OUKA
The University of Osaka
Institutional Knowledge Archive

Title	オフショアソフトウェア開発における中国語・日本語混在環境の知識共有に関する研究
Author(s)	蔡, 立
Citation	大阪大学, 2015, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/53940
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

オフショアソフトウェア開発における
中国語・日本語混在環境の知識共有に関する研究

提出先　　大阪大学大学院情報科学研究科

提出年月　2015 年 3 月

蔡　立

研究業績目録

A. 学術論文誌論文

1. L. Cai, Z. Wang, Y. Jiao, M. Akiyoshi, and N. Komoda, "Prototype of Knowledge Management System in Chinese Offshore Software Development Company", *WSEAS Transactions on Information Science & Applications*, Vol. 5, Issue 3, pp. 252–257 (2008).

2. L. Cai, Z. Wang, M. Akiyoshi, and N. Komoda, "BBS-based Information Management System in Chinese Offshore Software Development Company", *Intl. Journal of Systems Applications, Engineering & Development*, Vol.5, Issue 1, pp. 50–57 (2011).

3. L. Cai, Z. Wang, M. Akiyoshi, and N. Komoda, "A Knowledge Cards Classification Method with Conversion Loss Correction for Incomplete Translation Dictionary", *IEEJ Transactions on Electrical and Electronic Engineering*, Vol. 6, No. 6, pp. 566–570 (2011).

B. 国際会議

1. L. Cai, Z. Wang, Y. Jiao, M. Akiyoshi, and N. Komoda, "A Knowledge Sharing and Managing System for Offshore Software Development Company", in *Proc. of the 7th WSEAS Intl. Conf. on Applied Computer Science(ACS'07)*, pp. 340–345 (2007).

2. L. Cai, M. Akiyoshi, and N. Komoda, "A Classification Method of Knowledge Cards Represented in Japanese and Chinese at Offshore Software

Development Company", in *Proc. of the IADIS Intl. Conf. Applied Computing 2009*, pp. 63–67 (2009).

3. L. Cai, Z. Wang, M. Akiyoshi, and N. Komoda, "A Knowledge Cards Classification Method with Conversion Loss Correction for Incomplete Translation Dictionary", in *Proc. of the 3rd Japan-China Joint Symposium on Information Systems (JCIS2010)*, pp. 85–88 (2010).

4. L. Cai, Z. Wang, M. Akiyoshi, and N. Komoda, "Evaluation of BBS-based Information Management System in Chinese Offshore Software Development Company", in *Proc. of Intl.. Conf. of the Institute for Environment, Engineering, Economics and Applied Mathematics 2010(IEEEAM 2010)*, pp. 580–582 (2010).

5. X. Liu, M. Akiyoshi, N. Komoda, L. Cai, and Z. Wang, "Evaluation of Knowledge Cards Classification Method with Translation Disctionary", in *Proc. of the 4th Japan-China Joint Symposium on Information Systems(JCIS2011)*, pp. 61–64 (2011).

C. 学会講演

1. 蔡立, 曲義暁, 王平, 秋吉政徳, "オフショアソフトウェア開発組織における知識共有システムの構想", 電気学会情報システム研究会, IS-08-10, pp. 47–52 (2008).

2. 蔡立, 秋吉政徳, 薦田憲久, "オフショア開発組織内での中国語と日本語が混在する知識カードの分類管理システム", 電気学会情報システム研究会, IS-09-56, pp. 5–8 (2009).

3. 蔡立, 劉暁鵬, 秋吉政徳, 薦田憲久, "ドメイン辞書を用いた中国語と日本語が混在する知識カードの分類方式", 電気学会情報システム研究会, IS-11-89, pp. 69–73 (2011).

内容梗概

本論文は，筆者が2002年から現在まで済南凌佳科技有限公司，ならびに2007年から2012年まで大阪大学大学院情報科学研究科マルチメディア工学専攻在学中に行ってきたオフショアソフトウェア開発における中国語・日本語混在知識の共有に関する研究成果をまとめたものである．

近年，中国向けのオフショアソフトウェア開発が急速に増えている．このため，中国のオフショアソフトウェア開発企業では，多くの技術者は開発経験が浅く日本語レベルも低い若手が大半を占めており，継続的な技術者教育が必要である．また，情報技術が急速に発展するため，技術者としても勉強し続けることが要求される．このような状況で，教育をしつつスムーズかつ効率的に日常の開発業務を進めるためには，上級技術者の知識を開発組織内で共有するシステムが必要となっている．その際，オフショアソフトウェア開発企業における特有の問題として，日本語と中国語で書かれた知識の混在を許容しなければならず，このことを前提とした知識共有システムが必要となる．

本論文では，このような日本語と中国語が混在したソフトウェア開発作業における知識共有の試みとして，まず社内用の機能を組み入れた日常業務で用いることのできるBBS(Bulletin Board System)形式の知識共有システムを提案する．さらに，技術者の技術的興味の動向を把握するため，システムに蓄積された質問に対する回答のセット(以下，知識カード)に対して，これらの分類方式を提案する．本論文は全5章から構成される．

第1章では，オフショアソフトウェア開発企業の現状と知識共有の必要性について説明し，さらに本研究で取り上げる課題を述べ，関連研究を概観するとともに，本論文の目的と位置づけを明らかにする．

第2章では，オフショアソフトウェア開発企業における知識共有システムとして，筆者が所属する企業のBBS形式の知識共有システムである「凌佳知識共有システム」について述べ，利用者アンケートによる評価を示す．プロジェクトマネージャといった管理職やプログラマを含む技術者158名のアンケート

回答を分析し，特に回答者の 104 名を占めるプログラマの約 62%が日常的に提案する知識共有システムを利用し，開発作業の問題を解決していることが明らかになった．また，経営の視点からは，開発企業としての技術レベルとして「基礎技術」に関する教育・研修をまだまだ強化すべきということも明らかになった．知識共有システムに蓄積される知識カードとして,「日本語の知識カード」,「中国語の知識カード」,「日本語と中国語が混在して用いられた知識カード」があり，これらを活用するためには，知識共有システムに自動分類する仕組みが必要であることが明確化された．

第 3 章では，第 2 章での知識共有システムに蓄積された知識カードの中で，中国語のみで記述された知識カードの分類を目的に，日本語アンケートデータを対象に開発された分類方式を中国語の知識カード向けに拡張した分類方式を提案し，実データによる評価について述べる．中国語の知識カード 100 件をもとに，分類方式の特徴であるパラメータの「変換漏れ補正率」を決めた後に，中国語の知識カード 470 件を分類した結果，68.7%の分類精度が得られた．また,「中国語の知識カード」の分類方式として，“変換漏れ補正” を組み入れない場合には低い分類精度となり，“変換漏れ補正” を組み入れた判定処理が有効であることがわかった．

第 4 章では，第 3 章で提案する知識分類方式を，ひとつの知識カードの中に中国語と日本語が混在する文が含まれている場合にも処理できるように拡張した分類方式を提案し，実データによる評価について述べる．第 3 章と同様に，中国語の知識カード 100 件をもとに，分類方式の特徴であるパラメータの「変換漏れ補正率」を決めた後に，中国語の知識カード 470 件から作成した日本語と中国語の混在カード，人手で翻訳した日本語の知識カードの分類精度を比較し，いずれであってもＦ値が約 70%となる分類精度が得られたことから，同一内容の 3 種類の知識カードを分類する統一処理方式として機能することが明らかになった．

最後に，第 5 章では，本研究で得られた成果を結論としてまとめ，今後の課題を示す．

目 次

第 1 章　序論　1

1.1　研究の背景　. 1

1.2　関連研究　. 4

1.3　本研究における課題の解決方針　. 8

1.4　本論文の構成　. 10

第 2 章　BBS 上の日本語と中国語による知識共有システム　13

2.1　緒言　. 13

2.2　オフショアソフトウェア開発現場での BBS 形式の知識共有システム　. 15

2.2.1　BBS 形式の知識共有システムの概要　. 15

2.2.2　質問受付処理機能　. 17

2.2.3　回答受付処理機能　. 19

2.2.4　検索受付処理機能　. 20

2.3　アンケート調査による評価結果　. 23

2.4　結言　. 27

第 3 章　中国語知識カードの分類方式　29

3.1　緒言　. 29

3.2　カテゴリ辞書を用いた知識カード分類方式　. 30

3.2.1　カテゴリ辞書による分類方式の概要　. 30

3.2.2　中国語知識カードを処理する場合の課題　. 33

3.2.3　翻訳辞書と変換漏れ補正を用いた中国語知識カード分類方式　. 34

3.2.4　*Jaccard* 係数推定処理　. 34

3.3　実験結果ならびに分類精度の評価　. 37

3.3.1　実験条件　. 37

3.3.2 実験結果と評価 38
3.4 結言 41

第 4 章 日本語と中国語が混在する知識カードの分類方式 43
4.1 緒言 43
4.2 日中両言語が混在する知識カードの分類方式 44
4.2.1 形態素解析ツールの適用における課題 44
4.2.2 課題へのアプローチ 46
4.2.3 ドメイン辞書を用いた知識カード分類方式 47
4.3 実験結果ならびに分類精度評価 49
4.3.1 実験条件 49
4.3.2 実験結果と評価 50
4.4 結言 51

第 5 章 結論 53
5.1 本研究のまとめ 53
5.2 今後の研究課題 55

謝辞 57

参考文献 59

第1章

序論

1.1 研究の背景

近年の情報システム開発の規模の増大は，開発費用の増大という課題に直面し，この解決策の一つとして，これまでの国内ソフトウェアハウスへのアウトソーシングから，海外の人件費の安いソフトウェア開発企業へのアウトソーシングという開発体制を生み出してきた．このような開発体制は，オフショア開発と呼ばれ，1980 年代後半から実験的に始まり，韓国への委託を皮切りに中国，インドへと広がり，現在ではベトナムなどの東南アジア諸国，ロシアなどの多くの国へと展開されている．また，オフショア開発の実績を持つ日本企業として，オフショア開発の規模を拡大したい企業の割合も 2009 年の 58.5%から 2010 年には 71.1%まで増加し，情報システム開発におけるオフショア開発体制は，すっかり日本に定着している．

このような状況下で，日本から中国向けのオフショア開発は，中国人技術者の人件費値上げ及び人民元高などの影響があったにもかかわらず，一時期は年々増加の一途をたどり，2001 年時点で僅か 0.75 億ドルであったものが，2006 年には 8.7 億ドルまで達し，中国のソフトウェア開発企業では日本向けのオフショア開発が半分以上を占めていた [1]．このような日本と中国との間でのオフショア開発が拡大した大きな理由としては，主に以下の 3 点が考えられている．

1. 日中両国は漢字を使っているために，日本語がわからない中国人でも漢字を読むことで，日本語文書のおおよその意味を理解できる．

2. 日中間の地理的距離が近く，東京 (あるいは大阪) から北京 (あるいは上

海) まで飛行機で約 3 時間しかかからない．このため，対面で説明，打合せをすることが容易である．

3. 中国人 SE(Software Engineer) を低コストで確保するのが，比較的容易である．

一方，オフショア開発の拡大とともに顕在化してきたのは，オフショア開発特有の工程管理，品質管理の問題である．従来から，通常のソフトウェア開発においては，CMM(Capability Maturity Model)[2] がしばしば利用され，開発組織の成熟度を測ることがなされてきている．しかし，商慣習，コミュニケーション，社会慣習の異なる国境をまたがったオフショア開発においては，全体工程の中での委託元企業と受注側企業との関与度にばらつきもあり，CMM を適用しての管理は難しいことが指摘されている [3].

このようなオフショアソフトウェア開発における CMM について、表 1.1 は文献 [4] に示されたものであり，レベル 5 の企業は当初インドの企業だけであり，中国企業がより高いレベルの取得に熱心に取り組んでいることが述べられている．

これまでもオフショアソフトウェア開発をテーマとする研究 [5] や実践報告がなされてきているが，ほとんどが発注者側の企業の視点からのもの [6] であり，受注者側の企業からのものはなされてこなかった．これは，CMM のレベルを測る「開発プロセス」や「開発の指標」に関して，オフショア開発としての色々なノウハウを含めた “ベストプラクティス” が企業秘密にされていること，加えてレベルを測る「技術者」に関しては，受注者側の SE の技術レベルが低いこと，ならびに企業サイズが小さくその技術スキルの改善活動に割ける余力がないためでもある．

多くの中国オフショアソフトウェア開発企業では，急増した受注量をこなしていくために，開発経験の浅い若手中国人 SE を雇用し，プロジェクトに次々と投入していかなければならない実態がある [7]．そのために，開発委託元の日本企業が「開発プロセス」の改善 [8] や「開発ツール」の導入 [9],「開発の指標」につながる仕組みの整備 [10] といったことを発注先の中国企業とともに図っても，肝心の「技術者」の要因がいつまでたっても解決されない事態が続いている．中国人 SE の多くに対する継続的な指導教育が必要であり，また現場における「経験知」を活用することが重要であるといわれてきた [3].

表 1.1: オフショアソフトウェア開発企業への信頼性を保証する CMM

レベル	開発プロセス	技術者	開発の指標
1:初期	「とにかくやってみよう」という水準で未確立	成功は「スーパーマン」の存在	場当たり的
2:再現可能	プロジェクトごとに文書化	成功は個人に依存	プロジェクトごとに定まっている
3:定着段階	組織全体を通じて統合管理が実現	グループでの協力作業が実現	プロセスごとに収集・利用されている
4:管理段階	プロセスを定量的に把握し，個々の問題究明・除去が可能	プロジェクトでのチームワークが十分認識されている	組織全体で標準化されている
5:最適化	組織的かつ体系的に改善され，共通の問題究明・除去がなされる	組織全体でチームワークが十分に認識されている	プロセス改善を評価するために利用されている

例えば，中国人 SE は，日本語で書かれたドキュメントを理解すると同時に，日本語を用いて仕様書などのドキュメントを書かなければならない．それにもかかわらず，技術者の大部分は日本語レベルが低く，加えて情報技術は急速に発展するため，中国人 SE は日本語とともに技術用語を学び続けることを要求される．このような状況で，教育をしつつスムーズかつ効率的に仕事を進めるためには，経験豊かなベテランの中国人 SE の知識を開発現場で共有する仕組みとしての知識共有システムが重要となる．さらに，オフショアソフトウェア開発の企業経営の観点からは，日本語教育や技術教育の短期的・長期的計画や参考図書の整備等のため，現場の開発技術者の技術的興味の動向把握を目的として，知識共有システムに蓄積される知識の分析が必要となる．

中国人 SE に対する継続的教育には,「経験知」を社内現場で共有する「知識共有システム」の整備が必要であり，その際にオフショアソフトウェア開発作業における特有の問題として,「仕様書、日報、試験報告書」等の開発に際して大量の日本語文書を扱う必要があるため，日本向けのオフショアソフトウェア開発では，多くの場合，中国人 SE に日本語の使用を義務づけている．しかし，漢字文化を共有することから日本語で記載された内容をある程度推測できても、完全には理解できないという場合に，日本語レベルの低い中国人 SE が質問を発信する場合に真意を理解してもらうためには中国語で質問せざるをえないし，回答も中国語で書かざるをえない．このように，中国人 SE どうしのやり取りにおいて日本語と中国語が混在して用いられ，そのことから知識共有システムに蓄積される知識は日本語と中国語が混在したものを前提とした運用をしなければならない．

日本向けのオフショアソフトウェア開発作業におけるこのような「経験知」を開発現場で共有する「知識共有システム」としては，以上の対応が必須となるにも関わらず，従来の知識共有に関する研究ではほとんど取り上げられていない．

以上の状況を踏まえ，本論文では，効率の良い知識共有，技術者の自主的勉強及び管理者向けの必要な知識動向の把握手段の提供を目指して，日本語と中国語が混在したソフトウェア開発作業における知識共有システムの構築，さらに知識共有システムとして蓄積された質問に対する回答を一緒にしてセットにしたデータ (以下，知識カード) の分類方式の開発を課題として取り上げる．

1.2 関連研究

本節では，企業内部の「経験知」を活用するために取り組まれてきた「知識共有」の形態，ならびにその際に活用される「検索・分類」の技術，さらに「多言語処理」の現状を述べ，本論文が対象とするオフショアソフトウェア開発作業特有のものである，日本語と中国語が混在して用いられる際の「知識共有」や「分類」の課題との関連を明らかにする．

(1) 知識共有に関わる技術

「経験知」の活用は，野中が示した「暗黙知」と「形式知」の交換と知識移転のプロセスとして，“SECI (Socialization, Externalization, Combination, Internalization)” モデル [11] の中に位置づけられたことにより，情報技術を活用しての支援の方向性が明確になり，多くの研究がなされてきた．

ナレッジマネジメントとして,「知識資産」,「ナレッジワーカ間の知識共有/移転/再利用」,「ベストプラクティス共有」,「知識のライフサイクル管理」といった考え方に対して，文書などの一般的な資料/生産物の中の知識を共有する場合のシステムとして,「探している人にとって役立つ「知識」を含む資料なのかが，すぐに分かることである」という使い方のポイントがあげられている [12]．そのために，(1) 文書ディレクトリの自動生成，(2) エージェント基盤によるファシリテータの構築，(3) ワークフローシステムや文書管理システムの活用，ということが例示されている．これらは，統一的なデータ書式を前提としたシステムや使い方であり，オフショアソフトウェア開発作業でやり取りされる仕様理解やプログラミングに関わる自由記述形式のメモ的なデータでは，統一的なデータ書式をそもそも前提にすることが困難である．これに対して，共有フォルダを活用することが情報共有として企業内でしばしば行われていることに着目した上で，このようなボトムアップ的な情報蓄積は結果的に情報管理の属人化をひき起こし，組織が変わることによって，どこにどのような情報が保管されているかがわからなくなる “情報の死蔵” を招いているという指摘をもとに，長期にわたって情報を蓄積するためには，時間 (年度)，知識分類，案件の順番で共有フォルダを構成することが有効であるとした研究もなされている [13]．ただし，このような “知識分類” としては，次節にて関連研究とともに述べるが，多言語を前提とした知識の分類については触れられていない．

一方，統一的なデータ書式を前提としない，いわゆる非構造化文書データから知識を抽出するには，“テキストマイニング” [14][15][16][17][18] が多く用いられる．SECI モデルにおける「形式知から暗黙知への内面化のプロセス」を十分に回すために，“分析知” を間に挟むことで行おうとする試みもあり，この際に “形式知から分析知を生み出す” という点に，テキストマイニングが活用できる [19] と位置づけている．しかし，テキストマイニングは，そもそも膨大なデータを前提とした処理であり，文書データ数，あるいはそれぞれの文

書データに含まれる文の数として，類似した内容記述が必要となる．プロジェクト型開発のオフショアソフトウェア開発では，仕様理解やプログラミング作法といった点である程度は同じような内容記述があるが，メモ的なものとしてデータ数等はそれほど多くない．また，文書データのテキスト処理としては，自動要約 [20] や自動翻訳 [21][22][23] といった処理もあるが，テキストマイニングと同様に膨大な文書データが前提となっている．

そもそも，本論文の対象は “暗黙知” そのものを対象に「知識共有」を行おうとするものではなく，オフショアソフトウェア開発作業で多言語が用いられる中での “形式知” が整備できていない点を解決しようとするものである．

(2) 検索・分類に関わる技術

企業内での業務文書データの蓄積に伴い，それらの検索技術や応用の研究はさまざまになされてきた [24][25][26]．キーワード検索やディレクトリ検索に加えて，構文，同義語・同意語，文パターンなどの特徴を用いて類似文書や類似文を検索 [27] することで,「質問-応答」といった形式のデータ集合に対しても，該当する「質問-応答」を含むデータを抽出することが可能となってきている [28][29][30][31][32][33][34][35]．

「質問-応答」に関わる検索に関しては，1992 年に始まった TREC(Text REtrieval Conference) と呼ばれる国際会議 [1] においても主要研究トピックとして取り上げられてきている．構文解析や辞書構築，さらにそれらを前提とした「質問-応答」などのタスクに関わる研究が汎用的手法として取り組まれているが，本論文が対象としているオフショアソフトウェア開発といった分野依存の「質問-応答」に対する取り組みはなされていない．そもそも，自然言語に関する構文解析は文法に基づくことからも明らかなように言語依存であり，日本語と中国語が一つの文の中に混在する今回のタスクでは意味をなさない．辞書構築に関しては，翻訳辞書の自動生成 [36][37] が大きな研究トピックであるが，技術用語や業界用語といった分野依存の辞書に対してはあまり議論がなされてない，従って，本論文の対象とする「質問-応答」に関わる検索においては，不完全で分野依存の辞書を前提とした議論が必要と考える．

[1] http://trec.nist.gov/

蓄積された業務文書データを予め分類することができれば，ディレクトリ検索によって参照したい文書 (ある意味での知識) を容易に探すことができる．文書分類に関しては，文書内の単語に基づく統計情報をもとにした類似度による分類 [38][39][40][41][42][43][44][45]，同意語や固有表現といった単語の意味的情報あるいは文体表現や文パターンをもとにした類似度による分類 [46][47][48][49][50]，その他の自然言語処理技術を用いての分類 [51][52][53]，機械学習技術を用いての分類 [54][55][56][57][58][59] ，といったアプローチに大別される．これらは口語表現を含まない文書を前提としているが，オフショアソフトウェア開発現場での「質問-応答」においては，日本語能力の問題のために正確な表現が少なく，前述の意味的情報や自然言語処理技術に基づいたアプローチを適用できない．また，機械学習のようなかなりの数の学習データを必要とするアプローチも，対象データ数が十分多くない場合では適用が困難である．従来のアプローチの中で，統計情報を手がかりにした分類というアプローチがもっとも妥当と考えられるが，その際の「日本語と中国語が一つの文の中に混在する」といった場合を取り扱った研究はない．

(3) 多言語処理に関わる技術

複数の言語を対象にした研究の多くは,「翻訳」あるいは「検索」というタスクを取り扱っている．その際に，従来は構文解析や意味解析といった文法や辞書に基づくアプローチがなされていたが，最近では大規模記憶容量と計算機処理の高速化を背景に用例を活用したアプローチが多くなされてきている．用例を利用する場合には，言語非依存の N-gram による対応関係の利用 [60] や用例収集を効率的に実行する仕組みをもとに出来るだけ多くの対訳候補を作成した上で，それらの正確性判定を用いて対訳用の用例抽出 [61] などが行われている．しかし，これらはある言語で書かれたものを別の言語に翻訳するという点で,「日本語と中国語が一つの文の中に混在する」といった複数言語で書かれた文を対象としているわけではない．

「検索」に対しても同様に検索キーワードの翻訳に関して類似シソーラスの活用 [62] や単なる翻訳では多義性があることから曖昧性を排除できない点に対して，検索キーワードの共起情報を加味した検索 [63]，あるいはコーパスを活用して日英言語横断的に検索 [64] といったことがなされている．検索として，

対象分野を限定しない場合にはコストがかかったとしても，このようなシソーラス，コーパスを整備することも不可欠なためになされるが，今回の対象問題では分野が限定されている上に，通常のソフトウェア開発業務に加えてシソーラス，コーパスを整備するというのは現実的でない．

1.3 本研究における課題の解決方針

本研究の目的は，1.1 節で示したオフショアソフトウェア開発企業としての知識共有とその管理，そのような知識共有を効率的に実行するための中国語と日本語が混在する環境下での知識の分類に関する課題を解決することである．図 1.1 に，本研究における課題の解決方針についてまとめる．

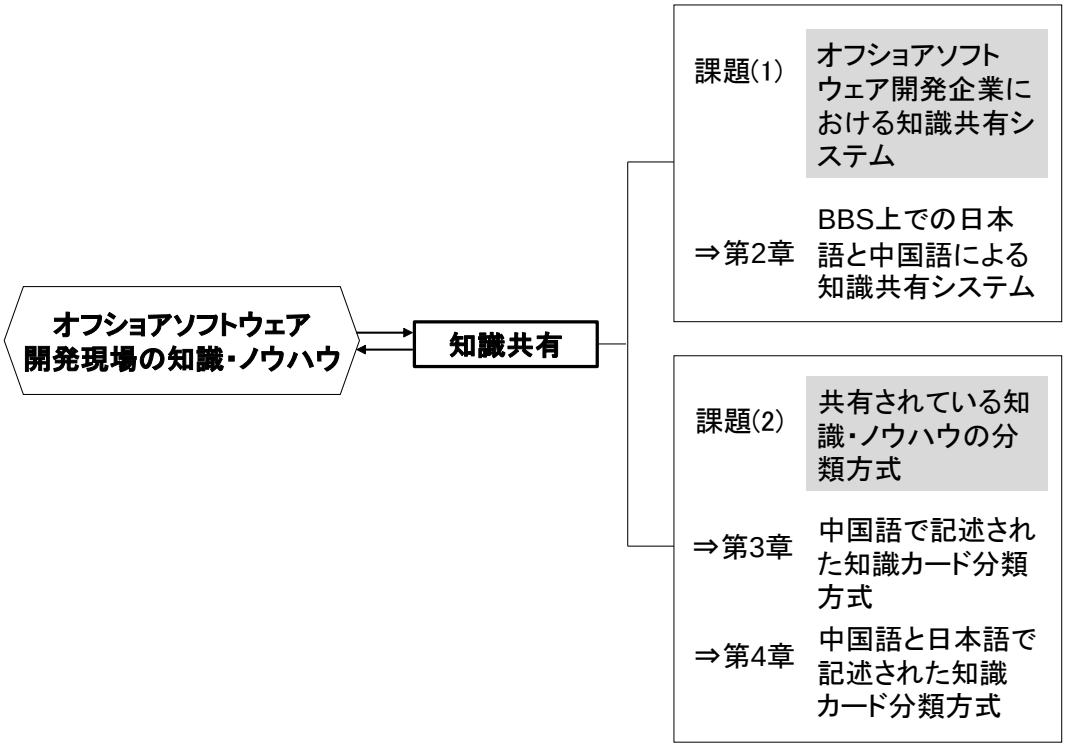

図 1.1: 本研究における課題の解決方針

(1) 知識共有システムの構築

中国における日本向けのオフショアソフトウェア開発現場では,「仕様書，日報，試験報告書」等の開発に際しての大量の日本語文書を，日本側発注元企

業とやり取りしなければならず，その際に中国人 SE による理解不足や表現不足を補う仕組みが，組織として必要である．ベテランとされる中国人 SE に直接口頭で尋ねることは，そのベテラン SE の通常業務を中断することとなり，メールによる問い合わせもそのベテラン SE が送受信する本来の業務メールに埋没して回答がタイムリーに得られないと考えられる．

そこで，開発現場での知識やノウハウをタイムリーに共有するために，BBS (Bulletin Board System) 形式による質問と回答を企業全体あるいはプロジェクトごとに管理し，共有する仕組みを構築し，その際に日本語レベルの低い SE であっても．質問ができるように，日本語及び中国語の両方を利用できるシステムを提案する．この際に収集された知識カードを蓄積，ならびに検索できることに加えて,「回答者指定」という機能を持たせて回答をできるだけ促すようにする．さらに，利用者のアンケート調査を行い，提案システムの有効性を検証する．

(2) 中国語と日本語が混在して利用される環境下での知識分類手法

知識共有システムが，日本語と中国語の使用を可能としていることから，蓄積される知識カードには,「日本語の知識カード」,「中国語の知識カード」,「日本語と中国語が混在している知識カード」の 3 種類がある．日本語での開発作業や商慣習に関わる知識の理解・修得を推進していることから,「中国語の知識カード」を，言及している知識内容が同じ「日本語の知識カード」と同じ分類カテゴリに保存して，閲覧利用できる必要があり，日々やり取りされるこれらのデータを自動分類する方法が必要である．ここで，自由記述によるメモ的データを分類する手法として，日本語を対象にした場合には，分類カテゴリに含まれる単語の統計的情報を利用した方式 [65][66][67] がある．

そこで，本論文が対象としている中国語で書かれている知識カードに対しても適用できるように拡張を行う．単語の統計的情報を算出する際には，形態素解析が必要であり，そこで中国語の形態素解析ツールを先ずは導入して，知識カードに書かれた文を単語列データに分解し，辞書を用いて日本語に翻訳することで，日本語のみで記述された知識カードとの類似性の判定を行う．しかし，辞書としてオフショアソフトウェア開発の特有な言い回しや技術用語に対して完全なものを準備することは現実的でないことから，翻訳漏れが発生する．そ

のために，翻訳漏れを前提とした類似性判定が必要であり，算出される類似度を補正する方式を提案する．

また，類似度を補正する方式では，知識カードが日本語か中国語かのいずれかの単一言語で書かれている場合には機能するが，知識カード内の一つの文中に両方の言語が混在する場合は，うまく分類ができない．そこで，類似度を補正する方式に加えて，知識カードの書かれている文が日本語なのか中国語なのか，あるいは混在しているのかを識別できる方式をもとに，知識カード内に両方の言語が混在する場合にも正しく分類する方式を提案する．

1.4 本論文の構成

本論文では，第 2 章以降を以下のように構成する．

第 2 章では，文献 [68][69][70][71][72] をもとに，オフショアソフトウェア開発企業における知識共有システムの必要性とそれに基づいて構築した BBS 形式で知識カードを蓄積した結果を述べる．利用者へのアンケート調査を実施して，構築したシステムを評価し，より有用なシステムとするための課題を明確にする．

第 3 章では，文献 [73][74][75][76] をもとに，知識共有システムに蓄積された中国語のみで記述された知識カードに対する分類方式を提案する．先ずは，日本語の知識カードを対象にした，カテゴリ辞書と呼ぶ分類カテゴリを特徴付ける辞書を用いた知識カード分類方式について述べ，その中で用いられている判定処理を中国語の知識カードに適用する際の課題を明らかにする．その上で，中国語の知識カードの分類に際して，翻訳用辞書 (中国語から日本語へと単語を変換) を用いる際の「変換漏れ」に対する処理を組み入れ，それをもとに分類する方式について述べる．さらに、実験による評価を行い，提案方式の有効性を示す．

第 4 章では，文献 [77][78] をもとに，中国語と日本語が知識カードの一つの文中に混在した場合の分類方式を提案する．中国語と日本語が混在する文をそれぞれの言語に対応した形態素解析ツールに入力して得た単語列データに対して，ドメイン辞書 (中国語から日本語へと単語を変換) をもとに “フィルタリング” と “変換漏れ補正” の処理を実行し，その結果としての単語列データに対する判定処理により分類する方式について述べる．さらに、実験による評価を

行い，提案方式の有効性を示す．

第 5 章では，結論として本研究で得られた成果をまとめ，今後の課題を述べる．

第2章

BBS上の日本語と中国語による知識共有システム

2.1 緒言

日本から中国向けのオフショア開発が年々増え，受注先の中国オフショアソフトウェア開発企業では，開発経験の浅い若手中国人SEを大量に雇用している[7]．その結果として，技術力，日本語能力，業務知識が不十分なために，工程や品質を守れないケースが多く発生している．本章では，このことを解決するために，日本語レベルの低い中国人SEが開発現場で日常的に活用できる知識共有システムを提案する．

一般のアウトソーシングにおいては，最終成果物をチェックすることで品質を管理しているが，開発プロセスの中で「中国企業側の仕様理解不足による日本側企業への設計への差し戻し」,「日本側企業の仕様説明不足や設計不備による中国側企業での工程の手戻り」といったことが，日本から中国向けのオフショア開発では多く発生している．このことにより，日常の開発段階で用いられている「仕様書、日報、試験報告書」等の大量の日本語文書まで含めた工程管理，品質管理が必要であり，このような文書作成やプログラム作成に際して，開発現場での中国人SEを日常的に支援し，かつ開発を継続しながらのSEのスキルアップを図ることが重要である．

ビジネスのグローバル化によるオフショアソフトウェア開発取り組みの目的として，表2.1は文献[79]に示された日本企業に関しての調査結果である．

表 2.1: オフショアソフトウェア開発取り組みの目的 (単位:%)

項目	割合
開発コストの削減	93.8
国内人材不足の補完	80.2
海外の高い技術力の活用	20.8
ソフトウェア関連の売り上げ拡大	18.8
開発のスピードアップ	9.4
相手先国市場の開拓	12.3
コア・コンピタンスへの経営資源の集中	7.3

中国側企業における日本向けのオフショアソフトウェア開発作業は，当初コーディングと単体テスト工程が大半であったが，表 2.1 に示される日本側企業のオフショア開発の目的にある開発コストの削減，国内人材不足に対応する形で，次第に詳細設計から結合テストまでと受注工程が広がってきた．最近では，日本側企業がシステムテストを実行できる開発サーバを専用線経由でオフショア側企業も利用でき，システムテストまでの工程を含めたオフショア開発案件が増えている．このような場合に，ソフトウェア開発の上流工程もスムーズかつ効率的に展開できるためには，業務知識が必須である．中国側企業内での類似した開発実績の業務内容やチームメンバーが纏めた反省点などの知識を共有するシステムを活用できるならば，特に経験の少ない若手技術者に対して有効と考えられる．また，このような知識共有システムを利用して，関連情報を勉強する際に，随時質問しながら，回答を受け取ることで，教育時間を減らすことが期待できる．

質問に対して回答を受け取る代表的な形態として，BBS と呼ばれるシステムが多く利用されている．情報交換や会話・議論などを非同期に行うことができ，オフショアソフトウェア開発現場に導入することで，開発途中での問題点について，プロジェクトの垣根を越えて質問し，回答をもらうことが可能と考えられる．ただし，オフショアソフトウェア開発企業に BBS を導入する際には，中国人 SE の日本語能力の低さを前提として，中国語と日本語を混在して利用できる環境を構築するすることが必要となる．

筆者が所属する済南凌佳科技有限公司では，日報管理，勤怠管理，トレーニング管理，ニュース，設備管理，といった企業としての日常管理機能を組み入れた「凌佳内部管理システム」を運用している．これの一部として，BBS 形式による知識共有機能を「凌佳知識共有システム」として開発した．この「凌佳知識共有システム」では，オフショアソフトウェア開発の技術者のために，オフショア開発の流れ，各分野の業務基礎知識，ソフトウェア開発技術 (フレームワーク，開発言語，データベースなど)，プロジェクトマネジメントにおける管理指標 (レビュー指摘率，バグ率)，といった事項に関しての様々な質問と回答からなる BBS 形式の Q&A(Question and Answer) 機能が組み入れられている．これらの Q&A は，現場では有用な知識として活用される．

以下，2.2 節では，BBS 形式の「凌佳知識共有システム」について述べる．2.3 節では，この「凌佳知識共有システム」に関して，実際の利用によるアンケート調査の結果を説明する．最後に 2.4 節で，本章のまとめを述べる．

2.2 オフショアソフトウェア開発現場での BBS 形式の知識共有システム

2.2.1 BBS 形式の知識共有システムの概要

提案する知識共有システムは，掲示板管理処理，質問受付処理，回答受付処理，検索受付処理といった典型的な BBS システムの四つの機能から構成される.「凌佳知識共有システム」の構成を図 2.1 に示す.

オフショアソフトウェア開発企業としては，中国人 SE が Q&A を書く時にも，日本語を用いることを原則的には義務付けている．しかし，日本語レベルの低い SE は中国語を用いないと真に質問したいことを表現できない上に，回答も中国語で書かれていないと理解できないために，システムとしては中国語を用いての Q&A も使えるようになっている．また，業務内容などに関する顧客の機密情報を漏洩させないためにログインパスワードでのアクセス制御が導入されている．さらに，プロジェクト特有の Q&A に関して，他のプロジェクトメンバに閲覧させないために，プロジェクトチーム，あるいはプロジェクトチームメンバーを指定できるようになっている．指定されたことに対してメー

ルでの通知もなされるが，その際には外部への発信を防ぐために，イントラネットに設置された社内向けメールサーバが用いられる．

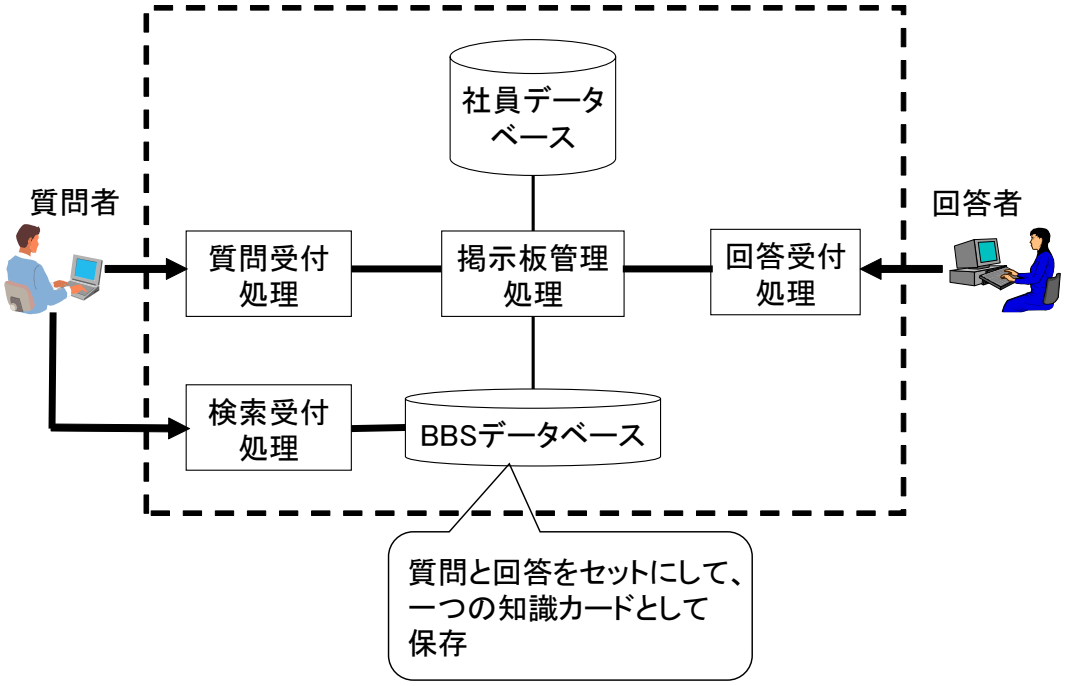

図 2.1: 「凌佳知識共有システム」の構成

BBS 形式による知識共有としての機能以外に，プロジェクトマネージャや経営幹部が現場 SE の技術的興味の動向を把握するために，どのような情報がやりとりされているかについて，プロジェクトごとの統計情報として，現時点での Q&A(知識) 合計数，未解決数，解決済み数を算出し，人気ジャンルランキングが何であるかを提示する機能もある．また，SE どうしで回答に対する評価を実施し，役に立った回答数を多く行った SE，Q&A が多いプロジェクトチーム，人気ジャンルランキングに関して，TOP5 の情報を取り出すことができる．また，回答を検索する際には検索対象情報の日付制限，質問を書き込む際には回答に対する期限の設定や回答者を指定することを通して，Q&A としての実用性や実効性を保証するようにしている．プロジェクトの垣根を超えて Q&A を増やすために，TOP5 に入った中国人 SE に対する奨励策も導入している．

各部署のリーダ達は，この知識共有システムから提示される情報をもとに，

中国人 SE の作業状況，プロジェクト開発等に関して，例えば「ある SE からは開発作業に関わる質問が多く出ている」,「未解決な Q&A がプロジェクト全体でいつまでもなくならない」といったことから，開発進捗が停滞しているなどの問題に気づくことができる．

この知識共有システムは，サブシステムとして「凌佳内部管理システム」に統合されていることから，毎日「凌佳内部管理システム」にログインする必要がある SE にとっては，業務中は随時 Q&A 機能を利用できるようになっている．

以下に，知識共有システムとしての質問受付処理機能，回答受付処理機能，検索受付処理機能の特徴をそれぞれ述べる．

2.2.2 質問受付処理機能

質問受付処理機能では，質問内容として日本語も中国語も使える形式となっており，やり取りされる Q&A は，質問とそれに関する一連の回答というような形の知識カードと呼ぶ形式で BBS データベースに保存される．

図 2.2 は，質問入力画面であり，表 2.2 に示すように必須入力項目とそれ以外の項目から構成される．

表 2.2: 質問入力項目

種別	項目名
必須入力項目	タイトル，問題キー，問題大分類，問題小分類
オプション入力項目	指定回答者，指定回答時間，問題内容，チェックボックス (メールでおしらせ)，ラジオボタン (全員共有/条件共有)

タイトル項目は質問のテーマであって，例えば「gcc コンパイラのバージョンによる特徴を教えて下さい.」などのように質問内容を十分に記述できるならば，問題内容項目に記入する必要がない．加えて，問題キー項目は，質問に関するキーワードであり，これら 2 つを併用して容易に質問を発信することができる．また，この問題キー項目は知識データとして保存されたデータベースへの後述する検索にも利用される．

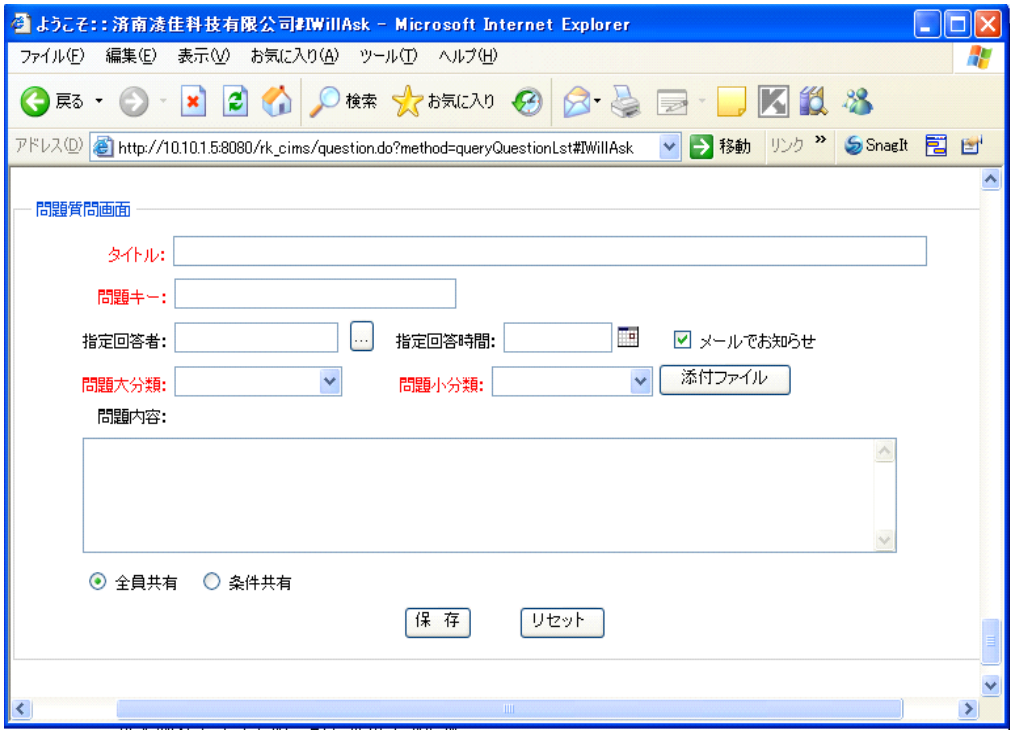

図 2.2: 質問入力画面

図 2.3: 回答者選択画面

さらに，先ほどの例のようにある「gcc コンパイラ」といった技術に熟知しているSE は誰であるかを質問者が知っている場合には，質問入力画面にあるように「指定回答者」を入力する，さらにその指定回答者に「質問が発信されたことをメールで通知する」といった機能を用いて，的確な回答を得ることもできるようになっている．図 2.3 は，指定回答者を検索，追加する画面であり，複数人を指定できる．

また，例えば「コンパイルエラー」の意味や対応の方法について質問を行いたいときには，補足説明のための電子ファイルを添付することもできる．日本語レベルの低い中国人 SE にとって，日本語版 C コンパイラの日本語のエラーメッセージを質問するにも, その見方が曖昧なままに行うことは困難であり，画面のスナップショットそのものを使って質問することで，質問へのハードルを下げることにつなげている．常に，このような日本語能力が不足していることに起因する問題を，機能設計において考慮している．

また，情報漏洩防止のためのアクセス制御として，質問入力画面の最下部には,「全員共有」と「条件共有」という二つのラジオボタンによる選択肢があり,「全員共有」では全従業員がこの質問に関しての閲覧も回答も行うことができる．一方,「条件共有」は，顧客の情報漏洩防止のために，プロジェクト体制で開発していることから質問者と同じ開発プロジェクトのチームメンバーしか質問ならびに回答にアクセスできないように制限される．

2.2.3 回答受付処理機能

図 2.4 は，回答入力画面であり，回答欄に記載される内容に回答者，回答日付，回答時刻を受付処理として自動的に作成して付与する．質問受付処理画面と同様に，必要な場合には「添付ファイル」機能を利用して，ファイルを添付して回答することができる．また，質問に対して複数の回答がスレッドとして記録され，質問者は回答スレッド確認して納得した場合には「採用」というボタンをクリックし，その結果，BBS としての Q&A 一覧画面では，ステータスとして「解決済」と表示される．

図 2.4: 回答入力画面

2.2.4 検索受付処理機能

Q&A データとして保存されたものは，開発レベルや日本語能力の低い中国人 SE にとって役立つ情報が多くあり，質問をする前に検索して調べることが行われる．質問入力画面で必須入力項目である「問題キー」を用いたキーワード検索以外に，以下に示すとおり質問と回答として記載された内容に対するキーワードマッチングによる全文検索も可能なようになっている．

図 2.5 に，検索画面の構成を示す．画面上部で一般的に用いられる検索キーワードを入力するだけでなく，検索オプションとして，所属チーム名，質問者，質問期間，検索対象として「問題キーのみ」,「問題キーと質問内容」,「問題キーと質問内容と回答内容」を指定できる．検索結果は，Q&A のステータスとともに表示され，リストに付与されたチェックボックスを選択して，図 2.6 に示すように内容を読むことができる．

図 2.5: 検索画面

図 2.6: 検索結果の内容詳細表示画面

ここで，中国人 SE による検索入力の不適切さや不正確さに対する検索効果を比較するために，同じ意味を持つ中国語，日本語をキーワードとして,「質問部分のみに対する検索」,「回答部分のみに対する検索」,「質問と回答の両部分に対する検索」の場合に分けて，10 個のキーワードを使って実験を行った．検索入力内容にヒットした Q&A の件数を，表 2.3 に示す．

表 2.3: 検索によるヒット数

検索実験	キーワード		質問部分のみ	回答部分のみ	両方部分
1	中国語	项目管理	3	3	5
	日本語	プロジェクト管理	21	13	30
2	中国語	安全	28	36	43
	日本語	セキュリティ	51	4	55
3	中国語	公司	21	45	61
	日本語	会社	70	10	79
4	中国語	开发	32	25	50
	日本語	開発	60	46	102
5	中国語	工作	35	33	58
	日本語	仕事	43	9	50
6	中国語	加班	2	3	1
	日本語	残業	5	1	6
7	中国語	学习	18	13	25
	日本語	勉強	77	37	97
8	中国語	测试	22	24	37
	日本語	テスト	48	29	68
9	中国語	中文 乱码	4	1	5
	日本語	中国語文字化け	0	1	1
10	中国語	品质 问题	4	3	6
	日本語	品質 問題	8	1	9
計	中国語検索		169	186	294
	日本語検索		383	151	497

質問部分と回答部分の両方に対する全文検索を行う場合，明らかにヒット総数が増えている．このことは，質問や回答において，現実に日本語と中国語が混在して用いられていることを示している．

2.3 アンケート調査による評価結果

済南凌佳有限公司の約 180 名の技術者によるこの知識共有システムの利用により，10,000 件以上の知識カードが蓄えられている．そこで，知識共有システムの評価として，アンケートを実施した．回答者 158 名の役職は，PM(プロジェクトマネージャ) が 5 名，PL(プロジェクトリーダー) が 16 名，SL(サブリーダー) が 33 名，PG(プログラマー) が 104 名となっている．なお，この役職認定は，表 2.4 の規則となっている．

表 2.4: 役職規則

役職名	規則内容
PM	開発経験年数 8 年以上，日本語 1 級レベル以上
PL	開発経験年数 5 年以上，日本語 2 級レベル以上
SL	開発経験年数 3 年以上，日本語 3 級レベル以上
PG	PM，PL，SL 規定に非該当

蓄積されている知識カードデータは,「フレームワーク，開発言語，データベース，開発ツール，環境構築」といったソフトウェア開発に関わる技術以外にも，品質，レビュー，見積りなどプロジェクト管理に関連する Q&A，さらに「日本語の勉強，社内環境，経営への提言」もやり取りされている．それらは,「開発技術，基礎技術，内製ツールの使い方，日常業務，一般ビジネス知識，プロジェクト管理，その他」といった 7 種類のカテゴリに分類されて保存されている．アンケート内容の項目を，表 2.5 に示す．

表 2.5: アンケート内容

項目	質問事項
1	あなたが知識共有システムを利用したいと思った理由は何ですか．(複数選択可)
2	あなたは知識共有システムへ登録するとき，どの言語を利用しましたか.
3	今後知識共有システムに Q&A を登録するときに，どの言語を利用したいですか.
4	あなたが質問したことに対する回答がありましたか．その回答内容はあなたにとって役立ちましたか.
5	あなたが質問したことは下記のどのカテゴリに属していますか. (複数選択可)
6	あなたが満足できない理由と推薦するカテゴリなどを記入してください.
7	あなたの入力したキーワードにより，期待したとおりに検索結果がよくなりましたか.
8	検索機能について，意見を記入してください.
9	あなたはしばしば知識共有システムにやり取りされている質問やアクセスの多い Q&A などを読みますか．(選択式)
10	あなたは未解決の質問を目にすると，できるだけ回答しようとしますか.
11	あなたは回答のあった質問 (解決済み) に対して，再アクセスしますか.
12	あなたは回答のない質問 (未解決) に対して，再アクセスしますか.
13	あなたの知識共有システムの機能に対する要望，意見を記入してください.

先ずは，“システムの利用頻度” について，アンケート内容の項目 9 の選択式回答から集計したアンケート結果を，表 2.6 に示す.

表 2.6: システムの利用実態の分布 (単位:%)

	役職名			
	PM	PL	SL	PG
毎日 2〜3 回	1.9	2.0	11.3	41.0
週に 4〜5 回	0.6	1.9	5.0	20.6
週に 2〜3 回	0.6	2.5	3.8	7.5
ほとんど利用しない	0.0	0.0	0.0	1.3

毎日 2〜3 回，知識共有システムを利用する技術者は 60%ぐらいであり，1 日の平均登録知識件数は，約 30 件〜50 件となっている．特に，PG の利用頻度が期待通りに高くなっている．

次に，“推薦カテゴリ” について，アンケート内容の項目 6 の記述式回答から集計したアンケート結果を，表 2.7 に示す．

表 2.7: 推薦カテゴリの割合 (単位:%)

	役職名			
	PM	PL	SL	PG
開発技術	3.1	9.4	18.0	56.3
基礎技術	3.1	9.4	15.6	65.8
内製ツールの使い方	1.3	3.1	9.4	62.5
日常業務知識	3.1	7.5	12.5	50.0
一般ビジネス知識	3.1	9.4	11.3	6.3
プロジェクト管理	3.1	9.4	15.0	9.4
その他	0.6	1.9	3.8	12.5

基礎技術に関しての評価がどの役職においても高く，これはまだまだ技術レベルの修得が企業全体として不十分であることを示している．

次に，“検索機能” について，アンケート内容の項目 8 の記述式回答から集計したアンケート結果を，表 2.8 に示す．

表 2.8: 検索機能の評価 (単位:%)

	役職名			
	PM	PL	SL	PG
かなり良い	0.6	3.1	6.2	30.0
良い	1.3	5.0	7.5	9.4
普通	1.3	0.6	2.5	12.5
悪い	0.0	0.6	1.9	15.6
かなり悪い	0.0	0.0	0.6	1.3

知識カードの登録数が 10,000 件以上となっているので，検索機能は重要であるにもかかわらず，表 2.8 に示すように,「かなり良い」と「良い」が 63%程度しかない．この理由としては，検索結果の重要度を判定する処理がなく，ランキング表示がなされていないために，結果的に 1 件ずつ読む必要があり，時間がかかる点があげられている．検索精度向上やランキングに関する機能改善が必要であるとともに，検索結果が適切でないと判断される場合は，Q&A データのステータスの状態を「未解決」に戻すことができる差戻し機能といったことも考慮しなければならない．

最後に，“システム全体の実用性” について，アンケート内容の項目 1 の選択式回答から集計したアンケート結果を，表 2.9 に示す．

表 2.9: システム全体の実用性評価 (単位:%)

	役職名			
	PM	PL	SL	PG
非常に実用的である	1.3	5.0	9.3	40.0
実用的である	1.9	1.9	5.6	16.3
普通	0.0	0.0	2.5	3.1
実用的でない	0.0	0.0	0.6	2.5
全く実用的でない	0.0	0.0	0.0	0.6

概して肯定的な意見が多く，表2.9のとおり,「非常に実用的である」と「実用的である」で81.3%が回答し，業務効率の向上を述べている．例えば,「JavaでDBにアクセス」に関する質問事項に関して,「ORACLE，HiRDB，SQLServerとのアクセス方法」や対応するソースコードがすぐに検索でき，そのまま利用できた事例もある．あるいは，上水道管理システムの開発プロジェクトの担当メンバが，水道事業の業務用語を勉強するため,「水道業務用語」をキーワード入力して検索し，検索結果から水道業務用語ファイルをダウンロードした事例もある．

また，アンケート項目のNo.13にある要望や意見を分析した結果，知識共有システムの分類カテゴリに対する細分化，類似した知識が中国語と日本語の両方が登録された場合に一括で検索できる機能，というものがあり，これらは知識データが追加された際の自動分類や中国語と日本語が混在した文を扱う必要性を示唆している．

2.4 結言

本章では，オフショアソフトウェア開発企業として構築した知識共有システム，ならびにそのシステムでのQ&A処理と検索処理について述べた．オフショアソフトウェア開発現場では，開発技術レベルの低さや日本語能力の不足といったことを前提として，SEの日常的な作業を支援することが必要であり,「日本語でも中国語でも利用できる，回答指定者を設定できる」といった点から，BBSを利用した知識共有システムを構築し，実アンケートをもとに評価を行った．

アンケート回答者は，PM(プロジェクトマネージャ)，PL(プロジェクトリーダ)，SL(サブリーダ)，PG(プログラマ)の計158名で，その中で104名のPGの約62%が日常的に知識共有システムを利用していると判断される回答を行っていた，このことは，開発した知識共有システムがオフショアソフトウェア開発現場での大きな課題として指摘した中国人SEの技術者としてのレベルの低い点を補う手段として機能していることを示している．また，そのような知識共有システムに蓄積された知識やノウハウの内容として,「基礎技術」のカテゴリをどの技術者層も役立つものと回答していることから，逆に経営の視点からは開発企業としての技術レベルとして「基礎技術」に関する教育・研修をま

だまだ強化すべきことが明らかにされた.

構築した知識共有システムは，PM といった管理職レベルから PG といった技術者にいたるまで利用され，さまざまな知識・ノウハウの提供・共有が行われているが，分類に関してはカテゴリそのものの整備と自動化が課題として認識された.

第3章

中国語知識カードの分類方式

3.1 緒言

本章では，第2章における「凌佳知識共有システム」に蓄積されている「知識カード」として,「日本語の知識カード」,「中国語の知識カード」,「日本語と中国語が混在している知識カード」があり，同じ内容のものは同じカテゴリに分類して活用することを目的に，先ずは「中国語の知識カード」の分類方式を提案する．

この背景には，知識カードの活用として，中国のオフショアソフトウェア開発企業として日本語での開発作業や商慣習に関わる知識の理解・修得を推進していることから,「中国語の知識カード」を言及している知識内容が同じ「日本語の知識カード」の知識分野 (以下，分類カテゴリ) に分類して，検索や閲覧に役立てる必要がある．従って,「日本語の知識カード」の分類カテゴリをもとに,「中国語による知識カード」を分類処理する必要がある．

1.2節で示した通り，知識カードに記された「質問-応答」においては，日本語能力の問題のために正確な表現が少なく，意味的情報や知識カードごとの単語に基づく統計情報をもとにした類似度による分類は適用できない．これに対して，短文かつ口語表現が含まれる自由記述アンケートデータを対象に，単一のアンケートデータの意味的情報や統計的情報を利用するのではなく，各々の分類カテゴリにある程度のアンケートデータが分類済みの状況で，それらの分類カテゴリに含まれるアンケートデータ群の単語の統計的情報を利用して，新たに追加されたアンケートデータを分類する,「カテゴリ辞書を用いた意見分

類方式」[65][66][67] が提案されている．

「凌佳知識共有システム」においても 7 種類の分類カテゴリが既に整備されており，アンケートデータと同様に短文かつ口語表現が含まれる知識カードの分類に，このような方式が適用できると考える．ただし，「カテゴリ辞書を用いた意見分類方式」では，あくまでも日本語のみで記述されたデータを対象としていることから，本研究が対象とする「日本語の知識カード」と「中国語による知識カード」といった 2 種類の知識カードを処理することはできない．

これを解決するために，まずは中国語のみで記述された知識カードを予め整備した翻訳辞書を用いて日本語の知識カードに変換し，その際に発生する単語の変換漏れに対する補正処理を行うことで，中国語で記述された知識カード群を自動的に分類する方式を提案する．

以下，3.2 節では，日本語知識カードを対象に，カテゴリ辞書を用いた知識カード分類方式について述べ，その分類に際しての特徴と中国語知識カードに適用する際の課題を明らかにする．その上で，中国語知識カードの分類に際して，変換漏れ補正率を定義し，それをもとに分類する方式について述べる．3.3 節では，実験による評価を行い，提案方式の有効性を示す．3.4 節では，本章のまとめを述べる．

3.2　カテゴリ辞書を用いた知識カード分類方式

3.2.1　カテゴリ辞書による分類方式の概要

日本語の自由回答式アンケートデータの分類用に開発された方式 ([67]) を本研究の日本語知識カードに適用した場合の処理の流れを図 3.1 に示す．

この分類方式では，予め設定した分類カテゴリに日本語知識カードを登録し，それぞれのカテゴリ内の単語とその単語のカテゴリ代表度の組み合わせから成る「単語代表度辞書」，加えて共起パターンと呼ぶカテゴリ内の 2 単語の組み合わせとその共起パターンのカテゴリ代表度の組み合わせから成る「共起パターン代表度辞書」を自動的に構築しておく．これは，カテゴリを単語の統計的情報を用いて特徴付けることになる．新たに登録したい日本語知識カードは，各々のカテゴリに紐づけられた「単語代表度辞書と共起パターン代表度辞書 (以下，カテゴリ辞書)」に含まれる単語をどの程度含んでいるかという “代

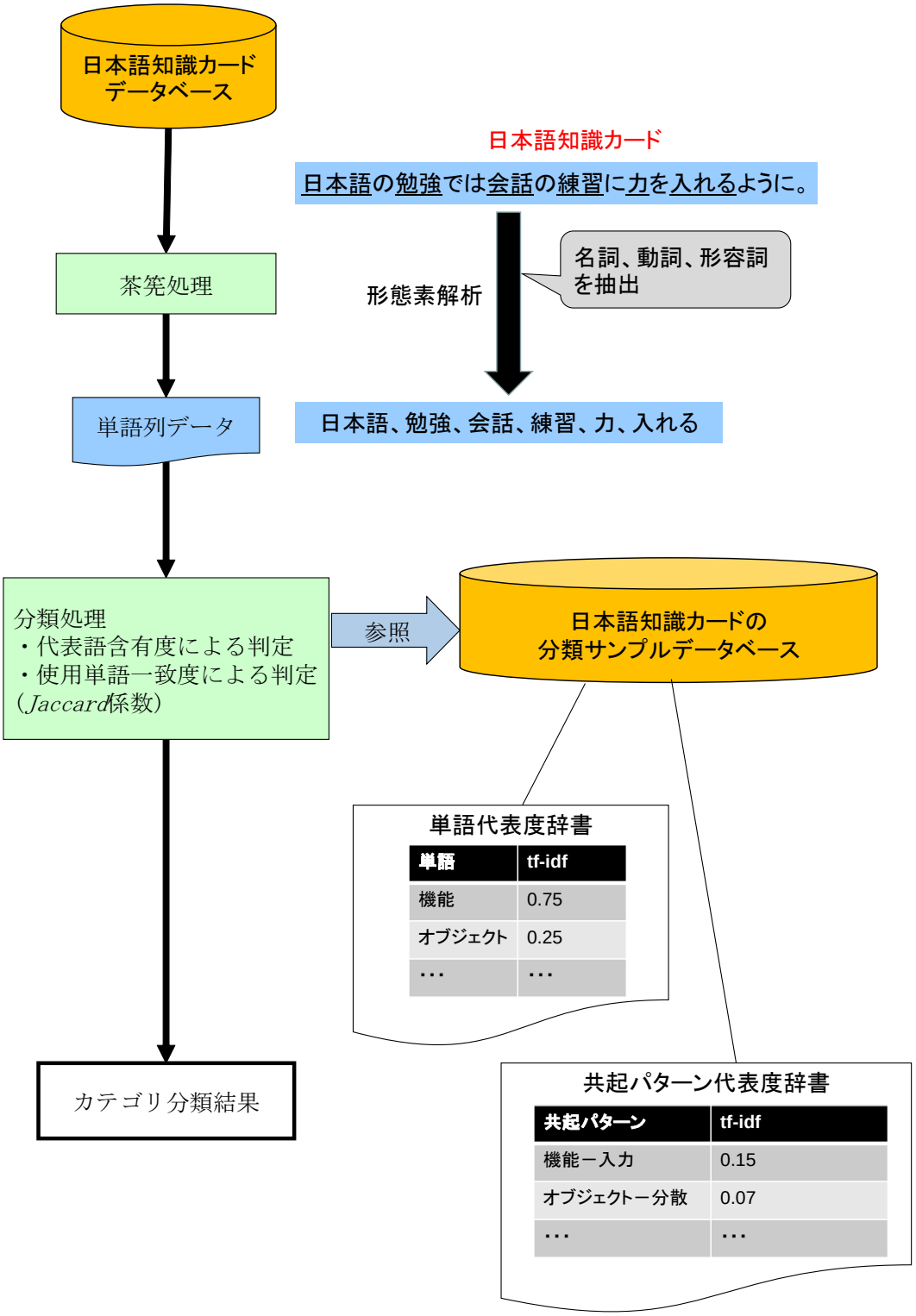

図 3.1: 日本語知識カードの分類方式

表語含有度” が閾値 Th_r を超えた場合には含まれる可能性のあるカテゴリ候補として選出し，その後にカテゴリ毎に含まれる知識カードに使われている単語との一致性を $Jaccard$ 係数という “使用単語一致度” をそれぞれの知識カードごとに算出し，その平均値が Th_j を超えるとそのカテゴリに含まれると判定する．

“代表語含有度” は，それぞれの単語で計算する $tf\text{-}idf$ だけでなく，共起パターンと呼ぶ2単語の組み合わせから算出する $co_tf\text{-}idf(co{-}occurrence\, tf\text{-}idf)$ も用いて計算される．ここで，$tf\text{-}idf$ は，“term frequency inverted document frequency” というもので，各文書に特徴的に出現する単語に関する指標として，テキストマイニング分野の研究では頻繁に活用されている．

$$tf\text{-}idf(w_i) = tf(w_i) \times log(\frac{N}{df(w_i)}) \tag{3.1}$$

$$co_tf\text{-}idf(w_j, w_k) = tf_{co}(w_j, w_k) \times log(\frac{N}{df_{co}(w_j, w_k)}) \tag{3.2}$$

式 (3.1) において，w_i は対象とする単語，$tf(w_i)$ は対象とする単語の対象カテゴリ内出現回数，N は全カテゴリ数，$df(w_i)$ は対象とする単語が含まれるカテゴリ数である．また，式 (3.2) において，w_j，w_k は対象とする共起パターンの単語，$tf_{co}(w_j, w_k)$ は対象とする共起パターンの対象カテゴリ内出現回数，N は全カテゴリ数，$df_{co}(w_j, w_k)$ は対象とする共起パターンが含まれるカテゴリ数である．

分類カテゴリに含まれる日本語知識カード (以下，知識カード S)，新たに追加される日本語知識カード (以下，知識カード X) のいずれであっても，それらに含まれる文を形態素解析ツール “茶筌” [1] を用いて，言語学的に意味を持つ最小単位である形態素に分解し，単語系列データとして計算を実行する．

“代表語含有度” の R は，次に示す式に単語ごとに算出した R_t と共起パターンで算出した R_{co} に対して，重み付け (式 (3.5) 中の α) を行って算出する．

[1] http://chasen-legacy.sourceforge.jp/

$$R_t = \frac{\sum_{w_i} tf\text{-}idf(w_i)}{n_t} \tag{3.3}$$

$$R_{cot} = \frac{\sum_{(w_j,w_k)} co_tf\text{-}idf(w_j,w_k)}{n_{cot}} \tag{3.4}$$

$$R = \alpha \times R_t + (1-\alpha) \times R_{cot},\ 0 \leq \alpha \leq 1 \tag{3.5}$$

n_t はカテゴリ辞書と知識カード X の共通単語数，n_{cot} はカテゴリ辞書と知識カード X の共通パターン数である．

“使用単語一致度” の J は，$Jaccard$ 係数であることから，以下のように算出される．

$$J = \frac{N_{sx}}{N_s + N_x - N_{sx}} \tag{3.6}$$

N_{sx} は知識カード S と知識カード X の共通単語数，N_x は知識カード X の単語数，N_s は知識カード S の単語数である．

3.2.2 中国語知識カードを処理する場合の課題

本研究で対象としている知識共有システムでは，技術者の日本語レベルが低いために，中国語で登録された知識カードが 29%を占めている．この際に，図 3.2 に示すように，“茶筅” 処理を中国語の文に適用すると，当然のことながら正しい単語列データが得られない．

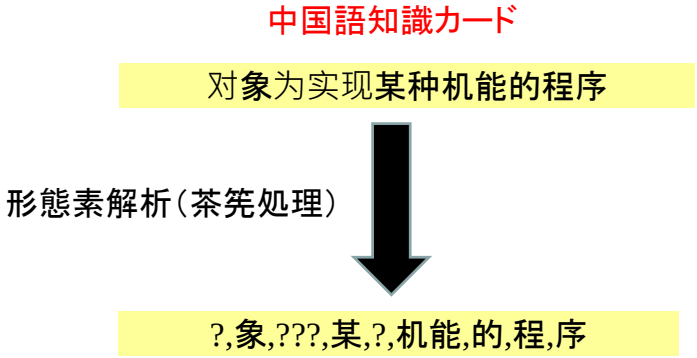

図 3.2: 中国語の文に対する茶筅処理の結果例

ここで，中国語の文を日本語に翻訳することも考えられるが，自動翻訳の変換レベルはまだまだ不十分である．この結果，図 3.1 の判定処理では正しい分類結果がえられない．また，3.2.1 節の方式は単語列データをもとにした処理であるため，名詞や形容詞といった品詞に対応した中国語の単語のみを日本語に変換させることも考えられるが，技術進歩の激しい情報分野で全ての IT 用語を登録し，完全な辞書を作ることは難しい．そこで，本研究では，定期的に IT 処理技術に関する書類，新聞，インタネット記事などから抽出した IT 用語をもとにした名詞を登録した不完全なながらも翻訳用辞書を用いることを前提とした分類方式を提案する．

3.2.3 翻訳辞書と変換漏れ補正を用いた中国語知識カード分類方式

中国語知識カードに対しては，中国語の形態素解析ツールである ICTCLAS[2] を用いて中国語の単語列データに分解する． 次に，翻訳辞書を使って変換できる単語のみを日本語に変換する．この際に変換漏れが発生するために，近似日本語単語列データとして扱う．従って，近似日本語単語列データによる “代表語含有度” の R ならびに “使用単語一致度” の J は変換漏れにより正確に算出されない．本来は R ならびに J に関する補正を必要とするが，式 (3.1)，式 (3.2)，式 (3.3)，式 (3.4)，式 (3.5) に示したように “代表語含有度” の補正は簡単にはいかないことから，本論文では先ずは “使用単語一致度” の J の補正を行う．J は，その定義から変換漏れがある場合には，本来の値よりも小さくなることをを考慮して，$Jaccard$ 係数をここで推定する．この推定 $Jaccard$ 係数を用いて，日本語知識カードと同様に分類処理を行う．

3.2.4 $Jaccard$ 係数推定処理

中国語の単語を日本語の単語に変換する際，翻訳用辞書に登録されていない中国語の単語は変換されない．そこで，中国語知識カードの単語数から式 (3.6) の分母の現れる N_x，ならびに分母と分子に現れる N_{sx} の単語数を推定するために，以下の変換漏れ補正率を用いて補正する．

[2]http://ictclas.org/

知識カードXの単語を翻訳した場合に，変換漏れが発生しているために，式3.6の分母にあるN_xは，知識カードXが完全に変換された場合の単語数よりは少なくなっている．しかし，変換漏れがなければ，知識カードXの変換前の単語数に等しいことからN_xには「知識カードXの変換前の単語数」を用いることとする．

一方，知識カードXの単語を翻訳した場合の知識カードSとの共通単語に関しては，変換漏れの影響が及ぶことから変換漏れ補正率$(w > 1)$を以下の式で定義する．

$$w = \frac{1}{1 - \text{共通単語変換漏れ率}} \tag{3.7}$$

この結果，式(3.6)の分母と分子にあるN_{sx}は,「知識カードXの単語を変換した結果と知識カードSとの共通単語数」にwを乗じて，推定することとする．

推定$Jaccard$係数は，以下の式で定義される．

$$\text{推定}\ Jaccard\ \text{係数} \quad = \quad \frac{w \times N_{sx}}{N_s + N'_x - w \times N_{sx}} \tag{3.8}$$

ここで，N'_xは,「知識カードXの変換前の単語数」である．

図3.3は，推定$Jaccard$係数を組み入れた「中国語の知識カード」を分類する方式の処理の流れを示している．図3.3中の破線で囲った部分が,「中国語の知識カード」を扱うために新しく追加された処理ブロックを示している．中国語の単語列データを日本語単語列データに変換する際に用いられる「翻訳用辞書」には，分類処理の際に参照される「分類サンプルである日本語知識カード」に含まれている単語を先ずは登録する．加えて，IT処理技術に関する書類，新聞，インタネット記事などから抽出したIT用語も登録する．

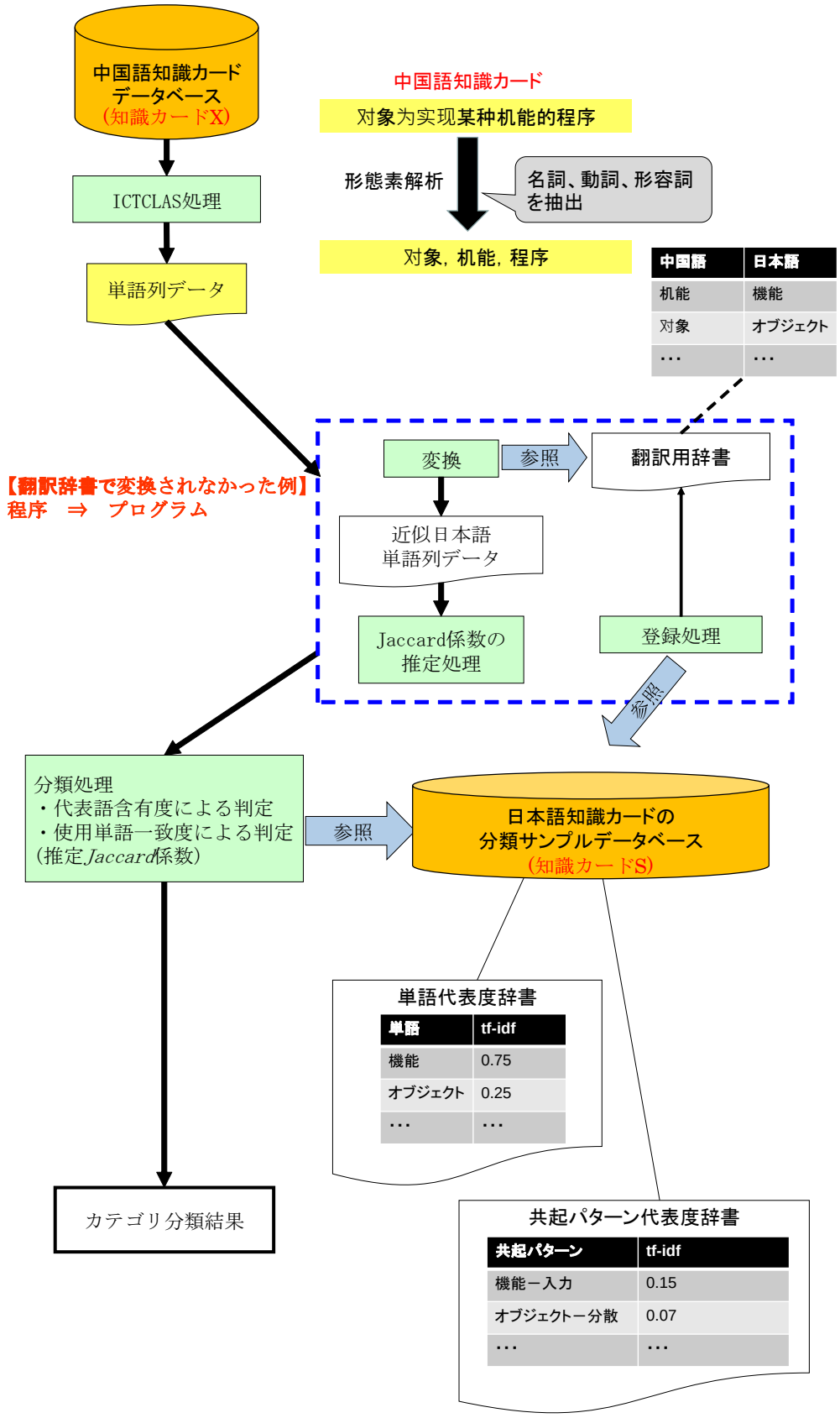

図 3.3: 中国語知識カードの分類方式

3.3　実験結果ならびに分類精度の評価

3.3.1　実験条件

「凌佳内部管理システム」に蓄積されている 570 件の中国語知識カードを実験に用いた．文として最も長い知識カードには 81 単語 (その中で，名詞は 38 単語) が含まれ，最も短い知識カードには 3 単語 (その中で，名詞は 2 単語) が含まれ，このことは内容だけでなく質問として簡潔なものもあれば，詳しく記述されるものがあることを示している．570 件の知識カードに含まれる文の平均単語数は 22 単語 (名詞に限れば，11 単語) である．翻訳用辞書としては，情報分野の専門用語を中心に 1587 個の単語を登録したものを使用した.「開発技術，基礎技術，内製ツールの使い方，日常業務，一般ビジネス知識，プロジェクト管理」の 6 種類のカテゴリに関して，人手により 36 件の日本語知識カードを予め登録した．

分類精度としては，再現率，適合率，さらに再現率と適合率の調和平均である F 値を用いる．変換漏れ補正率を決めるために，570 件の中国語知識カードから無作為で 100 件を抽出し，F 値が最大となる w を求めることとする．

以上をもとにした，分類精度の実験手順は，以下の通りである．

ステップ 1: 570 件の中国語知識カードから無作為に 100 件を抽出し，w の値を変えて，図 3.3 の処理から F 値を求め，F 値が最大となる w の値を算出する．

ステップ 2: 求めた w を用いて，ステップ 1 で用いた 100 件の中国語知識カードを除いた 470 件に対して，図 3.3 の処理から再現率，適合率，F 値を算出する．

ステップ 3: 比較するために，ステップ 2 で用いた 470 件の中国語知識カードを人手で翻訳した日本語知識カード，及びグーグル翻訳 [3] による日本語知識カードのそれぞれに対して図 3.3 の処理から再現率，適合率，F 値を算出する．

[3]https://translate.google.co.jp/

ステップ 1 からステップ 3 にいたる評価実験を 10 回繰り返し，得られた再現率，適合率，F 値の平均を分類精度評価の対象とする．

3.3.2 実験結果と評価

3.3.1 節で示した実験手順に従い，変換漏れ補正率 w を算出し，求めた w を用いて分類実験を行った．その結果に関し，以下に述べる．

変換漏れ補正率の算出

図 3.4 に，無作為に抽出した 100 件の中国語知識データに対して，w の値を変えたときの F 値の変化を示す．

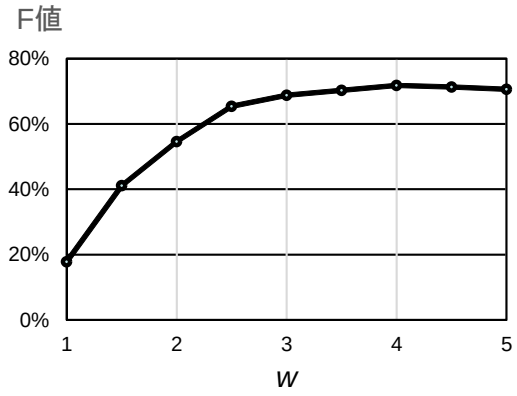

図 3.4: w の 1〜5 に対する F 値

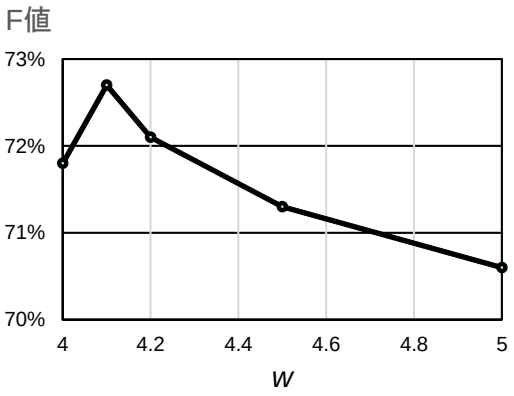

図 3.5: w の 4〜5 に対する F 値

$w = 1$ は，使用単語一致度の算出で “変換漏れ補正” をしないままに判定処理を行った場合を示している．図 3.4 に示すように，明らかに “変換漏れ補正” を行うことで，分類精度が向上していることがわかる．すなわち，「中国語の知識カード」の分類方式として，提案する “変換漏れ補正” を組み入れた判定処理が有効に機能していることとなる．

図 3.5 は，図 3.4 における w の 4〜5 の範囲に関して，刻み幅を細かく変化させたときのものである．図 3.5 から，変換漏れ補正率としての最適な値を求めることができる (今回の場合は，$w = 4.1$)．

分類結果

表 3.1 に 10 回の分類評価実験の結果を，各試行で用いた w とともに示す．

表 3.1: 各試行における分類精度 (単位:%)

試行	1	2	3	4	5	6	7	8	9	10
w	4.1	4.3	4.0	3.9	4.3	3.9	4.3	3.9	4.0	4.1
F 値	68.47	70.24	69.37	67.76	69.64	67.05	60.03	68.54	68.21	69.04
再現率	67.11	68.86	67.86	66.54	68.38	65.16	67.42	67.17	66.79	67.61
適合率	69.88	71.68	70.95	69.02	70.95	69.06	70.73	69.96	69.69	70.53

表 3.2 に，中国語知識カード，人手で翻訳した日本語知識カード，グーグル翻訳による日本語知識カードという 3 種類のデータセットに適用した分類実験の 10 回の平均値を示す．

表 3.2: 分類精度 (単位:%)

	中国語知識カード	人手で翻訳した日本語知識カード	グーグル翻訳による日本語知識カード
F 値	68.74	59.16	54.67
再現率	67.29	55.67	52.08
適合率	70.25	62.27	57.13

実験結果の分析

中国語知識カードに対して，再現率と適合率の平均値は約 67.3%と約 70.3%，F 値は約 68.7%となっている．この評価実験結果からは，提案方式の分類精度が人手あるいはグーグル翻訳で日本語に変換した知識カードの分類精度より高いことが明らかになり，提案方式は分類方式として有効であることがわかる．

提案方式が比較対象のデータセットよりもかなりいい結果が得られている要因として，以下のことがある．翻訳用辞書はそもそも IT 技術用語を中心に整備されていることから，分類カテゴリ内のサンプル知識カードに含まれる単語

が基本的には登録されている．そのために，変換後の共通単語数 N_{sx} を補正することで，分類カテゴリとの判定が適切に行われている．一方，人手による翻訳やグーグル翻訳では分類カテゴリ内のサンプル知識カードに含まれる単語に翻訳されない場合もあり，その結果 N_{sx} の値が低くなっている．

ただし，提案方式においても，本来とは異なるカテゴリへの誤分類が 149 件あり，これらの間違いパターンとその数は，以下のとおりである．

- 6 種類の分類カテゴリのいずれかに判定されるものを「その他」のカテゴリに分類 (58 件)
- 分類カテゴリの間違い (23 件)
- 複数のカテゴリに分類 (68 件)

これらの誤分類の原因を分析したところ，形態素解析ツール ICTCLAS に起因する以下に示す 2 つの要因が明らかになった．

(1) 複合名詞の扱いの違い**:** 予め分類カテゴリに登録している分類サンプルとしての日本語知識カードの文には，例えば,「販売システム」といった複合名詞と呼ばれる単語が存在し，これに対して “茶筅” を適用した場合に複合名詞は一つの単語として分解される．しかし，中国語知識カードでの同じ意味を持つに対する “ICTCLAS” の適用では,「销售系统」は「销售」と「系统」といった形で別々の名詞として分解されてしまう．この結果，中国語知識カードを変換した日本語知識カードの単語数が多くなることと，共通単語としてこれらが同定されないために推定 $Jaccard$ 係数の値としても誤った値が算出されている．

(2) 品詞の扱いの違い**:** “ICTCLAS” による分解された単語の品詞は，例えば「管理 (中国語でも同じ漢字が使われる)」は名詞と動詞の両方の意味を持つ単語であるが，“ICTCLAS” では動詞として出力される．提案方式の判定処理では，名詞を用いているために，中国語知識カードを変換した日本語知識カードでは，これらの単語が判定の計算処理で用いられなくなり，誤った分類カテゴリに判定されている．このような名詞と動詞の両方の意味を持つ中国語の単語としては，例えば「管理，更新，动作」といったように，比較的ソフトウェア開発現場で頻繁に用いられるものがある．

3.4 結言

中国オフショアソフトウェア開発企業で用いられる知識共有システムの機能として，日本語のみならず中国語で記述された知識カードを正しく分類管理することが求められている．これに対して，分類判定に用いられる $Jaccard$ 係数を不完全な「中国語－日本語翻訳辞書」を用いる場合に推定する方式を提案し，評価実験の結果より中国語知識カードの 68.7% が正しいカテゴリに分類可能となることが明らかとなった．人手による翻訳やグーグル翻訳による日本語知識カードの分類精度より提案方式では高い分類精度が得られており，これは翻訳用辞書によって内容を特徴付ける単語のみが変換されることで，判定処理が適切に行われていることがわかった．もちろん，68.7%の分類精度は，実用的な観点からは不十分である．ただし，使用単語一致度の算出で “変換漏れ補正” をしないままに判定処理を行った場合には，分類精度を示す F 値が約 20%と低く,「中国語の知識カード」の分類方式として，“変換漏れ補正” を組み入れた場合には判定処理が有効に機能していることがわかった．

そこで，今後の精度向上に向けての課題として，第一に実験結果の分析で述べたような中国語形態素解析ツール ICTCLAS に起因する「複合名詞」や「品詞」に対しての「特殊辞書」の構築があげられる．ただし,「辞書構築」に予め人手や時間をかけることはできないために，誤分類が中国語形態素解析ツールに起因するということの判定とその判定結果をもとにした事例からの「特殊辞書」の自動構築についての取り組みが必要となる．第二に，変換漏れ補正率 w は本論文では実験的に求めているが，人手や時間をかけずに求める必要がある．

第4章

日本語と中国語が混在する知識カードの分類方式

4.1 緒言

本章では，第2章における「凌佳知識共有システム」に蓄積されている「日本語と中国語が知識カードの一つの文章中に混在する」という状況に対する知識カード分類方式を提案する．実際,「凌佳知識共有システム」では，日本語のみを用いて登録された知識カードが66%，中国語のみを用いて登録された知識カードが29%，両言語が混ざった形で登録された知識カードが5%となっている．両言語が混ざった知識カードは，現状では全体に対する占める割合が少ないが，新たに雇用する中国人SEがいる限りは，日本語能力の不足から両言語を用いた知識カードは増え続けることになる．第3章で述べたとおり，言及している知識内容が同じ「日本語の知識カード」の分類カテゴリに，このような日本語と中国語が混在した知識カードも蓄積する必要があり，日本語のみの知識カードや中国語のみの知識カードの分類方式で用いた形態素解析による単語群の統計的情報を単純には適用できない．

このような多言語処理に対する研究としては，対訳用例や翻訳辞書を用いてある言語から別の言語への「翻訳タスク」に関してはなされているが[60][61]，一つの文の中に複数の言語が用いられているという言語学的には間違っている状況への取り組みはなされていない．

そこで，日本語と中国語が知識カードの一つの文中に混在している場合に，

該当する中国語の単語のみを翻訳用辞書によって日本語に正しく変換することで，第 3 章で提案した単語列データを使った分類処理を適用することを行う．この際にも不完全な翻訳用辞書を利用することを前提として，日本語と中国語が混在する知識カードを分類する方式を提案する．

以下，4.2 節では，日本語と中国語が一つの文中に含まれる場合に，形態素解析を実行して単語列データを得た場合の問題点を述べた上で，この課題に対するアプローチとそれに基づく分類方式について述べる．4.3 節では，実験による評価を行い，提案方式の有効性を示す．4.4 節では，本章のまとめを述べる．

4.2　日中両言語が混在する知識カードの分類方式

4.2.1　形態素解析ツールの適用における課題

「凌佳知識共有システム」に蓄積されている「日本語と中国語が一つの文中に混在する知識カード」の例を，表 4.1 に示す．なお，質問文に関しては，同じ意味の日本語を括弧内に示す．

図 4.1 は，“茶筅”，あるいは “ICTCLAS” に両言語が混在する文を入力し，形態素解析を行った結果の例を示す．

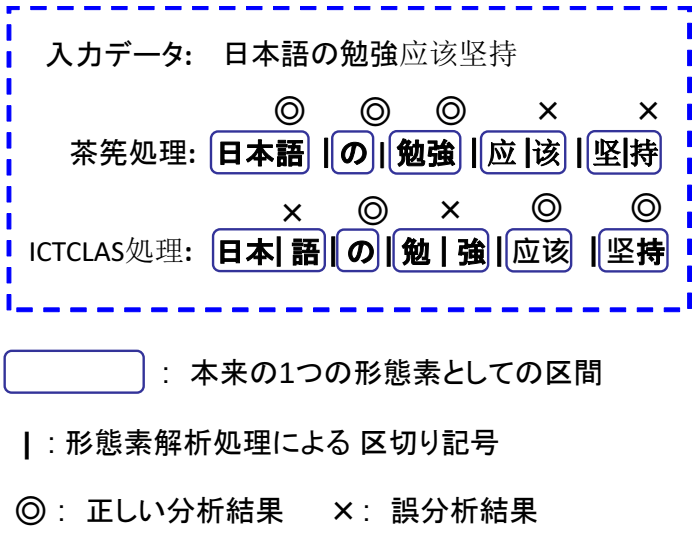

図 4.1: 両言語が混在する文の形態素解析結果の例

表 4.1: 日本語と中国語が混在する知識カードの例

分類カテゴリ	知識カード (上段：質問，下段：回答)
開発技術	请教 gcc 不同版本的编译方法 (gcc コンパイラのバージョンによる違いを教えてください．)
	GCC 4.3:Intel Core 2 や AMD Geode プロセッサーへの対応を強化したバージョン GCC 4.4:graphite ブランチで開発されてきた最適化機構が正式に取り込まれた．
基礎技術	请教スレッド排他制御的例子 (java スレッドの排他制御の例を教えてください．)
	synchronized(排他的に利用したいインスタンスの式)
内製ツールの使い方	JAVA 中的 Debug 方法を教えてください．(java のデバッガでのステップ実行の方法を教えてください．)
	双击你认为会出错的那部分代码之前断点这表示断点测试，使用 Debug 运行，当运行到断点，它就会停下来，然后 你确认进入 debug．可以一 步一 步的往下手动按 步骤走，同时查看属性值．这样就可以发现错误原因．
日常業務	B 票内容的填写，注意することは何ですか．(B 票記載内容記入で，注意することは何ですか．)
	直接的原因:プログラム的にどうなっていたかを具体的に記述すること．
一般ビジネス知識	请教日本語ビジネス文書作法要点．(日本語ビジネス文書作法のコツを教えてください．)
	和写中文商业资料一 样，内容要正 确，明 确，简洁．
プロジェクト管理	利用 WBS 能提高コスト見積もりの精度?(ワークブレクダウンストラクチャは，コスト見積もりの精度を高めますか．)
	估算方法有三类，WBS 估算是最准 确的 1.类似 项目估价算 (TopDown) 2. 系数模式估算 (功能点分析 (FP)，COCOMO) 3.BottomUp 估算 (WBS)

形態素解析ツールに両言語が混在する知識カードを入力すると，当然ながら文法的な分析は行わず，語彙だけを単語列データに分解しようとする．英語とは違い，日本語と中国語はともに漢字が用いられ，同じ漢字もかなりの数が含まれている．コンピュータ業界の文字列標準 Unicode では，日本，中国，韓国で使っている漢字を CJK(Chinese, Japanese & Korean) という文字セットに登録しているが，1 つの漢字に対して同一のコードが指定されているため，言語の判定に用いることができない．従って，漢字が含まれる知識カードの言語を見分けることができないために，両言語が混在する知識カードの形態素解析ツールによる分析結果は正しいものが得られない．

4.2.2 課題へのアプローチ

4.2.1 節で示した課題を解決するため，本研究では両言語の形態素解析ツールの特性を利用する．“茶筅” と “ICTCLAS” は，それぞれの言語に対応する文を正しく解析し，単語列データに分解できる．しかし，本来想定しているものと異なる言語の文を解析すれば，1 つの語彙をさらに小さい文字に分ける [80][81]．例えば，ICTCLAS が “勉強” という日本語の単語を解析すると，中国語にその単語がないので，“勉” と “強” という 2 つの文字に分解する．“茶筅” で中国語の文を解析するときにも，同様の結果が導かれる．

図 4.2 は，本来想定しているものと異なる言語で書かれた文を入力データとして，それぞれの形態素解析ツールを用いた際の出力結果を示したものである．ここで，2 つの形態素解析ツールの出力結果を文の先頭から比較し，単語列データとして長いものを順次採用することで，両言語が混在している文であっても，言語判定を行わずとも正しい単語列データが得られている．実際，図 4.3 に示すように，日本語と中国語が混在する文に適用した予備実験の結果では，正しい単語列データが導かれている．

従って，日本語知識カード，中国語知識カード，両言語が一つの文に混在する知識カードのいずれに対しても，2 つの形態素解析ツールで単語列データに分解し，それらの分解された 2 つの単語列データから正しい解析結果を生成するというアプローチを行う．

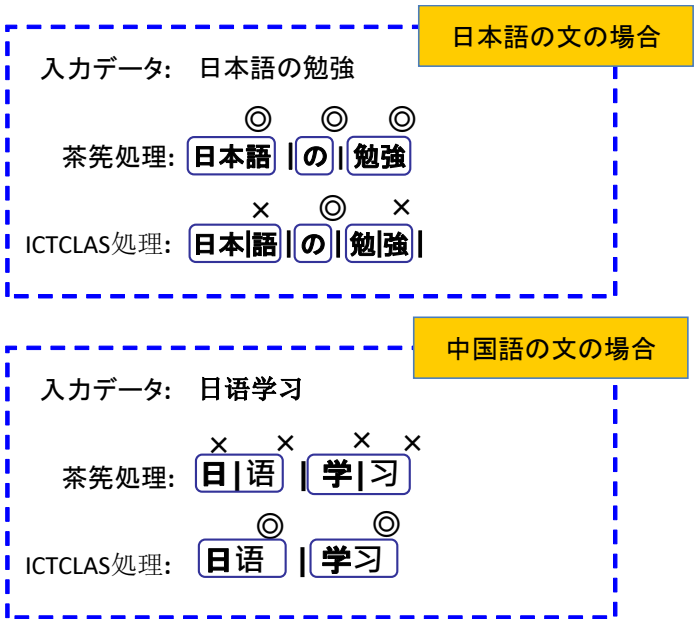

図 4.2: 言語ごとの形態素解析結果

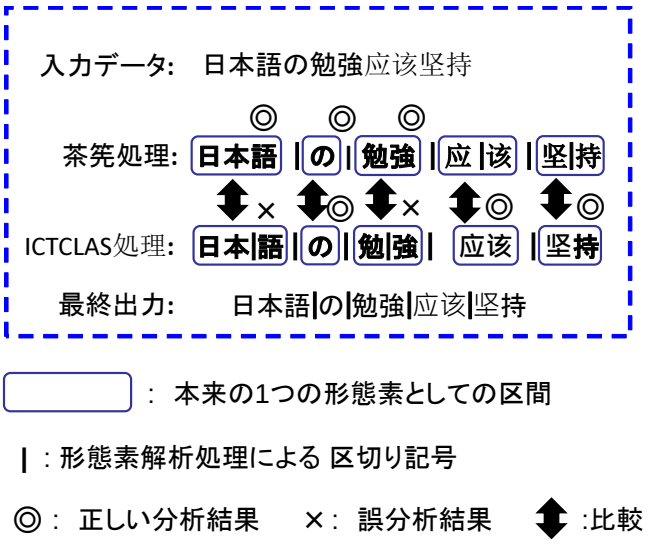

図 4.3: 言語ごとの形態素解析結果

4.2.3 ドメイン辞書を用いた知識カード分類方式

図 4.4 は，日本語，中国語或いは両言語が混在する知識カードを統一的に分類する方式の処理の流れを示している．

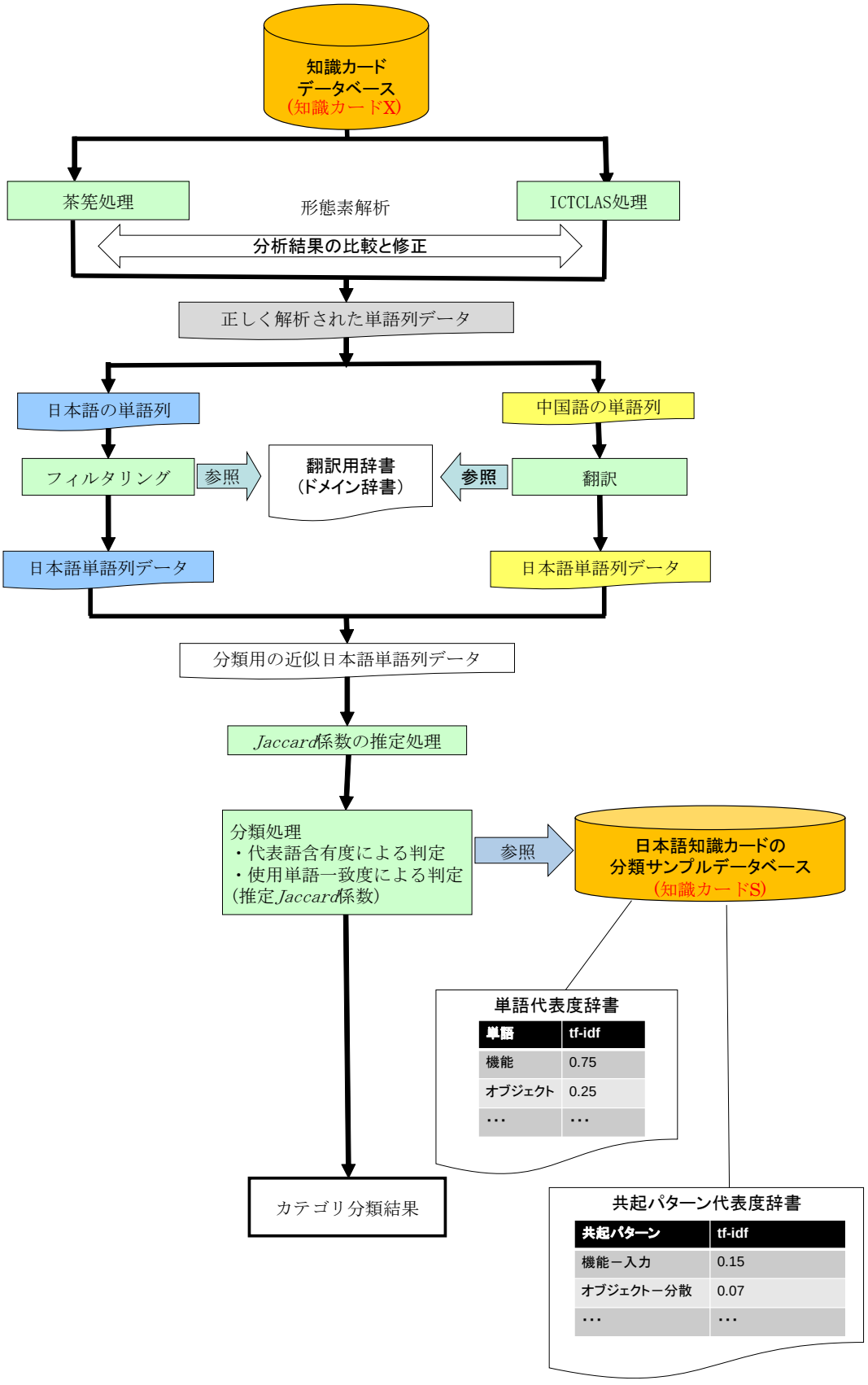

図 4.4: 知識カードの統一的分類方式

知識カードの言語を判定せず，日本語と中国語の形態素解析ツールを用いて，知識カードの内容を単語列データに分解する．日本語の形態素解析ツールとして “茶筅” を，中国語の形態素解析ツールとして “ICTCLAS” を用いて解析する．その後，2 つの形態素解析ツールで処理された結果を比較し，正しく解析された単語列データを生成する．解析された単語列データの中の中国語の部分のデータは，第 3 章で用いた翻訳用辞書を用いて，日本語の単語列データに変換する．一方，解析された単語列データの中の日本語の部分のデータに対しても，同じ辞書を用いてフィルタリングを行い，日本語の単語列データに変換する．これは，両言語を用いて書かれた文というのは，そもそも日本語能力のレベルが低いために発生していると考えられ，そのために日本語としても意味をなさない，あるいは不正確な単語が含まれるからである．このような日本語としても意味をなさない，あるいは不正確な日本語の単語列データをそのまま残して判定処理に用いると，必然的に推定 $Jaccard$ 係数の分母が大きくなり，判定処理に誤りを引き起こす．このことを回避するために，辞書に含まれる単語を専門分野を表す重要語として，解析された日本語単語列データにフィルタをかけることとする．このように第 3 章で用いた翻訳用辞書は，本章で提案する方式では,「翻訳」と「フィルタ」の用途で利用することから，改めて “ドメイン辞書” と呼ぶこととする．

ただし，このような “ドメイン辞書” によるフィルタ処理では，辞書に含まれない単語を全て取り除くので，“翻訳” の場合と同様に必要な単語が処理から漏れるという問題も発生する．すなわち，そのまま翻訳及びフィルタした日本語の単語列データをもとに $Jaccard$ 係数を計算するのではなく，第 3 章と同じく漏れ補正率をもとに推定 $Jaccard$ 係数を算出し，判定処理に用いる．

4.3 実験結果ならびに分類精度評価

4.3.1 実験条件

「凌佳内部管理システム」に蓄積されている「日本語と中国語が混在して用いられている知識カード」の内容は，6 つのカテゴリに万遍なく対応していない．そこで，第 3 章の実験で用いた「凌佳内部管理システム」に蓄積されている 570 件の中国語知識カードをもとに，人手で翻訳した日本語知識カード，人

手で作成した両言語が混在する知識カードを作成した．ドメイン辞書には，第3章と同じ情報分野の専門用語を中心に1587個の単語を登録した．また,「開発技術，基礎技術，内製ツールの使い方，日常業務，一般ビジネス知識，プロジェクト管理」の6種類のカテゴリに関して，人手により36件の日本語知識カードを予め登録した．漏れ補正率を決めるために，570件の中国語知識カードから無作為で100件を抽出し，F値が最大となるwを求めた．

以上をもとにした，分類精度の評価手順は，以下の通りである．

ステップ1: 570件の中国語知識カードから無作為に100件を抽出し，wの値を変えて，図4.4の処理からF値を求め，F値が最大となるwの値を算出する．

ステップ2: ステップ1で用いた100件の中国語知識カードを除いた470件の中国語知識カード(以下，CHNと表記)に対して，人手で翻訳した日本語知識カード(以下，JPNと表記)，人手で作成した両言語が混在する知識カード(以下，C&Jと表記)の3種類のデータセットを作成する．

ステップ3: ステップ1で求めたwを用いて，470件の“C&J”，“CHN”，“JPN”のそれぞれに対して，図4.4の処理から再現率，適合率，F値を算出する．

ステップ1からステップ3にいたる評価実験を10回繰り返し，得られた再現率，適合率，F値の平均を分類精度評価の対象とする．

4.3.2 実験結果と評価

4.3.1節で示した実験手順に従い，表4.2に3種類のデータセットに適用した分類実験としての分類精度の平均値を示す．

表 4.2: 分類精度 (単位:%)

	C&J	CHN	JPN
F 値	68.54	68.74	69.32
再現率	66.49	67.29	67.80
適合率	70.76	70.25	70.85

表 4.2 より，中国語知識カード及び両言語が混在する知識カードを分類した場合の F 値はそれぞれ 68.74%，68.54%となっており，日本語知識カードを分類した場合の F 値の 69.32%と比較しても遜色ない精度が得られた．評価実験の結果より，提案方式によって，日本語と中国語が混在する知識カードに対する分類方式の有効性を示すことができた．

4.4 結言

中国オフショアソフトウェア開発企業で用いられている知識管理システムの機能として，日本語のみならず中国語で記述された知識カードを正しく分類管理することが求められている．特に，日本語能力の低い中国人 SE が記述した知識カードには，中国語と日本語が一つの文に混在している場合があり，これまで単一言語を前提とした分類方式では処理できなかったことを，2 つの形態素解析ツールの特性を利用して単語列データを生成し，さらに “ドメイン辞書” による「翻訳」と「フィルタ」という処理を組み入れた分類方式を提案した．評価実験で用いた中国語知識カードから作成した日中混在カード，人手で翻訳した日本語知識カードのいずれであってもほぼ同程度の分類精度となっており，提案方式がこれら 3 種類の知識カードを分類する統一処理方式として機能することが明らかになった．

今後の精度向上に向けての課題として，“ドメイン辞書” の整備がある．評価実験が示しているように，“ドメイン辞書” によって，分類カテゴリ内のサンプル知識カードに含まれる単語に関係して，分類したい知識カードから「翻訳」と「フィルタ」により分類用の適切な単語列データが得られていることから，この “ドメイン辞書” を人手や時間をかけずに行う必要がある．定期的に

IT 処理技術に関する書類，新聞，インタネット記事などから抽出した IT 用語をもとに “ドメイン辞書” を構築していることから，Web マイニング [82][83] などの技術を援用した自動構築が可能と考える．

第5章

結論

5.1　本研究のまとめ

本論文では，オフショアソフトウェア開発企業における日本語と中国語を混在して利用できる知識共有システム，知識共有システムに蓄積された中国語知識カードの分類方式，さらに知識共有システムに蓄積された日本語及び中国語で記述された知識カードの分類方式の開発を目的として，BBS 形式での日本語と中国語による知識共有方式，翻訳用辞書を用いた中国語知識カードの変換処理を組み入れた知識分類方式，2 つの形態素解析ツール併用とドメイン辞書を用いた変換処理を組み入れた知識分類方式を提案した．本論文では，これらの研究成果を以下の 4 章に分けて述べた．

第 1 章では，日本から中国向けのオフショア開発が増大する中で，オフショア開発特有の工程管理，品質管理の問題の大きな要因として，開発経験の浅い若手中国人 SE の大量雇用があり，そのために継続的な教育や経営として開発現場の技術者レベル把握が求められている背景について述べた．日常的にプロジェクト横断で利用できる知識共有システムの構築が必要であるとともに，オフショアソフトウェア開発現場では「日本語の使用が推奨されているにもかかわらず，日本語と中国語が使われている」という現実に対して，知識共有システムの技術課題を関連研究とともに整理した．明らかとなった課題に対して，本研究での解決方針を述べた．

第 2 章では，BBS を用いて開発現場でやり取りされる Q&A のデータを共有する知識共有システムを開発した．技術者間の「質問-回答」という作業をス

ムーズかつ実行性のあるものにするために,「指定回答者の設定」や「回答者への通知」，さらに「質問が解決済みであるかのステータス表示」といった機能を組み入れ，日常的に利用を促す環境構築を行った．実利用による評価結果からは，プログラマによる利用が日常的に行われていること，継続的教育としてどのカテゴリの教育強化が必要であることなどオフショアソフトウェア開発企業としての実態が明らかになるとともに,「日本語の知識カード」,「中国語の知識カード」,「日本語と中国語が混在して用いられた知識カード」を活用するためには，知識共有システムとして，自動分類する仕組みが必要であることを明確化した．

第 3 章では，カテゴリ辞書を用いた知識カード分類方式として，日本語知識カードを対象にした分類方式を中国語知識カードに適用するために,「翻訳用辞書」の組み込みとその結果発生する「変換漏れ」に対する補正処理を組み入れた分類方式について述べた.「翻訳用辞書」は，IT 処理技術に関する用語を登録したものであり，IT 分野の技術革新が速いことから新規用語や名称変更となった用語などを考慮すると不完全な辞書を前提にした分類方式の必要性を述べた．その際に，不完全な辞書に対応するための「変換漏れ補正率」を定義し，この「変換漏れ補正率」を用いた判定処理を組み入れた中国語知識カードの分類方式を提案した，提案方式を第 2 章での知識共有システムに蓄積された実際の知識カードデータに適用し，人手による翻訳やグーグル翻訳による日本語知識カードの分類精度より高い分類精度であることを確認し，さらに誤分類に関する分析・考察を行った．また，使用単語一致度の算出で “変換漏れ補正” をしないままに判定処理を行った場合には，分類精度を示す F 値が約 20%と低く，最適な “変換漏れ補正” をもとに判定処理を行った場合には，分類精度を示す F 値が約 69%となることがわかった．このことにより,「中国語の知識カード」の分類方式として，“変換漏れ補正” を組み入れた判定処理が有効に機能していることがわかった．

第 4 章では，中国語と日本語が知識カードの一つの文中に混在する場合の分類について述べた．両言語が一つの文に混在する場合，形態素解析ツールのみを用いて文法的な言語判定を実施せずに，形態素解析結果として正しく単語列データを抽出する処理を組み入れ，さらに第 3 章での翻訳用辞書を中国語の単語の変換処理と日本語の単語のフィルタリングの両方に利用する分類方式を提案した．提案方式を第 3 章の実験で用いた中国語知識データ，さらにそれをも

とに人手で翻訳した日本語知識カード，人手で作成した日本語と中国語が一つの文に混在する知識カードを作成し，これら 3 種類のデータによる分類実験を行った．同じ意味内容を持つ 3 種類のデータセットの分類精度が，いずれも約 70%ほどとなり，提案方式の有効性を確認するとともに，精度向上への考察を行った．

5.2 今後の研究課題

最後に今後の課題について，3 点述べる．

1 点目の課題として，分類したカテゴリに蓄積されている知識カードを分析し，それらの内容からのキーワード抽出や要点整理があげられる．あるカテゴリに分類された知識カードの内容そのものは多岐である．カテゴリの細分化という要望も実利用評価ではあがっていたが，細分化しすぎることはかえって同じカテゴリの関連知識の派生学習を妨げることになる．ただし，カテゴリ内には極端な場合には 1 度しかやり取りされなかった内容の知識カードもあれば，文は異なっていても内容的には同じものが頻繁にやり取りされている知識カードもありえる．そこで，カテゴリ内の知識カードから頻繁に現れるキーワード抽出や内容的に同じもので頻繁にやり取りされている知識カードを一つの知識カードにまとめ直すといった要約が考えられる．キーワード抽出に関しては，第 3 章で述べた *tf-idf* をもとに拡張した方式が先ずは考えられるが，要約に関しては，これまで文書処理技術で培われた技術 [20] が口語的表現を対象にしていないことから，新しい視点での解決が必要と考える．

2 点目の課題として，分類サンプルデータベースの自動再構築である．本論文で用いた際にも，分類サンプルデータベースは予め 6 つの分野カテゴリと「その他」として整備されていたが，その結果として知識共有システムの運用開始後は 6 つの分野カテゴリに対応しない知識カードは全て「その他」に分類されてしまう．この「その他」カテゴリの知識カードには，技術者教育の面，企業経営の面から有用そうな内容が含まれている．人手で「その他」のカテゴリを分析するのでなく，クラスタリングといった技術を援用した分類サンプルデータベースの再構築が必要である．

3 点目の課題として，知識共有システムの利用環境として，端末非依存のシームレス化が考えられる．PC 上の Web ブラウザを利用した BBS 利用だけでな

く，スマートフォンやタブレット端末などのモバイル環境における BBS 利用を促進することで，自宅や移動時などでの学習にも役立てることができる．オフィスで見ていた知識カードそのものや分類カテゴリがモバイル端末でセッションを引き継ぐ形で再開できるなどの利用環境のシームレス化である．セキュリティ要件としては，アクセス制限だけでなく，閲覧内容のローカル保存 (コピー，ペースト操作) は不可といったことに留意しつつ，利便性の高い知識共有システムとして整備する必要がある．

謝辞

本研究の全過程を通じて，終始懇切丁寧なるご指導とご鞭撻，格別のご配慮を賜りました大阪大学大学院情報科学研究科マルチメディア工学専攻薦田憲久教授に深く感謝申し上げます．

本研究をまとめるにあたり，貴重なお時間を割いて頂き，丁寧なるご教示を賜りました大阪大学大学院情報科学研究科マルチメディア工学専攻藤原融教授，神奈川大学工学部情報システム創成学科秋吉政徳教授（元 大阪大学大学院情報科学研究科マルチメディア工学専攻准教授），工学院大学工学部情報通信工学科馬場健一教授（元 大阪大学 サイバーメディアセンター応用情報システム研究部門准教授）に深く感謝申し上げます．特に，秋吉政徳教授には，第3章，第4章の研究に関する共同研究者として，研究方針，研究方法，研究のまとめ方等，直接のご指導，ご議論を頂きました。

大学院博士後期課程において，情報工学全般に関して親切なるご指導とご助言を賜りました大阪大学大学院情報科学研究科マルチメディア工学専攻西尾章治郎教授，関西学院大学理工学部情報科学科岸野文郎教授（元 大阪大学大学院情報科学研究科マルチメディア工学専攻教授），大阪大学大学院基礎工学研究科システム創成専攻細田耕教授（元 大阪大学大学院情報科学研究科マルチメディア工学専攻教授），大阪大学サイバーメディアセンター応用情報システム研究部門下條真司教授に深く感謝申し上げます．

また，本研究全般について議論ならびにご支援を頂いた薦田研究室の前助教竹内亨博士 (現 NTT 未来ねっと研究所)，鮫島正樹助教に深く感謝します．

第2章の研究に関しては，共同研究者として様々なご討論，ご助言ならびに支援をいただきました済南凌佳科技有限公司元副総理 曲義暁氏，副総理王作埼氏，管理本部長張洪志氏，オフショア開発事業部長林祥華氏、張秀芬氏、株式会社日本凌佳システム伊藤優副社長，王建国副本部長をはじめ，済南凌佳科技有限公司と株式会社日本凌佳システムの各位に心から御礼申し上げます．第3章，第4章の研究に関しては，基礎となる日本語で書かれた知識分類方式の

プログラムを提供頂いた薦田研究室博士前期課程学生の飯田薫氏（現 (株) 日本 IBM），根来啓輔氏（現 (株) リクルート住まいカンパニー），岩井康一氏（現 (株) 野村総研），劉暁鵬氏をはじめとする各位に感謝します．特に，劉暁鵬氏には，知識カードの翻訳，各種実験の実施に際して，多大な支援を頂きました．

最後に，幼少の頃より健康な心身への成長を見守り，また学位取得に向け常に応援してくれた父 蔡志浩，母 王潔玉に感謝致します．また，本論文の執筆にあたり，応援してくれるとともに，執筆の時間の確保に配慮してくれた妻 王平と子供たちに心から感謝致します．

参考文献

[1] IPA 独立行政法人, 情報処理推進機構 IT 人材白書, (2009-2012).

[2] Carnegie Mellon University, "*The Capability Maturity Model: Guidelines for Improving the Software Process*", Addison Wesley Longman, Inc. (1994).

[3] 辻洋, 守安隆, 盛忠起, "オフショア・ソフトウェア開発の進化と技術者の経験知", 情報処理学会誌, Vol. 49, No. 5, pp. 551–557 (2008).

[4] 辻洋, 野々村琢人, 三部良太, "オフショア・ソフトウェア開発向けのシステムズ・アプローチ", システム/制御/情報学会誌, Vol. 52, No. 2, pp. 20–25 (2008).

[5] B. Meyer and M. Joseph (eds.), "*Software Engineering Approches for Offshore and Outsourced Development*", Lectures Notes in Computer Science 4716, Springer (2007).

[6] A. Tiwana, H. Bush, H. Tsuji, and K. Yoshida "Myths and Paradoxes in Japanese IT Outsourcing", *Communications of the ACM*, Vol. 51, No. 10, pp. 141–145 (2008).

[7] 北島義弘 監修, "オフショア開発 PRESS", 技術評論社 (2008).

[8] 齋藤邦夫, 門司太郎, 鈴木馨, "中国オフショア開発のプロセス改善への取り組み", 日立 TO 技報, No. 15, pp. 72–76 (2009).

[9] 祖国威, "中国でのオフショア仕様書チェックシステム", 東芝レビュー, Vol. 62, No. 1, pp. 70–71 (2007).

[10] 赤津雅晴, "成功するアウトソーシングの勘所", 情報処理学会誌, Vol. 46, No. 5, pp. 534–549 (2005).

[11] 野中郁次郎, 竹内弘高（梅本勝博訳）, 知識創造企業, 東洋経済新聞社 (1996)

[12] 幡鎌博, 津田宏, 益岡竜介, “ナレッジマネジメントへむけて：知識検索・整理および基盤技術”, 人工知能学会誌, Vol. 13, No. 6, pp. 912–919 (1998).

[13] 斉藤典明, 金井敦, “組織知識継承を実現する死蔵されない共有フォルダ構成法”, 情報処理学会論文誌, Vol. 54, No. 1, pp. 295–308 (2013).

[14] 那須川哲哉, 諸橋正幸, 長野徹, “テキストマイニング：膨大な文書データの自動分析による知識発見”, 情報処理学会誌, Vol. 40, No. 4, pp. 358–364 (1999).

[15] 那須川哲哉, 河野浩之, 有村博紀, “テキストマイニング基盤技術”, 人工知能学会誌, Vol. 16, No. 2, pp. 201–211 (2001).

[16] 工藤拓, 山本薫, 坪井祐太, 松本裕治, “言語情報を利用したテキストマイニング”, 情報処理学会研究報告 自然言語処理研究会報告 2002(20), pp. 65–72 (2002).

[17] 保田明夫, “テキストマイニングの概要”, 電気学会論文誌C, Vol. 125, No. 5, pp. 682–689 (2005).

[18] C. Dozier and P. Jackson, “Mining Text For Expert Witnesses”, *IEEE Software*, Vol. 22, No. 3, pp. 94–100 (2005).

[19] 松井くにお, 渡部勇, 内野寛治, “ナレッジマネジメントにおけるテキストマイニング”, 情報処理学会誌, Vol. 47, No. 8, pp .893–899 (2006).

[20] 奥村学, “テキスト自動要約”, 情報処理学会誌, Vol. 45, No. 6, pp. 574–579 (2004).

[21] 宮崎正弘, 白井諭, 林良彦, “日英翻訳システム ALT-J/E における日本語解析技術”, 情報処理学会 全国大会講演論文集, pp. 1751–1752 (1986).

[22] 隅田英一郎, “機械翻訳のいま　統計的手法を中心に”, 情報管理, Vol. 57, No. 1, pp. 12–21 (2014).

[23] 祖国威, 吉村裕美子, 加納敏行, “構文的特性に着目した可読性診断技術”, 東芝レビュー, Vol. 66, No. 4, pp. 51–55 (2011).

[24] 辻洋, 間瀬久雄, 津原進, 衣川一久, “ヘルプデスクにおける類似文書検索システムの構成と機能について”, 情報処理学会研究報告 デジタルドキュメント研究会報告 97(116), pp. 23–30 (1997)

[25] 仲川こころ, 高田喜朗, 関浩之, “可変なカテゴリ構造を用いた文書検索支援手法”, 情報処理学会論文誌, Vol. 42, No. 10, pp. 2441–2453 (2001).

[26] 森本由起子, 間瀬久雄, 平井千秋, 衣川一久, “問合わせ事例を活用したヘルプデスクオペレータ支援機能の開発”, 情報処理学会論文誌, Vol. 44, No. 7, pp. 1731–1739 (2003).

[27] 市川宙, 橋本泰一, 徳永健伸, 田中穂積, “テキスト構文構造類似度を用いた類似文検索手法”, 情報処理学会研究報告 データベース・システム研究会報告 2005(42), pp. 39–46 (2005).

[28] 村田真樹, 内山将夫, 井佐原均, “類似度に基づく推論を用いた質問応答システム”, 情報処理学会研究報告 自然言語処理研究会報告 2000(11), pp. 181–188 (2000).

[29] 佐々木裕, 磯崎秀樹, 平博順, 平尾努, 賀沢秀人, 鈴木潤, 国領弘治, 前田英作, “SAIQA：大量文書に基づく質問応答システム”, 情報処理学会研究報告 自然言語処理研究会報告 2001(86), pp. 77-82 (2001).

[30] 日高直哉, 桝井文人, “質問応答における回答絞り込み手法の比較”, 人工知能学会全国大会論文集, pp. 1–3 (2003).

[31] 福本淳一, “質問応答技術”, 情報処理学会誌, Vol. 45, No. 6, pp. 580–585 (2004).

[32] 陳亮, 徳田尚之, 候平魁, 永井明, 陳若愚, “良くある質問（FAQ）のコンテンツ・構文検索を組み合わせた自然言語質疑応答システム”, 情報処理学会研究報告 自然言語処理研究会報告 2005(11), pp. 69–79 (2005).

[33] W. Song, M. Feng, N. Gu, and L. Wenyin., “Question Similarity Calculation for FAQ Answering”, in *Porc. of the third Intl. Conf. on Semantics, Knowledge and Grid*, pp. 298–301 (2007).

[34] H. Kim and J. Seo, “Cluster-based FAQ Retrieval Using Latent Term Weights”, *IEEE Intelligent Systems*, Vol. 23, Issue 2, pp. 58–65 (2008).

[35] Z. M. Juan, “An Effective Similarity Measurement for FAQ Question Answering System”, in *Proc. of 2010 Intl. Conf. on Electrical and COntrol Engineering*, pp. 4638–4641 (2010).

[36] 張玉潔, 馬青, 井佐原均, “英語を介した日中対訳辞書の自動構築”, 自然言語処理, Vol.12, No.2, pp. 63–85 (2005).

[37] 安田圭志, 隅田英一郎, “日中特許対訳コーパスを用いた対訳辞書の自動構築”, 言語処理学会 第 19 回年次大会, pp. 306–309 (2013).

[38] 岩山真, 徳永健, “自動文書分類のための新しい確率モデル”, 情報処理学会研究報告 情報学基礎研究会報告 94(37), pp. 47–52 (1994).

[39] 湯浅夏樹, 上田徹, 外川文雄, “大量文書データ中の単語間共起を利用した文書分類”, 情報処理学会論文誌, Vol. 36, No. 8, pp. 1819–1827 (1995).

[40] 久光徹, 丹羽芳樹, “組み合わせ的確率モデルに基づく特徴単語選択方法”, 情報処理学会研究報告 自然言語処理研究会報告 2000(107), pp. 85–90 (2000).

[41] 高村大也, 松本裕治, “文書分類のための共クラスタリング”, 情報処理学会論文誌, Vol. 44, No. 2, pp. 443–450 (2003).

[42] 正田備也, 高須淳宏, 安達淳, “混合ディリクレ分布を用いた文書分類の精度について”, 情報処理学会論文誌 データベース 48, pp. 14–26 (2007).

[43] 岡野原大輔, 辻井潤一, “全ての部分文字列を考慮した文書分類”, 情報処理学会研究報告 自然言語処理研究会報告 2008(90), pp. 59–64 (2008).

[44] 鈴木誠, 平澤茂一, “単語と N-gram の各カテゴリにおける出現頻度の比の和を用いたテキスト自動分類手法”, 電気学会論文誌 C, Vol. 129, No. 1, pp. 118–124 (2009).

[45] 斉藤和巳, 上田修功, 金田有二, “確率モデルを用いた文書分類体系間の構造マッチング”, 情報処理学会研究報告 自然言語処理研究会報告 2004(47), pp. 33–38 (2004).

[46] 乾裕子, 内元清貴, 村田真樹, 井佐原均, “文末表現に着目した自由回答アンケートの分類”, 情報処理学会研究報告 自然言語処理研究会報告 98(99), pp. 181–188 (1998).

[47] 国定美佐代, 平松綾子, 能勢和夫, “自由記述アンケート分類のための限定的同意語特定手法”, 第 50 回自動制御連合講演会講演論文集, pp. 367–370 (2007).

[48] 福本文代, 鈴木良弥, “WordNet の同義語クラスとその上位関係を利用した文書の自動分類”, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1852–1865 (2002).

[49] 久保淳人, 鷲崎弘宜, 高須淳宏, “文書中のパターン間の文書類似度による関連分析”, 日本データベース学会 letters 3(3), pp. 13–16 (2004).

[50] 上嶋宏, 三浦孝夫, 塩谷勇, “同義語、多義語の考慮による文書分類の精度向上”, 電子情報通信学会論文誌 J87-D-I(2), pp. 137–144 (2004).

[51] 間瀬久雄, 辻洋, 絹川博之, 石原正博, “特許テーマ分類方式の提案とその評価実験”, 情報処理学会論文誌, Vol. 39, No. 7, pp. 2207–2216 (1998).

[52] 川谷 隆彦, “文書集合間の差異検出法と文書分類への応用”, 情報処理学会研究報告 自然言語処理研究会報告 2002(20), pp. 1–8 (2002).

[53] 柴田知秀, 姜ナウン, 黒橋禎夫, “同一文抽出に基づく類似ページの検出と分類”, 人工知能学会論文誌, Vol. 25, No. 1, pp. 224–232 (2010).

[54] 桂田浩一, 小山誠, 大原剛三, 馬場口登, 北橋忠宏, “文書分類システムの分類誤りに着目した分類ルール修正法”, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1880–1889 (2002).

[55] 石田栄美, “テキストの自動分類に関わる諸要素”, 日本図書館情報学会誌, Vol. 49, No. 2, pp. 65–78 (2003).

[56] 福重貴雄, 菅野祐司, “対応分析とベイジアンネットワークを用いた文書分類”, 情報処理学会研究報告 データベース・システム研究会報告 2003(51), pp. 167–174 (2003).

[57] 高村大也, 松本裕治, “SVM を用いた文書分類と構成的帰納学習法”, 情報処理学会論文誌. データベース 44, pp. 1–10 (2003).

[58] 鈴木大介, 内海彰, “Support Vector Machine を用いた文書の重要文節抽出”, 人工知能学会誌, Vol. 21, No. 4, pp. 330–339 (2006).

[59] 前田康成, 吉田秀樹, 鈴木正清, 松嶋敏泰, “学習データが少量しかない場合の文書分類に関する一考察”, 電気学会論文誌C, Vol. 131, No. 8, pp. 1459–1466 (2011).

[60] 坂本廣, 北村泰彦, 福島拓, 吉野孝, “N-gram に基づく多言語用例検索手法の評価”, 電子情報通信学会技術研究報告 AI, 人工知能と知識処理 110(428), pp. 51–56 (2011).

[61] 福島拓, 吉野孝, “Web データを用いた多言語用例対訳候補の抽出手法の検討”, 電子情報通信学会技術研究報告 NLC 言語理解とコミュニケーション 111(427), pp. 59–64 (2012).

[62] サダトファティア, 前田亮, 吉川正俊, 植村俊亮, “言語横断情報検索における辞書ベースと統計ベースのアプローチの統合”, 情報処理学会研究報告 データベース・システム研究会報告 2000(44), pp. 61–68 (2000).

[63] 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 41, pp. 12–21 (2000).

[64] 金沢輝一, 相澤彰子, 高須淳宏, 安達淳, “関連性の重ね合わせモデルを用いた日英言語横断検索”, 情報処理学会研究報告 情報学基礎研究会報告 2001(74), pp. 73–80 (2001).

[65] 根来啓輔, 大礒洋明, 秋吉政徳, 薦田憲久, “自由回答形式アンケートデータに基づいたアンケートカテゴリ辞書によるユーザ意見分類方式”, 電気学会情報システム研究会, IS-07-16, pp. 1–4 (2007).

[66] 中島基晶, 根来啓輔, 大礒洋明, 秋吉政徳, “アンケート意見の分類結果サンプルから抽出した分類基準による意見分類方式”, 電気学会情報システム研究会, IS-07-36, pp. 23–26 (2007).

[67] K. Negoro, K. Iida, and M. Akiyoshi, “Analysis Support System of Open-ended Questionnaires by Using Category Dictionary with Co-occurrence TF-IDF”, in *Proc. of the 1st Japan-China Joint Symposium in Information Systems*, pp. 71–74 (2009).

[68] L. Cai, Z. Wang, Y. Jiao, M. Akiyoshi, and N. Komoda, “Prototype of Knowledge Management System in Chinese Offshore Software Development Company”, *WSEAS Transactions on Information Science & Applications*, Vol. 5, Issue 3, pp. 252–257 (2008).

[69] L. Cai, Z. Wang, M. Akiyoshi, and N. Komoda, “BBS-based Information Management System in Chinese Offshore Software Development Company”, *Intl. Journal of Systems Applications, Engineering & Development*, Vol.5, Issue 1, pp. 50–57 (2011).

[70] L. Cai, Z. Wang, Yufeng Jiao, M. Akiyoshi, and N. Komoda, “A Knowledge Sharing and Managing System for Offshore Software Development Company”, in *Proc. of the 7th WSEAS Intl. Conf. on Applied Computer Science(ACS'07)*, pp. 340–345 (2007).

[71] L. Cai, Z. Wang, M. Akiyoshi, and N. Komoda, “Evaluation of BBS-based Information Management System in Chinese Offshore Software Development Company”, in *Proc. of Intl.. Conf. of the Institute for Environment, Engineering, Economics and Applied Mathematics 2010(IEEEAM 2010)*, pp. 580–582 (2010).

[72] 蔡立, 曲義暁, 王平, 秋吉政徳, “オフショアソフトウェア開発組織における知識共有システムの構想”, 電気学会情報システム研究会, IS-08-10, pp. 47–52 (2008).

[73] L. Cai, Z. Wang, M. Akiyoshi, and N. Komoda, “A Knowledge Cards Classification Method with Conversion Loss Correction for Incomplete

Translation Dictionary", *IEEJ Transactions on Electrical and Electronic Engineering*, Vol. 6, No. 6, pp. 566–570 (2011).

[74] L. Cai, M. Akiyoshi, and N. Komoda, "A Classification Method of Knowledge Cards Represented in Japanese and Chinese at Offshore Software Development Company", in *Proc. of the IADIS Intl. Conf. Applied Computing 2009*, pp. 63–67 (2009).

[75] L. Cai, Z. Wang, M. Akiyoshi, and N. Komoda, "A Knowledge Cards Classification Method with Conversion Loss Correction for Incomplete Translation Dictionary", in *Proc. of the 3rd Japan-China Joint Symposium on Information Systems (JCIS2010)*, pp. 85–88 (2010).

[76] 蔡立, 秋吉政徳, 薦田憲久, "オフショア開発組織内での中国語と日本語が混在する知識カードの分類管理システム", 電気学会情報システム研究会, IS-09-56, pp. 5–8 (2009).

[77] X. Liu, M. Akiyoshi, N. Komoda, L. Cai, and Z. Wang, "Evaluation of Knowledge Cards Classification Method with Translation Disctionary", in *Proc. of the 4th Japan-China Joint Symposium on Information Systems(JCIS2011)*, pp. 61–64 (2011).

[78] 蔡立, 劉暁鵬, 秋吉政徳, 薦田憲久, "ドメイン辞書を用いた中国語と日本語が混在する知識カードの分類方式", 電気学会情報システム研究会, IS-11-89, pp. 69–73 (2011).

[79] 総務省, "オフショアリングの進展とその影響に関する調査研究報告書", (2008).

[80] H-P. Zhang, H-K. Yu, D-Y. Xiong, and Q. Liu, "Hhmm-based Chinese Lexical Analyzer ICTCLAS", in *Proc. of the second SIGHAN workshop on Chinese language processing*, pp. 184-187 (2003).

[81] Y. Dapeng, S. Min, J. Peilin, R. Fuji, and S. Kuroiwa, "Rule-based Translation of Quantifiers for Chinese-Japanese Machine Translation", in *Proc. of the 10th WSEAS Intl. Conf. on COMPUTERS*, pp. 559–564 (2006).

[82] 坂本比呂志, 有村博紀, “Web マイニング”, 人工知能学会誌, Vol. 16, No. 2, pp. 233–238 (2001).

[83] 佐藤理史, 佐々木靖弘, “ウェブを利用した関連用語の自動収集”, 情報処理学会研究報告 自然言語処理研究会報告 153(8), pp. 57–64 (2003).