5Q-5

産学連携マッチング支援システムの研究*
-日英二ヶ国語から構成される専門用語の抽出-

木浪孝治†，池田哲夫†，高山毅†
岩手県立大学ソフトウェア情報学部†

1．はじめに

今日,大学は社会に貢献することが求められているようになっている．特に,産業界と関係の深い学部においては産学連携が強く求められるようになってきている.そのような産学連携を可能にするためには大学側のシーズを簡単に検索するシステムが必要になってきている.

そこで著者らは,産学連携の専門家(コーディネータ等)が産業側の要請に対応可能な研究者(集団)・研究のシーズを専門用語によって簡単に検索することができる DB システムの構築を目標とした研究を開始している.その第一歩としてソフトウェア分野の専門文献によく見られる日本語と英語両方の単語で構成されている専門用語の抽出を目的として研究を行ったので報告する.

2．先行研究

湯本,中川らが作成した日本語/英語/中国語からの専門用語抽出を行う言選 Web[言選]がある．このシステムは名詞と一部の特殊な形容詞を単名詞として扱い,それらを連結して複合名詞を作成している．その結果得られた複合名詞を専門用語としている[湯本01]．さらに,ある程度の日本語・英語が混在した場合でも専門用語抽出が可能である.その他にこれら言選 Web の機能を Perl モジュール化した TermExtract モジュール[TE]がある.

3．研究目的

本研究の目的は,日本語・英語が混在した文章から専門用語を抽出することである.

先行研究では日本語･短文の英語が混在している文章からの専門用語抽出は存在するものの,長文の英語が存在した場合はうまく抽出することができない[1].言選ではアルファベットを全て結合しているため,英文の後ろに日本語の名詞が続いた場合複合語として結合してしまう．以下に例文を示す.

例文)This system is …．情報処理学会論文誌.
→ 英文である「This system is….」はアルファベットであるため全て結合される．その後に続いた「情報処理学会論文誌」は名詞であるため英文と連結される．結果,例文全てが専門用語として抽出される.

4．提案方式

現在,単一の形態素解析器では日本語と英語が混在した文章の形態素解析を正しく行うことができない.例えば,日本語の形態素解析器では英語はアルファベットや未知語として扱われ,英語の形態素解析器では日本語を認識することができない．そのため,日本語文と英文を分割して処理する必要がある.

分割された英語部分の処理において,専門用語として抽出された部分の品詞を日本語の固有名詞に変更する工夫により,前後の日本語名詞と連結した複合名詞を抽出可能とする方式を考案した.

考案した専門用語抽出の手順を以下に示す．(図 1)

(1) 日英が混在した文章を,文章間の前後関係を維持したまま分割する．(図 1(ア))
(2) 日本語部分は日本語形態素解析器を用いて形態素解析を行う．(図 1(イ),(ウ))
(3) 英語部分は英語形態素解析器によって形態素解析を行い,湯本らの手法によって専門用語抽出を行う.抽出された専門用語の品詞を固有名詞に変更する．(図 1(エ))
(4) (2),(3)の出力結果を結合した後,湯本らの手法によって専門用語抽出を行う．(図 1(オ)(カ))

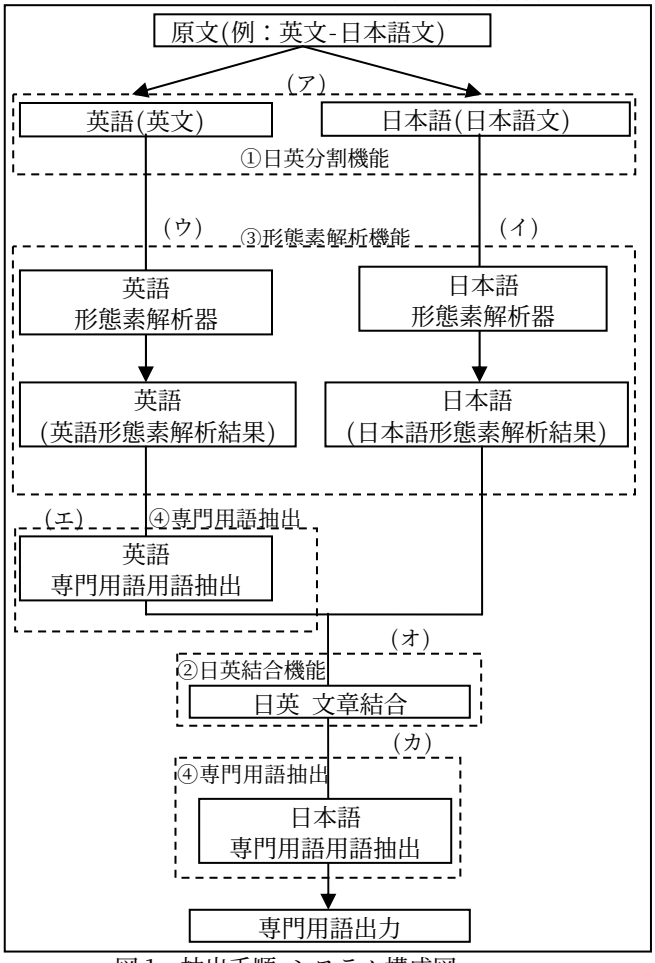

図 1　抽出手順,システム構成図

---

* Search System for University Researchers to Facilitate Industry-University Cooperation.

† K. Kinami, T. Ikeda, T. Takayama, (Faculty of Software and Information Science, Iwate Prefectural University)

[1]本研究開始時点では言選 Web では適切に抽出できなかった文章でも,本稿執筆時点で抽出可能になっているものがある.言選 Web は定常的に改良を続けているものと思われる.

提案方式による処理例を以下に示す.
例文)Disc Array システムを利用したシステムの構築
日本語文,英文への分割を行う.(図 1(ア))
英文：Disc Array
日本語文：システムを利用したシステムの構築

日本語文の形態素解析結果は以下の通り(図 1(イ))

原文	品詞-細分類
システム	名詞-一般
を	助詞-格助詞
利用	名詞-サ変接続
し	動詞-自立
た	助動詞
システム	名詞-一般
の	助詞-連体化
構築	名詞-サ変接続

英文の形態素解析結果は以下の通り(図 1(ウ))

原文	品詞
Disc	NNP(固有名詞,単数)
Array	NNP(固有名詞,単数)

形態素解析結果を用いて専門用語抽出を行う.(図 1(エ))
→"Disc Array"という専門用語が抽出される
抽出された専門用語の品詞を"名詞-固有名詞"に変更する.

日本語の形態素解析結果に英語の専門用語抽出結果を適切な場所へ挿入し専門用語抽出を行う.(図 1(オ))

原文	品詞-細分類
Disk Array	名詞-固有名詞
システム	名詞-一般
を	助詞-格助詞
…略…	…略…

専門用語抽出の結果,"Disk Array システム"という専門用語が抽出される.(図 1(カ))

5. 試作

4.で提案した方式の試作を行った.

5.1 試作環境

試作システムの環境を以下に示す.
OS ：FedoraCore2
日本語 形態素解析器：Chasen[Chasen]
英語 形態素解析器：BrillsTagger[BT]
日本語辞書 ：ipadic2.6.3-20
英語辞書 ：BrillsTagger 標準辞書

5.2 システム構成

本システムは大きく日英分割(①)/結合機能(②),形態素解析機能(③),専門用語抽出機能(④)から構成される.図 1 に構成を示す.

6. 評価

6.1 評価

現段階の試作システムで日英混在文章からの専門用語抽出が可能となった.約 100 の日英混在文章で適切に用語抽出可能なことを確認した.従来研究との比較という観点から以下の評価を今後行う予定である.
(1) NTCIR-1[NTCIR]を用いた精度評価
(2) 言選 Web との比較

6.2 今後の課題

湯本らの研究は主に情報処理分野の文章を対象にしている.筆者らは現在情報処理分野以外の分野にも対象を広げている.そのような分野では「名詞以外の品詞によって専門用語が構成されることがある」が判明している.例を以下に挙げる.

	対象単語	抽出結果	非抽出語	品詞
例 1	低分子物質	分子物質	低	接頭詞
例 2	粘弾性発現	弾性発現	粘	形容詞

例1) 低分子物質
→ 専門用語として「分子物質」が抽出され,接頭詞である「低」が専門用語の一部として抽出されない.
例2) 粘弾性発現
→ 専門用語として「弾性発現」が抽出され,形容詞である「粘」が専門用語の一部として抽出されない.

これらのケースも抽出可能にすることが今後の課題である.

7. まとめと今後の展望

本論文では日本語・英語が混在した文章からの専門用語抽出について述べた.今後は 6.で述べたように,試作システムの評価を行う予定である.また,評価が完了次第,今後の課題に取り組む予定である.

参考文献

[湯本 01] 湯本 紘彰, 森 辰則, 中川 裕志：出現頻度に基づく専門用語抽出, 情報処理学会情報学基礎自然言語処理研究会報告, No.065, pp.111-118, 2001.
[Chasen] 奈良先端科学技術大学院大学自然言語処理学講座, 日本語形態素解析器 Chasen, http://chasen.naist.jp/hiki/ChaSen/
[BT] Eric Brill：英語形態素解析器 Brill's Tagger, http://research.microsoft.com/%7Ebrill/
[言選] 中川裕志, 森辰則：言選 Web, http://gensen.dl.itc.u-tokyo.ac.jp/index.html
[TE] 中川 裕志, 前田 朗, 小島 浩之：専門用語自動抽出用 Perl モジュール TermExtract, http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html
[NTCIR] 情報検索システム評価用テストコレクション構築プロジェクト, http://research.nii.ac.jp/ntcir/index-ja.html