E-015

複合語翻訳による異言語で記述された書誌情報の同定
Cross-Lingual Identity Judgement of Bibliographic Information using Compound Word Translation

谷口 裕子 † 難波 英嗣 ‡ 相沢 輝昭 ‡
Yuko Taniguchi Hidetsugu Nanba Teruaki Aizawa

1. はじめに

引用文献索引データベースとは、論文とその論文が引用している文献との関係を記述したデータベースである。近年では CiteSeer § や PRESRI¶といったような、WWW 上で提供されている大量の論文データを収集して自動的に構築されたサービスも始まっている。このようなデータベースを作る際、書誌情報の自動同定という処理が必要である。これまでに複数の同一書誌情報を同定する研究が行われてきたが [3, 4, 5, 6]、これらの研究は同一言語で記述された書誌情報が対象で、異なる言語で記述されている場合に対応できない。そこで本研究では異言語で記述された書誌情報の同定を試みる。

以下は、英語論文中で引用された日本語論文の書誌情報と、その元の情報である。

- 奥村学, 難波英嗣: **テキスト自動要約に関する研究動向**, 自然言語処理 vol.6,No.6,pp.1-26(1999).
- M. Okumura and H. Nanba: **Automated Text Summarization: A Survey**,*Journal of Natural Language Processing*, 6(6):1-26,1999.(in Japanese)

この例から分かるように、英語論文中で日本語論文を引用する場合には題目、著者名等は英語で表記し、書誌情報の最後に (in Japanese) を付与する。このような書誌情報の対が同一であるかどうかを判定するには、題目や著者名等を翻訳する技術が必要不可欠である。しかし、題目には多くの専門用語が含まれるため、専門用語の翻訳技術が必要である。藤井ら [1] は、専門用語を語基と呼ばれる要素の組合わせによって漸進的に作られる複合語であると考え、EDR 日英専門用語対訳辞書 [7] から語基の対訳を抽出し、専門用語を自動的に翻訳する手法を提案している。

本研究では、まず藤井らの手法に基づいて題目中の専門用語を翻訳し、次に同一言語を対象にした書誌情報同定技術を用いて異なる言語で記述された書誌情報の同定を試みる。

本論文の構成は以下の通りである。次節で関連研究について述べる。3 節では本研究で提案する書誌情報同定の手法について説明する。提案手法の有効性を調べるため、実験を行った。4 節では実験方法と結果について述べる。

2. 関連研究

これまでに同一言語で記述された書誌情報を自動的に同定する数多くの手法が提案されている [3, 4, 5, 6]。これまでの研究では、ヒューリスティックスや機械学習により獲得された規則を用いて、題目、著者名、ページ数等のフィールドを特定、比較することで書誌情報を同定している。しかし 1 節で示した例のような場合、著作年、ページ、巻号等の数値的な情報はそのまま同定処理に利用できるが、題目等は「翻訳」技術を用いて 2 つの書誌情報の言語を統一しなければならない。そこで、本研究では、先行研究 [1, 2] に基づいて論文題目を複合語翻訳し、その結果を用いて異言語で記述された書誌情報の同定を試みる。

藤井らは (1) 語基辞書の作成と、(2) 訳語曖昧性の解消の 2 つの手順で複合語翻訳を実現している。これらを順に説明する。

(1) 語基辞書の作成

語基とは、例えば “associative memory(連想メモリ)” の場合、“associative(連想)” と、“memory(メモリ)” である。このような語基の訳語対を藤井らは EDR 日英専門用語対訳辞書 [7] を用いて作成している。ここで、日本語には語基の区切りがなく、英語語基との対応付けが困難である。そこで、藤井らは以下のルールを用いて日本語の語基を抽出している。

1. 字種の切れ目 (複数ある場合は最左の位置) で分割する。
2. 同一の字種だけで構成される場合は、中央 (奇数文字列の場合は中央文字の左側) で分割する。
3. 単語の先頭に現れない文字 (「ァ」「ン」「ー」など) の直前では分割しない。このような場合は、分割位置を右にずらす。

例えば、“連想メモリ” の場合、ルール 1 を適用することで、“連想” と “メモリ” の 2 つの語基に分割できる。

(2) 訳語曖昧性の解消

EDR 日英専門用語対訳辞書には、“associative learning(相関学習)” という用語対があり、ここからルール 2 により “associative(相関)” と “learning(学習)” が語基として抽出される。ここで、先に示した例と合わせると、“associative” には「連想」と「相関」という 2 つの訳語が存在するため、翻訳の際には訳語の曖昧性の解消が必要になる。藤井らは、頻度情報に基づく統計的な尺度を用いてこれを解消している。

本研究では、藤井らの手法に基づき予備実験を行った結果、(1) に関していくつか不具合があることが分かった。そこで、(1) を改良し、これを書誌情報の同定に用いる。次節では書誌情報同定の手順及び、藤井らの手法の改良方法について述べる。

†広島市立大学 情報科学研究科
‡広島市立大学 情報科学部
§http://citeseer.ist.psu.edu/
¶http://www.presri.com

表 1: 先行研究の問題点

分割前	分割後	英語表記
2進数	2　進数	binary number
1次不等式	1　次不等式	linear inequality
3倍精度	3　倍精度	triple precision
一意性の定理	一意性　の定理	uniqueness theorem
パターン系情報	パターン　系情報	pattern information
アナログ型リレー	アナログ　型リレー	analog relay

3. 書誌情報同定の手順

本研究では以下に示す手順で書誌情報の同定を行う。

1. **内容語の抽出**: 題目から専門用語を抽出する。
2. **複合語翻訳**: 抽出した専門用語を翻訳する。
3. **書誌情報の比較**: 翻訳結果を含む書誌情報を探す。

以下に手順 1 と 2 について説明する。

手順 1: 内容語の抽出

本研究では、まず論文データベースから (in Japanese) を含んだ英語の書誌情報を抽出し、次にこれらに対応する日本語の書誌情報を同定する。英語論文の題目を Apple Pie Parser[‖] を用いて構文解析し、名詞句を抽出する。これらの名詞句は次のステップで複合語翻訳される。

手順 2: 複合語翻訳

基本的には藤井らの手法に基づいて翻訳を行うが、先にも述べた通り語基辞書の作成について改良を行う。予備実験の結果、表 1 に示したように、英語語基と対応しないような分割が行われてしまうことが分かった。

このような問題を解決するために、次の 3 つのルールを新しく追加する。

- **ルール 1** 「・」がある場合、「・」を削除しそこで分割する。
- **ルール 2** 接尾辞 (単位) :
  以下の文字にマッチし、1 字前が数字の場合、その文字の後ろで分割する。
  「進, 分, 項, 次, 角, 階, 倍, 欄」
- **ルール 3** その他 :
  以下の文字にマッチし、1 字前の字種と異なる場合、その文字の後ろで分割する。
  「重, 値, 種, 局, の, 列, 系, 万, 型」

4. 実験

3 節で述べた提案手法の有効性を調べるため、実験を行った。

- 実験に用いるデータ
  藤井らと同様、EDR 日英専門用語対訳辞書中の 2 語基から構成される英単語とその日本語対訳 50,348 対を用いる。
- 評価尺度

$$\text{成功率 [\%]} = \frac{\text{成功した語数}}{\text{調べた語数}} \times 100$$

- 結果
  結果を表 2 に示す。

[‖] http://nlp.cs.nyu.edu/app/

表 2: 追加ルールの精度

	調べた語数 [語]	成功した語数 [語]	失敗した語数 [語]	成功率 [%]
1)	500	489	11	97.8
2)	128	103	25	80.5
3)	200	174	26	87.0

- 考察

分割の失敗例をいくつか挙げる。まずルール 1 では「アップ・ダウン・フリップ・フロップ (updown flip-flop)」や「アド・オン・メモリー (add-on memory)」のように、「・」が 2 回以上あるときに失敗していた。また、「アナログ・デジタル変換器 (analog-to-digital converter)」や「異種データ・ベース (heterogeneous database)」のように、字種が 2 種類以上で「・」が使われているときにも失敗していた。ただしこれらの場合には、「アップダウン・フリップフロップ」や「異種データベース」など複数の表記があり、そのどれかで語基の対応がとれる分割ができていた。

次にルール 2 では、「16 進 数定数 (hexadecimal constant)」、「8 角 形距離 (octagonal distance)」や「2 次元信号 (two-dimensional signal)」のように、1 文字でない接尾辞（この場合、〜進数、〜角形、〜次元）にマッチしてしまったとき失敗していた。

5. おわりに

本論文は、複合語翻訳を用いた日英間の書誌情報の同定手法を提案した。本研究では先行研究 [2, 1] で提案されている複合語翻訳法を用いて題目を翻訳し、異言語で記述された書誌情報を同定する手法を提案した。本論文では特に複合語翻訳手法の改良を試み、その有効性ついて調査した。その結果、80%以上の精度で先行研究の手法を改良できることが分かった。今後は、論文題目の翻訳結果を用いて実際に書誌情報の同定を試みる。

参考文献

[1] 藤井敦, 石川徹也, “技術文書を対象とした言語横断情報検索のための複合語翻訳,” 情報処理学会論文誌, Vol.41, No.4, pp.1038-1045, 2000.

[2] Fujii, A. and Ishikawa, T., “Japanese/English Cross-language Information Retrieval: Exploration of Query Translation and Tranliteration,” *Computers and the Humanities*, Vol.35, pp.389-420, 2001.

[3] 伊藤敬彦, 堀部史朗, 新保仁, 松本裕治, “複数尺度を用いた参考文献の同定, 情報処理学会研究会報告, 2003-DBS-130, pp 181-188, 2003.

[4] Lawrence, S., Giles, C. L. and Bollacker, K., “Digital Libraries and Autonomous Citation Indexing,” *IEEE Computer*, Vol.32, No.6, pp.67-71, 1999.

[5] McCallum, A., Nigam, K., Rennie, J. and Seymore, K., “Building Domain-specific Search Engines with Machine Learning Techniques,” *AAAI Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[6] Nanba, H., Abekawa, T., Okumura, M., and Saito, S., “Bilingual PRESRI: Integration of Multiple Research Paper Databases,” *Proceedings of RIAO 2004*, pp.195-211, 2004.

[7] 日本電子化辞書研究所 : 専門用語辞書 (情報処理), 1995.