ

英語	日本語
Arts & Humanities	芸術と人文
Business & Economy	ビジネスと経済
Computers & Internet	コンピュータとインターネット
Education	教育
Entertainment	エンターテインメント
Government	政治
Health	健康と医学
News & Media	メディアとニュース
Recreation & Sports	趣味とスポーツ
Reference	各種資料と情報源
Science	自然科学と技術
Social Science	社会科学
Society & Culture	生活と文化

それぞれの言語版の最上位のインデックスページに登録されている 13 カテゴリに統合した．ただし，英語版の “Regional” および日本語版の “地域情報” のカテゴリは除いている．表 3.2 は英語版および日本語版 Yahoo! カテゴリにおける最上位の 13 カテゴリのリストである．また，最上位のインデックスページに登録されているカテゴリはどの言語版においても同じ構造であるため，その対応をそのまま言語間におけるカテゴリの対応付けに用いた．

Web 文書から単語を抽出する際に，英語版では単語の活用形を原形にしたのち，ストップワードを取り除いた．日本語版では，“茶筌” を用いて名詞，動詞，形容詞，未知語を抽出した．特徴語の翻訳には，EDR 電子化辞書の対訳辞書を用いた．また，各カテゴリの特徴語数は，特徴語の重みの上位 5,000 語とした．

本手法ではまず，抽出したい概念がどのカテゴリに適合しているかを推定する．今回の実験では，上述の抽出したい概念が，英語版のカテゴリ “Business and

Economy” に適合するものと仮定した．よって，このカテゴリに対応する日本語のカテゴリ “ビジネスと経済” が選択される．

次に，対訳辞書から英語の “bank” という単語の訳語候補を抽出する．“bank” という単語の訳語候補は以下のとおりであった．

峰, 灰をかぶせる, 銀行, 貯蔵所, 傾斜する, バンクさせる, 漕ぎ手席, 銀行に預ける, 副見出し, 親元, 岸, 外側を高くさせる, 傾いて飛行する, 堤で囲む, 積み上げる, ゴム縁, 袖見出し, 堆積, 湯銭, 土手, 堤, 配列する, 坑口, 銀行業を営む, 銀行へ預金する, 横に傾ける, 胴元の用意金, 漕ぎ手, バンク, 〜バンク, 列に並べる, 土手を築く, 左右に傾ける, 堆積する, 片勾配, 堆, 縦坑口, テーブル, カウンター, 小見出し, 胴元になる, 浅瀬, 層をなす, クッション, 縦坑の入口付近, 貯金箱, 堤防, 堤のようになる, 傾斜させる, 横に傾く, 古代船のオールの列, 胴元, バンクする, 横傾斜, 隆起する, 積み重なる, タイプライターのキーの列, 砂州, 親元になる

そして，上記の訳語候補が日本語のカテゴリ”ビジネスと経済”の特徴語として含まれているかを調べ，含まれていた訳語のみを抽出する．その結果抽出された単語は，“峰”, “銀行”, “バンク”, “カウンター” であった．

3.4 2言語オントロジ構築における考察

3.3 における実験の結果，“銀行”, “バンク” のように日本語のカテゴリ”ビジネスと経済”に関連する単語が抽出されているおり，全ての訳語候補のうちから目的の概念を表した単語にある程度絞り込むことができた．しかし，同時にあまり関連の無いと思われる，“峰”, “カウンター” といった単語も抽出されている．前者 2 単語と後者 2 単語では明らかに異なった概念の単語であり，これらを同じ概念の単語とする結果となったことは，2 言語オントロジの構築において十分に語彙の曖昧性解消を行えていない．このままでは，言語横断情報検索に 2 言語オントロジを利用しても，訳語の曖昧性解消は十分には行うことが出来ない．

今回の実験で目的以外の概念を表した単語も抽出された原因は，カテゴリの統合によりカテゴリ数が少なくなったため，一つのカテゴリが対象とする分野の範

囲が大きくなったからであると考えられる．よって，一つのカテゴリが対象とする分野の範囲をより小さく限定することができれば，語彙の曖昧性解消の精度が向上する可能性がある．また，茶筌では，“銀行に預ける” という語は “銀行” と “預ける” という二語に分割されるため，“銀行に預ける” という特徴語は存在しない．よって，こういった複数の単語に分割される訳語は抽出されないということも問題である．さらに，WordNet.OWL では類似の概念であっても品詞が異なる場合は別の概念としているため，名詞だけを用いることで抽出精度が向上する可能性がある．

しかしながら，2 言語オントロジの構築において語彙の曖昧性解消の精度を大幅に向上することは非常に困難であると思われる．なぜならば，異なる言語間においてシソーラスやオントロジの構造が明らかに異なるからである．ある言語では存在する単語が，別の言語ではその単語に相当する単語が存在しないことも少なくない．場合によっては概念自体が，文化的な相違により一方では存在するが，他方では存在しないこともある．一方に存在しない概念を結びつけることは，少なくとも単一の概念だけでは不可能である．このような概念に対して，もう一方の言語における複数の概念集合とを結びつけることは可能かもしれない．それができたとしても，複数の概念集合を表現する語彙を決定することは困難である．複数の概念集合を表現するには，それぞれの概念から単語を一つずつ取り出し，それらをつなぎ合わせることになるが，そうして作られた語彙が一般に使用される語彙であるとは限らないからである．

それゆえ，訳語の曖昧性解消のために 2 言語オントロジを構築することを検討するよりも，その際に利用した Web ディレクトリを直接言語横断情報検索システムの構築のために利用するほうが効率がよいと思われる．

4. Web ディレクトリの言語横断情報検索への利用

言語横断情報検索において問合せを翻訳する際に生じる訳語の曖昧性の問題に対して，事前に曖昧性解消がなされた辞書を構築し，それを問合せの翻訳に利用することで解決を図ろうとしたが，そのような辞書を構築することは困難であることは 3.4 において述べた．3.2 において提案した 2 言語オントロジの構築手法において，オントロジを翻訳する際に Web ディレクトリを利用することを提案した．2 言語オントロジの構築は達成できなかったが，3.3 の実験の結果，Web ディレクトリを訳語の曖昧性解消に活用できる可能性があることが示された．

それならば，問合せの翻訳における曖昧性解消のために Web ディレクトリそのものを活用することも考えられる．すなわち，Web ディレクトリをコーパスの代わりとして言語資源として活用することである．従来のコーパスを利用した訳語の曖昧性解消の手法では，問合せおよびコーパスが対象としている分野の範囲が一致している必要があるが，Web 文書の検索のように問合せが対象とすると想定される分野の範囲が広範囲である場合，これらの全てに対応するには数種類のコーパスを用意する必要がある．しかし数種類ものコーパスを入手することは容易ではない．そもそも，Web 文書が対象としている分野の全てに対してそれぞれにコーパスが存在すること自体非現実的である．それに対して，Web ディレクトリをコーパスの代わりに言語資源として活用することは，従来のコーパスを利用した訳語の曖昧性解消の手法におけるこの問題を解消することが可能である．

Web 文書を分類し整理するために構築されたのが Web ディレクトリである．言い換えれば，Web 文書は Web ディレクトリのいずれかのカテゴリに分類される．つまり，Web ディレクトリを言語資源として活用することは，Web 文書が対象とする全ての分野を網羅した言語資源を利用していることとほぼ同義である．それゆえ，従来のコーパスを利用した訳語の曖昧性解消の手法における，問合せとコーパスが対象としている分野の範囲の一致の問題を考慮する必要が無い．そこで本章では，言語横断情報検索に Web ディレクトリを言語資源として活用する手法について研究を行った．

4.1 提案する手法

Web ディレクトリには，収集された Web 文書のリンクが，それぞれの文書の内容の分野ごとに分類されている．分類された Web 文書のリンクは，分野ごとに “カテゴリ” に格納される．同一のカテゴリの中には，同一の分野を対象とした Web 文書のリンクが集まっている．カテゴリは階層構造になっており，カテゴリの分野の上下関係や包含関係が表現されている．それにより，Web ディレクトリの階層構造を上に行けばカテゴリの分野は一般化され，逆に階層構造を下に行けばカテゴリの分野は詳細化されることになる．

提案手法では，それぞれのカテゴリに登録されている Web 文書をカテゴリごとに収集し，その Web 文書から得られる統計情報を抽出する．こうして得られた統計情報を利用することにより，言語横断情報検索における訳語の曖昧性解消を行う．このとき，統計情報はカテゴリごとに抽出する点が，本手法の特徴である．Web ディレクトリの全てのカテゴリを対象とすることで Web 文書が対象とする分野を網羅することができるが，カテゴリごとに統計情報を抽出しなければ分野ごとの統計情報の偏りは平均化され，特定の分野においてよく使用されるような単語は軽視されることになる．しかし，問合せが対象とする分野が特定される場合においては，そのような単語のほうが重要である可能性が高い．そこで本手法では，問合せが対象とする分野を推定し，その推定されたカテゴリから抽出された統計情報を利用することにより，訳語の曖昧性解消を行う．これにより，問合せが対象としている分野においてより適切な訳語を選択することが可能となる．

Lee ら [39] は，検索対象を分野ごとにクラスタリングすることにより言語横断情報検索の問合せ翻訳における訳語の曖昧性の解消の精度が向上すると報告されている．提案手法は，検索対象の分野を限定するのではなく，曖昧性解消に用いる言語資源を問合せが対象とする分野に絞り込むという違いはあるものの，Lee らの報告からもわかるように，対象分野を特定することにより，訳語の曖昧性解消の向上が見込まれる．

言語横断情報検索において問合せを翻訳する際に曖昧性解消を行うには，翻訳後の言語，すなわち検索対象の言語の統計情報が必要である．よって本手法でどのカテゴリの統計情報を用いるか決定する際にも，検索対象の言語でのカテゴリ

を推定する必要がある．しかし，言語横断情報検索では問合せと検索対象で用いられる言語が異なっている．それゆえ，検索対象の言語でのカテゴリを問合せから直接特定するのは困難である．そこで本手法では，二種類の Web ディレクトリを利用する．一つは問合せと同じ言語版の Web ディレクトリ，もう一つは検索対象と同じ言語版の Web ディレクトリである．事前に，この二つの Web ディレクトリのそれぞれのカテゴリに対して，対応するカテゴリをもう一方の言語版の Web ディレクトリの中から決定しておく．問合せが対象とするカテゴリを推定する際には，まず問合せと同じ言語版の Web ディレクトリで対象となるカテゴリを推定する．推定されたカテゴリには，検索対象と同じ言語版における対応するカテゴリが事前に決定されているので，その対応するカテゴリを選択する．こうすることにより，検索対象と同じ言語版における問合せが対象とするカテゴリを推定することができる．こうして推定されたカテゴリの統計情報を利用することで，言語横断情報検索の問合せ翻訳における訳語の曖昧性解消を行う．

図 4.1 は提案手法のシステムの概要を表している．本システムは，問合せおよび検索対象のそれぞれと同じ言語版の Web ディレクトリ，それぞれの言語の特徴語データベース，対訳辞書，検索対象となる文書群から構成される．図 4.1 において点線で囲まれている部分は，問い合わせの翻訳処理の構成を表している．

本システムは，Web ディレクトリの各カテゴリから特徴語を抽出してそれを特徴語データベースに事前に格納しておく前処理と，与えられた問合せを翻訳して検索を行う検索処理の 2 つの処理に分けられる．

4.1.1 前処理

図 4.2 は前処理の流れを示したものである．前処理として事前にそれぞれのカテゴリにおいて，特徴語の抽出と異言語のカテゴリとの対応付けを行う．前処理の手順を以下に示す.

1. 特徴語の抽出
   各言語版の Web ディレクトリの全てのカテゴリに対して

   (a) そのカテゴリに属する Web 文書から単語を抽出し，重み付けを行う．

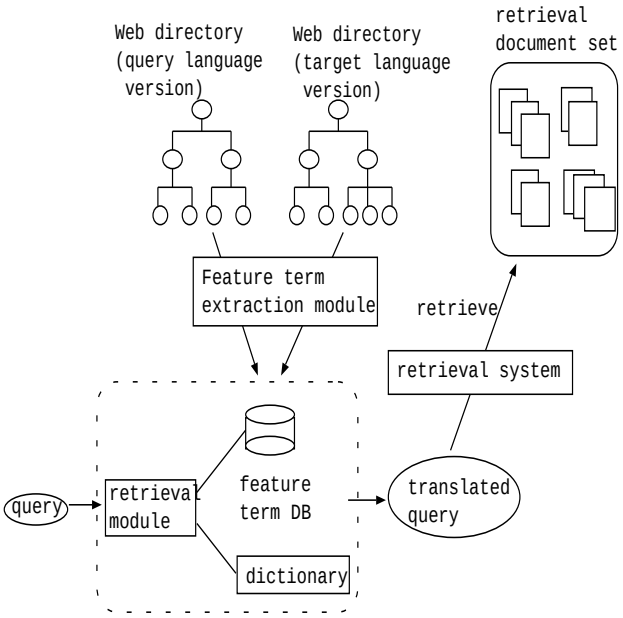

図 4.1 提案システムの概要

(b) 重みの大きい上位 n 語の単語をそのカテゴリの特徴語として抽出する．

(c) 抽出された特徴語を特徴語データベースに格納する．

2. 言語間でのカテゴリの対応付け

全てのカテゴリに対して対応する異言語のカテゴリを推定し，対応付ける．

例えば図 4.2 の query language version のカテゴリ a に対する対応付けでは，まずカテゴリ a に属する文書群から単語を抽出し，それらのカテゴリ a における重みを計算する（1）（a）．次に，抽出された単語のうちから重みの大きいものから n 語を特徴語として抽出し，特徴語集合 f_a を得る（1）（b）．こうして得られた特徴語集合 f_a を特徴語データベースに格納する（1）（c）．得られた特徴語集合 f_a に最も類似していると思われる target language version のカテゴリを探し，カテゴリ a とそのカテゴリに対して対応付けを行う（2）．なお，対応付けの方法はどのような方法によって行ってもよい．例えば，カテゴリの特徴語を比較することにより対応付けを行うことが考えられる．また，人手により直接対応付けを行

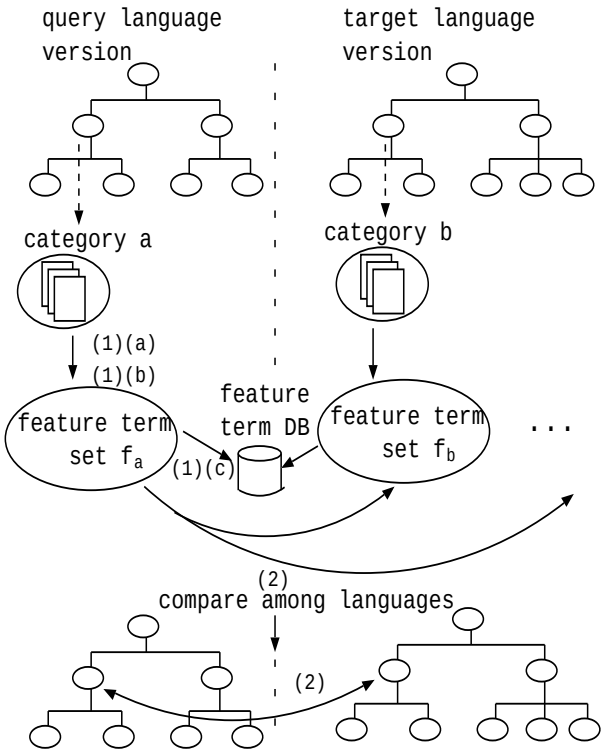

図 4.2 前処理における処理の流れ

うことも考えられる．こうして得られたカテゴリの対応は，文書の検索を行うときに利用する．

特徴語の抽出　前処理においてまず行われるのが，カテゴリの特徴語の抽出である．各カテゴリは特徴語集合によりその特徴を表現される．特徴語集合は，そのカテゴリの特徴を表現していると思われる単語の集合である．特徴語を抽出するために，まず各カテゴリに属する Web 文書から単語を抽出する．次に，抽出された単語をカテゴリごとに集計し，その単語がカテゴリの内容を表現する程度を表す重みを計算する．抽出された単語のうち，重みが大きいものをそのカテゴリの特徴語として抽出する．

Web 文書から抽出された単語の重みは，TF·ICF（term frequency · inverse cat-

egory frequency）[36] により計算する．これは．一般に良く知られた単語の重み付けの手法の 1 つである TF·IDF（term frequency · inverse document frequency）[57] を発展させたものである．TF·IDF は単語の出現頻度（TF）と文書頻度の逆数 (IDF) との積により求められる．TF は単語の網羅性を表し，IDF は単語の特定性を表しており，これらの積である TF·IDF は網羅性と特定性がともに高い単語の重みが大きくなるようになっている．TF·IDF は，次の式で求められる．

$$tf \cdot idf(t_i, d) = \frac{f(t_i)}{N_d} \cdot \log(\frac{N}{n_d} + 1)$$

ここで，$f(t_i)$ は特徴語 t_i の出現頻度，N_d は文書 d の全単語数，N は全文書数，n_d は特徴語 t_i が出現する文書数を表す．

TF·IDF では文書を単位として重みを計算するが，文書のかわりにカテゴリを単位として重みを計算するのが TF·ICF である．本手法では，カテゴリごとに特徴語の抽出を行うが，各カテゴリには同じ分野について記述された Web 文書が分類されており，Web 文書同士の関連が高い．そのため，ページ単位の重みづけである TF·IDF を用いるより，カテゴリ単位に重みの計算を行う TF·ICF を用いるほうが適切であると考える．TF·ICF により重みを計算することで，文書単位で計算する TF·IDF より，カテゴリの内容をより反映した重み付けを行うことができる．TF·IDF は，次の式で求められる．

$$tf \cdot icf(t_i, c_j) = \frac{f(t_i, c_j)}{N_{c_j}} \cdot \log(\frac{N_c}{n_c} + 1)$$

ここで，$f(t_i)$ は特徴語 t_i の出現頻度，N_{c_j} はカテゴリ c_j に属する全ての文書における全単語数，N_c は全カテゴリ数，n_c は特徴語 t_i が出現するカテゴリ数を表す．

表 4.1 は，英語版 Yahoo! カテゴリにおけるトップレベルカテゴリ “Government” の特徴語の上位 10 語を表している．表 4.1 の単語は他の 12 カテゴリの特徴語にも含まれているが，カテゴリごとで重みが異なっているため，同じ特徴語でもカテゴリによって重要度も異なっている．

4.1.2 検索処理

本システムにおける検索の流れを図 4.3 に示す．まず，問合せの適合カテゴリを選択し，続いて適合カテゴリに対応付けられている異言語のカテゴリを選択し，

表 4.1 英語版 Yahoo! カテゴリにおけるカテゴリ “Government” の特徴語（上位 10 語）

特徴語	重み
law	0.001908
font	0.001560
court	0.001381
var	0.001233
information	0.001142
document	0.001136
war	0.001124
px	0.001014
time	0.000958
government	0.000938

そのカテゴリの特徴語集合を利用して問合せの翻訳を行い，最後に翻訳された問合せを用いて文書群に対して検索が行われる．検索における処理の手順は次のようになる．

1. 問合せと同じ言語版の全てのカテゴリに対して
   問合せとカテゴリの特徴語集合との適合度を求める．
2. 最も適合度の高いカテゴリを問合せの適合カテゴリと決定する．
3. 検索対象の言語版のカテゴリから，適合カテゴリに対応付けられているカテゴリを選択する．
4. 選択された対応カテゴリの特徴語集合を利用して問合せを翻訳する．
5. 翻訳された問合せにより，検索対象の文書群を検索する．

問合せの適合カテゴリの選択　検索処理において最初に行われるのは，問合せの適合カテゴリの選択である．本システムにおける問合せは文章ではなく，数語の単語から構成されていることを前提としている．ここで，$t_1, t_2, \cdots, t_n$ の単語から構成される問合せ q に対する問合せベクトル $\vec{q}$ を次のように定義する．

$$\vec{q} = (q_1, q_2, \ldots, q_n)$$

なお，q_k は問合せの k 番目の単語 t_k に対応しており，その値は 1 である．

与えられた問合せについて，まず同言語間において，問合せと各カテゴリとの適合度を計算し（図 4.3（1）），そのうちから最も適合度が高くなるカテゴリを，問合せが適合する同言語のカテゴリと決定する（図 4.3（2））．問合せとカテゴリの適合度は，問合せベクトルとカテゴリの特徴語集合のベクトルの内積にこの 2 ベクトルのコサイン距離を掛けることにより計算する．ここで，カテゴリ c の特徴語集合のベクトル $\vec{c}$ を，次のように定義する．

$$\vec{c} = (w_1, w_2, \ldots, w_n)$$

なお，w_k は，単語 t_k のカテゴリ c における特徴語の重みを表す．

問合せとカテゴリの適合度 $rel(q, c)$ は次のように求められる．

$$rel(q, c) = (\vec{q} \cdot \vec{c}) \frac{\vec{q} \cdot \vec{c}}{|\vec{q}| \cdot |\vec{c}|}$$

二つのベクトルの内積のみから適合度を求めると，次のような場合に問題が生じる．幾つかある問合せ語のうちの一つだけしか特徴語集合に存在していないが，その重みが大きいという場合である．この場合，それ以外の問合せ語がその特徴語集合に存在していなくても，その存在している問合せ語の重みが大きいため，適合度の値が大きくなることがある．特徴語集合に存在している問合せ語の数が多く，かつそれらの重みが高い，という二つの要求の両方を満たしている度合いが高いほど，$rel(q, c)$ の値も高くなるのが理想である．しかし，一つの問合せ語の重みのみが大きい場合では，上記の要求の前者を満たしていない．そこでベクトルのコサイン距離を掛けることで，特徴語集合に存在している問合せ語の数を考慮する．適合度の算出式については 4.1.3 で詳細に述べる．

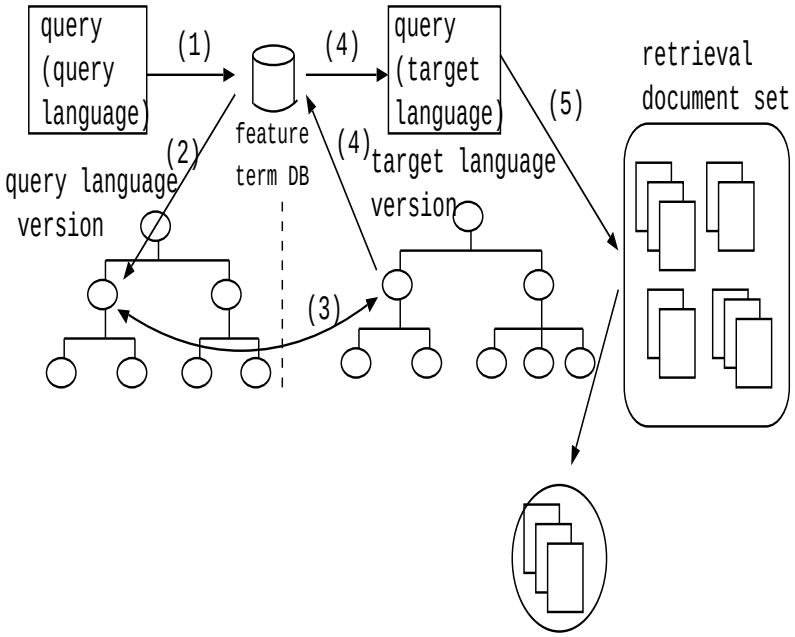

図 4.3 検索の流れ

こうして求めた適合度が最も高いカテゴリを，問合せに対する適合カテゴリとする．本論文では適合カテゴリを一つだけ選択したが，適合度が閾値以上となるカテゴリを適合カテゴリとする方法も考えられる．このとき適合度が閾値以上となるカテゴリが複数ある場合は，これらを全て適合カテゴリとして選択する．

次に，前処理で得られた対応付けからそのカテゴリに対応する異言語のカテゴリが決まる（図 4.3（3））．こうして得られた異言語のカテゴリの特徴語集合および対訳辞書を利用して，次の問合せの翻訳の項で述べる方法により問合せを翻訳する（図 4.3（4））．こうして得られた問合せを用いて検索対象文書に対して検索を行う．以上の処理を経て得られた文書群が検索結果となる（図 4.3（5））．

問合せの翻訳　検索処理において次に行われるのは，問合せの翻訳である．問合せの翻訳の流れを図 4.4 に示す．まず，問合せ中の各単語 q に対する対訳辞書の全ての訳語 $t_1, t_2, t_3, \cdots$を，訳語の候補として抽出する．抽出された全ての訳語候補について，適合カテゴリに対応付けられている異言語のカテゴリ（以下: “対応

カテゴリ”）bの特徴語に含まれているかを調べる．適合カテゴリの決定およびその対応カテゴリの決定方法については，4.1.2 において述べた方法で行う．含まれていた訳語のうち，特徴語の重みが最も大きい訳語を，その問合せ語の訳語と決定する．このとき，対応カテゴリの特徴語集合の中にいずれの訳語候補も存在しない場合，その問合せ語は使用しない．しかし，例えば，日本語で書かれた Web 文書中において英単語が使われるといったことも頻繁にあるため，翻訳を行わないほうが良い場合もある．そこで，いずれの訳語候補も比較している対応カテゴリの特徴語に含まれていない場合，翻訳する前の問合せの単語そのものが，比較している対応カテゴリの特徴語に含まれているかを調べる (図 4.4 点線)．もし含まれていれば，翻訳前の単語そのものをこの問合せ語の訳語とみなす．

例えば，英語のカテゴリ “Computers and Internet” が問合せの適合カテゴリであるときに英語の “system” という単語の訳語を決定する場合を考える．“system” の訳語の候補として，“宇宙”，“方法”，“組織”，“器官”，“システム”，“系統”，…

などが得られる．この訳語の候補の全てに対して，適合カテゴリの対応カテゴリである日本語のカテゴリ “コンピュータとインターネット” の特徴語集合に存在するどうかを調べる．そのうち重みが最も高いもの，今回は “システム” を，英単語 “system” の訳語と決定する．もし，“system” のいずれの訳語候補も対応カテゴリの特徴語集合に存在しない場合は，“system” という単語そのものが対応カテゴリの特徴語集合に存在するか調べ，存在していれば “system” という単語そのものを訳語とみなす．

文書の検索　検索処理の最後に行うのは，文書の検索である．提案した問合せの翻訳手法により翻訳された問合せを用いて検索対象文書群に対して検索を行う．検索対象文書群は，必ずしも Web ディレクトリに登録されている文書でなくてもよい．検索システムは既存のシステムを使用することができる．こうして得られた文書群が，問合せに対する検索結果となる．

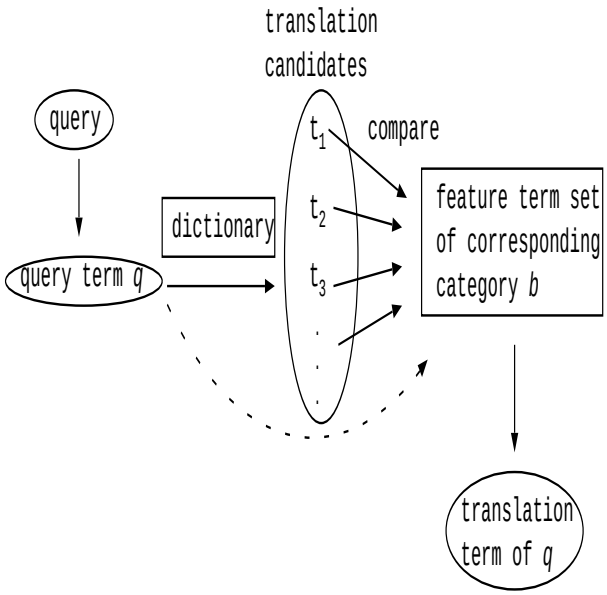

図 4.4 問合せの訳語の決定

4.1.3 問合せの適合カテゴリの選択の実験

4.1 で述べた問合せとカテゴリの適合度の計算方法が有効であるかどうかについて，本節において検証する．

本実験では，2003 年に収集した，英語版 Yahoo! カテゴリにおけるカテゴリ「Computers and Internet」以下の 559 カテゴリに対して行った．これらのカテゴリを，4.1.4 で述べる方法により統合し，「Computers and Internet」から 3 階層目までに統合した．その結果，統合後のカテゴリは 342 カテゴリとなった．

本実験における問合せは，“encryption”(=q_1), “security”(=q_2), and “system”(=q_3) の 3 語から構成されている．この問合せに対して，どのカテゴリが適合カテゴリとして選択されるかについて実験を行う．

まず，問合せとカテゴリの適合度を単純に内積をもちいて計算した場合について実験を行った．そのときの問合せとカテゴリの適合度 $rel_{inner}(q, c)$ は次の式で表される．

$$rel_{inner}(q,c) = \vec{q} \cdot \vec{c}$$

表 4.2 は，内積による適合度を用いた場合の適合カテゴリの上位 10 件を表している．表 4.2 に挙げられているカテゴリのほとんどが，ある程度問合せと適合しているように思われる．しかし，この方法では問題もある．表 4.2 の 2 番目や 10 番目に挙げられているカテゴリのように，一つの問合せ語の重みは高いが，残りの 2 つの問合せ語に関しては全く関連していないことである．適合カテゴリは，問合せの全ての単語と関連が大きい事が望ましい．

そこで，多くの問合せ語との関連が高い場合により適合度が高くなるように改良したのが，4.1 で述べた問合せとカテゴリの適合度の計算方法である．問合せとカテゴリの適合度 $rel(q,c)$ の式は，以下のように表される．

$$rel(q,c) = (\vec{q} \cdot \vec{c}) \frac{\vec{q} \cdot \vec{c}}{|\vec{q}| \cdot |\vec{c}|}$$

この式から分かるように，改良版の適合度は，問合せベクトルとカテゴリの特徴語ベクトルとの内積に，両ベクトルのコサイン距離を掛けることで求められる．両ベクトルのコサイン距離は，一つの問合せ語だけの影響が大きい場合よりも，多くの問合せ語からの影響が大きいほうが，値は大きくなる．よって，両ベクトルのコサイン距離を掛け合わせることにより，多くの問合せ語から影響を受ける場合の補正効果が得られる．これにより，より多くの問合せ語と関連の高いカテゴリの適合度を高く評価できる．この式では問合せベクトルとカテゴリの特徴語ベクトルとの内積の 2 乗も計算していることにもなるが，これにより内積部分が強調されることにより，問合せに適合しているカテゴリの適合度をより強調する効果がある．また，他手法よりも比較的単純な計算で求められることは，本式の利点である．

表 4.3 は，提案手法による適合度を用いた場合の適合カテゴリの上位 10 件を表している．表 4.3 の 3 番目のカテゴリのように一つの問合せ語だけに影響を受けている場合もあるが，表 4.2 の場合に比べると，より多くの問合せ語と関連があるカテゴリが上位に上げられている．この結果は，提案手法による問合せとカテゴリの適合度の計算方法が有効であることを示している．

表 4.2 内積による適合度を用いた場合の適合カテゴリ (上位 10 件).

category name	relevance	weight($q_1/q_2/q_3$)
Computers and Internet/Security and Encryption/Challenges/	0.1668	0.1126/0.0542/0.0000
Computers and Internet/Security and Encryption/Conferences/	0.1269	0.0000/0.1269/0.0000
Computers and Internet/Security and Encryption/Web Directories/	0.1062	0.0125/0.0937/0.0000
Computers and Internet/Security and Encryption/Organizations/	0.0891	0.0066/0.0765/0.0060
Business and Economy/Business to Business/Computers/Security and Encryption/	0.0873	0.0063/0.0746/0.0062
Computers and Internet/Security and Encryption/Encryption Policy/	0.0862	0.0751/0.0110/0.0000
Computers and Internet/Security and Encryption/Mailing Lists/	0.0753	0.0172/0.0581/0.0000
Computers and Internet/Software/Operating Systems/File Systems/	0.0750	0.0276/0.0249/0.0224
Computers and Internet/Internet/World Wide Web/Security and Encryption/	0.0731	0.0056/0.056/0.0113
Computers and Internet/Software/Operating Systems/Inferno/	0.0709	0.0000/0.0000/0.0709

表 4.3 提案手法による適合度を用いた場合の適合カテゴリ (上位 10 件).

category name	relevance	weight($q_1/q_2/q_3$)
Computers and Internet/Security and Encryption/Challenges/	0.1285	0.1126/0.0542/0.0000
Computers and Internet/Software/Operating Systems/File Systems/	0.0748	0.0276/0.0249/0.0224
Computers and Internet/Security and Encryption/Conferences/	0.0733	0.000/0.1269/0.0000
Computers and Internet/Security and Encryption/Web Directories/	0.0689	0.0125/0.0937/0.0000
Computers and Internet/Security and Encryption/Organizations/	0.0595	0.0066/0.076/0.0060
Business and Economy/Business to Business/Computers/Security and Encryption/	0.0585	0.0063/0.0746/0.0062
Computers and Internet/Security and Encryption/Encryption Policy/	0.0565	0.0751/0.0110/0.0000
Computers and Internet/Security and Encryption/Mailing Lists/	0.0541	0.0172/0.0581/0.0000
Computers and Internet/Internet/World Wide Web/Security and Encryption/	0.0536	0.0056/0.0561/0.0113
Computers and Internet/Programming and Development/Languages/Java/Security/	0.0464	0.0000/0.0542/0.0127

4.1.4 カテゴリの統合

2003 年に英語版 Yahoo! カテゴリにおけるカテゴリ「Computers and Internet」以下の 559 カテゴリと，日本語版 Yahoo! カテゴリにおけるカテゴリ「コンピュータとインターネット」以下の 654 カテゴリに登録されている Web 文書の収集を行った．HTML タグ除去後の各カテゴリにおける Web ページのバイト数の総計は，英語版では平均 45,905 バイト，最小 476 バイト，最大 1,084,676 バイトであり，日本語版では平均 22,770 バイト，最小，467 バイト，最大 409,576 バイトであった．この結果からも分かるように，カテゴリによっては一つのカテゴリに属している Web 文書が少ないため，十分な統計情報が得られない可能性が高い．このような十分な統計情報が得られないカテゴリは，ノイズとなりシステムに悪影響を及ぼすこととなる．

一つのカテゴリに属する Web 文書が少なくなる原因は，カテゴリが細分化されすぎていることにあると考えられる．それにより同一のカテゴリに属するべき Web 文書群が複数のカテゴリに分散するため，文書数が不足するカテゴリが生じていると考えられる．それゆえ，各カテゴリに属する文書を増やすためには，カテゴリ数を減らす必要がある．

そこで，Web ディレクトリの構造を利用し，下位の階層のカテゴリを上位の階層に統合することにより，カテゴリの細分化の問題を解決する．下位の階層のカテゴリは上位のカテゴリの内容を特化したものであることから，これらが対象とする分野は同じとみなせることが多い．また，直接繋がっていないカテゴリ間でも，それほど階層の離れていない共通の上位のカテゴリを持つならば，互いが対象としている分野に大きな差異はないといえる．よって，このように下位のカテゴリを上位に統合し，一つのカテゴリとして扱うことを考える．カテゴリを統合すると，一つのカテゴリに属する総文書数も増加する．このようにカテゴリを統合することにより，カテゴリ数を減少することができ，一つのカテゴリに属する Web 文書数が増加し，その総バイト数も増加するため，有意な統計量が得られると考えられる．

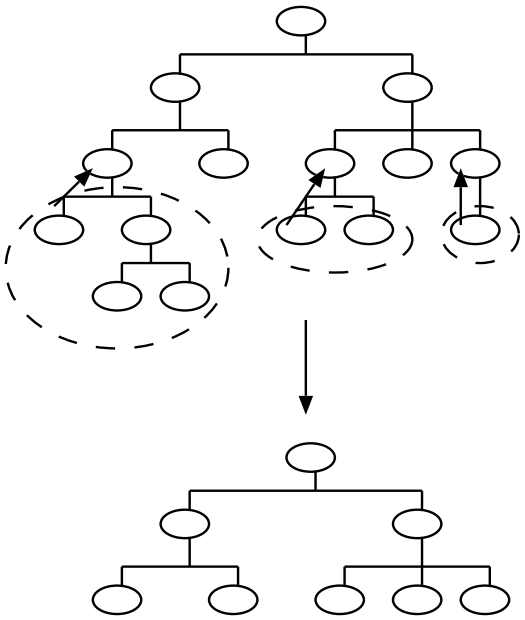

図 4.5 カテゴリの統合

4.2 評価実験

提案手法の有効性を検証するために，日本語の問合せから英語の文書群に対して検索する実験を行った．今回の実験の目的は，Web ディレクトリを利用した本手法が言語横断情報検索に対して検索精度の向上が得られるか，について検証することである．今回の実験では，提案手法だけではなく，比較対象として訳語の曖昧性解消を行わない場合についても，同様の実験を行った．比較対象における問合せの翻訳は，対訳辞書から得られる問合せ語の訳語をすべてを翻訳された問合せ語とすることで行う．ただし，複数の単語からなる訳語は，翻訳された問合せから省いている．問合せ翻訳後の処理については，提案手法と同様の方法で検索を行う．

本実験では，国立情報学研究所が作成した，第 3 回 NTCIR ワークショップ[1] の言語横断検索タスクで用いられた文書群と検索課題 (以下 NTCIR3 テストコレク

[1] http://research.nii.ac.jp/ntcir/index-ja.html

ション) を使用した．このテストコレクションのうち，1998-1999 年に台湾で発行された英字新聞各種からなる EIRB010，および同年に日本で発行された英字新聞である “毎日デイリー 1998-1999” の二つを検索対象として用いた．また，このテストコレクションの日本語の検索課題を本実験における問合せとして用いた．NTCIR3 の日本語検索課題は 50 の問合せが用意されており，この全ての問合せを用いて実験を行った．図 4.6 は，NTCIR3 言語横断検索タスクの日本語検索課題を抜粋したものである．

また，訳語の曖昧性解消のために用いる Web ディレクトリとして，Yahoo の英語版と日本語版を用いた．本実験では，英語のトップレベルカテゴリ “Regional”，および日本語のトップレベルカテゴリ “地域情報” 以下のカテゴリを除いた全てのカテゴリから Web 文書を収集し，曖昧性解消に用いた．英語のトップレベルカテゴリ “Regional”，および日本語のトップレベルカテゴリ “地域情報” 以下のカテゴリを除いたのは，これらのカテゴリには世界各地の地域に関する文書が属しているため，英語および日本語の翻訳に用いるのには適さないからである．英語版ではカテゴリ数は 84,835 カテゴリ，文書数は 800,000 文書，日本語版ではカテゴリ数は 3,175 カテゴリ，文書数は 34,443 文書であった．今回の実験では，下位のカテゴリを上位のカテゴリに統合し，最終的には各言語版のトップページに登録されている 13 のカテゴリに統合した．カテゴリの統合を行った理由は，カテゴリによっては属している Web 文書が少なく十分な統計情報が得られない場合もあるためである．統合後のカテゴリから抽出された単語数は，英語版では 1 カテゴリあたり 322,672 語であった．

Web 文書から単語を抽出する際に，英語版では単語の活用形を原形にしたのち，ストップワードを取り除いた．ストップワードのリストは “Information Retrieval: Data Structures and Algorithms” の chapter 7[15] に掲載されているものを用いた．図 4.7 にストップワードの一覧を示す．日本語では，英語のように単語の区切りが明確でないため，“茶筌”[2] などの形態素解析ツールを用いる必要がある．本実験では “茶筌” を用いて単語に分割した後，名詞，動詞，形容詞，未知語を抽出した．また，問合せの翻訳のための対訳辞書には，“EDR 電子化辞書” の “日英対

[2] http://chasen.aist-nara.ac.jp/hiki/ChaSen/

<TOPIC>
<NUM>001</NUM>
<SLANG>CH</SLANG>
<TLANG>JA</TLANG>
<TITLE>展覧会「漢代の芸術と文化」</TITLE>
<DESC>
故宮博物館で行われた「漢代の芸術と文化」という展覧会についての情報を探す.
</DESC>
<NARR>
故宮博物館は，中国コレクションがすぐれていることでよく知られている．漢代のコレクションは，紀元前 206 年から西暦 220 年までの中国の繁栄期を表すものである．芸術的な展示物における文化的・歴史的遺物の種類，展覧会の日程，故宮博物館がどのように展覧会の準備をしたか，展覧会の協賛パートナー，展覧会に対する市民の反応などに焦点を当てている文書を関連文書とみなす．展示されていない漢代の芸術や文化や他の展覧会の紹介は，不適合とする．
</NARR>
<CONC>
漢代，漢代の芸術と文化，展覧会，故宮博物館，歴史
</CONC>
</TOPIC>

図 4.6 NTCIR3 言語横断検索タスク 日本語検索課題 (抜粋)

訳辞書”[3] を用いた．単純に対訳辞書で翻訳した場合，1 単語に対して平均で 5.17 語の訳語候補が得られた．カテゴリの特徴語の抽出において，各カテゴリの特徴語数は 10,000 語とした．特徴語数の決定の詳細は 4.3 において述べる．また，言語間におけるカテゴリの対応付けは人手により行った．Yahoo では，各言語版のトップページに登録されている 13 のカテゴリの構成は，いずれの言語でも同じであるので，対応が明らかであったためである．

前処理が済んだのち，NTCIR3 テストコレクションを用いて検索の実験を行った．問合せは，NTCIR3 言語横断検索タスクの日本語検索課題から “TITLE” フィールドを抽出したもの，および “DESC” フィールドを抽出したものの二とおりを用いた．表 4.2 以下五つの表は，NTCIR3 の日本語検索課題の TITLE フィールドの問合せの一覧，および DESC フィールドの問合せの一覧である．一般に，Web 検索エンジンの一般的な利用者が投入する問合せは平均 2 単語程度であると言われている [31]．そのため，NTCIR3 テストコレクションの検索課題のフィールドのうちから比較的単語数の少ない “TITLE” および “DESC” フィールドを今回の実験で用いた．それぞれの問合せに対して “茶筌” により形態素解析を行い，名詞，動詞 (サ変動詞の一部除く)，形容詞，未知語のみを抽出し，これらを問合せの単語として用いた．

問合せを翻訳したあとで行う検索処理においては，“SMART”[4] [57] を用いて検索を行った．検索対象文書群の索引は augmented tfidf により重み付けを行った．SMART は一般的な英語文書を検索対象とする検索システムとしては最も代表的なシステムの一つであり，30 年以上にもわたり改良が続けられている．SMART は検索モデルにベクトル空間モデルを採用しており，このモデルを最初に採用したシステムである．ベクトル空間モデルの特徴は，文書および検索質問を多次元のベクトルにより表現し，ベクトル間の類似度を計算することにより文書の類似検索を実現していることである．ベクトル空間モデルでは各文書に対して検索質問との類似度が計算できるため，この類似度を比較することにより適合文書の順位付けが可能である．

[3] http://www2.crl.go.jp/kk/e416/EDR/J_index.html
[4] ftp://ftp.cs.cornell.edu/pub/smart/

different n necessary need needed needing newest next no nobody non noone
not nothing now nowhere of off often new old olderoldest on once one only
open again among already about above against alone after also although
along always an across b and another ask c asking asks backed away a
should show came all almost before began back backing be became because
becomes been at behind being best better between big showed ended ending
both but by asked backs can cannot number numbers o case few find finds
cases clearly her herself come could d did here beings fact far felt
become first for four from full fully furthers gave general generally get
gets gives facts go going good goods certain certainly clear great greater
greatest group grouped grouping groups h got has g have having he further
furthered had furthering itself faces highest him himself his how however
i if important interests into is it its j anyone anything anywhere are
area areas around as seconds see seem seemed seeming seems sees right
several shall she enough even evenly over p part parted parting parts per
down place places point pointed pointing points possible present presented
presenting ends high mrs much must my myself presents down problem problems
put puts q quite will with within r rather really room rooms s said same
right showing shows side sides since small smaller smallest so some
somebody someone something somewhere state states such sure t take taken
than that the their then there therefore these x thought thoughts three
through thus to today together too took toward turn turned turning turns
two still u under until up others upon us use used uses v very w want
wanted wanting wants was way we well wells went were what when where
whether which while who whole y year years yet you everyone everything
everywhere young younger youngest your yours z ever works every everybody
f face other our out just interesting high might k keep keeps give given
higher kind knew know known knows l large largely last later latest least
less needs never newer let lets like likely long high longer longest m
made make making man many may me member members men more in interest
interested most mostly mr opened opening new opens or perhaps order
ordered ordering orders differ differently do does done downed downing
downs they thing things think thinks this those ways why without work
worked working would during e each early either end though still whose saw
say says them second any anybody

図 4.7 ストップワード一覧.

問合せ番号	問合せ語
1	展覧会「漢代の芸術と文化」
2	WTO への加入
3	大学学術追求卓越発展計画
4	E コマース
5	中国の経済改革
6	ノーベル物理学賞
7	中華航空機墜落
8	オスカー
9	人工衛星 ST1
10	逆エルニーニョ現象
11	阿里山の森林鉄道
12	バンコクでのアジア競技大会
13	台湾省の再組織化
14	コンピュータウィルスによる被害
15	クローン牛の誕生
16	佐々木主浩投手のマリナーズ入団
17	北野武監督作品
18	終末思想
19	欧州通貨統合の経済的影響
20	日産とルノーの資本提携
21	トルコの大地震の被害や救援活動
22	ポル・ポト氏の戦争犯罪
23	金大中大統領の対アジア政策
24	国境なき医師団
25	対人地雷全面禁止条約

表 4.4 NTCIR3 日本語問合せ（TITLE）–前半.

問合せ番号	問合せ語
26	世界の人口 60 億人突破
27	マカオの返還
28	日本の北朝鮮訪問団の派遣
29	日本の火山の噴火
30	天皇のデンマーク訪問
31	京都のもみじ
32	1998 年に中国で起こった洪水に対する支援
33	クリントンのスキャンダル
34	米の輸入
35	戦争犯罪訴訟
36	原子力に対する抗議
37	人間のクローンの禁止
38	アフリカの大使館爆破の反応
39	大学入試政策
40	新年の休暇の間のテレビ番組
41	世界 NGO 会議
42	EU とアジア諸国との関係
43	世界的な自然災害
44	大統領になる以前の金大中
45	環境問題に対する国際協力活動
46	飲酒運転についての法規と損害
47	韓国と日本の貿易
48	青年のためのカウンセリング
49	科学キャンプ
50	ティーンエイジャーのファッション

表 4.5 NTCIR3 日本語問合せ（TITLE）–後半.

問合せ番号	問合せ語
1	故宮博物館で行われた「漢代の芸術と文化」という展覧会についての情報を探す。
2	台湾が WTO に加入した後、産業界が直面するだろうと思われる問題を探す。
3	大学学術追求卓越発展計画の内容について探す。
4	Ｅコマースとは何か、および、その内容について探す。
5	首相着任以後の、朱鎔基の経済改革について探す。
6	1998 年のノーベル物理学賞に関する文書を検索する。
7	桃園国際空港で着陸中の中華航空機が墜落した事故についての文書を検索する。
8	1998 年のアカデミー賞受賞作品「タイタニック」に関する文書を検索する。
9	人工衛星 ST1 に関する報道やコメントを探す。
10	逆エルニーニョ現象とは何であるか、および、逆エルニーニョ現象との比較について探す。
11	阿里山の蒸気機関車の歴史と, 阿里山の蒸気機関車と森林や観光との関係について探す。
12	バンコクで開催されたアジア競技大会のニュースを探す。
13	台湾省の再組織化に関する法令と再組織化後の宋楚愉氏の態度についての文書を探す。

表 4.6 NTCIR3 日本語問合せ（DESC）問 1–13.

問合せ番号	問合せ語
14	コンピュターウィルスによって被害を受けた事件について報じた記事が読みたい。
15	体細胞技術を使って、クローン牛をつくることに成功した記事を読みたい。
16	佐々木主浩投手の米大リーグ、シアトル・マリナーズ入団に関する記事が読みたい。
17	北野武が監督した映画に関する記事が読みたい。
18	宗教的な終末思想に関連して起きた事件について知りたい。
19	ヨーロッパの通貨統合の経済的影響についての記事を読みたい。
20	日本の日産自動車とフランスのルノーの資本提携に関連する記事が読みたい。
21	1999 年に起きたトルコ西部の大地震の被害状況や被害者への救助・救援活動などについて知りたい。
22	カンボジアのポル・ポト元首相の戦争犯罪について述べている記事が読みたい。
23	金大中大統領の対アジア政策に関する記事を読みたい。
24	ノーベル平和賞を受賞した国境なき医師団の活動についての記事を読みたい。
25	対人地雷全面禁止条約の各国の批准についての記事が読みたい。
26	世界の人口が 60 億人を突破したことで提起される、世界的な人口増加の問題に関する記事が読みたい。

表 4.7 NTCIR3 日本語問合せ（DESC）問 14–26.

問合せ番号	問合せ語
27	ポルトガル領マカオの中国への返還
28	日本の村山元首相を団長とする北朝鮮への訪問団派遣
29	日本では、これまでにどのような火山の噴火が起こったのか。
30	もしも行ったことがあるとすれば、日本の天皇がデンマークに行ったのはいつか。
31	紅葉したもみじを京都で見るとき、最適の場所はどこか。
32	1998 年に中国で起こった洪水の際の、さまざまな人道的支援について述べている文書。
33	クリントンのスキャンダルに対するアジアでの反応はどのようなものであったか。
34	アジア諸国の米の輸入政策はどういうものか。
35	第二次世界大戦中に日本が犯した戦争犯罪から生じた、日本における訴訟についての情報がほしい。
36	原子力に対する抗議についての情報がほしい。
37	人間のクローン作成に対する、政府、あるいは国際的な禁止についての情報を探す。
38	ケニアとタンザニアの米国大使館がテロリストに爆破されたことに対するアジアの反応について述べている文書がほしい。
39	大学入試政策（制度）、および、親、学生、教員の意見について述べている文書を適合とする。

表 4.8 NTCIR3 日本語問合せ（DESC）問 27–39.

問合せ番号	問合せ語
40	年末年始の休暇中に、テレビ局が放送する特別番組はどんなものか。
41	NGO に関する世界会議の開始について報じている文書、あるいは女性の権利の改善に関する会議で何が議論されたか述べている文書を適合とする。
42	EU（ヨーロッパ連合）とアジア諸国との間の経済関係はどのようなものか。
43	洪水、地震、飢きんのような、世界的に発生する異常な自然現象によって引き起こされる自然災害とはどんなものか。
44	例えば民主化運動のような、大統領になる以前の反対派指導者としての金大中の活動や、彼のノーベル平和賞受賞の背景について述べている文書にはどのようなものがあるか。
45	大気や水、土壌、自然に対する汚染のような環境問題についての国際的な協力活動にはどのようなものがあるのか。
46	飲酒による交通事故で引き起こされる人命の損失や物的な損害、飲酒防止に適用される法律について述べた文書にはどのようなものがあるか。
47	韓国と日本の間の貿易の種類についての説明や予測を述べた文書を検索する。
48	青年の悩みに対するカウンセリングをおこなう機関の名称や、カウンセリングのサービスを受けるための案内を含んだ文書にはどのようなものがあるのか。
49	科学への青年の関心を高めたり、彼らの好奇心を満たすための科学プログラムにはどのようなものがあるのか。
50	服装や髪型、化粧におけるティーンエイジャーのファッションを記述した文書を検索する。

表 4.9 NTCIR3 日本語問合せ（DESC）問 40–50.

検索システムの評価　情報検索システムは一般に次の二つの観点から評価される.

- 完全性: 問合せに適合する文書をもれなく検索しているかどうか.
- 正確性: 問合せに適合する文書だけを検索しているかどうか.

これらを評価する尺度として最も一般的に用いられるのが再現率と適合率である．前者を評価するのが再現率であり，後者を評価するのが適合率である.

- 再現率: 完全性を評価する尺度であり，検索対象となる文書集合の中の問合せに適合する文書のうち，実際に検索された文書の割合を示す．検索漏れの少なさを示す尺度である.
- 適合率: 正確性を評価するための尺度であり，検索された文書集合の中で，問合せに適合する文書の割合を示す．検索ノイズの少なさを示す指標である.

再現率および適合率の双方とも，取り得る値は 0 から 1 の範囲である．理想的な検索システムでは，再現率および適合率の双方ともに 1 に近い値をとることである．しかし実際には，再現率と適合率との間にはトレードオフの関係が成立しており，一般的には，再現率を上げようとすれば適合率が低下し，逆に適合率を上げようとすれば再現率が低下するという現象が起こる．それでもできる限り再現率と適合率の双方ともの値を高くしようとするのが，検索システムの改善の目標である.

検索システムには，問合せに対する適合度が高いと判断した順に順位をつけて検索結果を出力するものがある．このようなシステムでは，検索結果の上位何番目まで検索結果として採用するかにより，適合率と再現率は変化する．上位から i 番目までの検索結果を用いて算出した再現率および適合率を R_i, P_i と表し，適合率 P_i を再現率 R_i の関数とみなして 2 次元の座標上にプロットすると，図 4.8 のようなグラフが得られる．このグラフを再現率・適合率曲線 (recall-precision curve) と呼ぶ.

一般に再現率・適合率曲線は右下がりのグラフとなるが，この曲線が上に位置するほどその検索システムは性能が良いと評価されることとなる．図 4.9 は，図

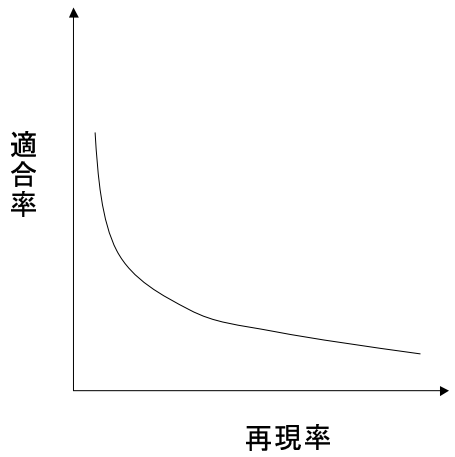

図 4.8 再現率・適合率曲線 (a)

4.8 の場合よりも右上のほうに位置しており，図 4.9 の検索システムほうが性能が良いと評価することができる．しかし，図 4.10 のように途中でグラフが交差する場合は，どちらの検索システムが優れているとは一概に断定することはできない．このような場合は，再現率と適合率のどちらを重視するかにより，検索システムの性能の優劣の判断を行う必要がある．

再現率・適合率曲線が交差する場合には検索システムの性能の優劣を決めることは困難であるが，再現率と適合率曲線を総合的な観点から一つの値により評価することができれば，検索システムの性能の評価がしやすくなる．そのような指標として，平均適合率 (average precision) がある．平均適合率は，n 個の再現率における適合率を求め，その平均を算出することにより求められる．この場合の平均適合率は，n 点平均適合率と呼ばれる．情報検索システムの評価でよく用いられるのは，11 点平均適合率である．11 点平均適合率では再現率を 0.0, 0.1, $\cdots$, 0.9, 1.0 と 0.1 刻みに 11 点取り，それぞれの再現率において適合率を求めこれらの平均を求める．この値を情報検索システムの性能を評価する指標として用いる．

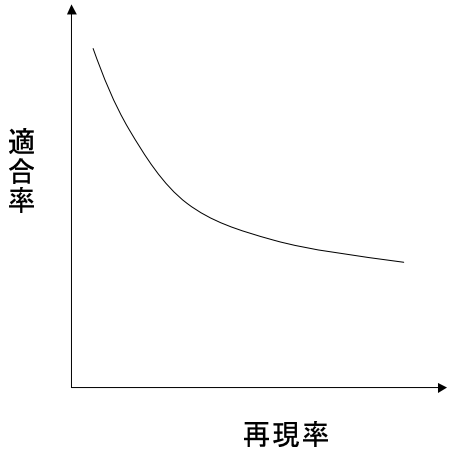

図 4.9 再現率・適合率曲線 (b)

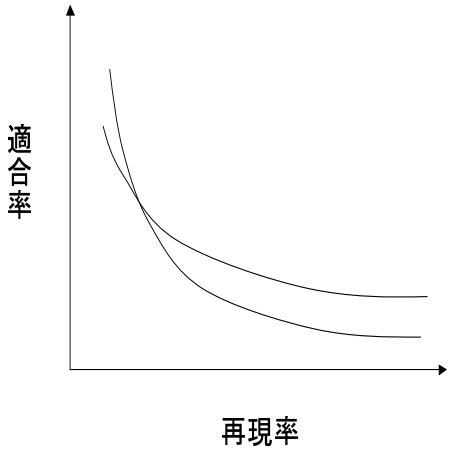

図 4.10 再現率・適合率曲線 (c)

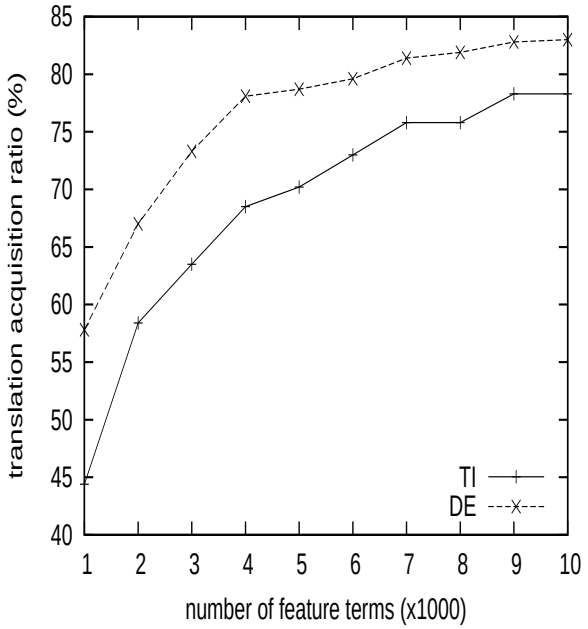

図 4.11 カテゴリの特徴語数と得られた訳語獲得率との関係

4.3 問合せの翻訳

提案手法ではまず，問合せと同言語である日本語のカテゴリから適合カテゴリを推定する．このとき問合せと比較するカテゴリの特徴語集合は，各カテゴリにつき 1,000 語使用した．次に，推定された適合カテゴリに対応付けられている英語のカテゴリを選択する．そして，対訳辞書を用いて問合せ単語の訳語候補を全て抽出する．最後に選択した英語の対応カテゴリの特徴語と問合せの訳語候補とを 4.1 で述べた方法で比較することにより，訳語を決定する．

このとき，各カテゴリの特徴語数を事前に決定する必要がある．本システムにおいて，適切な特徴語数を調べるために，特徴語数の変化に対して訳語獲得率がどのように変化するかについて実験を行った．訳語獲得率とは，全問合せ語数に対して，何らかの訳語が得られた問合せ語数の割合である．今回の実験では，英語の対応カテゴリの特徴語数は，1,000 語から 10,000 語の間で，1000 語ごとに実験を行った．

図 4.11 は実験の結果を示している．“TI”，“DE” はそれぞれ “TITLE” および “DESC” フィールドから得られた問合せを用いた場合の結果を示している．カテゴリの特徴語数が 1,000 語から 4,000 語の間では，訳語獲得率の上昇の度合いが大きい．さらに特徴語数を増加させていくと徐々に訳語獲得率の上昇の度合いは小さくなってゆき，9,000 語から 10,000 語ではほぼ変化はなくなった．カテゴリの特徴語数が 1,000 語の場合，TITLE では 44.4%，DESC では 57.0%と，全問合せ単語数の半分程度しか訳語が得られていない．特徴語数を 10,000 語にした場合，それぞれ 77.0%，83.0%の獲得率であり，特徴語数が 1,000 語の場合に比べて 30%程度獲得率が上昇している．

問合せの訳語が得られない原因として，対訳辞書に訳語候補が存在しない場合とカテゴリの特徴語に全ての訳語候補が存在しない場合の 2 点が挙げられる．このうち前者の原因は，カテゴリの特徴語数に関係なく，訳語の獲得率の低下を引き起こす．この実験で用いた問合せ語に対して，何らかの訳語が対訳辞書に存在している割合は，“TITLE” フィールドでは 88.8%，“DESC” フィールドでは 92.8%であった．ともに，特徴語数が 10,000 語の場合よりも約 10%高くなった．この差は，上記の原因の後者から生じていると考えられる．この後者の原因は，カテゴリの特徴語数がそのまま獲得率に影響する．よって，カテゴリの特徴語数を 10,000 より多くすることにより訳語の獲得率が上昇する可能性はある．しかし，特徴語数をあまり多くすると，そのカテゴリの特徴をあまり表していない単語が含まれてしまい，問合せに対して適切でない訳語を選択してしまうという問題が生じる可能性が考えられる．特徴語数が 10,000 語で訳語獲得率の上昇が小さいことも考慮すると，さらに特徴語数を増やすのはあまり有効ではないと言える．

4.3.1 検索実験の結果

翻訳された問合せにより，NTCIR3 言語横断検索タスクの英語文書群に対して検索を行った．本実験では，提案手法における問合せの翻訳に用いる各カテゴリの特徴語数は 10,000 語とした．その結果を図 4.12 および 表 4.10 に示す．なお，文書が適合しているかどうかの判定は，NTCIR テストコレクションの Relax 正解集合を用いた．NTCIR3 テストコレクションには，“Rigid” と “Relax” の 2 種

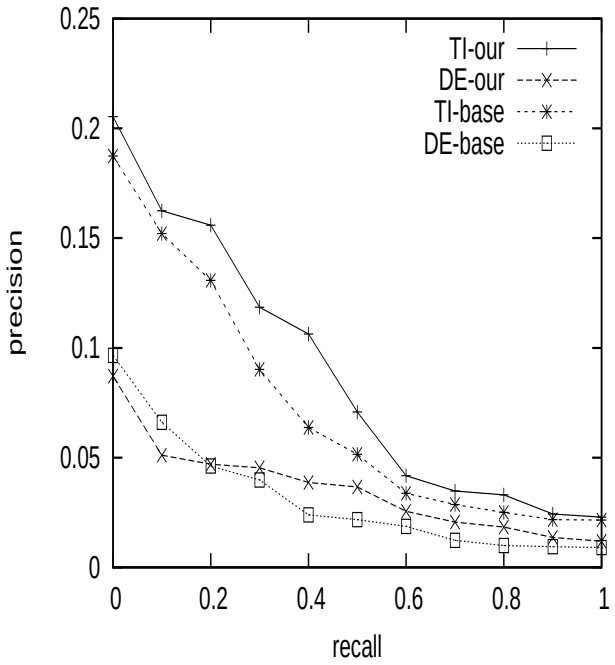

図 4.12 検索結果の適合率・再現率グラフ

類の正解集合が用意されている．Rigid 正解集合では文書が適合しているかどうかの基準はやや厳格であり，それに比べて Relax 正解集合ではその基準をやや緩和している．

図 4.12 は検索結果の評価を適合率・再現率グラフで表したものである．この図における “TI-our”，“DE-our” はそれぞれ，“TITLE” フィールドを抽出した問合せ (以下: 問合せ TI)，および “DESC” フィールドを抽出した問合せ (以下: 問合せ DE) を用いて提案手法により検索を行った結果である．“TI-base”，“DE-base” はそれぞれ，問合せ TI および問合せ DE を，対訳辞書のみによる翻訳を行った場合 (以下: ベースライン) の検索結果である．適合率とは，検索結果として得られた文書のうち，実際に問合せに適合している文書の割合のことである．再現率とは，文書群全体に含まれている適合文書のうち，検索結果に含まれている割合のことである．また，表 4.10 は，上記 4 種類の問合せでの検索結果の 11 点平均適合率を表している．11 点平均適合率とは，再現率が 0.0, 0.1, 0.2, … , 0.9, 1.0 と

表 4.10 検索結果の平均適合率

	TITLE	DESC
提案手法	0.0851	0.0311
ベースライン	0.0677	0.0254

なる 11 点における適合率の平均値であり，その検索システムの検索精度を評価する指標の一つである．

提案手法の総合的な検索性能の比較のため，上記で得られた実験結果に対して考察を行なう．表 4.10 から分かるように，問合せ TI においては提案手法がベースラインより 1.74 ポイント，問合せ DE では 0.57 ポイント平均適合率が上昇しており，問合せ TI，問合せ DE のいずれにおいても，本手法の有効性が認められる．図 4.12 においても，ほとんどの再現率の区間で，本手法のほうが良い結果が得られた．ただし，問合せ DE においては再現率が 0.00〜0.20 の区間においてのみ，本手法が下回る結果となった．この区間において本手法が下回った原因は，問合せの翻訳において訳語が一部得られなかったことにあると考えられる．一方ベースラインの場合，訳語候補は全て用いるため，重要な問合せ語が含まれる確率も提案手法よりも高くなる．よって，再現率が低い段階では，重要な問合せ語が含まれていることの効果が強く現れたと考えられる．しかし，再現率が高くなってくると，重要な問合せ語が含まれることよりもむしろ，不要な問合せ語の影響が大きくなってくると考えられる．そのため，不要な語を出来るだけ排し重要な語のみを訳語として用いる提案手法が，再現率が高くなる区間では良い結果が得られた．

問合せ TI と問合せ DE で比較すると，提案手法とベースラインのいずれにおいても，問合せ TI のほうが良い結果が得られた．“TITLE” フィールドは，重要な単語を列挙したものと言ってよく，そこに現れる単語のほとんどが重要な問合せ語であると言える．しかし “DESC” フィールドは文章として完結した形となっているため，問合せには不要な語も含まれる．例えば “〜に関する記事を探したい”，“〜について記述された文書を検索する” といった記述が多くあり，このよ

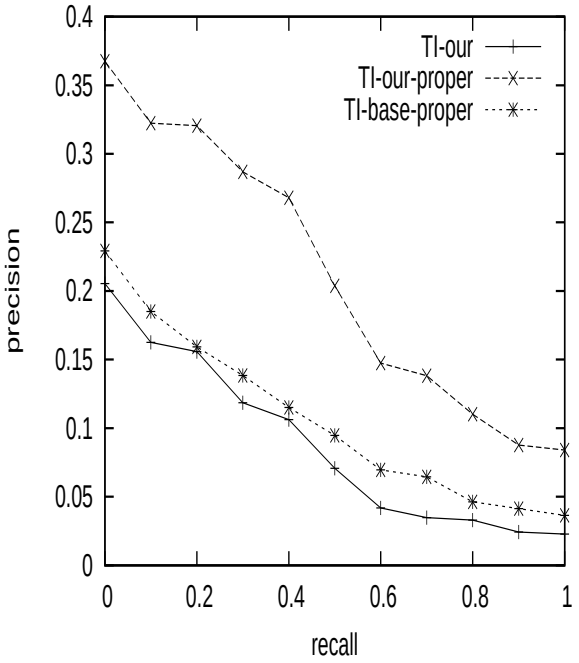

図 4.13 固有名詞を人手で翻訳した場合の適合率・再現率グラフ

うな記述から抽出された単語が不要な問合せ語となった．このような不要な問合せ語が，検索精度の低下の原因となったと考えられる．このことからも，問合せから不要な単語を排除することが重要であると言える．

NTCIR3 の CLIR タスクに参加した手法の多くは，日本語 → 英語では 0.2000 程度の 11 点平均適合率が得られている [8]．これと比較すると，今回の実験結果の 11 点平均適合率はかなり低い．その原因は，固有名詞に対する対策をほとんど取らなかったことにある．固有名詞は適合文書を特定する重要な手がかりとなる．しかし固有名詞は対訳辞書に載っていないことが多いため，適切な訳語を得るのは困難である．そのため，今回の実験では多くの固有名詞の訳語が得ることができず，適合率が低くなったと考えられる．

そこで，提案手法により問合せを翻訳するときに，固有名詞のみ人手により翻訳した場合についても，同様の検索の実験を行った．図 4.13 は，その結果を適合率・再現率グラフで表したものである．問合せ TI に対して，提案手法で翻訳し

表 4.11 固有名詞を人手により翻訳した場合の検索結果の平均適合率

TI-our	TI-our-proper	TI-base-proper
0.0851	0.2291	0.1022

た場合が “TI-our”，固有名詞を人手により翻訳した場合が “TI-our-proper” である．また，“TI-base-proper” は，ベースラインにおいて固有名詞を人手により翻訳した場合の検索結果である．固有名詞を人手により翻訳することにより，適合率が格段に向上している．また，表 4.11 は，この実験における 11 点平均適合率を表している．固有名詞を人手により翻訳した場合の 11 点平均適合率は 0.2291 となった．この結果は，他の言語横断情報検索の手法と比較しても遜色ない値であるといえる．一方，表 4.11 の “TI-base-proper” における 11 点平均適合率は 0.1022 であり，提案手法により検索精度が向することが確認された．また，NTCIR3 の CLIR タスクにおける日本語 → 日本語の単言語検索では，最もよい結果が得られた場合で 11 点平均適合率は 0.4000 程度であった．これは日本語 → 英語の言語横断情報検索において実現可能な上限であると考えられる．

5. 問合せ翻訳手法の改良

4.1 において提案した問合せ翻訳手法では，問合せが対象としていると推定されるカテゴリを一つだけ選択し，そのカテゴリの特徴語集合に含まれている訳語候補の中から最も適切であると思われるものを一つだけ選択することで問合せ語の翻訳を行った．しかしこの方法では，問合せ語を翻訳したときに，訳語の得られない問合せ語が生じることもあった．提案手法では，問合せの対象分野を限定することによりその分野において適切な訳語を得やすくすることを図っているのであるが，それゆえにもし問合せが対象としている分野の推定で過誤が生じてしまうと，適切な訳語が得られにくくなる可能性が高くなってしまう．これは，問合せの対象分野を限定することの副作用といえる．利用者が入力する問合せ語は 2 語から 3 語程度とあまり多くないため，一つ一つの問合せ語が適合文書を探し出すための重要な情報源である．訳語が得られないことは，その数少ない重要な情報源を失うことを意味する．よって，この問題の改善を行うことは検索精度の向上のためにも重要となる．

また，問合せ語を適切に翻訳できたとしても，適合文書の検索漏れが起こらなくなるかというと，実際はそうはならない．言語表現には多様性があるからである．ある概念を表現するために使用可能な単語は一つとは限らない．一般には，複数の表現方法すなわち複数の単語が使用可能であることのほうが圧倒的に多い．例えば,「金」,「現金」,「通貨」などはいずれも「おかね」という同一の概念を表している．文書の作成者は,「おかね」の概念を表現したい場合は，これらのいずれの単語使用してもかまわないため，語選択の多様性が生じる．一方，文書を検索する際には，問合せ語として選択された単語が検索対象文書に含まれていなければ，適合文書として抽出することはできない．例えば問合せにする文書には「金」と記述されている場合に，検索者が「通貨」という語を問合せ語として用いて問合せを行うと，実際にはその文書が問合せに適合していたとしても,「金」と「通貨」は別の単語として認識されるため，その文書が検索されることはない．こういった問題を改善する方法として，検索質問拡張が行われる．検索質問拡張とは，元の問合せ語と同じ概念を表現している別の単語のいくつかを問合せ語に加え手拡張し，この拡張された問合せを用いて検索対象文書群を検索する [55]．これに

より，先ほどの例のような場合でも，適合文書として検索することが可能となる．このように，検索質問拡張を行うことで検索結果の再現率を改善する．しかしその一方で，むやみに問合せ語を追加しすぎると適合率の低下を招くため，注意が必要である．4.1 における提案手法では問合せを翻訳するとき，一つの問合せ語につき選択する訳語は一つだけであった．よって，検索質問拡張と同様に，問合せ語を追加する，すなわち一つの問合せ語につき選択する訳語の数を増やすことにより，検索精度の向上が得られる可能性がある．

以上のことを踏まえ，提案手法の検索精度を向上図るため，本章においてはの以下の 2 点について検討した．

- 適合カテゴリを複数選択する
- 一つの問合せ語に対する訳語を複数選択する

5.1 適合カテゴリの選択

4.1 では適合カテゴリを一つしか選択しなかったため，問合せ翻訳の曖昧性解消を行うときには，一つの対応カテゴリの特徴語集合のみについてしか調べない．しかし，その特徴語集合に訳語候補が含まれているとは限らないため，訳語が得られないことも多かった．そこでそのような場合，2 番目以降の適合カテゴリについても調べるように問合せ翻訳の手法を改良した．対応カテゴリの特徴語集合に含まれているうちの重みが最も大きい訳語候補を選択するというのは変わらないが，最初に調べたカテゴリに訳語候補が一つも存在していない場合は，その次に適合している対応カテゴリの特徴語集合も同様に調べるように変更し，最大三つのカテゴリについて調べるように改良した．これにより，訳語が得られない可能性を減少させることができる．

5.1.1 実験（適合カテゴリの選択）

適合カテゴリを複数選択する手法の有効性を検証するため，適合カテゴリを複数選んだ場合について 4.2 と同様の実験を行った．

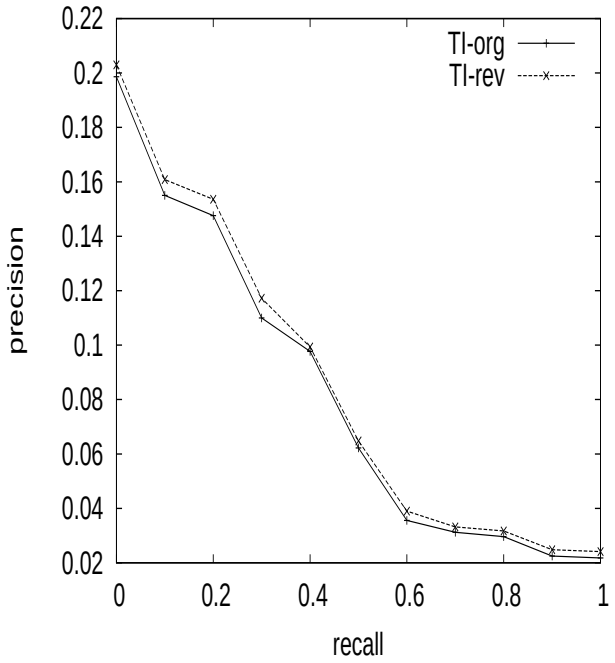

図 5.1 適合カテゴリを複数選択した場合の検索結果の適合率・再現率グラフ

その結果，50 件の問合せのうち，本論文の改良版手法により新たに訳語が追加されたのは 6 件であった．これらの問合せのうち，1 件は平均適合率に変化なかったが，3 件で向上した．なお，残り 2 件はもともと適合文書が存在しないため，本テストコレクションでは評価されていない．

図 5.1 は，実験結果を適合率・再現率グラフに表したものである．“TI-org” が改良前の手法，“TI-rev” が本論文で提案した手法である．わずかではあるが，改良版のほうが以前の手法よりも上回っている．11 点平均適合率においても，“TI-org” が 0.0803，“TI-rev” が 0.0829 と，0.26 ポイント上回った．

5.2 複数の訳語の選択

4.1 では一つの問合せ語に対して訳語を一つだけ選択していたが，この制約のもとでは検索精度を向上させることは困難である．例えば，同じ物事を指してい

るにもかかわらず，文書によって異なった単語を用いるという場合がある．このような場合，一方の単語しか問合せに含めないならば，含められなかったほうの単語を用いている文書は検索されない可能性がある．

そこで，訳語を一つだけという制約を取り払い，複数の訳語を選択してもかまわない翻訳手法を提案する．4.1 では，対訳辞書から抽出した訳語候補のうち，対応カテゴリの特徴語に含まれていてかつ重みの最大のものを訳語として用いるとしていた．これを，対応カテゴリの特徴語に含まれている訳語候補のうち重みが高いものから数単語を訳語として用いることに変更する．

本論文では，以下の三つの手法を提案する．

1. 選択する訳語の数の上限を設定する (以下 “訳語数限定手法”)
2. ある閾値以上の重みの訳語を選択 (以下 “固定閾値手法”)
3. 最も重みの高い訳語の重みに対して，一定以上の重みを持つ訳語を選択 (以下 “変動閾値手法”)

訳語数限定手法は，一つの問合せ語に対して選択できる訳語の数の上限を設定する手法である．仮に訳語数の上限を 3 としていたなら，対応カテゴリの特徴語に含まれている訳語候補のうち，重みが上位 3 語を訳語として用いる．もし対応カテゴリの特徴語に訳語候補が 2 語以下しか含まれていない場合は，それら全てを訳語として用いる．

固定閾値手法では，対応カテゴリの特徴語に含まれている訳語候補のうち，重みが事前に設定された閾値以上の単語はすべて訳語として用いる．閾値の値は全ての問合せ語に対して共通である．

変動閾値手法は，閾値を対応カテゴリの特徴語に含まれている訳語候補のうち最も重みの高い訳語の重みに対するある割合に設定する．この手法の特徴は，閾値は固定ではなく，翻訳前の問合せ語ごとに変化することである．例えば，閾値を重みが最大のものの 1/5 としたとする．ある問合せ語に対して対応カテゴリの特徴語に含まれていた訳語候補が “a”,“b”, “c” であったとし，その重みがそれぞれ “0.085”, “0.023”, “0.015” であったとする．このとき重みが最大である a の重み

表 5.1 訳語数限定手法の 11 点平均適合率.

訳語の上限数	1	2	3	4	5
11 点平均適合率	0.0829	0.0838	0.0774	0.0741	0.0752
訳語の上限数	6	7	8	9	10
11 点平均適合率	0.0747	0.0745	0.0743	0.0739	0.0738

の 1/5，つまり “0.017” がこの問合せ語を翻訳する場合の閾値となる．よって，a, b は訳語として用いられるが，c は今回の閾値を下回るので訳語として用いない.

5.2.1 実験（複数の訳語の選択）

提案した上記の三つの手法のうち，どの手法が優れているかを検証するために，5.1.1 と同様の実験を行った．なお，選択する適合カテゴリは三つとした.

表 5.1 と図 5.2 は，訳語数限定手法の実験結果である．この実験結果は，訳語数の上限と適合率の関係を示している．訳語数の上限が 1 のときの適合率の値は，5.1.1 の “TI-rev” と同じである．訳語数の上限が 2 のとき，適合率は最大となり，“TI-rev” を上回った．訳語数の上限が 3 以上では，上限が大きくなるにしたがって適合率が低下した.

表 5.2 と図 5.3 は，固定閾値手法の実験結果である．この結果は，閾値の値と適合率の関係を示している．閾値が 2×10^{-4} のとき，適合率が最大となった．しかし，“TI-rev” よりも適合率が低くなった．また，閾値が 2×10^{-4} 以上では，閾値が大きくなるにしたがって適合率が低下した.

表 5.3 と図 5.4 は，変動閾値手法の実験結果である．この結果は，訳語候補の最大の重みに対する閾値の比率と適合率の関係を示している．比率の分母が大きくなるほど閾値は小さくなり，より多くの訳語候補を訳語として用いることになる．図 8 の横軸の値は，訳語候補の最大の重みに対する閾値の比率の分母を示している．比率の分母が大きくなるにしたがって適合率も徐々に高くなってゆき，分母が 18 または 19 となったとき適合率が最大となった．しかし，適合率が最大

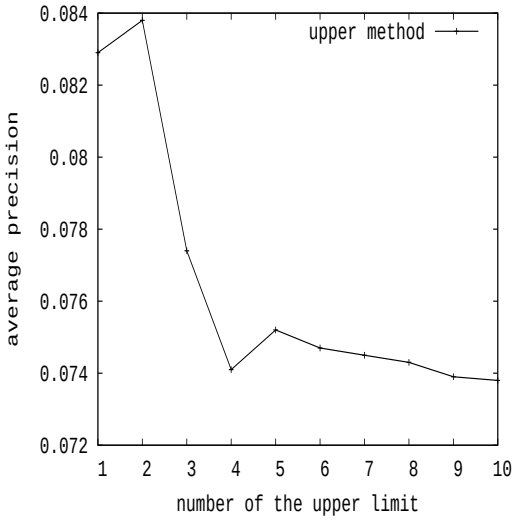

図 5.2 適合カテゴリを複数選択した場合の検索結果の適合率・再現率グラフ

表 5.2 固定閾値手法の 11 点平均適合率.

閾値 ($\times 10^{-4}$)	1	2	3	4
11 点平均適合率	0.0739	0.0753	0.0616	0.0489
閾値 ($\times 10^{-4}$)	5	6	7	8
11 点平均適合率	0.0479	0.0426	0.0382	0.0383
閾値 ($\times 10^{-4}$)	9	10	1.5	1.8
11 点平均適合率	0.0383	0.0379	0.0737	0.0746

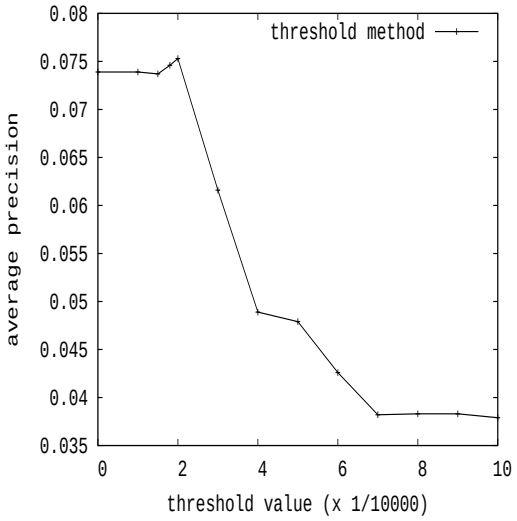

図 5.3 適合カテゴリを複数選択した場合の検索結果の適合率・再現率グラフ

の場合でも “TI-rev” よりも適合率が低くなった．また，分母が 19 よりも大きくなると，分母が大きくなるにしたがって適合率は低下した．

5.2.2 考察

ここで提案した三つの手法は，検索精度を向上させるために問合せの語の訳語を追加している．これには利点と欠点の両方の要因がある．利点は，翻訳後の問合せに新たな問合せ語を追加でき，検索漏れを減らすことが可能となることである．一方欠点は，目的の文書とは無関係な訳語を問合せに加えてしまい，検索精度が低下してしまう恐れがあることである．三つの手法のいずれにおいても，訳語として用いる訳語候補が多くなると平均適合率が低下する傾向にあることが分かった．これは，比較的欠点の影響のほうが現れやすいことを示唆している．よって，訳語の追加は限定的にする必要がある．

三つの手法のうち最も効果的であったのは訳語数限定手法であり，訳語数の上限が 2 のときに最も平均適合率が高くなった．しかし，いずれの手法においても利点と欠点の両方の影響を受けており，効果が相殺されることとなった．

表 5.3 変動閾値手法の 11 点平均適合率.

閾値 (最大の重みに対する比率)	1/50	1/100	1/2	1/3	1/4	1/5
11 点平均適合率	0.0739	0.0739	0.0673	0.0751	0.0752	0.0735
閾値 (最大の重みに対する比率)	1/6	1/7	1/8	1/9	1/10	1/11
11 点平均適合率	0.0737	0.0766	0.0770	0.0770	0.0788	0.0788
閾値 (最大の重みに対する比率)	1/12	1/13	1/14	1/15	1/16	1/17
11 点平均適合率	0.0753	0.0753	0.0800	0.0800	0.0800	0.0800
閾値 (最大の重みに対する比率)	1/18	1/19	1/20			
11 点平均適合率	0.0804	0.0804	0.0746			

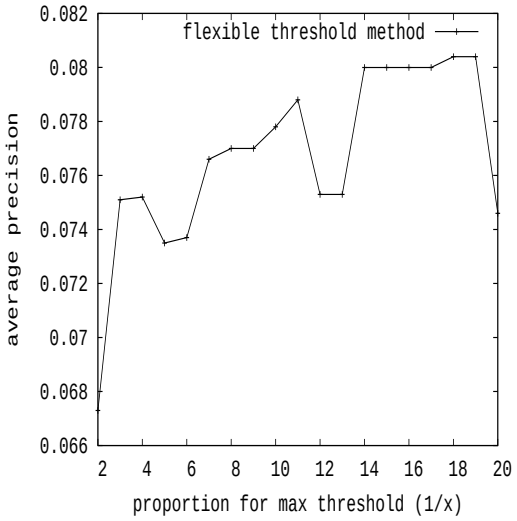

図 5.4 適合カテゴリを複数選択した場合の検索結果の適合率・再現率グラフ

訳語数限定手法の上限数が 1 と 2 の場合において，問合せ語がどのように翻訳されたか詳細を比較する．問合せ番号 9 の日本語の問合せ語に “衛星” という語がある．上限数 1 の場合，この問合せ語を “secondary” と翻訳した．しかし，ここでは “satellite” と翻訳されることが望ましい．上限数 2 の場合，“secondary”，“satellite” の 2 語が訳語として用いられた．訳語として “satellite” が追加されることで，適合率は 0.0 から 0.3790 に増加した．一方で問合せ番号 26 のように，訳語を追加することで適合率が低下する場合もあった．上限数 1 では，“世界” および “突破” という問合せ語がそれぞれ “world”，“breakthrough” という適切であると思われる訳語が得られた．しかし上限数 2 では，“life”，“penetration” というあまり適切ではない訳語を追加することとなった．それにより適合率は 0.1517 から 0.1080 に低下した．

固定閾値手法と変動閾値手法は両方とも閾値を設ける手法であるが，全体的に変動閾値手法のほうが適合率が高くなった．また，固定閾値手法では閾値が 2×10^{-4} を超えた後の適合率の低下の幅が大きかったのに対して，変動閾値手法では比較的変動の幅が小さかった．これは，対応カテゴリによって問合せ語の重要度は異なるが，変動閾値手法では閾値を問合せ語ごとに柔軟に設定することができることにより，対応カテゴリの特徴語集合中で重みが上位にあろうと下位にあろうと影響を受けないため，より適切に訳語を選択することができるからであると考えられる．

6. Web ディレクトリのカテゴリ構造を利用した訳語の曖昧性解消

4 章において，言語横断情報検索の問合せ翻訳における訳語の曖昧性解消に Web ディレクトリを言語資源として利用する手法を提案した．しかし，カテゴリによっては属する Web 文書が少ないことなどにより，そのようなカテゴリの特徴語集合からは訳語の曖昧性解消を行うために必要な統計情報が十分に得られなかった．そこで，4.1.4 において述べたように，下位にあるカテゴリを上位のカテゴリに含めることによってカテゴリの統合を行った．4.2 では，下位のカテゴリを最終的に最上位の 13 のカテゴリに統合することにより，統計情報不足のカテゴリを解消した．これにより提案手法が有効に機能することが 4.2 の検証実験により実証された．

4.2 では下位のカテゴリを最終的に最上位の 13 のカテゴリに統合しているが，これは結果的に最上位の一階層だけしか利用していない．下位のカテゴリを上位のカテゴリに統合する際に Web ディレクトリのカテゴリの階層構造を利用しているのであるが，それだけでは Web ディレクトリのカテゴリの階層構造を十分に利用しているとは言えない．また，最上位の 13 のカテゴリに統合したことにより，一つのカテゴリが対象とする分野の範囲が大きくなることが新たな問題となった．統合後のカテゴリ数が少ないことは，言語間でのカテゴリの対応付けが容易であるという利点はある．特に，4.2 においては Yahoo! カテゴリの最上位の構造はいずれの言語版においても全く同じ構造となっているため，確実に一対一の対応が言語間で取ることができた．しかし，それと同時に一つのカテゴリが対象とする分野の範囲が大きくなることは，問合せの翻訳における訳語の曖昧性解消が十分に行えない可能性が生じる．本手法では，問合せが対象としていると推定される分野を特定し，その分野に適切であると思われる訳語候補を選択することにより，訳語の曖昧性解消を行う．それゆえ，問合せが対象としていると推定される分野を十分に絞り込むことができなければ，訳語候補の選択の幅が広くなり，訳語の曖昧性解消も十分に行えなくなる．4.2 では，Yahoo! カテゴリで構築されている Web ディレクトリの最上位の構造，すなわち 13 のカテゴリから問

合せに関連のあるカテゴリを一つ選択し，そのカテゴリ以下に含まれている全ての Web 文書から抽出された統計情報を用いて，問合せに適切な訳語を選択している．この場合，Web ディレクトリに登録されている Web 文書数は膨大であるため，それらを 13 のカテゴリに分類したとしても，一つのカテゴリに属している Web 文書数は大量となる．すなわち，問合せに適合すると判断されたカテゴリの選択をうまく行えたとしても，大量の Web 文書がそのカテゴリに属しているため，訳語の選択の際に利用される分野情報の特定がうまく行えず，問合せに適切な訳語を選択できない可能性が高くなる．

そこで本章では，一つ一つのカテゴリが対象とする分野の範囲があまり大きくなりすぎないようにし，カテゴリを一つ選択すればある程度対象とする分野を限定できるようにする．具体的には，カテゴリの統合を 4.2 の場合よりも低い階層でとどめておくことにより，一つのカテゴリが対象とする分野の範囲が大きくなりすぎないようにする．

検索モデルの一つであるクラスタモデルは検索の効率性を改善するだけでなく，検索の有効性も改善できる可能性を持っている．クラスタモデルは,「文書が類似していれば，同じ検索質問に対する適合性も同様に類似している」というクラスタ仮説を前提にして提案されたモデルである．一般に，利用者が入力する問合せ語の数は 2, 3 語程度とあまり多くないため，問合せ語の表現の違いによって文書中の索引語と問合せ語が一致しない可能性も生じる．そこでクラスタ仮説では，あらかじめ類似した文書をグループ化しておくことにより，このような不一致の可能性を減少させることを狙っている．つまり，検索対象をグループ化することにより，一種のスムージングの効果を狙っているといえる．提案手法では Web ディレクトリを用いて言語横断情報検索の問合せ翻訳における訳語の曖昧性解消を行うが，Web ディレクトリのカテゴリは階層構造を構成していることから，提案手法はクラスタモデルによる検索の利点を享受できる可能性がある．

中村ら [46] は，対話的に利用者の要求に近いクラスタを適切に選択することを提案している．これにより，利用者に役立つクラスタを抽出することを目指している．若木ら [66] は，索引語をクラスタリングしておくことにより問合せ語の曖昧性解消を行っている．この手法は，単語の共起頻度をもとに特定のトピックに

強く関係する語を索引語として抽出することにより，索引語の問合せ語に対する曖昧性を排除している．この手法は索引語の曖昧性解消を行っており，提案手法とは異なっているが，索引語の対象を限定することで問合せが対象としているトピックを推定できることを示している．このことは，本手法で提案した問合せの対象とする分野を特定することにおいても同様の有効性がある可能性があることを示唆している．

6.1 カテゴリの細分化

本章の冒頭に述べたように，4.2 での問題は一つのカテゴリが対象とする分野が広く，適切な訳語を選択できなかったためである．したがって，Web ディレクトリの構造を最上位だけではなくより低い階層まで利用し，検索対象となる Web 文書の分野を限定することで，問題を解決することにした．一つ一つのカテゴリが対象とする分野の範囲があまり大きくなりすぎないようにし，カテゴリを一つ選択すればある程度対象とする分野を限定できるようにする．具体的には，カテゴリの統合を 4.2 の場合よりも低い階層でとどめておくことにより，一つのカテゴリが対象とする分野の範囲が大きくなりすぎないようにする．

このようにカテゴリをより細分化しておくことにより，一つのカテゴリが対象とする分野の範囲をより限定することができる．しかしその一方で，カテゴリが対象とする分野の範囲をより限定すればするほど，そのカテゴリに割当てられる Web 文書の数は段々と減少していく．そのため，カテゴリをあまり細分化しすぎると，そのカテゴリに属する Web 文書が不足し，十分な統計情報を得ることができないカテゴリが多くなってしまう．

よって，カテゴリをできるだけ細分化することにより，一つのカテゴリが対象とする分野の範囲をできるだけ限定することが望ましいが，それと同時に，各カテゴリに割当てられる Web 文書ができるだけ不足することのないように，どの程度までカテゴリを細分化できるか見極める必要がある．

そこで本章では，どの程度までカテゴリを細分化しておくのがよいか検証を行った．4.1.4 で述べた手法によりカテゴリを統合するのであるが，このとき最上位の 13 カテゴリに統合した場合（これを「1 階層」と呼ぶ），その一つ下の階層

のカテゴリに統合した場合（これを「2 階層」と呼ぶ），最上位の階層から二つ下の階層にあるカテゴリに統合した場合（これを「3 階層」と呼ぶ）について検証実験を行い，どの程度までカテゴリを統合しておくのがよいかについて調査した．

ここで，検索対象となる Web 文書の分野を限定することでカテゴリから抽出される統計情報に問題が生じ，適切な訳語を抽出できない問題が起こり得るが，この問題への対処のために，サブカテゴリに属する Web 文書から抽出される特徴語数が 10,000 語に満たないものは，分野推定の対象から除外している．

6.1.1 実験結果

本節では，どの程度までカテゴリを細分化しておくのがよいか検証を行った．カテゴリの細分化の程度を 1 階層，2 階層，3 階層の三つの場合について実験を行った．1 階層の場合は，Yahoo! カテゴリのトップページにリンクされているカテゴリを用い，2 階層の場合はさらにもう 1 階層下のカテゴリまでを用いている．3 階層では 2 階層の場合よりもさらにもう 1 階層下のカテゴリを用いている．また，1 階層，2 階層，3 階層のいずれの場合においても，そのカテゴリの下位に属しているサブカテゴリはすべてそのカテゴリに属しているものとした．ただし，各カテゴリに属している Web 文書が少ないと，十分な統計情報が得られず，システムに悪影響を及ぼす可能性が高い．よって，統合後のそれぞれのカテゴリにおいて，サブカテゴリに属する Web 文書から抽出される特徴語数が 10,000 語に満たないカテゴリについては，分野推定の対象から除外している．

表 6.1 は，階層ごとのカテゴリ数を示している．除外前というのは，カテゴリの統合を行った段階での各階層ごとのカテゴリ数である．つまり除外前のカテゴリの中には，Web 文書の不足により十分に統計情報が得られないカテゴリも含まれている．除外後というのは，十分に統計情報が得られないカテゴリを除外した場合，つまり，特徴語数が 10,000 語未満となるカテゴリを省いたカテゴリ数を示している．1 階層では，全てのカテゴリがそれぞれ対象としている分野の範囲が大きいため，全てのカテゴリで十分な統計量が得られており，除外の前後にかかわらず全てのカテゴリが利用されている．2 階層，3 階層においてはそうはいかず，統計量が十分に得られないカテゴリが存在している．除外の前後における

表 6.1 階層ごとのカテゴリ数

		1 階層	2 階層	3 階層
英語	除外前	13	397	4066
	除外後	13	255	644
日本語	除外前	13	391	2953
	除外後	13	154	153

カテゴリ数の変化は，2 階層では英語が 397 カテゴリ →255 カテゴリ，日本語が 391 カテゴリ →154 カテゴリであり，3 階層では英語が 4066 カテゴリ →644 カテゴリ，日本語が 2953 カテゴリ →153 カテゴリであった．特徴語数が 10,000 語以上であるカテゴリの割合は，2 階層では英語 64%，日本語 39%であった．それに対して 3 階層では英語 16%，日本語にいたっては 5%であり，3 階層においてはそれぞれのカテゴリが対象とする分野の範囲がかなり細分化されていることを示唆している．

表 6.2 は 1 階層，表 6.3 は 2 階層，表 6.4 は 3 階層における問合せと適合カテゴリの適合度が高い上位 3 カテゴリの適合度を示している．また，表 6.5 は，1 階層，2 階層，3 階層における問合せと適合カテゴリの適合度が高い上位 3 カテゴリの適合度の平均を示している．1 階層と 2 階層を比較すると，1 階層では適合率の平均が 0.001461，2 階層では 0.003745 と顕著な差が出ている．この結果より，1 階層よりも 2 階層のカテゴリを利用して言語横断情報検索における問合せ翻訳を行うことにより，訳語の曖昧性解消の効果がより得られる可能性が高いことを示している．

2 階層と 3 階層を比較した場合，適合率の平均 2 階層では 0.003745，3 階層では 0.003973 であり，3 階層のほうが多少適合率の平均が高くなっているが，1 階層と 2 階層の場合に比べそれほど差が顕著に現れてはいない．この結果からは，2 階層でも 3 階層でも訳語の曖昧性解消の効果もそれほど差が無いように見える．しかし，2 階層と 3 階層とでは選択される適合カテゴリに大きな違いが生じている．今回の実験では，日本語の適合カテゴリに対応する英語のカテゴリの選択を

表 6.2 1 階層までカテゴリを統合した場合の問合せとカテゴリの適合度

問合せ番号	1 番目	2 番目	3 番目
2	0.000122	0.000121	0.000080
5	0.004249	0.003295	0.002482
9	0.000582	0.000520	0.000519
12	0.003355	0.002511	0.002051
13	0.001042	0.000669	0.000482
14	0.001128	0.000962	0.000864
18	0.000379	0.000258	0.000257
19	0.001188	0.000902	0.000887
20	0.000701	0.000492	0.000454
21	0.003107	0.002988	0.002930
23	0.002804	0.001426	0.001267
24	0.001206	0.000369	0.000326
26	0.002291	0.002282	0.002225
27	0.000070	0.000056	0.000040
28	0.004155	0.002714	0.002533
29	0.001886	0.001643	0.001602
31	0.000412	0.000368	0.000364
32	0.002130	0.001823	0.001559
33	0.000029	0.000019	0.000017
34	0.001202	0.000911	0.000617
35	0.003813	0.001405	0.001196
36	0.000578	0.000338	0.000274
37	0.000623	0.000605	0.000598
38	0.000661	0.000490	0.000372
39	0.004306	0.002837	0.001714
42	0.001938	0.001352	0.001197
43	0.001933	0.001893	0.001874
45	0.008369	0.006891	0.006666
46	0.000418	0.000293	0.000225
50	0.000261	0.000230	0.000204

表 6.3 2 階層までカテゴリを統合した場合の問合せとカテゴリの適合度

問合せ番号	1 番目	2 番目	3 番目
2	0.000472	0.000294	0.000278
5	0.005571	0.005571	0.004002
9	0.007014	0.007014	0.004433
12	0.011697	0.009799	0.005489
13	0.001394	0.001394	0.001293
14	0.004931	0.001355	0.001242
18	0.001323	0.000810	0.000810
19	0.002093	0.001577	0.001544
20	0.003166	0.001648	0.001127
21	0.009766	0.008791	0.008791
23	0.004482	0.004482	0.003500
24	0.001396	0.001271	0.000634
26	0.003262	0.003210	0.003201
27	0.000289	0.000256	0.000224
28	0.008614	0.007384	0.005524
29	0.009801	0.006129	0.006129
31	0.001069	0.000708	0.000697
32	0.003722	0.002761	0.002746
33	0.000849	0.000782	0.000198
34	0.002257	0.002257	0.002257
35	0.012377	0.005046	0.004335
36	0.001208	0.001208	0.000879
37	0.001289	0.001233	0.001106
38	0.001548	0.001251	0.000996
39	0.011530	0.009531	0.003791
42	0.002851	0.002647	0.002637
43	0.005964	0.005964	0.005506
45	0.014636	0.014636	0.011142
46	0.000501	0.000484	0.000429
50	0.004781	0.001439	0.001356

表 6.4 3 階層までカテゴリを統合した場合の問合せとカテゴリの適合度

問合せ番号	1 番目	2 番目	3 番目
2	0.000472	0.000472	0.000409
5	0.007570	0.006369	0.003394
9	0.004433	0.001446	0.000692
12	0.015418	0.012305	0.012056
13	0.004632	0.001853	0.001293
14	0.004931	0.004931	0.004330
18	0.000961	0.000810	0.000810
19	0.007264	0.002105	0.001465
20	0.001644	0.001097	0.000695
21	0.012070	0.008791	0.008791
23	0.002856	0.002856	0.002856
24	0.001601	0.001328	0.000634
26	0.006706	0.003696	0.003349
27	0.000468	0.000289	0.000289
28	0.005524	0.005524	0.005318
29	0.006129	0.006129	0.003645
31	0.001818	0.000969	0.000727
32	0.007266	0.005206	0.003700
33	0.000096	0.000068	0.000056
34	0.004511	0.002257	0.002257
35	0.012377	0.012377	0.003250
36	0.001208	0.001208	0.000620
37	0.001100	0.001016	0.000942
38	0.001744	0.001251	0.000930
39	0.005495	0.005329	0.003833
42	0.004351	0.002637	0.002637
43	0.005964	0.005964	0.005219
45	0.023967	0.014636	0.014636
46	0.000542	0.000542	0.000416
50	0.000745	0.000490	0.000484

表 6.5 問合せと適合カテゴリの適合率の平均

	1 階層	2 階層	3 階層
平均適合率	0.001461	0.003745	0.003973

手動で行ったが，適合カテゴリに対応する英語のカテゴリが見つからない場合はその適合カテゴリを利用することを断念し，次善のカテゴリを適合カテゴリとして選択した．このように適合カテゴリに対応する英語のカテゴリが見つからない原因として，以下の二点が考えられる．

- Web ディレクトリの言語版によってカテゴリの構造が異なるため．
- 特徴語数が 10,000 語未満のカテゴリを対象から排除したため．

前者の問題は，2 階層，3 階層のいずれにおいても起こりうる．しかし，階層が深くなれば深くなるほど言語によるカテゴリ構造の違いは顕著になる．実際 Yahoo! カテゴリでは，最上位の 13 カテゴリはどの言語版であっても全く同じ構造をとっている．2 階層目からは各言語版で独自にカテゴリ構造を設計しているため，カテゴリの階層が深くなるほど言語によるカテゴリ構造の違いが顕著となっている．それゆえ，2 階層に比べて 3 階層のほうが適合カテゴリに対応する英語のカテゴリが存在しない可能性が高くなる．

また，後者の問題においては，カテゴリが細分化されればされるほどそのカテゴリに属する Web 文書が少なくなり，特徴語数が十分に得られない可能性も高くなる．表 6.1 でも示したように，階層が深くなるほど除外されるのカテゴリ数が増加している．よって 3 階層では適合カテゴリに対応する英語のカテゴリが除外されている可能性が 2 階層の場合よりも高くなっている．

上記の二つの原因のいずれにおいても，3 階層のほうが適合カテゴリに対応する英語のカテゴリが見つからない可能性が高くなっている．表 6.6 は，実際に適合カテゴリの上位何番目が選択されたかを示している．また，表 6.7 は選択された適合カテゴリの順位の平均を示している．これらの結果も 3 階層のほうが適合カテゴリに対応する英語のカテゴリが見つからないことを示している．

さらに，2階層では日本語の適合カテゴリが問合せに対してほとんど適切であった．適切でない場合は2件のみであった．しかし，3階層の場合では日本語の適合カテゴリの選択の時点で問合せと適合していないと思われる場合が見うけられた．全体の半数近くで，問合せと適合しているとは言い切れない，あるいは完全に適合していない場合が起こった．

以上の点を踏まえると，2階層までカテゴリの統合を行い，これらのカテゴリを言語横断情報検索の問合せの翻訳における訳語の曖昧性解消に用いることが最も有効であると考えられる．

6.1.2 2階層まで利用した検索実験

本節では，カテゴリの細分化の程度を2階層にした場合について言語横断情報検索の検索実験を行った．本実験では，カテゴリの細分化の程度を1階層にした場合と2階層にした場合において比較を行った．1 階層の場合は，Yahoo! カテゴリのトップページにリンクされているカテゴリを用い，2 階層の場合はさらにもう 1 階層下のカテゴリまでを用いている．実験環境は，4.2 と全く同じものを使用している．

表6.8は，提案手法において利用する Web ディレクトリのカテゴリの階層を 1 階層の場合と2階層にした場合についての検索結果である．11 点平均適合率においては，1階層の場合に比べて2階層のほうが0.0031ポイント向上している．また，表6.9 および表6.10 には，1 階層と 2 階層で差異の見られた問合せにおいて，それぞれの階層の場合で選択した訳語のリストを示している．

表6.8 を見ればわかるように，2階層の場合で Web ディレクトリの階層構造を利用して検索対象となる Web 文書の分野を限定することで，11 点平均適合率 (以下，平均適合率と呼ぶ) に変化が見られる問合せが 31 個の問合せ中 11 個存在する．このうち，提案手法を適用することで平均適合率が高くなった問合せは9個，逆に低くなった問合せは2個であった．

平均適合率が高くなった問合せの訳語数に着目すると，訳語数が多くなった問合せは3個，少なくなった問合せは6個であった．訳語数が多くなった問合せでは，問 34 (import → import, importation) や問 50 (fashion → fashion, fashinable

表 6.6 2 階層および 3 階層における選択された適合カテゴリ

	2 階層			3 階層		
問合せ番号	1 番目	2 番目	3 番目	1 番目	2 番目	3 番目
2	1	2	3	2	3	5
5	1	2	5	1	2	3
9	1	2	3	1	2	7
12	1	4	5	1	2	3
13	1	2	5	1	2	4
14	1	3	5	1	2	3
18	1	2	3	1	2	3
19	1	3	4	1	2	3
20	1	2	3	3	4	5
21	2	4	5	2	3	4
23	1	2	3	1	2	3
24	1	3	4	1	2	3
26	2	3	4	1	2	3
27	2	3	4	1	2	3
28	1	2	5	3	8	9
29	1	2	3	2	3	4
31	1	2	3	1	2	4
32	2	3	5	1	2	3
33	1	2	4	2	4	6
34	2	3	4	1	3	4
35	1	3	4	1	2	4
36	1	2	3	1	2	3
37	3	5	6	5	6	7
38	1	2	3	1	2	3
39	1	2	4	1	2	3
42	1	2	4	1	2	3
43	1	2	3	1	2	4
45	2	3	5	1	2	3
46	4	6	9	2	7	10
50	1	2	3	2	3	4

表 6.7 選択された適合カテゴリの順位の平均

	2 階層	3 階層
順位の平均	2.722	2.822
標準偏差	1.512	1.852

closes, vogue) のように派生語や類義語が多く得られる傾向があった．また，その逆の訳語数が少なくなった問合せでは，問 9 に見られるように誤訳 (「衛星」の訳語として “dependency”) が無くなったり，問 20 (joint business, cooperation → cooperation) のように，問合せが対象としている分野にふさわしい訳語を選択できるなどの効果が見てとれる．したがって，提案手法の Web ディレクトリの階層構造を利用することは，訳語数の変化につながり，結果的に平均適合率の向上につながったと考えることができる．

これに対し，平均適合率が低くなった二つの問合せの訳語に着目すると，問 19 の「経済的」という訳語が “economic” から，“economic” と “economical” の二語に増えていたり，問 28 では「派遣」という訳語 “dispatch” が欠落したりという現象が起こっている．前者は Web ディレクトリの階層構造を利用することによる検索対象 Web 文書の分野の限定がうまく働いた例，後者は分野の限定がうまく機能せず訳語を抽出できなかった例であるということができる．前者は正確な訳語を得ることとしては意図した結果となったのであるが，検索語としてはあまり適切でない訳語が選択される結果となった．1 階層では特徴語集合に含まれない訳語であるが，分野を限定することにより，その訳語の重要度が相対的に上がり，2 階層では訳語として選択されることとなったと考えられる．

一方，平均適合率に変化が見られなかった問合せも 20 存在していた．これは，Web ディレクトリの階層構造を利用しても訳語に違いが見られなかったためである．これは，2 階層目のカテゴリであっても対象とする分野が十分に限定できていないため，問合せが対象としている分野に適切な訳語を選択する効果が得られていないからであると考えられる．このことはさらに低い階層のカテゴリを利用することにより，検索対象とする Web 文書の分野をより限定し，適切な訳語を

表 6.8 各問合せごとの 11 点平均適合率

問題番号	1 階層	2 階層
2	0.0406	0.0419
4	0.0472	0.0472
5	0.0433	0.0433
9	0.1079	0.1086
12	0.0112	0.0112
13	0.0118	0.0118
14	0.1996	0.1996
18	0.0000	0.0000
19	0.1133	0.0757
20	0.0240	0.0485
21	0.0112	0.0112
23	0.1043	0.1390
24	0.0070	0.0169
26	0.0079	0.0109
27	0.6743	0.6821
28	0.1945	0.1446
29	0.0832	0.0832
31	0.0919	0.0919
32	0.0053	0.0053
33	0.2854	0.2854
34	0.0477	0.0833
35	0.0191	0.0191
36	0.3693	0.3693
37	0.3221	0.3221
38	0.0105	0.0105
39	0.0598	0.0598
42	0.0067	0.0067
43	0.0003	0.0003
45	0.0009	0.0009
46	0.0168	0.0168
50	0.0756	0.1407
平均	0.0965	0.0996

表 6.9 各問合せごとの訳語候補 (その 1)

問題番号	1 階層の訳語	2 階層の訳語
9	human labor	human labor
	human skill	human skill
	artificial	artificial
	artificial heart	artificial heart
	artificial satellite	artificial satellite
	satellite	satellite
	moon	moon
	secondary	secondary
	dependency	
	ST1	ST1
19	Europe	Europe
	currency	currency
	money	money
	synthesis	synthesis
	integration	integration
	economic	economic
		economical
	influence	influence
	effect	effect
	consequence	consequence

表 6.10 各問合せごとの訳語候補 (その 2)

問題番号	1 階層の訳語	2 階層の訳語
20	Nissan	Nissan
	Renault	Renault
	funds	
	capital	capital
	fund	fund
	investment money	investment money
	joint business	
	cooperation	cooperation
28	Japan	Japan
	North Korea	North Korea
	visit	visit
	call	call
	dispatch	
	send	send
34	rice	rice
	meter	meter
	American	American
	America	America
	import	import
		importation
	introduction	introduction
50		fashionable clothes
		vogue
	fashion	fashion
	mode	mode
	style	style

選択できる可能性が残されていることを意味する．この場合，訳語の曖昧性解消に Web ディレクトリのどの階層のカテゴリまでを利用するのかさらに精査が必要となる．しかしそのためには，3 階層以下のカテゴリも利用する必要があるが，各階層のカテゴリを画一的に扱うと逆に検索精度が低下してしまう．それを避けるには一つ一つのカテゴリを精査する必要があるが，階層が深くなればなるほど精査する必要のあるカテゴリも増大するため，現状ではあまり現実的ではないと思われる．

7. 現状と今後の課題

7.1 本研究により得られた知見

多言語情報アクセスシステムを実現するためには，特定の言語しか理解できない利用者，例えば母国語しか理解できないような利用者であっても言語の違いを意識することなくシステムを利用できるようにする必要がある．言語横断情報検索の研究は，それを実現するための手段の一つとして挙げられる．言語横断情報検索では，問合せを翻訳することにより言語の違いを吸収することが多いが，この際に問題となる訳語の曖昧性解消を行うことが重要である．一般的にはコーパスを用いて訳語の曖昧性解消を行うことが多いが，Web 文書の検索などのように検索対象の分野が広範囲にわたる場合は，コーパスの対象分野との不一致により十分に訳語の曖昧性解消が行うことができない可能性がある．

この問題を解消するために，本研究では特定の分野に限定されない言語資源の構築および言語資源の活用について研究を行った．言語資源の構築については 2 言語オントロジの構築，言語資源の活用については Web ディレクトリの言語資源としての活用を行った．

2 言語オントロジの構築の目的は，事前に曖昧性解消が行われている対訳辞書として利用することであった．このような対訳辞書が活用できれば，訳語の曖昧性の問題が解決できるからである．本研究では，ある言語で構築されたオントロジを翻訳することで 2 言語オントロジの構築を試みた．オントロジの各概念の単語を翻訳する際，訳語候補のうちからその概念を表現している訳語候補のみを抽出する必要がある．しかしながら，実際に 2 言語オントロジを構築しようとすると，その概念を表現していない訳語候補も抽出される結果となった．特定の概念を表現している訳語候補のみを抽出する際に Web ディレクトリを利用したのであるが，このときに利用したのは Web ディレクトリの上位の階層のみであった．しかし，それぞれの概念が表現する意義は非常に細分化されているため，十分に訳語候補の分類が十分に行うことができなかったことが失敗の一因であったことが分かった．また，言語によって語彙体系が異なるためオントロジの構造も異なることが，2 言語オントロジ構築が困難である要因である．

Web ディレクトリの言語資源としての活用は，従来のコーパスによる訳語の曖昧性解消手法が抱えていた，コーパスと問合せとの対象分野の不一致の問題を解消することが目的であった．Web ディレクトリを言語資源として活用した言語横断情報検索システムを構築し評価した結果，Web ディレクトリが，Web 文書が対象とする分野を網羅した言語資源として活用できることが示された．また，問合せが対象とする分野を推定し，その分野により関連のある訳語候補を訳語として選択することが有効であることが分かった．さらに，問合せと最も適合しているカテゴリを利用しただけでは訳語が得られない場合でも，次善のカテゴリを利用することで訳語が得られる可能性が高くなることが分かった．また，Web ディレクトリのカテゴリ構造を利用してより深い階層のカテゴリを用いることにより，問合せが対象とする分野をより特定できることが分かった．これにより，訳語の曖昧性解消の精度が向上し，言語横断情報検索の性能も上がることが確認された．さらに，利用する階層は深すぎると問題が多くなるため，最上位から 2 階層までのカテゴリを利用するとよいことも分かった．

7.2 今後の課題

多言語情報アクセスのための言語資源として理想的なのは，訳語の曖昧性が生じないことである．究極的には，問合せの単語群が決定されればそれに対する訳語群も一意に決定できればよい．しかし，問合せ語が 2 単語の場合だけでもその組合せは単語数の 2 乗のオーダーとなり，問合せ語が 3 単語以上の場合も考慮するとその組合せは無数に存在する．それゆえ，そのような言語資源の構築は不可能である．また，本研究でも述べたように，言語によって語彙体系が異なるため，訳語の曖昧性が生じないような言語資源の構築は非常に困難である．

よって次善の方法を検討することになるが，現段階で有力な方法として，特定の分野において適切な訳語を選択することが挙げられる．検索対象が対象とするあらゆる分野のそれぞれにおいて，訳語の曖昧性解消を行うのに十分な統計情報が得られるような言語資源が構築できれば，この方法を実現するための言語資源として理想的である．本研究で提案した Web ディレクトリを言語資源として活用することは，この理想を目標としていたのであるが，残念ながらそこまでには

いたっていない．また，一般のコーパスに比べ，Web ディレクトリにはノイズとなる情報が含まれていることが多いことも，検討すべき問題である．よって，言語資源を構築する際には，悪影響を及ぼす情報を可能な限り排除することにも注意を払う必要がある．それにより，より精度の高い言語資源の構築が可能となる．

謝辞

本研究を行うにあたり，懇切丁寧なご指導を賜りました植村 俊亮教授に心より感謝の意を表し，厚く御礼申し上げます．ご多忙にも関わらず，本研究の主指導教官になっていただき，研究の全過程において的確なご助言をいただきました．また，研究発表に関する旅費等のサポートなど，研究指導以外においても格別のご配慮を賜りましたことを心より感謝いたします．

本研究の副指導教官になっていただきました関 浩之教授に深く御礼申し上げます．関教授には本論文の草稿に対して多数のコメントをいただき，誠に感謝いたしております．

本研究の副指導教官になっていただきました宮崎 純助教授に深く御礼申し上げます．宮崎助教授には，研究ミーティング等において厳しい意見をいただくことも多かったのですが，それが本研究の成果を得るための糧となりました．誠に感謝いたしております．

本研究に対して有益なご助言，ご協力をいただきました京都大学の吉川 正俊教授に深く御礼申し上げます．吉川教授には，本研究の初期よりご助言をいただき，本研究を行う上での基礎となりました．心より御礼申し上げます．

本研究に対して有益なご助言，ご協力をいただきました立命館大学の前田 亮助教授に深く御礼申し上げます．前田助教授には，言語横断情報検索について初歩から懇切丁寧にご指導していただきました．本研究の初期から本論文の執筆にいたるまで，全過程においてご指導，ご協力いただきました．心より御礼申し上げます．

本研究に対して有益なご助言，ご協力をいただきました同志社大学の波多野 賢治専任講師に深く御礼申し上げます．波多野専任講師には，本研究に対するご助言，ご協力をいただくとともに，研究の遂行にあたり多大な支援を賜りました．誠に感謝いたしております．

本研究に対して有益なご助言，ご協力をいただきました筑波大学の天笠 俊之講師に深く御礼申し上げます．天笠講師には，研究室の全体研究会等において貴重なご意見を数多く賜りました．心より御礼申し上げます．

ここに挙げさせていただきました教員のみなさまには，研究内容に対するご指

導はもとより，研究者として必要なことを数多く学ばせていただきました．重ねて御礼申し上げます．

本研究に対するご助言，ご協力いただいた奈良先端科学技術大学院大学情報科学研究科マルチメディア統合システム講座，奈良先端科学技術大学院大学情報科学研究科バイオ情報学領域データベース学分野研究室の学生，秘書の皆様に感謝申し上げます．私が本研究室に在籍している間に様々な方々にご協力いただいたことにより，本研究を遂行することができました．研究以外においても多大な支援を賜り，誠に感謝しております．心より御礼申し上げます．

幸運にも上記のような恵まれた環境において本研究に取り組むことができたのですが，そのきっかけを与えてくださった大阪教育大学の永田 元康教授に深く御礼申し上げます．永田教授には，私が学部および修士において情報工学の基礎知識を授けていただきました．また私にとって最初の研究指導を賜り，研究者としての基礎を学ばせていただきましたことを深く感謝いたします．

最後に，最も永きにわたって暖かく見守り続けてくれた両親に感謝します．

参考文献

[1] L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. the 21st ACM SIGIR conference (SIGIR ' 98), 1997.

[2] S. Bechhofer, F. V. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. OWL Web Ontology Language Reference, 2003.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web, 2001.

[4] A. Bharati, K. Varanasi, C. Kamisetty, R. Sangal, and S.M. Bendre. A Document Space Model for Automated Text Classification based on Frequency Distribution across Categories. In *Published in the proceedings of ICON2002*, pp. 18–21, 2002.

[5] P. Borst, H. Akkermans, and J. Top. Engineering Ontologies. *International Journal of Human-Computer Studies*, Vol. 46, No. 2/3, pp. 365–406, 1997.

[6] B. Chandrasekaran, A. K. Goel, and Y. Iwasaki. Functional Representation as Design Rationale. *COMPUTER*, pp. 48–56, 1993.

[7] L. R. Chapman. *Roget's International Thesaurus (Fourth Edition)*. Harper & Row, 1999.

[8] K-H. Chen, H-H. Chen, N. Kando, K. Kuriyama, Lee S., S.H. Myaeng, K. Kishida, K. Eguchi, and Kim H. Overview of CLIR Task at the Third NTCIR Workshop. *Working Notes of the Third NTCIR Workshop Meeting*, pp. 1–38, 2002.

[9] CLEF (Cross Language Evaluation Forum). http://clef.iei.pi.cnr.it/.

[10] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Col-

lections. In *the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 318–329, 1992.

[11] M. Davis. New experiments in cross-language text retrieval at NMSU's computing research lab. In *the fifth text retrieval conference (TREC-5)*, 1996.

[12] S. Deerwester, S. T. Dumais, T. K. Furnas, G. W.and Landauer, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, Vol. 46, No. 1, pp. 391–407, 1990.

[13] S. T. Dumais. Using LSI for information filtering: TREC-3 experiments. In *The Third Text Retrieval Conference (TREC3)*, pp. 219–230, 1995.

[14] D. Eichmann, M. E. Ruiz, and P. Srinivasan. Cross-language information retrieval with the UMLS metathesaurus. the 21st ACM SIGIR conference (SIGIR ' 98), 1998.

[15] W. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms, chapter 7*. Prentice-Hall, 1992.

[16] 藤井敦. 言語横断情報検索への入門 – 翻訳と検索の統合が生み出す新たな可能性–. 情報処理, Vol. 42, No. 3, pp. 327–329, 2001.

[17] A. Fujii and T. Ishikawa. Cross-Lauguage Information Retrieval ad ULIS, 1999.

[18] J. Gao, M. Zhou, J.-Y. Nie, H. He, and W. Chen. Cross-language information retrieval: resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relation. 25th ACM SIGIR conference (SIGIR ' 01), 2001.

[19] E. Glover, D.M. Pennock, S. Lawrence, and R. Krovetz. Inferring hierarchical descriptions. In *Proc. 11th International Conference on Information and Knowledge Management (CIKM'02)*, pp. 507–514, 2002.

[20] J. Gonzalo, F. Verdejo, C. Peters, and N. Calzolari. Applying EuroWordNet to cross-language text retrieval. *Computers and the Humanities*, Vol. 32, pp. 185–207, 1998.

[21] Google. Google. http://google.co.jp.

[22] G. Grefenstette. *Cross-Language Information Retrieval*. Kluwer Academic Publishers, 1998.

[23] Miniwatts Marketing Group. Internet World Stats. http://www.internetworldstats.com/stats7.htm.

[24] 林良彦, 横尾昭男, 古瀬蔵. 多言語情報アクセスシステム. NTT 技術ジャーナル, Vol. 14, No. 1, pp. 76–80, 2002.

[25] 林良彦, 松尾義博, 永田昌明, 古瀬蔵. クロス言語情報検索と多言語情報アクセスシステム. *NTT R&D*, Vol. 52, No. 2, pp. 92–99, 2003.

[26] M.A. Hearst. Automatic Acuisition of Hyponyms from Large Text Corpora. In *14th International Conference on Computational Linguistics*, pp. 539–545, 1992.

[27] D. A. Hull and G. Grefenstette. Querying across languages: a dictionary-based approach to multilingual information retrieval. the 19th ACM SIGIR conference (SIGIR ' 96), 1996.

[28] D. A. Hull. Using structured queries for disambiguation in cross-language information retrieval. *Electronic Working Notes of the AAAI Symposium on Cross-Language Text and Speech Retrieval*, March 1997.

[29] 岩山真, 徳永健伸. 確率的クラスタリングを用いた文書連想検索. 自然言語処理, Vol. 5, No. 1, pp. 101–118, 1998.

[30] M. G. Jang, S. H. Myaeng, and S. H. Park. Using mutual information to resolve query translation ambiguities and query term weighting. In *the 37th annual meeting of the association for computational linguistics*, 1999.

[31] B. J. Jansen, A. Spink, and T. Saracevic. Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web. *Information Processing and Management*, Vol. 36, No. 2, pp. 207–227, Jan. 2000.

[32] K. Kishida and N. Kando. A Hybrid Approach to Query and Document Translation Using a Pivot Language for Cross-Language Information Retrieval. In *Working Notes for the CLEF 2005 Workshop*, pp. 93–101, Sep. 2005.

[33] K. Kishida, N. Kando, and K.-H. Chen. Two-Stage Refinement of Transitive Query Translation with English Disambiguation for Cross-Language Information Retrieval: A Trial at CLEF 2004. In *Working Notes for the CLEF 2004 Workshop*, pp. 135–142, Sep. 2004.

[34] 來村徳信, 溝口理一郎. オントロジー工学に基づく機能的知識体系化の枠組み. 人工知能学会誌, Vol. 17, No. 1, pp. 61–72, 2002.

[35] Information Knowledge and Data Processing Group. WordNet.OWL. http://taurus.unine.ch/GroupHome/knowler/wordnet.html.

[36] Y. Ko and J. Seo. Automatic Text Categorization by Unsupervised Learning. In *the 17th conference on Computational linguistics(COLING-2000)*, pp. 453–459, 2000.

[37] T. Kohonen, S. Kaski, Krista. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive document colletion. *IEEE Transactions on Neural Networks*, Vol. 11, pp. 574–585, 2000.

[38] 国分智晴, 田中崇, 森辰則. 空間分割型 CL-LSI による大規模言語横断情報検索. 情報処理学会論文誌: データベース, Vol. 43, No. SIG 2, pp. 27–36, 2002.

[39] K.-S. Lee, K. Kageura, and K.-S. Choi. Implicit ambiguity resolution using incremental clustering in cross-language information retrieval. *Information Processing and Management*, Vol. 40, pp. 145–159, 2004.

[40] C.-J. Lin, W.-C. Lin, G.-W. Bian, and H.-H. Chen. Description of the NTU Japanese-English cross-lingual information retrieval system used for NTCIR workshop. *First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 145–148, August 1999.

[41] M. L. Littman, S. T. Dumais, and T. K. Landauer. *Automatic cross-language information retrieval using latent semantic indexing*, pp. 51–62. Kluwer Academic Publishers, 1998.

[42] 前田亮, 吉川正俊, 植村俊亮. 言語横断情報検索における Web 文書群による訳語曖昧性解消. 情報処理学会論文誌：データベース, Vol. 41, No. SIG 6 (TOD 7), pp. 12–21, 2000.

[43] A. Maeda, F. Sadat, M. Yoshikawa, and S. Uemura. Query Term Disambiguation for Web Cross-Language Information Retrieval using a Search Engine. *The 5th International Workshop on Information Retrieval with Asian Languages (IRAL2000)*, pp. 25–32, 2000.

[44] G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

[45] 溝口理一郎. オントロジー研究の基礎と応用. 人工知能学会誌, Vol. 14, No. 6, pp. 45–56, 1999.

[46] 中村朋建, 上土井陽子, 若林真一, 吉田典可. クラスタリング結果の特徴抽出を用いる高次元データの対話的クラスタリング. 情報処理学会論文誌: データベース, Vol. 47, No. SIG 19 (TOD 32), pp. 28–41, 2006.

[47] 中野洋. 「分類語彙表」形式による語彙分類表（増補版）, 1996.

[48] National Institute of Informatics. the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering, 2003.

[49] 小熊淳一, 内海彰. 語の共起情報を用いた文書クラスタリング. 人工知能学会第 19 回全国大会論文集, 2005.

[50] 大野晋, 浜西正人. 角川類語新辞典, 1981.

[51] 奥村明俊, 石川開, 佐藤研治. コンパラブルコーパスと対訳辞書による日英クロス言語検索. 自然言語処理, Vol. 5, No. 4, pp. 77–98, Oct. 1998.

[52] Global Reach. Global Internet Statistics (by Language) . http://global-reach.biz/globstats/index.php3.

[53] B. Rehder, M. L. Littman, S. Dumais, and T. K. Landauer. Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing. *the sixth text retrieval conference (TREC-6)*, pp. 233 –239, 1997.

[54] P. Resnik and N. A. Smith. The Web as a Parallel Corpus. *Computational Linguistics*, Vol. 29, No. 3, pp. 349–380, 2003.

[55] S. E. Robertson. On term selection for query expantion. *Journal of Documentation*, Vol. 46, No. 4, pp. 359–364, 1990.

[56] T. Sakai, T. Manabe, A. Kumano, M. Koyama, and T. Kokubu. Toshiba BRIDJE at NTCIR-5 CLIR: Evaluation using Geometric Means. In *NTCIR-5 Workshop Meeting*, pp. 56–63, Dec. 2005.

[57] G. Salton and M. J. McGill. *Instruction to Modern Information Retrieval*. McGraw-Hill, 1983.

[58] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *22nd ACM SIGIR Conference (SIGIR'99)*, pp. 206–213, 1999.

[59] H.-C. Seo, S.-B. Kim, H.-C. Rim, and S.-H. Myaeng. Improving Query Translation in English-Korean Cross-Language Information Retrieval. *Information Processing and Management*, Vol. 41, No. 3, pp. 507–522, May 2005.

[60] K. Shinzato and K. Torisawa. A Simple WWW-based Method for Semantic Word Class Acquisition. In *Recent Advance in Natural Language Processing (RANLP05)*, pp. 493–500, 2004.

[61] K. Shinzato and K. Torisawa. Acquiring Hyponymy Relations from Web Documents. In *Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL04)*, pp. 73–80, 2004.

[62] 徳永健伸. 情報検索と言語処理. 東京大学出版会, 1999.

[63] 梅田靖, 冨山哲男, 吉川弘之. 機能設計支援のための FSB モデリングの提案. 精密工学学会誌, Vol. 63, No. 6, pp. 795–800, 1997.

[64] International Telocommunication Union. Free statistics – Maps and graphs –. http://www.itu.int/ITU-D/ict/statistics/maps.html.

[65] C. J. van Rijsbergen. Further experiments with hierarchic clustering in document retrieval. *Information Storage and Retrieval*, Vol. 10, pp. 1–14, 1974.

[66] 若木裕美, 正田備也, 高須淳宏, 安達淳. 検索語の曖昧性解消のためのトピック指向単語抽出および単語クラスタリング. 情報処理学会論文誌: データベース, Vol. 47, No. SIG 19 (TOD 32), pp. 72–85, 2006.

[67] P. Willett. Recent trends in hierarchic document clustering: A cretical review. *Information Processing & Management*, Vol. 24, No. 5, pp. 577–597, 1988.

[68] WordNet. http://www.cogsci.princeton.edu/ wn/.

[69] Y. Yang, J. G. Carbonell, R. D. Brown, and R. E. Frederking. Translingual information retrieval: learning from bilingual corpora. *AI Journal Special Issue*, pp. 323–345, 1998.

[70] 酒井哲也, 梶浦正浩, 住田一男, G. Jones, N. Collier. 機械翻訳を用いた英日・日英言語横断検索に関する一考察. 情報処理学会論文誌, Vol. 40, No. 11, pp. 4075–4086, nov 1999.

[71] 神崎正英. http://www.kanzaki.com/docs/sw/jwebont.html.

研究業績

学術論文

1. 木村 文則, 前田 亮, 宮崎 純, 吉川 正俊, 植村 俊亮: “Web ディレクトリを言語資源として利用した言語横断情報検索”, 情報処理学会論文誌: データベース, Vol. 45, No. SIG 7 (TOD 22), pp. 208-217, 2004 年 6 月.

国際会議 (査読あり)

1. Fuminori Kimura, Akira Maeda, Masatoshi Yoshikawa, and Shunsuke Uemura: “Cross-Language Information Retrieval Based on Category Matching between Language Versions of a Web Directory”, The 6th International Workshop on Information Retrieval with Asian Languages (IRAL2003) in conjunction with the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), pp.153-159, Sapporo, Japan, July 7, 2003.

2. Fuminori Kimura, Akira Maeda, Masatoshi Yoshikawa and Shunsuke Uemura: “Cross-Language Information Retrieval using Web Directories”, IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM ’03), pp. 911-914, Victoria, B.C. Canada, August 28-30, 2003.

3. Fuminori Kimura, Akira Maeda, Jun Miyazaki, Masatoshi Yoshikawa and Shunsuke Uemura: “Cross-Language Information Retrieval Using Web Directories as a Linguistic Resource”, the first Asia Information Retrieval Symposium (AIRS2004), pp.297-300, Beijing, China, October 18-20, 2004.

4. Fuminori Kimura, Akira Maeda, Jun Miyazaki, and Shunsuke Uemura: “Query Disambiguation for Cross-Language Information Retrieval Using Web Directories”, International Workshop on Challenges in Web Information Retrieval and Integration (WIRI2005), pp. 154-159, Tokyo, Japan, April 2005.

国際会議 (査読なし)

1. Fuminori Kimura, Akira Maeda, and Shunsuke Uemura: “CLIR using Web Directory at NTCIR4”, the 4th NTCIR Workshop Meeting of Evaluation of Information Retrieval, Question Answering, and Summarization, pp.123-127, June 2004.

国内発表

1. 木村文則, 前田亮, 吉川正俊, 植村俊亮. “ディレクトリ型検索エンジンのカテゴリ間対応付けによる言語横断検索”， 電子情報通信学会第 13 回データ工学ワークショップ (DEWS2002), 2002 年 3 月.

2. 木村文則, 前田亮, 吉川正俊, 植村俊亮. “ディレクトリ型検索エンジンを利用した言語横断情報検索”， 第 1 回情報科学技術フォーラム論文集, 第 2 分冊 pp.69-70, 2002 年 9 月.

3. 木村 文則，前田 亮，吉川 正俊，植村 俊亮 “Web ディレクトリを利用した言語横断情報検索における特徴語抽出”, 「情報アクセスのためのテキスト処理」シンポジウム， pp.1-8, 2003 年 2 月.

4. 木村 文則，前田 亮，吉川 正俊，植村 俊亮 “Web ディレクトリの階層構造を利用した言語横断情報検索”， 電子情報通信学会第 14 回データ工学ワークショップ (DEWS2003)，2003 年 3 月.

5. 木村 文則, 前田 亮, 越田 高志, 宮崎 純, 植村 俊亮: “Web ディレクトリを用いた 2 言語オントロジーの構築”, 第 24 回ディジタル図書館ワークショップ, No.24, pp.3-10, 情報処理学会研究報告, No.73, 2003-FI-73, pp.25-32, 2003 年 11 月.

6. 木村 文則, 前田　亮, 宮崎　純, 吉川 正俊, 植村 俊亮: “言語横断情報検索における Web ディレクトリを利用した訳語の曖昧性解消”, 第 3 回情報科学技術フォーラム (FIT2004), 2004 年 9 月.

レターズ

1. 木村 文則, 前田 亮, 吉川 正俊, 植村 俊亮: “Web ディレクトリの階層構造を利用した言語横断情報検索”, 日本データベース学会 Letters, Vol.2, No.1, pp.71-74, 2003 年 5 月.