様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月　５日現在

機関番号：13901
研究種目：挑戦的萌芽研究
研究期間：2010～2011
課題番号：22650047
研究課題名（和文）
選択型翻訳による言語横断検索の実現
研究課題名（英文）
Cross-Language Headword-Search by using Non-Productive Machine Translation
研究代表者
佐藤　理史（SATO SATOSHI）
名古屋大学・工学研究科・教授
研究者番号：30205918

研究成果の概要（和文）：本研究では、選択型翻訳とよぶ新しいターム翻訳の方式を提案するとともに、それを実行する効率的なアルゴリズムを開発した。さらに、この方法の日英ターム翻訳への適用において、７種類の拡張機構を実現し、翻訳の性能向上を図った。実現した方式を英語ウィキペディアに対する日英横断見出し語検索に適用し、実験により、その有効性と既存の機械翻訳を上回る性能を確認した。

研究成果の概要（英文）：We have proposed a new framework of term translation—non-productive machine translation—and an efficient algorithm of the execution. We have also developed seven extension techniques for this framework to improve translation performance in the application to Japanese-English term translation. We have implemented a Japanese-English headword search system for the English Wikipedia, which can retrieve English articles from Japanese terms. In the experiments, we have confirmed the effectiveness and high performance of the system.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,600,000	0	1,600,000
2011 年度	1,300,000	390,000	1,690,000
総　計	2,900,000	390,000	3,290,000

研究分野：　情報学
科研費の分科・細目：　情報学・図書館情報学・人文社会情報学
キーワード：　情報検索、言語横断検索、ターム翻訳、日英翻訳

１．研究開始当初の背景

現代社会において、英語は、事実上の国際語として定着している。そのため、英語で書かれた情報の量は、他の言語で書かれた情報を圧倒的に凌駕している。このような状況において、英語非母国語話者は、情報収集という点において、英語母国語話者に対して大きなハンディキャップを負っており、それを軽減する機能の実現が強く求められている。

このような機能を実現する有力な方法の一つに、情報の記述に用いられている言語以外の言語で、その情報へのアクセスを可能とする言語横断検索がある。この方法は、情報を見つけ出すという作業を支援するものである。

たとえば、「反結合性オービタル」という用語について調べたい場合を考えよう。日本語で書かれた百科事典に該当する項目が存在しなかった場合でも、英語で書かれた百科事典には、これを説明する項目が存在するかもしれない。このとき、「反結合性オービタル」という日本語で英語百科事典を引くことができれば、その訳語を知らない場合でも、英語百科事典を活用することが可能となる。これが言語横断検索の一例である。

言語横断検索の標準的な方法は、検索クエ

リ（検索キーワード）を翻訳するという方法である。たとえば、日本語で英語文書を検索する場合、システムは、まず入力された日本語クエリを英語に翻訳し、次に、その翻訳結果で英語文書を検索する。当然のことながら、その性能は翻訳精度に強く依存することになるが、単に「日本語クエリを英語に翻訳する」という一般的な問題設定では、それほど高い翻訳精度は望めない。

ここで、言語横断検索を、先ほどの例のように「見出し語を探す場合」に限定することを考える。この場合、まず、翻訳対象はターム（語、複合語）に限定できる。次に、ここでの翻訳の目的は、百科事典を引くことであるため、百科事典が引けないターム（つまり、見出し語以外のターム）に翻訳できたとしても、検索には役に立たない。すなわち、「日本語タームの翻訳先を、英語の見出し語となっているタームに限定することができる」ことになる。

このようなアイディアは、ターム翻訳の性能を向上させることができると考えられるが、これまで提案されたこともなければ、確かめられたこともなかった。機械翻訳の研究において、ターム翻訳の問題は軽視されおり、タームの対訳を対訳辞書に登録する以外、有効な方法論は提案されてこなかった。

２．研究の目的

上記のアイディアは、次のように要約される。「タームの翻訳問題を、ターゲットリスト（正解訳が含まれるリスト＝言語横断検索の対象となる情報源の索引語リスト）からの選択問題として解く」。これを選択型翻訳と呼ぶ。

本研究の目的は、この選択型翻訳のアイディアを実現するとともに、これを英語ウィキペディアへの日英横断見出し語検索に適用することを通して、その有効性を検証することにある。

具体的には、以下に示す目標を設定し、初年度に目標１から３を、次年度に目標４を達成することを目指す。

目標１：カタカナ語の選択型翻訳（翻字）アルゴリズムを確立する。

目標２：複合語の選択型翻訳アルゴリズムを確立する。

目標３：上記２つのアルゴリズムを組み合せ、日本語タームの選択型翻訳法を確立する。

目標４：選択型翻訳法を英語ウィキペディアに適用し、日英横断見出し語検索を実現する。

本研究が狙う直接的な効用は、英語非母国語話者への情報探索支援（情報アクセスのチャンスの拡大）である。本研究で実現される方式は、英語索引リストを持つあらゆる情報源（辞書、辞典、データベース等）に適用可能であり、これらの情報源に対する日本語タームによる検索が可能となる。特に、画像や映像データベースなどの、到達した情報の解釈に英語力が不要のものに対しては実用性・有用性が高く、多くの人々にとってニーズがある。

３．研究の方法

(1) 選択型翻訳は、(a)ターゲットリスト（正解訳が含まれるリスト＝言語横断検索の対象となる情報源の索引語リスト）、(b)対訳辞書（部分翻訳規則集合）、(c)選択型翻訳アルゴリズム、の３つのコンポーネントから構成される。このうち、(a)は、検索対象とする情報源から得られることを仮定する。すなわち、実際に作成する必要があるのは(b)と(c)の２つである。この枠組自体は、任意の言語対の翻訳に適用可能であるが、本研究では、日英翻訳に焦点を絞る。

(2) 単純化した選択型翻訳の枠組は、次のようになる。いま、翻訳したい日本語ターム(n個の部分要素から構成されるものとする)を$J=j_1 j_2 \ldots j_n$とする。ここで、対訳辞書Dにおいて、それぞれのj_iに対して、英訳e_i（一般に複数個）が与えられているとする。このとき、この対訳辞書によって得られるJの英訳候補Eは、$E=e_1 e_2 \ldots e_n$と書ける。正解訳はターゲットリストTに含まれているので、そのようなEをTから見つけ出せばよい。

実際には、(i)日本語タームJをいくつの部分要素に分解すれば正しい訳が得られるかは、あらかじめ与えられない(nは未知)ので、可能な分解をすべて試す必要がある。また、(ii) 対訳辞書Dとして完全なものが用意できるとは限らないので、Tに含まれる英訳候補Eを構成できるとは限らない。このような場合には、対訳辞書Dを拡張したり、Eに最もよく似たTの要素を探したりするように、枠組みを拡張する必要がある。これらの拡張を検討し、選択型翻訳の枠組を確定させる。

(3)日本語タームを文字に分解して上記の枠組みを適用すると、カタカナ語の翻訳（逆翻字）が実現できる。この場合、対訳辞書は文字単位の対応規則となる。日本語タームを単語に分解して上記の枠組みを適用すると、複合語の翻訳が実現できる。この場合、対訳辞書は語あるいは複合語の対訳となる。これら２つを組み合わせ、タームの選択型翻訳方式を実現する。

(4)単語・複合語の対訳辞書は、既存の英和辞書を利用して作成する。この対訳辞書には、複合語の構成要素間の対訳も含める。ターゲットリストは、英語ウィキペディアの見出し語リストから作成する。

(5) 以上を組み合わせて、英語ウィキペディアを日本語で引くシステムを実現する。このシステムを評価するためのテストセット（正解対訳集合）を作成し、それを用いて、他の翻訳システム等と性能を比較する。

４．研究成果

(1) 選択型翻訳(Non-Productive Machine Translation, 非生産型機械翻訳とも呼ぶ)の枠組みを明らかにした。

まず、翻訳文法 G を次のように定義する。

$$G=(A,B,R)$$

ここで、A は原言語の単語の集合、B は相手言語の単語の集合、Rは翻訳規則集合(対訳辞書)を意味する。翻訳規則 r は、次のような形式とする。

$$r=(a,b) \text{ where } a\in A^*,\ b\in B^*$$

このような翻訳文法において、翻訳規則の列 d は、複数の部分要素から構成される対訳を表す。すなわち、…bn)

$$\begin{aligned} d&=r_1r_2\cdots r_n \\ &=(a_1,\ b_1)(a_2,\ b_2)\cdots(a_n,\ b_n) \\ &=(a_1a_2\cdots a_n,\ b_1b_2\cdots b_n) \end{aligned}$$

翻訳規則の列 d の原言語側を src(d)、相手言語側を tgt(d)と表現すると、翻訳文法 G は、次のような対訳の集合 $L(G)$を規定することとなる。

$$L(G)=\{(\mathrm{src}(d),\ \mathrm{tgt}(d))\mid d\in R^*\}$$

このような翻訳文法 G を用いて、選択型翻訳の枠組みは、次のように定義できる。

与えられるもの：

翻訳文法 $G=(A,B,R)$

ターゲットリスト $T\subset B^*$

入力 $s\in A^*$

求めるもの：

$$F=\{\mathrm{d}\mid \mathrm{d}\in R^*,\ \mathrm{src}(d)=s,\ \mathrm{tgt}(d)\in T\}$$

出力：

$$W=\{\mathrm{tgt}(d)\mid d\in F\}$$

この枠組みにおいて、翻訳出力 tgt(d)は必ずターゲットリスト Tの要素であり、新たな訳を作り出すことはない。言い換えるならば、既知の訳の中から適切なものを選ぶということである。

(2) 選択型翻訳を実行する効率的なアルゴリズムを明らかにした。

出力 Wを求めるためには、次のようにすればよい。まず、与えられた入力 sをいくつかの部分要素に分割する。次に、そのそれぞれを翻訳規則集合 Rに含まれる規則を用いて翻訳する。その後、得られた部分訳を組み合わせて全体訳を構成する。ここまでで、部分要素への分割に複数の可能性が存在したり、一つの部分要素に複数の訳が存在したりするので、一般に複数の全体訳が得られる。最後に、得られた全体訳の中から、ターゲットリスト Tに含まれるものを選び、それらを出力する。

上記の方法は、動的計画法とプレフィックス・フィルタリングを用いることで、効率的に実行できる。まず、入力 sの前から１単位(単語または文字)毎に、そこまでの部分列の翻訳（部分訳）を求め、その結果を記憶し再利用する（動的計画法)。その際、その部分訳は、かならず、ターゲットリスト Tのいずれかの要素のプレフィックス（前方文字列）となっていなければならない（そうでないと全体訳が T の要素となることはない）ので、その条件に合格するもののみを部分訳として採用する。これがプレフィックス・フィルタリングである。

このアルゴリズムは、実用的な設定（翻訳規則 160 万件、ターゲットリスト 550 万件）において、標準的なパーソナルコンピュータ(iMac: Core 2 Duo 2.16GHz)で、1 翻訳あたり、平均数十 ms で動作する。

(3) 単語を単位とした選択型翻訳方式に、文字を単位とした選択型翻字方式を拡張辞書として統合した。

本研究が対象とする日英ターム翻訳では、カタカナ語の翻訳が頻繁に必要となる。カタカナ語のほとんどは、英語から翻字（音訳）によって輸入された語であり、カタカナ語の日英翻訳は、逆翻字問題となる。

本研究の申請時から交付決定の間に行った研究において開発した選択型翻字方式を対訳辞書の一部として組み込むことにより、もともとの対訳辞書がカバーしないカタカナ語の翻訳が可能となった。

(4) 選択型翻訳の拡張機構を開発した。

選択型翻訳が正しい翻訳結果を出力するためには、(a)正しい翻訳結果がターゲットリストに含まれる、(b)入力と正しい翻訳結果の対が、対訳辞書から合成できる、という２つの条件を満たす必要がある。言語横断見出し語検索では、全見出し語をターゲットリストとするため、(a)の条件は問題とならない。一方、実際に入手可能な対訳辞書は、翻訳という現象の一部しかカバーしないため、しばしば、(b)の条件が満たされず、その結果、何も出力されないことが起こる。その主たる原因は、表記のゆれ、翻訳のゆれ、複合語の翻訳における構成要素の品詞変化である。

日本語は表記ゆれが多い言語と言われている。たとえば、数字の表記ゆれ（「二名式命名法」と「２名式命名法」)、カタカナの表記ゆれ（「モホロビチッチ」と「モホロヴィチッチ」)、漢字の表記ゆれ（「浸蝕」と「浸食」)、漢字とカタカナのゆれ（「蛋白」と「タンパク」）などがある。一般に、辞書では一

つの表記が採用され、他の表記は記述されていないことが多い。そのため、文字列の一致に基づく機械的な辞書引きでは、対訳辞書の検索に失敗する。

翻訳のゆれは、「proprioceptor」に対して、「自己受容器」または「自己受容体」のように複数の訳語が存在する以外に、「orbital」が「軌道」と訳したり、「オービタル」と翻字（音訳）されたりするゆれが存在する。これらも、辞書には一方しか記述されていないことがほとんどである。

複合語（タ－ム）の翻訳では、上記の２種類のゆれが組合わさった形で現れるとともに、「multifactorial inheritance」の「multifactorial」を形容詞ではなく名詞的に訳し、「多因子遺伝」という訳語をあてる現象もしばしば観察される。このような品詞をまたぐ訳語対も、辞書には記述されていない。

これらの問題を解決するために、上記の(3)の拡張を含む、７つの拡張機構を考案した。

① 無翻訳：日本語側の文字列がアルファベット等の場合は、それ自身も訳語として採用する。
② 逆翻字の導入：上記の(3)。日本語側の文字列がカタカナ列で、かつ、長さが５以上の場合に、その逆翻字結果も、訳語として採用。
③ 表記ゆれに対応した辞書引き：表記ゆれ変換規則を導入し、その適用結果を用いて辞書引きを行う。
④ 読みで辞書を引く：③でも辞書が引けなかった場合は、読みで辞書を引く。
⑤ 辞書に対して類似文字列検索を実行する。
⑥ 接尾辞・付属語の削除・追加：辞書引きの際、接尾辞や付属語を削除・追加して辞書を引く。
⑦ 「X の Y」のパターンの際に、語順変化「Y of X」を許すように拡張。

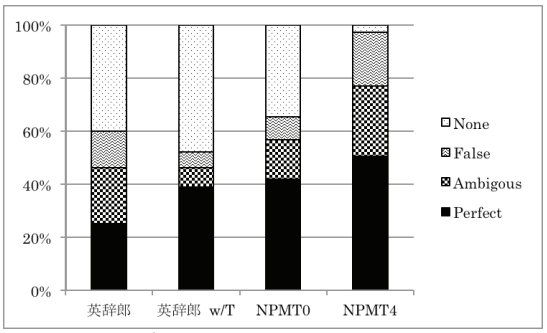
(5) 性能評価実験の用いるための３組のテストセットを作成した。テストセットに含まれる対訳は、すべて、(a)日本語側は日本語ウィキペディアの見出し語集合に含まれていない、(b) 英語側は英語ウィキペディアの見出し語集合に含まれている、(c) 日本語側はウェブ上に 10 回以上出現する、という３つの条件を満たす。

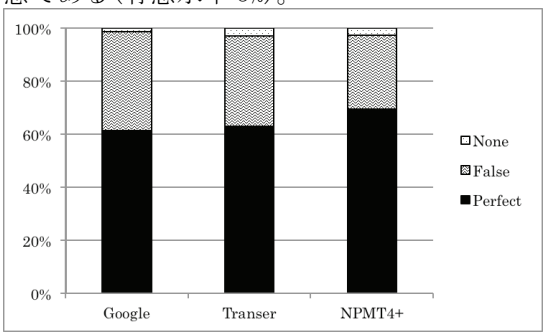
① オックスフォード・テストセット：『オックスフォード科学辞典』から抽出した 2534 対。
② 技術用語セット：『日・英・西 技術用語辞典』から抽出した 6036 対。
③ 分野混合テストセット：23 分野の『学術用語集』および『経済・法律 英和・和英辞典』から抽出した 2500 対。25 分野それぞれに対する 100 対の対訳から構成されている。

(6) 性能評価実験を行い、本方式の有効性を確認するとともに、市販の翻訳システムを超える翻訳性能を有することを確認した。

次のグラフは、単に対訳辞書を引いた場合(英辞郎)、対訳辞書とターゲットリストを組み合わせた場合(英辞郎 w/T)、選択型翻訳で拡張機構を導入しない場合(NPMT0)、選択型翻訳で拡張機能を導入した場合(NPMP4=提案方式)のそれぞれにおける翻訳性能を表している(オックスフォード・テストセットを使用、対訳辞書 236 万件、ターゲットリスト 590 万件)。Perfect は正しい訳のみ出力した場合、Ambiguous は出力に正しい訳が含まれる場合、False は出力に正しい訳が含まれない場合、None は出力がなかった場合を示す。このグラフから、(a)タームの翻訳は対訳辞書を用いるだけでは不十分であること、(b)ターゲットリストが正しい訳語の選択に寄与すること、(c)拡張機構が有効に機能していること、がわかる。英語ウィキペディアにおいて適切な見出し語に到達できる割合は、Perfect と Ambiguous の和であり、対訳辞書を引く方法は 48%にとどまるのに対し、提案方式(NPMT4)では 77%となっている。

次のグラフは、Google 翻訳、市販ソフト Mac Transer との性能比較を示している。これら２つの翻訳システムは訳語を１つしか出力しないため、選択型翻訳にウェブヒット数を利用した訳語選択を追加した場合(NPMT4+)と比較した。２つの翻訳システムの翻訳精度は 61-63%であるのに対し、選択型翻訳の精度は 70%である。この差は統計的に有意である(有意水準 5%)。





(7) 本研究では、英語ウィキペディアを日本語で引くシステムを実現したが、提案方式は、

有限の見出し語集合を持つ英語辞書や英語百科事典などの言語横断見出し語検索に広く適用できる。また、枠組み自体は、日英翻訳に限定されるものではなく、対訳辞書を準備することができれば、あらゆる言語対の翻訳に適用可能である。

(7) これまで、タームの翻訳は、機械翻訳研究において、中心的な課題として捉えられてこなかった。しかしながら、最先端の市販の機械翻訳システムにおいても、ターム翻訳の精度はそれほど高くなく、十分に解けている問題とは言いがたい。本研究は、ターム翻訳の難しさの原因の一部を明らかにし、それに対する解決策を示した点に、大きな貢献がある。

(8) 本研究のもう一つの貢献は、「翻訳を選択問題として解く」という新しい問題設定の提案である。タームの翻訳では、新たな訳を作り出すことが必要な場合は非常に限られており、ほとんどの場合は、既知の訳を探すことで適切な訳に到達できる。このような考え方は、機械翻訳研究にこれまでとは異なる方向性をもたらす可能性がある。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計４件）

① 岡田昌也,佐藤理史,駒谷和範. 英語ウィキペディアを日本語で引く：性能向上の検討. 言語処理学会第 18 回年次大会,2012.3.15,広島市立大学(広島県).
② Satoshi Sato and Masaya Okada. Japanese-English Cross-Language Headword Search. 9th International Conference on Terminology and Artificial Intelligence, 2011.11.8, Paris, France.
③ 岡田昌也,佐藤理史,駒谷和範. 英語ウィキペディアを日本語で引く. 2011 年度人工知能学会全国大会(第 25 回), 2011.6.2, アイーテいわて県立情報交流センター(岩手県).
④ 岡田昌也,佐藤理史. 大規模訳語候補集合を利用した専門用語翻訳. 2010 年度人工知能学会全国大会(第 24 回), 2010.6.10, 長崎グリックホール(長崎県).

６．研究組織
(1)研究代表者

佐藤　理史（SATO SATOSHI）
名古屋大学・工学研究科・教授
研究者番号：30205918

(2)研究分担者なし

(3)連携研究者なし