既存辞書・シソーラス・単言語コーパスを用いた 機械翻訳用対訳辞書の新規連語獲得

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

†シャープ(株)　‡龍谷大学　††情報通信研究機構

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、二言語コーパスがよく用いられるが[Kaji et al. 2005;柴田他 2005; 宇津呂他 2005]、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。そこで本研究では、既存の対訳辞書と、単言語シソーラスに加え、利用しやすい単言語コーパスを用いて、連語の対訳を拡充する方法を示す。

２．提案方法の概要

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」及び「前方不一致・後方一致」の連語を処理対象とする方法を提案する。ただし、本稿では「前方一致・後方不一致」についてのみ説明する。

上記の「前方一致・後方不一致」の連語は、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新規連語の適切な全体訳は、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元の「前方一致・後方不一致」の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” の場合にも “shaker” を「シェーカ」ではなく「入れ」と訳す必要がある。

このような考えに基づいて、辞書に登録されていない新たな連語とその全体訳をまず獲得し、優先順位づけする[Kutsumi et al. 2006]。更に、単言語コーパスへの出現頻度情報を用いて、頻度に基づく優先度の高い対を抽出する。提案方法は次の４段階から成る。

(1) 対訳辞書から「前方一致・後方不一致」の連語(生成元連語)を抽出する。
(2) 「前方一致・後方不一致」の連語の第一単語の類義語をシソーラスから抽出し、その類義語と「前方一致・後方不一致」の連語第二単語を結合したものを新たな連語(新規連語)とし、さらにその全体訳を生成する。
(3) 新規連語とその全体訳の対へ優先順位を付ける。
(4) 新規連語及び全体訳のコーパス中に出現する頻度の上位の対を抽出する。

３．新規連語とその全体訳の生成

提案方法では、既存の対訳辞書に収録されている生成元連語について、その第一単語の類義語をシソーラスから抽出し、第一単語を類義語で置き換えたものを新規連語とする。

シソーラスとしてWordNet2.0 版の階層構造[Miller 1998]を用いる。与えられた語Xに対して、WordNetに存在する語Yとの間で、式(1)で計算される類似度$SIM(X,Y)$ [黒橋 1996]がある閾値以上になる場合、YをXの類義語とみなす。[1]

[1] 本研究の実験では閾値を 0.7 とした。

$$SIM(X,Y) = \frac{2 \times d_C}{d_X + d_Y} \quad (1)$$

ただし、d_X と d_X はそれぞれ WordNet における根節点から X までの深さと Y までの深さであり、d_C は X と Y に共有される節点までの深さである。

生成元連語：	salt shaker
生成元全体訳：	塩 入れ
新規連語：	spice shaker
新規全体訳：	スパイス 入れ

表 2 新規連語とその全体訳の例

新規連語の全体訳を生成するために、第一単語の類義語を機械翻訳して得られる訳と生成元連語の第二単語に対応する訳とを結合する。例えば生成元連語が“salt shaker”である場合、“salt”との類似度が閾値以上である語として“spice”がシソーラスから抽出されるので、新規連語として“spice shaker”が得られる。“spice shaker”の全体訳としては、“spice”を機械翻訳して得られる「スパイス」と“salt shaker”において“shaker”に対応する「入れ」とを結合した「スパイス入れ」を与える。

４．新規連語とその全体訳の対の順位付け

前記の方法で生成された新規連語とその全体訳の対を、翻訳品質向上への貢献度の観点から選別するために、各対に優先順位を付け、順位が高い対から順に出力する方法について述べる。

４．１　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 $SimW_E$ について式(1)で計算される類似度 $SIM(W_E, SimW_E)$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。

$$Score_{SeedPair}(NewPair) = SIM(W_E, SimW_E) \quad (2)$$

４．２　日本語シソーラスによる類似度の利用

英語シソーラスや機械翻訳システムに由来する不適切な語義選択や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 3 節の式(1)で計算する。日本語シソーラスとして EDR 電子化辞書を用いる。

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(3)による値 $Score_{SeedPair}(NewPair)$を各対 *NewPair* に与える。

$$Score_{SeedPair}(NewPair) = SIM(W_E, SimW_E) \times SIM(W_J, SimW_J) \quad (3)$$

４．３　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。

たとえば、新規連語“spice shaker”は、前記した方法によって、“salt shaker”からも“pepper shaker”からも生成可能である(“salt shaker”も“pepper shaker”も辞書登録語である場合)。一方、新規連語“carbonate shaker”は、辞書登録語のうち“salt shaker”からのみ生成可能である。このような場合、新規連語“spice shaker”は“carbonate shaker”よりも優先する。

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 $ScoreSeedPair_i(NewPair)$ の和を求め、それを *NewPair* の最終的な優先度 *AggScore*(*NewPair*)とする。

$$AggScore(NewPair) = \sum_{i=1}^{n} Score_{SeedPair_i}(NewPair) \quad (4)$$

５．評価実験

５．１　類似度と生成元連語数による順位付け

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二語で構成される連語を抽出した。

抽出された連語は 25351 語である。これらを、2 節の表 1 に示した 4 種類に分類し、このうち「前方一致・後方不一致」の連語 2148 語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。これらの対に対し、4 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、上位 500 位までを抽出した。

抽出された 500 対に対し、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ネイティブが行なった。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ネイティブが比べ、「有益」、「有害」、「同等」のいずれかであるかを判断した。この結果を表 4 に示す。

有益	不適切連語	同等	有害	計	改善率
297 (59.4%)	150 (30.0%)	31 (6.2%)	22 (4.4%)	500 (100.0%)	275 (55.0%)

表 4 4 節での順位付け方法での性能

５．２　単言語コーパスでの出現頻度による判定

本実験では、前記の方法で生成された新規連語やその全体訳について、その適切性をネイティブが人手で判断したが、この手法の実用化のためには適切性の判断の自動化を目指す必要がある。よって、人手判定の代替となる指標として、コーパス中の出現頻度を利用することを考える。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 検索サイト Google を利用し、新規連語あるいは全体訳の検索結果として Google が示すヒット件数を、英語及び日本語 web コーパス中の出現頻度と見なした。

5.1 節で抽出した 500 件の新規連語と全体訳の対を対象として、次の各方法で順位付けし直した。

(ア) 英語での類似度と日本語での類似度と生成元連語数とを考慮した優先度(5.1 節の方法)
(イ) 新規連語(英語)の出現頻度
(ウ) 全体訳(日本語)の出現頻度

上記の(ア)～(ウ)の方法で順位付けし、それぞれ上位 81 位までを抽出して、「有益」「無害(不適切連語)」「無害(同等)」「有害」の分布を調べた。81 位までとした理由は、上記(イ)の順位付け方法(新規連語の出現頻度)で、出現頻度 10,000 件で足切りした際に残るのが 81 位までだからである。[2]

[2](イ)以外の順位付け方法で、81 位に相当する対が同一スコアで複数ある場合は、それらすべてを抽出した。

分布結果から次の指標を求めた。

・改善率：足切り範囲内の「有益」と「有害」の差の割合
・不適切率：足切り範囲内の「無害(不適切連語)」の割合
・有益切り捨て数：頻度調査対象とした 500 件中の「有益」全数(297 件)から、足切り範囲内の「有益」件数を引いた数

不適切率を求めた理由は、これらの順位付け法の上位を適切連語とする判定が、人手判断による英語新規連語の適切性を示す指標になりうるかを知ることにある。また、ある順位で足切りした際に「有益」な対が捨てられることは避けられないが、その際に頻度を考慮した順位付けをすれば、同じ順位で足切りしても「有益」対の切り捨て数がどの程度改善(減少)されるかを知るために、「有益」切り捨て数を求めた。

順位付け法	ア	イ	ウ
有益	53	57	54
不適切連語	28	8	20
同等	2	9	4
有害	2	7	3
計	85	81	81
改善率(%)	60.0	61.7	63.0
不適切率(%)	32.9	9.9	24.7
有益切り捨て数	244	240	243

表 5 頻度を考慮した順位付け法での性能

更に、足切り件数を次のように変えて分布の変化を調べた。

(a) 上位 81 件 ((イ)の方法で出現頻度 10000 件相当)
(b) 上位 117 件 (イ)の方法で出現頻度 1000 件相当)
(c) 上位 185 件 ((イ)の方法で出現頻度 400 件相当)
(d) 上位 266 件 ((イ)の方法で出現頻度 100 件相当)

順位付け法ごとの、足切り件数の変化による改善率・不適切率・「有益」切り捨て数の変化をそれぞれ図 1・図 2・図 3 に示す。改善率は高いほど好成績で、不適切率及び「有益」切り捨て数は低いほど好成績となる。

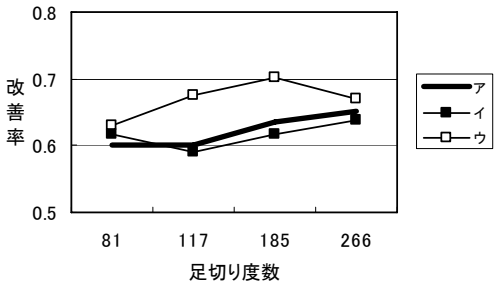

図 1 改善率の足切り度数による変化

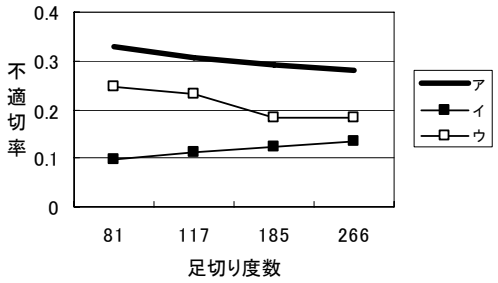

図 2 不適切率の足切り度数による変化

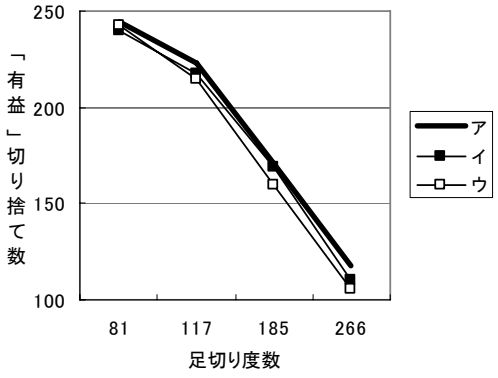

図 3 「有益」切り捨て数の足切り度数による変化

改善率は、(イ)の方法では(ア)(頻度を考慮しない順位付け)と大差ない。(ウ)は(ア)(イ)より改善されており、中でも、足切り度数を上位 185 件とした場合の(ウ)の改善率は 70.3%となった。

一方、不適切率は、足切り度数全般にわたり(イ)の方法が大きく改善されている。(ウ)は、(ア)よりは改善されているが、(ア)と同様に足切り度数との逆連関傾向(上位に限るほど成績が悪い)がそのまま出ている。

「有益」切り捨て数は、頻度を考慮した方法はいずれも(ア)より改善されていて、中でも全体訳(日本語)の頻度情報を利用する(ウ)が良い傾向を示す。

以上から、順位付けに頻度情報を利用した結果は、これを利用しない結果より全般に改善されていることがわかる。

６．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、新規連語や全体訳の、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が向上することが期待される対が最大約 70%の改善率で獲得できた。この精度は、コーパス中の出現頻度を考慮した処理を行なわない場合の改善率 55.0%を上回るものである。

参考文献

Kaji, H. (2005). “Extracting Translation Equivalents from Bilingual Comparable Corpora.” *IEICE Transactions on information and systems*, E88-D(2), 313-323.

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.

Miller, G. (1998). “Nouns in WordNet.” In Fellbaum, C. (Ed.), *WordNet: An Electronic Lexical Database*, pp.23-46. The MIT Press.

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.

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.

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.*