2言語コーパスからの複合語の対訳曖昧さ解消

鈴木正史　　中川裕志
横浜国立大学 大学院 工学研究科　　東京大学 情報基盤センター

1. はじめに

近年の急激なインターネット利用人口の増加によって、我々は電子化された文書を気軽に利用できるようになった。また、これによって英語で書かれた文書を目にすることも多くなってきた。それゆえ、機械翻訳の利用も増してきていると言えるだろう。そこで、それらの文書を翻訳するための対訳辞書を作成することが重要になって来ている。ところが、進展の早い技術分野では専門用語が日々増加しているため、人手での辞書整備が追いつかない。したがって、分野毎に対訳辞書を自動抽出する必要がある。

従来、対訳辞書の自動抽出は、ソース言語とターゲット言語の2言語に渡るコーパスを用いてきた。2言語コーパスは大きく分けて、

- 平行コーパス
- 非平行コーパス

の2つに分類される。

平行コーパスとは対訳されているコーパスのことで、通常対訳は文単位で行われているものである。また、非平行コーパスとは対訳されていないコーパスのことであり、特にある1つの分野の文書から構成されるものをコンパラブルコーパスという。対訳ないし平行コーパスはカナダや香港のような極限られた2言語社会の公文書などで見られる以外は、一般に存在しない。まして特定の技術分野では希少である。したがって、本研究のように専門分野の対訳辞書構築にあたっては、平行コーパスを前提にすることができない。よって、本研究では、コンパラブルコーパスからの対訳辞書の自動抽出を目的にする。

2. 従来の対訳表現抽出の研究

コンパラブルコーパスでは分野特有の表現を含むが、対訳となる文などは基本的に存在しないので、パラレルコーパスの場合より細かい単位で文脈を利用する必要がある。以下にいくつかの手法を述べる。

1.**Context Heterogeneityを使う手法**[Fung95]

[Fung95]では訳すべき単語の左右に連接する単語の種類数(context heterogeneity)に着目し、単語 w_1 と w_2 の context heterogeneity のユークリッド距離が近いものを対訳と判定する。

2.**既存の対訳辞書および文脈を利用する手法**
[Rapp95,99]、[Fung98]、[Tanaka96]

[Rapp95,99]ではある語 w の前後2単語の共起ベクトルを用いて w の対訳を見つける。

[Fung98]は[Rapp95,98]と同様のアイデアであるが、共起ベクトルはある語 w の存在する文内の語で、既存の対訳辞書に載っている語すべてに関して計算する。

[Tanaka96]では、既存の対訳辞書(EDICT)を用い、それに対訳が曖昧な語を、その語と同じ文に共起する既知単語を統合した辞書をターゲット言語とソース言語における共起関係を満足するような制約の元に最適化する。

これらの方法はいずれも既存の対訳辞書を利用して未知語の対訳を求めるという方法であり、精度も70%以上になってきているが、計算量が大きい。また、基本単語での誤訳もある。

3. 複合語抽出による曖昧性解消方法

現在でも語基(単名詞、単動詞など)の対訳辞書、例えば EDICT、は存在する。したがって、専門用語に多い複合語の対訳を行うための最も単純な方法は、その複合語を構成する語基を既存の対訳辞書で翻訳する方法である。ただし、既存の対訳辞書には対訳に曖昧性がある。例えば、EDICT では、ひとつの日本語の単語に対して複数の訳語が存在することが頻繁にある。したがって、複合語の対訳には大きな曖昧性が生じる。仮に単語につき3個の訳語があったなら、2単語からなる複合語ですら9通りの曖昧さが生じる。したがって、この曖昧さを解消することが重要な課題となる。

我々は、この課題に対して、専門分野の専門用語であることを利用する曖昧性解消方法を以下で提案する。専門分野においては、意志の疎通を明確にする必要

から、次の仮定が成立するものと考えられる。

仮定1:専門分野の用語は概念と一対一に対応する

一方、2言語コーパスが十分大きくて専門用語が漏れなく捕捉されているなら、ソース言語で抽出された用語にはターゲット言語でも対応する訳語が存在するはずである。つまり、上記の仮定によれば、ターゲット言語での一意的訳語が高い確率で見つかるはずである。

このようなアイデアに基づくと、対訳の曖昧性解消は以下の手順で行えばよいことが分かる。

対訳の曖昧性解消手順

Step1 ソース言語のコーパスから用語抽出を行う。
Step2 ターゲット言語のコーパスからも用語抽出を行う。
Step3 ソース言語の複合語から既存の対訳辞書を利用してターゲット言語における対訳を生成する。ただし、一般に対訳は複数生成される。
Step4 Step3 で生成された対訳のうち、ターゲット言語で抽出された用語に一致するものがあれば、それを対訳として選択する。

以上をまとめ、例えば英日対訳辞書を2言語コーパスから自動生成するシステムは図1のようになる。

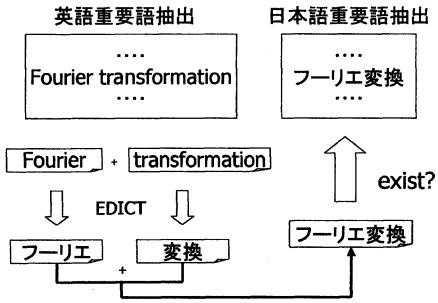

図1　英日対訳辞書の自動生成システム

ところが、この方法で求めた対訳では、特に英→日の対訳を生成する際、次に示すような誤訳を生成してしまう可能性が残る。

誤訳例 "character space"という語はそれぞれ、"character→人"、"space→間"という訳が既存の対訳辞書から得られるため、日本語の"人間"という語に対応付けられる。これは英語から日本語に翻訳する際、ターゲット言語に存在する "人間"という単語が"人"と"間"による複合語ではないということがこの手法ではわからないために起こる。

このような誤訳を削除することは、"人間" から "character space" に翻訳されるかどうかを調べることによって可能となる。なぜなら、形態素解析システムによって "人間" が単語であり、"人" + "間" に分解されることはありえないということがわかるからである。

これらを踏まえて、次の処理では日→英、英→日それぞれで逆引きできなかったものを削除する。つまり、英語 A が抽出された英→日英辞書によって日本語 α に訳されたとき、日本語 α が抽出された日→英辞書によって英語 A に戻らなければ、英語 A に対する日本語訳語 α を辞書から削除する。これによって上記のような誤訳を削除することができる。

したがって、前記の**対訳の曖昧性解消手順**に続いて次の Step を追加する。

Step5 Step4 で抽出された辞書から、互いに逆引きできないものを削除する。

以上までの 5Step で得られた対訳辞書にはまだ曖昧性が残っている。たとえば、この段階では以下のような曖昧性を含むものがあった。

memory system:	メモリシステム
	記憶システム
	メモリ方式
	記憶方式
	脳組織

これは既存の対訳辞書によって、"memory→脳"、"system→組織" と翻訳され、それらから生成された "脳組織" という複合語が "memory system" の対訳であると対応付けられたからである。この例の様な場合、"脳組織" が "脳" + "組織" の複合語であるため、前述の Step5 では曖昧性解消ができない。

そのための解決策として以下のことを考えた。まず、用語の重要度に応じた順位付けを以下のような方法で行う。ただし、順位付けは以下に述べる方法以外の方法も使えるであろう。

語基の連接数による重要度と順位付け

まず、語基 N の *Pre(N)*、*Post(N)* という以下で定義される重要度を導入する。

Pre(N): コーパスから得られた用語において名詞 N に前接し複合名詞を作る名詞の種類数

Post(N) コーパスから得られた用語において名詞 N に後接し複合名詞を作る名詞の種類数

次に語基 N_1 N_2 … N_k からなる複合語の重要度を、各語基の *Pre(N)*、*Post(N)*の相乗平均で定義する。このようにして定義した重要度の高い順に抽出した用語を順位付けておく。

このような状態で、上記の問題の解決策を仮定1から次のように考えた。

(1) **用語抽出で得られた用語それぞれに、一対一の対訳が存在すると考える**
(2) **その用語を構成する語基から得られた *Pre(N)*および *Post(N)*は、対訳となる語基同士で同じような値を持つことが予想される**
(3) **よって対訳となる用語対は *Pre(N)*、*Post(N)* を用いた順位付けで得られた順位が近いはずである**

この順位付けの近さを用いた曖昧性解消によって、先の例では次のような結果が得られた。数字は、全抽出用語を 100 とみなした場合の正規化順位の差である。

memory system:	メモリシステム	0.051493
	記憶システム	0.956459
	メモリ方式	1.234347
	記憶方式	3.809609
	脳組織	63.498688

上記からわかるように、明らかに違う概念を持つ脳組織という用語が大きな順位の差を持っていることから、曖昧性の解消を行うことができると実証された。

したがって以上までに述べた、**対訳の曖昧性解消手順**に最後の Step を追加する。

Step6 Step5 までで抽出された辞書に順位の差による重み付けを行う。

以上の 6Step によってコンパラブルコーパスから辞書を抽出し、その曖昧性を解消する手法を提案する。また、本手法で用いた既存の対訳辞書は EDICT[Breen95]である。

4. 実験結果

実験には NTCIR-1 のコーパスの、電子情報通信学会の論文アブストラクト(日本語 54854 記事、英語 33311 記事)を用いた。英→日辞書、と日→英辞書を抽出した。記事数と抽出された訳語数の関係を表2に示す。

表2
抽出元記事数と抽出された対訳辞書の大きさ

記事数	日 → 英	英 → 日
100	637	361
500	1796	1030
1000	2727	1620
5000	6907	4865
10000	10556	7792
20000	15981	12417
30000	22660	17293
全記事	32376	22784
Comparable	15895	11331

ここで、表2の Comparable は完全に対訳ではない記事、日英約 16000 記事を用いて対訳辞書を抽出した結果である。

表2から、抽出元記事数が増えると抽出される対訳辞書の量も増えていることがわかる。また、完全にコンパラブルなコーパスからの抽出結果が、1万～2万記事からなるコーパスからの抽出結果の間にあることから、本手法がパラレルコーパスであるかどうかによらず、コンパラブルコーパスからの抽出でも有用であるということが確かめられた。

5. 対訳辞書の評価

本稿で提案した手法によって抽出された対訳辞書全体の評価は正解データがないため困難である。そこでサンプルで評価することにした。サンプルの正解としては、インターネット上で閲覧できる辞書である“情報・通信辞書 e-Words”を用いて評価した。

e-Words は登録語数が 2132 語(2001 年 1 月 31 日

現在)とあるが、そのうち対訳とみなせるものは表3の通りである。

表3　対訳語数

	日本語見出し語	英語見出し語
(1)	1167	1355
(2)	956	1123

ここで、(1)は対訳とみなせるものの数であり、(2)は(1)から本手法では扱っていなかった、"数詞＋名詞"というものを取り除いた後の数である。

評価対象の対訳辞書はそれぞれ、

(a)電子情報通信学会論文アブストラクト全記事(英語33311記事、日本語54854記事)

(b)情報処理学会論文アブストラクト全記事(英語12119記事、日本語15039記事)

(c)(a)および(b)全記事

からの抽出結果である。結果は表4、5に示す。

表4

日本語→英語方向

	(a)	(b)	(c)
(d)	193	177	232
(e)	154 79.8	150 84.7	186 80.2
(f)	163 84.5	157 88.7	196 84.5
(g)	-	-	-
(h)	-	-	-

表5

英語→日本語方向

	(a)	(b)	(c)
(d)	217	208	258
(e)	127 58.5	139 66.8	154 59.7
(f)	165 76.0	164 78.8	195 75.6
(g)	173 79.7	171 82.2	208 80.6
(h)	176 81.1	172 82.7	211 81.8

ここで

(d)抽出した辞書の語で、e-Wordsに登録されていた語数

(e)第一位の訳語の正解個数とその割合(%)

(f)第二位までの訳語の正解個数とその割合(%)

(g)第三位までの訳語の正解個数とその割合(%)

(h)第七位(本実験での訳語の最下位)までの訳語の正解個数とその割合(%)

である。また、日本語→英語方向ではこの場合、第二位の訳語までしか存在しなかったため、(g)、(h)については表記していない。

6.　おわりに

本研究では、日本語→英語方向で第二位までの訳語の正解カバー率が85%前後、英語→日本語方向で第三位までの訳語の正解カバー率が80%という結果が得られた。

本手法の計算量の少なさを考えると、2.節で述べた他の共起情報を利用する手法と比べて、**少ない計算量で高い精度を持つ**手法であるということがわかった。

以上の結果から、本手法での曖昧性解消が有用であると立証された。

参考文献

[Breen95] James W. Breen, Edict, freeware Japanese/English dictionary, 1995, ftp://ftp.cc.monash.edu.au/pub/nihongo/00INDEX.html

[Fung95] Pascale Fung, "Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus", *Workshop on Very Large Corpora*, pp.173-183, 1995

[Fung98] Pascale Fung, "A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora", *Lecture Notes in Artificial Intelligence*, Springer Publisher, vol. 1529, pp.1-17, 1998

[Rapp95] Reinhard Rapp, "Identifying Word Translations in Non-Parallel Texts", *33rd ACL*, pp.320-322, 1995

[Rapp99] Reinhard Rapp, "Automatic Identification of Word Translations from Unrelated English and German Corpora", *37th ACL*, pp.519-526, 1999

[Tanaka96] Kumiko Tanaka and Hideya Iwasaki, "Extraction of Lexical Translation from Non-aligned Corpora", *the 16th International Conference on Computational Linguistics*, pp.580-585, 1996