E-021

大規模単言語コーパスの利用による機械翻訳用対訳辞書の新規連語獲得性能の向上

Improving Performance of Acquisition of Bilingual Pairs for MT Using Large Monolingual Corpora

九津見 毅[†]　吉見 毅彦[‡/††]　小谷 克則[††]　佐田 いち子[†]　井佐原 均[††]

T. Kutsumi[†]　T. Yoshimi[‡††]　K. Kotani[††]　I. Sata[†]　H. Isahara[††]

1. はじめに

機械翻訳システムにおいて、連語を翻訳した際の訳として、連語を構成する個々の単語の訳語から構成的に得られた翻訳結果が、適切でない場合がある。従って、連語の的確な対訳を拡充していくことは重要である。

従来研究では、対訳辞書のための情報を獲得する際に、文アライメントのついた二言語コーパスがよく用いられる([北村他 2006]など)が、アライメントがよくできた良質な二言語コーパスは不足している。このような認識から、我々の先行研究[Kutsumi et al. 2006]では、既存の対訳辞書とシソーラスを用い、コーパスを使用せずに連語の対訳を拡充する方法を示した。しかし、この方法には、翻訳品質にはほとんど悪影響を与えないが、現実の文中にはまず出現しないために翻訳処理で使用されないような連語とその対訳も比較的高い割合で生成されてしまうという問題がある。

そこで、本稿では、二言語コーパスに比べて利用しやすい単言語コーパスを追加して利用することで、適切な連語対訳獲得の効率を高める方法を示す。

2. 新規連語とその全体訳の生成

本研究では、二語で構成される英語の連語 (名詞句) を、連語全体としての日本語訳(以下、全体訳と呼ぶ)と個々の単語の単独訳をそのまま結合した日本語訳とがどのように一致するかという観点から、次の表 1 のように分類する。

分類	英語連語	日本語訳	上段:全体訳 下段:単語訳
両方一致	toy box	おもちゃ箱 おもちゃ/箱	
前方一致・後方不一致	salt shaker	塩入れ 塩/シェーカ	
前方不一致・後方一致	member company	会員会社 メンバー/会社	
両方不一致	bullet train	新幹線 弾丸/列車	

表 1 連語の分類とその例

上記の分類のうち、本研究では、「前方一致・後方不一致」の連語を処理対象とする方法を提案する[1]。

上記の「前方一致・後方不一致」の連語(以下「生成元連語」と呼ぶ)は、その全体訳とその第二単語の単独訳が後方一致しない。このような連語の第一単語をその類義語で置き換えて得られる連語(以下「新規連語」と呼ぶ)の適切な全体訳は、第一単語の類義語の単独訳と第二単語の単独訳をそのまま結合したものではなく、第一単語の類義語の単独訳と生成元連語で第二単語に対応する訳を結合したものである可能性が高い。例えば、“salt shaker” において “shaker” が「入れ」に対応していることから、“shaker” と “salt” の類義語 (“spice” など) 結合した連語 “spice shaker” とその対訳が辞書に登録されていない場合、“shaker”を「シェーカ」ではなく「入れ」と訳し、“spice shaker”と「スパイス入れ」の組を対訳辞書に新たに登録する。

† シャープ株式会社, Sharp Corporation
‡ 龍谷大学, Ryukoku University
†† 独立行政法人情報通信研究機構, NICT

[1]提案手法は、「前方不一致・後方一致」の連語にも適用可能であると考えられるが、今回の実験では「前方一致・後方不一致」の連語のみを対象とした。

3. 従来の順位付け手法

我々の先行研究では、前記の方法で生成された新規連語とその全体訳の対を、以下に述べる方法で優先順位を付けている。詳細は文献[Kutsumi et al. 2006]を参照されたい。

3. 1 英語シソーラスによる類似度の利用

新規連語は生成元連語の第一単語をその類義語で置き換えることによって生成される。このため、新規連語が適切な英語名詞句である可能性は、第一単語とその類義語が意味的に近いほど高いと考えられる。

この考えに基づき、生成元連語の第一単語 W_E とその類義語 $SimW_E$ についての英語シソーラス(WordNet)上での類似度 $SIM(W_E, SimW_E)$を、生成元連語とその全体訳の対 *SeedPair* から生成される新規連語とその全体訳の対 *NewPair* の優先度 $Score_{SeedPair}(NewPair)$とする。

$$Score_{SeedPair}(NewPair) = SIM(W_E, SimW_E)$$

W_E と $SimW_E$ の類似度 $SIM(W_E, SimW_E)$は、標準的な計算式の一つである次の式で求める [黒橋 1996][2]。

$$SIM(W_E, SimW_E) = \frac{2 \times d_C}{d_{WE} + d_{SimWE}}$$

ただし、d_{WE} と d_{SimWE} はそれぞれシソーラスにおける根節点から W_E までの深さと $SimW_E$ までの深さであり、d_C は W_E と $SimW_E$ に共有される節点までの深さである。

$SIM(W_E, SimW_E)$が閾値以上である $SimW_E$ を用いて新規連語を生成する。閾値は実験的に決定する。

3. 2 日本語シソーラスによる類似度の利用

英語シソーラスにおける不適切な語義選択や、生成元連語の第一単語の類義語の翻訳に利用する機械翻訳システムにおける不適切な訳語選択を抑制するために、生成元連語の第一単語の訳 W_J とその類義語の訳 $SimW_J$ の日本語シソーラス(EDR 辞書)上での類似度を利用する。日本語シソーラスでの類似度は英語シソーラスの場合と同様に計算する。

ある生成元連語とその全体訳の対 *SeedPair* から得られる新規連語とその全体訳の対 *NewPair* の優先度は英語シソーラスでの類似

[2]本稿は、類似度の計算式を提案することが目的ではないため、計算式の比較検討は今後の課題とする。

度 $SIM(W_E, SimW_E)$ と日本語シソーラスでの類似度 $SIM(W_J, SimW_J)$ によって決まると考え、次の式(2)による値 $Score_{SeedPair}(NewPair)$ を各対 *NewPair* に与える。

$$Score_{SeedPair}(NewPair) = SIM(W_E, SimW_E) \times SIM(W_J, SimW_J)$$

例えば、対訳辞書に登録されている "pop vocal"（「ポップ歌手」）という生成元連語の第一単語 "pop" の類義語として "soda" があるが、"pop" を "soda" で置き換えた "soda vocal"（「ソーダ歌手」）は適切な連語ではない。このような問題への対策として日本語シソーラスでの類似度を採用すれば、生成元連語の第一単語 "pop" の訳「ポップ」と "pop" の類義語 "soda" の訳「ソーダ」との類似度は低くなり、このような不適切な連語の優先度を下げることができる。

3．3　生成元連語数の考慮

人手で構築された対訳辞書に登録されている生成元連語は、辞書登録することで翻訳品質が向上すると辞書開発者によって判断された連語である。このため、より多くの生成元連語から生成される新規連語ほど翻訳品質の向上に貢献する可能性が高いとみなす。

この考えに基づいて、第二単語が同じである生成元連語 $SeedPair_1, \ldots, SeedPair_n$ から一つの新規連語とその全体訳の対が生成される場合、$SeedPair_i$ から *NewPair* が生成されるときの各優先度 $Score_{SeedPair_i}(NewPair)$ の和 $AggScore(NewPair)$ を求め、それを *NewPair* の優先度とする。

$$AggScore(NewPair) = \sum_{i=1}^{n} Score_{SeedPair_i}(NewPair)$$

例えば、第二単語が同じである連語として、"salt shaker" と "pepper shaker" がある。また、"salt" の類義語としても "pepper" の類義語としても "spice" があり、"salt" の類義語ではあるが "pepper" の類義語ではない語に "carbonate"（「炭酸塩」）がある。このとき、2 個の生成元連語から生成される新規連語 "spice shaker" の優先度は "salt shaker" 由来の優先度 0.556 と "salt shaker" 優先度 0.648 とを合計した 1.204 であるのに対し、"salt shaker" のみから生成される新規連語 "carbonate shaker" の優先度は 0.762 となり、"spice shaker" の優先度を下回る。

4．提案手法による順位付け

3 節で述べた従来手法による新規連語とその全体訳の組の順位付けの結果、順位が上位の組の中にも、対訳辞書に登録するには不適切であるものが出現する。たとえば "telephone set tree" のように不適切な新規連語や、"schoolhouse management" と「校舎経営」の組のように新規連語は適切であるが全体訳が不適切なもの[3]である。

このような意味的に妥当でない不適切な名詞句は、コーパスでの出現頻度が低いはずである。従って、従来手法の問題点に対処するために、本稿では、新規連語やその全体訳の適切性の指標として大規模な単言語コーパスでのそれらの出現頻度を用いて、新規連語とその全体訳の組への順位付けを改善する。より大規模なコーパスを利用するため、英語・日本語それぞれの単言語コーパスを用いる。本研究では web 空間を大規模コーパスとみなし、新規連語あるいは全体訳の web 検索サイトにおける検索結果のヒット件数を、英語及び日本語 web コーパス中の出現頻度と見なし、これらを優先度に加味することを考える。Web を利用した研究の例としては、文献[Grefenstette1999]や[柴田他 2005][外池他 2006]などがある。

新規連語とその全体訳の組 *NewPair* の、web コーパス中での出現頻度を考慮した優先度 $FreqAggScore(NewPair)$ を、次の式により与える。

$$FreqAggScore(NewPair) = AggScore(NewPair) + a_E \times \log_{10}(freq_E + 1) + a_J \times \log_{10}(freq_J + 1)$$

ここで、$freq_E$ は新規連語(英語)の出現頻度、$freq_J$ はその全体訳(日本語)の出現頻度、a_E および a_J は各々の重み係数である。上の式で、出現頻度の対数を採用した理由は、web 検索エンジンが示すヒット件数が 1 桁～百万のオーダと大変広い範囲にわたっており、5 節で述べる実験の範囲内でのその分布は表 2 に示す通り、英語においても日本語においても対数的分布を示している(比率を等間隔にした各階級の度数がほぼ同じである)ので、頻度値の影響を適切に反映させるには対数処理が適していると考えたためである。

重み係数 a_E と a_J は、新規連語とその全体訳の出現頻度をどの程度重視するかを表す。5 節の評価実験では、a_E と a_J をそれぞれどのような値に設定すれば新規連語とその全体訳の組に対して適切な優先度を与えることができるのかを検証する。

英語頻度	個数	日本語頻度	個数
0	2011	0	2380
1－9	338	1－9	346
10－99	301	10－99	782
100－999	899	100－999	921
1000－9999	380	1000－9999	97
10000－99999	740	10000－99999	328
100000－999999	303	100000－999999	120
1000000－9999999	37	1000000－9999999	35

表 2 3 節の方法での順位で上位 5009 件の新規連語の頻度の分布

5．評価実験

5．1　方法と結果

実験には、シャープ(株)で開発している英日翻訳エンジンとその辞書データを用いた。対訳辞書の一部分から「前方一致・後方不一致」の連語を 2148 語抽出し、これらを生成元連語として新規連語とその全体訳の対を生成した。このとき、3.1 節で述べた類似度が 0.7 以上のものを新規連語とした。この結果得られた新規連語とその全体訳の対は、759,704 件であった。これらの対に対し、まず 3 節で述べた、英語シソーラスでの類似度と日本語シソーラスでの類似度と生成元連語数を考慮した順位付けを行い、優先度が上位 5009 位[4]までを抽出した。次に、抽出された 5009 対に対し 4 節で述べた優先度計算式を適用し、頻度情報を加味した優先度 $FreqAggScore(NewPair)$ を求めた。係数 a_E および a_J は、それぞれ 0, 0.1, 0.2, 0.5, 1.0 に変化させて調べた。ここで、a_E および a_J が共に 0 なら、頻度情報を加味しない 3 節の方法であり、これをベースラインとする。そして、係数 a_E と a_J の組み合わせごとに、$FreqAggScore(NewPair)$ の上位 500 件[5]を抽出し、その結果を次のよ

[3] この例での全体訳は、この例での単語訳の組み合わせである「校舎管理」の方がまだ適しているといえる。

[4] 上位 5000 位までを抽出しようとした際、3 節の方法での 5000 位に相当する優先度の組が同一優先度で複数組存在したので、それらをすべて含めた結果、5009 組となった。

[5] これについても、500 位に相当する優先度の組が複数存在する場合は、それらすべてを含めた。その結果、抽出件数が 500 件を超えた場合もある。

うな手順で評価した[6]。まず、新規連語が適切な英語名詞句であるか否かの判断を英語母語話者が行い、このとき不適切な英語名詞句であるとされたものを「不適切連語」に分類した。次に、適切な英語名詞句であると判断された新規連語について、その全体訳とその構成単語の単独訳をそのまま結合した訳とを日本語母語話者が比べ、その結果、新規連語とその全体訳を辞書に登録した場合の効果が「有益」、「有害」、「同等」のいずれかであるかを判断した。「有益」は、新規連語の全体訳のほうが新規連語の構成要素の単独訳を結合した訳よりも良く、対を対訳辞書に登録することによって翻訳品質の向上が期待できる場合を意味する。これに対して、「有害」は、新規連語の全体訳のほうが悪く、対を対訳辞書に登録すれば翻訳品質が低下する恐れがある場合を意味する。どちらの訳も新規連語の訳として適切である場合あるいは両方とも不適切である場合は「同等」とした。

集計結果を図 1～図 3 に示す。ここで、「有害+不適切」は、「有害」及び「不適切連語」に分類された対の全体に対する割合である。図 1 は、日本語頻度係数 a_J を 0 に保ったまま英語頻度係数 a_E を 0 ～ 1 に変化させたもの（a_E が 0 であるのがベースライン）、図 2 は、英語頻度係数 a_E を 0 に保ったまま日本語頻度係数 a_J を 0～1 に変化させたもの（a_J が 0 であるのがベースライン）、図 3 は、英語頻度係数 a_E と日本語頻度係数 a_J の両方を変化させたもの（a_E 及び a_J が 0 であるのがベースライン）である。

5．2 考察

図 1 から、英語連語の出現頻度を導入した影響をみると、ベースラインでの「不適切連語」の割合が 30.0%に対し、a_E を 0.1 にした際の「不適切連語」の割合が 17.8%（ベースラインからの減少率 41%）、a_E を 0.2 にすると 11.6%（ベースラインからの減少率 61%）となり、期待通りの効果が得られていることがわかる。たとえば不適切連語の例である“horseless carriage plant”と「自動車工場」の組[7]は、ベース ライン手法では“horseless carriage”と“auto”との類似度などのため優先度が上位 500 位に入っている（6 位）が、コーパス中の英語連語の出現頻度が 0 であるので、a_E が 0.2 を超えると上位 500 位から外れる（600 位）。この例のように、英語において単語単位で意味が類似していても連語としての用法が奇異であるような連語候補を、英語頻度の導入により排除することができるため、英語頻度を適切な重みで導入すると「不適切連語」の優先度を効果的に下げることができると考えられる。ただし係数 a_E が 0.2 を超えてから「不適切連語」の割合はわずかに増加に転じる。一方で「有益」が増加しているがこれも a_E が 0.2 を超えると減少に転じる。この結果、「有益」は a_E が 0.2 の場合に極大（最良）、「有害+不適切」も a_E が 0.2 の場合に極小（最良）を示している。a_E が 0.2 を超えるとそれ以上 a_E を増加しても傾向があまり変わらない原因は、a_E が 0.2 を超えると類似度に基づくベースライン手法の影響が相対的にほとんど効かなくなり、英語頻度のみに基づく結果になっているためと考えられる。

次に、図 2 から、日本語連語の出現頻度を導入した影響をみると、ベースラインでの「有害」の割合が 4.4%に対し、a_J を 0.1 にすると 2.2%（減少率 50%）となり、a_J が 0.2 の場合に極小（最良）を示している（「有害」割合 1.8%、ベースラインからの減少率 59%）。たとえば「有害」の例である“remuneration bill”と「報酬法案」の組[8]（単語訳の組み合わせ「報酬請求書」の方がより適した訳語で

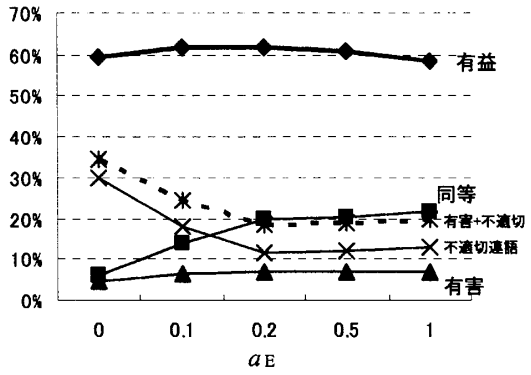

図 1 英語頻度への重み係数を変化させた結果

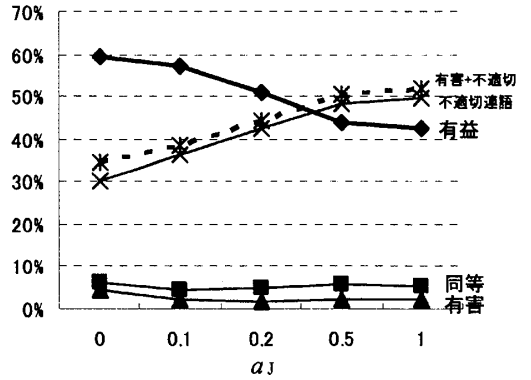

図 2 日本語頻度への重み係数を変化させた結果

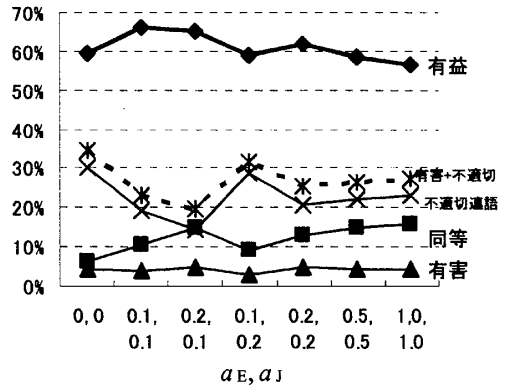

図 3 英語・日本語両方の頻度を加味し重み係数を変化させた結果

[6] つまりここでは、*AggScore(NewPair)*による上位 5000 件の抽出の後、*FreqAggScore(NewPair)*による上位 500 件の抽出という、2 段階抽出を行っている。理想的には、後者のみの 1 段階抽出にて評価することが望ましいが、そのためには新規連語とその全体訳のすべての組（約 76 万対）の Web 出現頻度を求めることが必要になる。

[7] 辞書登録連語“auto plant”「自動車工場」の組と、単語対訳“horseless carriage”「自動車」とから生成された。

[8] 辞書登録連語“spending bill”「支出法案」の組と、単語対訳“remuneration”

ある)は、ベースライン手法では"remuneration"と"spending"との類似度の近さなどのため優先度が上位 500 位に入っている(219 位)が、コーパス中の日本語連語の出現頻度が 0 であるので、a_Jが 0.1 を超えると上位 500 位から外れる(778 位)。

ただし、図 2 から、a_Jを増加するにつれ「有益」が減少し「不適切連語」が増加していることがわかる。前者の理由としては、たとえば辞書登録連語"sodium compound"「ナトリウム化合物」の組と、単語対訳"americium"「アメリシウム」とから生成された「有益」な新規連語"americium compound"「アメリシウム化合物」の組のように、日本語として不自然ではないが実際の用例がほとんどない(調査したコーパス中「アメリシウム化合物」の出現頻度は 0)ような連語は、日本語頻度を加味すると優先度が下がるためと考えられる。後者の理由は、たとえば辞書登録連語"antenna shop"「アンテナ・ショップ」の組と、単語対訳"aerial"「アンテナ」とから生成された「不適切」な新規連語"aerial shop"「アンテナ・ショップ」の組のように、生成された連語の日本語が既存の連語と同じ(ただし英語連語としては不適切)例が多く見受けられ、このような日本語連語の日本語コーパス中の出現頻度は当然高いので、日本語頻度を加味すると優先度が上がるためと考えられる。

さらに、図 3 から、英語連語・日本語連語両方の出現頻度を組み合せて導入した影響をみる。この結果、「有益」は、a_E=0.1,a_J=0.1 の組み合わせの際に本実験中で最高(最良)の 66.2%、a_E=0.2,a_J=0.1 の組み合わせの際にもこれに次ぐ 65.4%を示した。「有害+不適切」は a_E=0.2,a_J=0.1 の組み合わせの際に本実験中で最低(最良)に近い 19.6%を示した。この結果から、英語連語の出現頻度と日本語連語の出現頻度の組み合せが適切に機能していると考えられる。

以上のことから、選別結果が最良になるのは、英語連語の出現頻度と日本語連語の出現頻度をそれぞれ、英語頻度への重み係数 0.1～0.2 程度、日本語頻度への重み係数 0.1 で導入した場合であることがわかった。

6．関連研究

提案手法を二言語コーパスからの対訳獲得手法と比較する。コーパスから適切な語彙知識を獲得するためには、(1) 二言語コーパスにおいて英語表現と日本語表現を正しく対応付ける処理と、(2) 対応付けられた表現対を辞書に登録するか否かを判定する処理が必要である。なぜならば、対応付けられた表現対には、辞書に登録することによって翻訳品質が向上することがほぼ確実かどうかによって表現対を選別する必要があるからである。二言語コーパスからの対訳獲得研究では、対応付けに焦点が当てられていることが多く、選別については検討が不十分である。すなわち、獲得される表現対の品質よりも量を重視していると言える。これに対して、提案手法は、人手による判断を行なわなくても、辞書登録による翻訳品質の向上がほぼ確実な連語を(たとえ少量であっても）確実に獲得することを目的としている。既存の対訳辞書に登録されている連語は、辞書登録することで翻訳品質が向上すると辞書開発者によって判断された連語であることに着目して、これらを生成元連語として利用することにより、この目的の達成を目指している。具体的には、1 節で述べたように、対訳辞書に"salt shaker"と「塩入れ」の組が既に登録されている場合、この組から生成される"spice shaker"と「スパイス入れ」の組や"pepper shaker"と「コショウ入れ」の組は、辞書に未登録であれば、登録することによって翻訳品質の向上に貢献する可能性が高いと考えられる。

提案手法を二言語コーパスからの対訳獲得手法と組み合わせて用いることもできる。二言語コーパスから連語とその対訳を獲得する際に生じる問題として、連語の構成語とその対訳候補の構成語の間で素直な対応関係が成り立たないものは対応付けが困難であるということが挙げられる。田中ら[田中・松尾 2001]は、連語とその対訳の間の対応関係として、対訳辞書による素直な対応の他に、類義性による対応、対訳共起による対応を設定することで、本来対応関係にある連語とその対訳の間の対訳確信度(類似度)が低くならないように対処している。対訳共起による対応は、本稿でいう「前方一致・後方不一致」や「前方不一致・後方一致」の連語における不一致部分の単語とその訳語との対応に相当する。従って、本稿の提案手法を利用すれば、田中らの方法で対訳共起による対応付けのために利用された(専門用語)辞書を二言語コーパスからの対訳獲得処理に用いる前に拡張しておくことができる。このことによって、対訳共起による対応付けがより有効に働くようになることが期待される。

7．おわりに

本稿では、対訳辞書に登録されている既存の連語の構成要素をその類義語で置き換えることによって、辞書に登録されていない新たな連語を獲得し、それをその全体訳と共に辞書に登録することによって辞書を拡充する方法を示した。提案方法では、生成された新規連語とその全体訳の対への優先順位付けにおいて複数のシソーラスを参照する処理と生成元連語数を考慮した処理を行ない、その結果に対して更に、新規連語と全体訳のコーパス中の出現頻度を考慮した優先順位付けを行なって、上位の対を抽出した。評価実験の結果、辞書登録によって翻訳品質低下の恐れがある対の生成を 19.6%の割合に抑え、逆に辞書登録によって翻訳品質向上が期待される対を 65.4%の割合で獲得することができた。この精度は、コーパスでの出現頻度を考慮した処理を行なわない場合の精度(34.4%と 59.4%)を上回るものである。

参考文献

Grefenstette, G. (1999): The World Wide Web as a Resource for Example-based Machine Translation Tasks, In Procs. of the 21st International Conference on Translating and the Computer.

北村美穂子, 松本裕治(2006). "言語資料を活用した実用的な対訳表現抽出." 自然言語処理 Vol.13, No.1, pp.3–25.

黒橋禎夫 (1996). "辞書とコーパス." 長尾真(編), 自然言語処理, pp.231-264. 岩波書店.

Kutsumi, T., Yoshimi, T., Kotani, K., Sata, I. and Isahara, H. (2006). "Expansion of Machine Translation Bilingual Dictionaries by Using Existing Dictionaries and Thesauruses." In *Procs. of 21st International Conference on the Computer Processing of Oriental Languages.*

柴田雅博, 冨浦洋一, 田中省作 (2005). "Web 上の語の共起性に基づいたコロケーションの翻訳支援." 情報処理学会論文誌, 46(6), 1480-1491.

田中貴秋, 松尾義博 (2001). "対訳関係のないコーパスからの複合名詞対訳表現の獲得," 電子情報通信学会論文誌, Vol.J84-D2, No.12, 2605—2614.

外池昌嗣, 宇津呂武仁, 佐藤理史(2006). "ウェブと要素合成法を用いた専門用語訳語推定." 言語処理学会第 12 回年次大会発表論文集, pp. 412-415.

宇津呂武仁, 日野浩平, 堀内貴司, 中川聖一 (2005). "日英関連報道記事を用いた訳語対応推定." 自然言語処理, 12(5), 43-69.

「報酬」とから生成された。