複合語翻訳による異言語で記述された書誌情報の同定

谷口 裕子[†] **難波 英嗣**[‡] **相沢 輝昭**[‡]

[†]広島市立大学 情報科学研究科

[‡]広島市立大学 情報科学部

1 はじめに

近年，CiteSeer[1]をはじめ，Cora[4]，Google Scholar[2]，PRESRI[3][7] 等，WWW 上にある大量の論文データを自動的に収集してデータベースを構築するサービスが提供されている．これらは，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索引のデータベースであり，引用論文データベースと呼ばれている．このようなデータベースを作る際，ある論文が引用している論文 A と別の論文が引用している論文 B が同一の書誌であるかどうかを判定する「書誌情報の同定」という処理が必要になる．これまでにも，書誌情報を同定する研究は行われてきた [1, 3]．しかし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．日本語の論文を英語論文中で引用する場合，一般に英語論文は英語圏の研究者を対象に書くため，必要最低限の日本語論文しか引用されない．そのため総引用数の中で日本語論文が占める割合は高くないが，それらは概して重要な引用であることが多い．そこで本研究では，異なる言語で記述された書誌情報の同定を試みる．

異言語で記述された同一書誌情報の例

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 Survey**, *Journal of Natural Language Processing*, 6(6):1-26, 1999.

CiteSeer や PRESRI は，ヒューリスティックスや機械学習を用いて，題目，著者名，ページ等のフィールドを特定・比較し同一言語で記述された書誌情報を同定している．上例の場合，ページや巻号等の数値情報はそのまま同定処理に利用できるが，題目や出典の同定には翻訳技術を用いる必要がある．しかし，一般的な翻訳器では論文の題目を正確に翻訳出来るとは限らない．なぜならば，一般的な翻訳器は文中の主語や動詞を前提にしているが，題目中には一般に動詞が含まれないからである．また本研究では技術論文を対象にしているので，専門用語の翻訳が同定の性能を大きく左右する．そこで，本研究では題目全体を翻訳するかわりに次節で述べる手順で部分的に題目の翻訳を行い，書誌情報の同定を行う．

[1]http://citeseer.ist.psu.edu/
[2]http://scholar.google.com/
[3]http://www.presri.com

2 書誌情報の同定

本研究では以下の手順で書誌情報の同定を行う．

1. セグメンテーション

2. 翻訳

- **2-1. 表題解析:** 手掛かり語を基に題目を解析.
- **2-2. 専門用語の抽出:** 解析結果を基に題目から専門用語を抽出.
- **2-3. 複合語翻訳:** 抽出した専門用語を翻訳.
- **2-4. 著者名翻訳:** 著者名をローマ字表記に翻訳.

3. 書誌情報の比較

4. 同定

以下では，このうち表題解析，専門用語の抽出，複合語翻訳，及び著者名翻訳について述べる．

2.1 表題解析

題目を効率良く翻訳するためには，不要な語句を除く必要がある．また題目に使用されている語句の中でも，専門用語はその論文の特徴を表現する重要な語句である．本研究では，関連研究 [5, 8] に基づき，手掛かり語を用いて表題解析を行う．その結果から題目中の専門用語を抽出し，その意味的な役割を明らかにする．以下に解析例を示す．

解析前：stuck-at-faults testing in asynchronous logic circuits based on module partitioning

解析後：stuck-at-faults testing
<RESTRICT cue="in">asynchronous logic circuits<RESTRICT>
<METHOD cue="based on">module partitioning<METHOD>

上例では，“stuck-at-faults testing”，“asynchronous logic circuits”，“module partitioning” が専門用語として抽出できる．また，“module partitioning” の直前に “based on” という手掛かり語があるた

め，“module partitioning” という用語は “stuck-at-faults testing” の要素技術になっていることが分かり，“METHOD” のタグが付与される．

このような題目の構造解析を行うため，英語題目用に 31 個，日本語題目用に 165 個の手掛かり語を使用した．また名詞句間の関係を表すタグは英語題目用に 11 種類，日本語題目用に 10 種類設定した．表 1 に手掛かり語とタグの一部を示す.

表 1: 表題解析用手掛かり語とタグの例

タグ	手掛かり語 (英，日)	
METHOD	by	による
	based on	に基づく
	using	を用いた
RESTRICT	on	に関する
	of	の
	in	における
GOAL	for	のための
	towards	に向けて
CONJ	and	と，や
	or	及び

2.2 専門用語の抽出

表題解析の結果だけでは十分に専門用語だけに絞り込めない場合がある．例えば，“study of the dialogue model for intelligent support system of group learning” という題目からは，表題解析の結果 “study”，“dialogue model”，“intelligent support system”，“group learning” が翻訳用専門用語候補として得られる．しかし “study” という単語は使用頻度が非常に高く，この訳語を用いて題目対を検索した場合，不要な題目候補を数多く拾ってきてしまう可能性がある．そこで本研究では，表題解析によって抽出した専門用語候補を，中川ら [6] が構築した専門用語抽出器 Termex [4] を用いて作成した専門用語データと照合することで，候補の中からより専門性の高い語句のみ抽出する．上述の例では “study” を翻訳用名詞句候補から除外し，残りの “dialogue model”，“intelligent support system”，“group learning” を翻訳器に渡す.

2.3 複合語翻訳

多くの専門用語は，既存の語基の組合わせによって漸進的に作られた複合語であるが，それらの対訳を網羅的に辞書に記述するのは困難である．例えば，情報処理関連の専門用語 12 万語を収録した EDR 日英専門用語対訳辞書には，「知識抽出 (knowledge extraction)」や「特徴抽出アルゴリズム (feature extraction algorithm)」などの複合語は定義されていても，「抽出 (extraction)」という語基は定義されていない．従って，「情報抽出 (information extraction)」のような新たな語基の組み合わせを翻訳することができない．専門用語の翻訳には様々な手法があるが [2, 10]，本研究では藤井ら [2] の手法に基づき，以下の 2 つの手順で翻訳を行う.

(1) 語基辞書の作成: 既存の対訳辞書から，2 語基の訳語対を抽出し，ヒューリスティックスにより日本語を分割して英語語基と対応づける.

(2) 訳語曖昧性の解消: 複数の訳語候補がある場合，共起情報をもとに最適な訳語を決定する.

なお (1) について予備実験を行った結果，先行研究 [2] で提示されていたルールだけではいくつか不具合が生じたので，1)「・」は削除しそこで分割，2)「進, 分, 項」等にマッチし 1 字前が数字の場合，その文字の後ろで分割，3)「重, 値, の」等にマッチし 1 字前の字種と異なる場合，その文字の後ろで分割，という 3 ルールを新たに追加した．それぞれのルールが適応される語句について，ルールを適応した結果が正確に分割されているかを人手で調べ評価を行なった [9] 結果を表 2 に示す.

表 2: 複合語翻訳の追加ルールの精度

	精度 [%]
追加ルール 1)	97.8 (489/500)
追加ルール 2)	80.5 (103/128)
追加ルール 3)	87.0 (174/200)

2.4 著者名翻訳

本研究では，日本語書誌情報中の著者名をローマ字表記に翻訳し，英語書誌情報中の著者名との比較に使用した．著者名翻訳用の辞書には，カタカナ表記と漢字表記の対応をとるために「茶筌[5]」の人名辞書を使用した．この辞書には約 33,000 件の人名が収録されている.

漢字表記をカタカナ表記に変換すると，次にカタカナを一字ずつそれに対応するローマ字の表記に置き換える．例えば「今中」という名前は，人名辞書より「イマナカ」というカタカナ表記に変換できる．この 4 つのカタカナに対応するローマ字はそれぞれ “i”，“ma”，“na”，“ka” であるので，これらを並

[4] http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/resource/termext/atr.html

[5] http://chasen.naist.jp/hiki/ChaSen/

べて “imanaka” というローマ字表記に変換する．また,「チ」(“chi”，“ti”) や「オオノ」(“ohno”，“ono”，“oono”) のように表記に揺れがある場合も考慮した．

3 実験

NTCIR[6]の言語横断検索タスクに用いられたデータセット 33 万件の中から日英対訳となっている書誌情報をランダムに 750 件抽出し，これに対して以下の 4 つの手法で，それぞれ書誌情報の同定を行った．

手法 1) 専門用語の複合語翻訳結果

手法 2) 手法 1 + 著者名翻訳の結果

手法 3) 手法 2 + 著作年

手法 4) 手法 3 + 専門用語のタグ情報

また，これらの手法と比較するためのベースライン手法として，著者名の翻訳と著作年のみ用いた手法で書誌情報の同定を行う．

各手法は，以下に示す精度と再現率で評価する．

$$\text{精度} = \frac{\text{システムが検出した正解題目数}}{\text{システムが検出した題目候補数}} \times 100$$

$$\text{再現率} = \frac{\text{システムが検出した正解題目数}}{\text{総正解題目数}} \times 100$$

なお翻訳結果を用いた比較は，得られた訳語の全組合わせの AND 検索と OR 検索の 2 通りで行う．結果を表 3 に示す．

表 3: ベースライン手法と提案手法の結果

		精度 [%]	再現率 [%]
ベースライン手法		1.7	98.3
AND	手法 1	2.6	33.6
	手法 2	15.2	33.6
	手法 3	91.5	33.6
	手法 4	94.5	29.2
OR	手法 1	1.4	55.2
	手法 2	12.7	55.2
	手法 3	87.3	55.2
	手法 4	89.0	49.6

この結果から，比較に使用するフィールドを増やすことで，再現率を下げることなく精度が向上していることが分かる．しかし AND，OR 検索ともに手法 4 の再現率が低下していた．

4 考察

同定に失敗した原因を項目別に説明する．

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

専門用語抽出の失敗

今回作成した表題解析器は，いくつかの手掛かり語を用いてタグを付与した．しかし，題目 “digital type multi function protective relaying system” のように，題目中に手掛かり語が出現しない題目があった．この場合題目全体が 1 つの名詞句となるが，本研究で作成した複合語翻訳器の処理能力の関係で，5 語以上から成る名詞句は翻訳しない．従って翻訳される専門用語が抽出できず，同定に失敗していた．

訳語の不足

以下のような題目対があった．

- large scale problem and building block method of neural network
- ニューラルネットワークの大規模問題とビルディングブロック法

この英語題目から抽出された “building block method” の翻訳は「建築構法」となっていたが，日本語題目では「ビルディングブロック法」が使用されている．

また，次の例は日本語題目では略語が使用されているが，英語題目ではフルスペルが使用されている例である．対応する箇所を下線で示す．

- optimum mfd of thermally-diffused expanded core fiber
- TEC ファイバにおける MFD の最適拡大率の検討

今回使用した複合語翻訳の辞書では略語を考慮していないので，このような場合は同定に失敗する．

名詞句の不一致

以下の題目対を例に説明する．

- study of the dialogue model for intelligent support system of group learning
- 知的グループ学習支援システムのための対話モデルの研究

この英語題目からは，専門用語抽出の結果として “dialogue model”，“intelligent support system”，“group learning”，その翻訳結果としてそれぞれ「対話モデル 対話型」,「知的支援システム 知的支援機構」,「グループ学習 集団学習」等が得られた．しかし，日本語題目中では “intelligent support system” と “group learning” に相当する部分が「知的グループ学習支援システム」のように 1 つの専門用語として出現している．

このように，日本語題目では 1 つの専門用語として出現するが，英語題目の方では分離している場合や，日本語題目中にある語が英語題目に存在しない

場合，逆に英語題目中にある語が日本語題目の方には存在しない場合は，正しく翻訳できたとしてもその結果と実際に題目で使用されている専門用語との対応が取れず，題目対の同定に失敗していた．

タグの不一致

ここでは，題目，著者名，著作年のフィールドを使用した比較では同定に成功したが，タグまで比較すると失敗した例を示す．

- priority control for a photonic self-routing circuit using vstep
- VSTEP を用いた光セルフルーチング回路における優先制御法の検討

英語題目からは “priority control(タグ無し)”，“photonic self-routing circuit(GOAL)” が専門用語として抽出でき，その翻訳結果として「優先制御 プライオリティ制御 (タグ無し)」,「回路 回線 (GOAL)」が得られた．しかし日本語題目を解析すると,「優先制御」,「回路」ともに RESTRICT タグが付与され，英語題目の解析結果から得たものと異なる．

“priority control” と「優先制御」のタグが異なっている理由は，日本語題目中の「の検討」にあたる部分が英語題目中に存在しないため，日英題目間のタグの付与位置にずれが生じたからである．

また，“photonic self-routing circuit” と「回路」のタグが異なっている理由について説明すると，一般的に ‘for’ の訳としては「〜のための」といった目的を示すようなものが多く使用されるが，この例では「〜における」と限定を示すものが使用されていたため，別々のタグが付与されたのである．

5 おわりに

本研究では，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．初めに先行研究の問題点を指摘し，改善を試みた．提案手法の精度は 8 割以上であった．更に書誌情報同定システムを構成し実験を行った結果，29.2%の再現率，94.5%の精度を得ることができ，複合語翻訳法を書誌情報の同定に用いることは有効な手段であることが確認できた．

6 今後の課題

書誌情報同定が失敗する原因の 1 つに語基辞書の不備がある．実験の結果，翻訳したい英単語に対して語基辞書中の訳語が不足していたり，英単語自体が辞書に載っていない場合があった．この問題の解決法として語基辞書作成における HMM を用いた日本語の複合語分割法の洗練が挙げられる．また，2 言語コーパスから対訳を抽出する手法による語基辞書を拡張や，略語とその展開形の対応が取れる略語辞書の導入も，本研究に有効であると考える．

また実験の結果で，比較に使用するフィールドの数を増やすことで書誌情報同定の精度が向上したことを示したが，今回作製した表題解析器によって題目中の専門用語に付与されるタグを使用した場合，再現率の低下が見られた．この原因の 1 つは，表題解析に使用した手掛かり語の前後の専門用語間の関係だけを考慮していたことが考えられる．対処法として，構文解析等による題目全体の構造，係り受け情報を利用し，専門用語間の関係をより厳密に調べてタグを付与することが挙げられる．

謝辞

NTCIR コレクションは国立情報学研究所 (NII) の許可を得て使用させて頂きました．本研究は，NEDO 平成 16 年度産業技術研究助成事業の支援を受けて行われました．

参考文献

[1] 相澤彰子, 大山敬三, 高須淳宏, 安達淳 : **レコード同定問題に関する研究の課題と現状**, 電子情報通信学会論文誌, Vol.J88-D-I, No.2, pp.576-589(2005).

[2] 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).

[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, DBS-130, pp.181-188(2003).

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

[5] 長尾真, 辻井潤一, 矢田光治, 柿元俊博 : **科学技術論文表題の英和機械翻訳システム**, 情報処理学会論文誌, Vol.23, No.2, pp.202-210(1982).

[6] 中川裕志, 湯本紘彰, 森辰則 : **出現頻度と連鎖頻度に基づく専門用語抽出**, 自然言語処理, Vol.10, No.1, pp.27-45(2001).

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

[8] 佐藤理史 : **論文表題を言い換える**, 情報処理学会論文誌, Vol.40, No.7, pp.2937-2945(1999).

[9] 谷口裕子, 難波英嗣, 相澤輝昭 : **複合語翻訳による異言語で記述された書誌情報の同定**, 情報科学技術フォーラム (FIT 2004).

[10] M. Tonoike, M. Kida, T. Takagi, Y. Sasaki, T. Utsuro, and S. Sato : **Translation Estimation for Technical Terms using Corpus collected from the Web**, *Proceedings of the Pacific Association for Computational Linguistics(PACLING05)*, pp.325-331(2005).