対訳コーパスから対応する表現対の自動抽出[1]

山田節夫　中岩浩巳　池原悟
ＮＴＴコミュニケーション科学研究所

1. はじめに

近年、ルールベース型機械翻訳システムが実用化されてきているが、全ての言語現象を考慮した辞書が作成されているわけではない。現状では、翻訳対象分野によらない一般的な表現に対してはシステム辞書を、また、翻訳対象分野に依存した表現に対しては翻訳対象分野に応じた辞書を人が追加修正している。しかし人手に頼っている以上、大規模な辞書を作成するには費用的、時間的なコストが高いという問題がある。

この問題を解決するために、対訳コーパスから対訳関係にある語句を自動的に抽出する様々な研究がなされている。これらは大きく分けると統計情報を利用するもの[1,2]、言語情報を利用するもの[3]、統計情報と言語情報の両方を利用するもの[4,5]が提案されている。

統計情報を利用するものは、あらかじめ人手によって作られた辞書を利用しないので、辞書の精度によらず一程度の頻度のあるものは抽出される。しかし、翻訳対象分野毎に十分な量のコーパスを用意するのは実際には困難なため、この方法のみでは辞書作成には不十分である。

言語情報を利用するものは、すでに構築されている辞書を使って対訳関係を推定しているので比較的小規模な対訳コーパスから有効な辞書が構築される。推定される対訳関係の抽出は利用する辞書の能力に依存しているが、[3]では対訳辞書のみを利用するのではなく、シソーラスから取り出した同義語も訳語の候補として利用しているので、訳語辞書にない場合でも対訳関係が抽出可能な場合もある。しかし、あらかじめ決められた専門用語の形態のみにしか対応できない。

統計情報と言語情報の両方を利用するものは、上述した利点を活かしているので、統計情報のみを利用するものから比べると小規模な対訳コーパスから未知語も含めて有効な対応関係が抽出される。しかし、統計情報を利用している限り、ある程度の頻度がないと未知語は抽出されない。さらに、抽出対象である専門用語の認定も辞書に依存しているので、決まった形の品詞列しか取り扱えない。

本稿では、比較的小規模な対訳コーパスから言語情報、構文情報を利用して、機械翻訳システムで意味がある対訳関係のある表現対を抽出する方法について述べる。ここで、対訳関係のある表現対には、既存の辞書に依存した単語そのもの以外に句全体として対応が取れたものも含んでいる。また、言語情報、構文情報は既存の機械翻訳システムを利用して得ている。

2. 対応の取れる表現対の抽出

本章では、対訳コーパスから対応する表現対を抽出する処理について述べる。以下にこの処理の流れを示す。

1. 対訳コーパス中の目的言語の文を構文解析し、構文構造を保存する。（これを理想訳の構文構造とする。）
2. 対訳コーパス中の原言語の文を機械翻訳システムに入力して目的言語である機械訳の構文構造を得る。
3. 上記2で得られた機械訳の構文構造と上記1で保存された理想訳の構文構造を比較して、対応関係が単語レベルと認定された表現対（単語辞書に登録できる表現対）を抽出する。
4. 上記3で抽出された表現対とその周りの文節情報から対象分野用辞書を作成する。
5. 対象分野用辞書を使って同じ対訳コーパスを再翻訳し、再度機械訳の構文構造を得る。
6. 対応関係の認定された表現対が増えなくなるまで、上記3、4、5を繰り返す。

[1] Automatically Extracting Aligned Expressions using a Bilingual Corpora

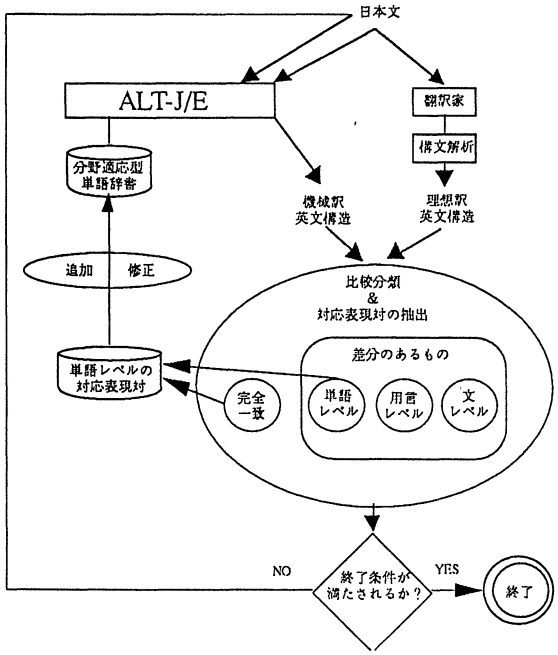


図 1：対象分野の単語辞書を自動作成する実現例

この処理を、我々が開発中の日英機械翻訳システムALT-J/E [6] に適用したのが、図 1である。次節より、各項目について詳しく述べる。

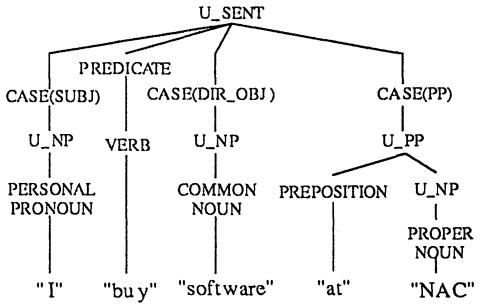
2.1. 理想訳と機械訳の比較

対訳コーパス中の目的言語（理想訳）の構文構造はALT-J/Eで採用している結合価文法に基づいて現在は人手で作成している。機械訳の構文構造はALT-J/Eに原言語を入力した結果得られる。「私は日本ＡＢＣ社でソフトを買う。」に対する理想訳の構文木と機械訳の構文木を図 2に示す。

理想訳と機械訳の構文構造を比較しその差分から対応関係が認定されたところを抽出するのは、基本的に我々が提案した分野適応型翻訳機構 [7] に基づいて行う。この方法では、まず差分のあったものを機械翻訳システムでの解決方法別（単語辞書で解決すべきもの、パターン対辞書（用言を中心とした単文構造を変換するための辞書）で解決すべきもの、ルールで解決すべきもの）に分類し、それぞれに対して分野適応型辞書やルールを構築するような枠組みになっている。本稿では、単語辞書に関するもののみを取り扱うので、単語辞書で解決すべきであると分類されたもの（機械訳と理想訳の構文構造の差分が単語レベルであるもの）が対象となる。

例「私は日本ＡＢＣ社でソフトを買う」

理想訳: I buy software at NAC.



機械訳: I buy a soft hat in a Japanese ABC company.

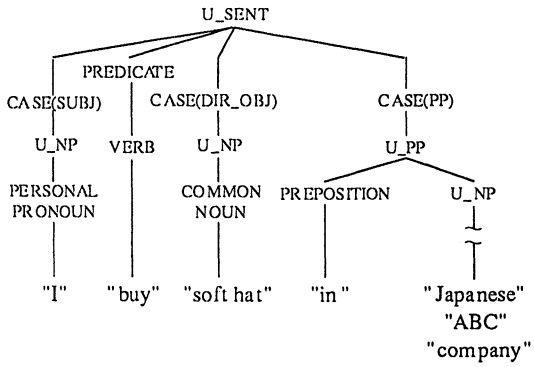


図 2：英語構文木の例

[7] で提案した方法では、格要素を一つの単位として対応が取れるかどうかを辞書を用いて決めている。図 2の例の場合、機械訳と理想訳の格要素の対応の取れ具合は表 1のようになる。これはALT-J/E が翻訳に使用するシステム辞書を用いただけの結果であるので、ここから得られる対応の取れた表現対は“私”に対して“I”、“ソフト”に対して“software”のみである。これでは、システム辞書に依存したものしか抽出できないので、我々は以下に挙げる方法（拡張対応法）を用いることで、未知語も抽出できるようにした。

(A) 語尾変形（派生語、ing 形等）の違いは、英語辞書の情報を利用して対応を取る。
（例　edit と editing の対応を取る）

(B) 前置詞は数に限りがあり、対象分野用に新規に登録する必要がない上に、動詞に依存して前置詞が変わる可能性もある。そこで、格要素内の単語の対応が取れない場合は、その格要素内の

表 1：格要素の対応の取れ具合の例

		理想訳の格要素		
		"I"	"software"	"at NAC"
機械訳の格要素	"I"	一致	不一致	不一致
	"soft hat"	不一致	他訳語一致	不一致
	"in Japanese ABC company"	不一致	不一致	不一致

一致：字面一致
他訳語一致：機械翻訳システムが選ばなかった他の訳語候補と一致
不一致：一致も他訳語一致もしない

機械訳と理想訳の両方から前置詞を除いて再度対応を取る。

(C) 格要素の中で対応が取れない単語が機械訳と理想訳どちらかが1単語で残っている場合はそれらの対応を取る。
(例 機械訳：
electronic character generating unit
理想訳：
electronic character generator
"generating unit" と "generator" の対応を取る)

(D) 未対応の格要素が機械訳、理想訳にそれぞれ1つずつ残っている場合はそれらの対応を取る。(例 図 2では残っている未対応の格要素がそれぞれに1つずつあるので、理想訳の格要素 "at NAC" と機械訳の格要素 "in Japanese ABC company" の対応を取る。)

以上の方法を取り入れると、図 2の例の場合、(D)の方法によって "at NAC" と "in Japanese ABC company" の対応が取れ、(B)の方法によって前置詞が取り除かれるので、"NAC" と "Japanese ABC company" の対応が取れる。その結果、未知語である "日本ＡＢＣ社" と "NAC" が対応付けられる。

2.2. 抽出された表現対の辞書登録

抽出された表現対は、そのままでは機械翻訳システムでは役に立たないので、抽出された前後の語と抽出された頻度を一緒に辞書に登録した。表現対に付随されたこれらの情報をもとに訳語選択が行われる。さらに、対応の取れた表現対が未知語あるいは格要素全体であった場合、形態素解析でそれらを一つの単語として認識してほしいので、翻訳対象用日本語辞書にも登録した。これにより、その翻訳対象分野である決まった表現をするような語や句が一単語として正しく解析される。また、辞書に登録する前に機械訳と理想訳の字面が一致したものに関しては、すでにシステムが持っている辞書で十分翻訳できるものであるので、除外した。例えば、図2の例から抽出されるのは "私" と "I"、"ソフト" と "software"、"日本ＡＢＣ社" と "NAC" であるが、辞書に登録するのは、"私" と "I" 以外のものである。

こうして新しく構築された翻訳対象分野の辞書を利用して同じ対象文集で再翻訳を行う。再翻訳された機械訳の構文構造と理想訳の構文構造を再度比較する。これは、一回目の抽出時には未知語が多くて対応が取れなかったもののなかに、すでに未知語が登録されたために抽出が可能になるかもしれないからである。つまり、翻訳対象分野の辞書をも使って対応する表現対を再度抽出するのである。このプロセスを繰り返し、抽出される表現対が増加しなくなった時に終了する。

3. 実験

[7] で作った単語対応事例抽出のプロトタイプに2章で述べた拡張対応法を組み込んで、ALT-J/E を用いて実験をした。このとき対象とした文は、比較するために [7] で用いたものと同じ技術マニュアル文から単文（５７７文、理想訳の格要素数は約７００）を選び、再現率、適合率で評価した。評価式は下記の通りである。

$$\text{再現率} = \frac{\text{プロトタイプで抽出された正しい格要素対の数}}{\text{人手で対応の取れる格要素対の数}}$$

$$\text{適合率} = \frac{\text{プロトタイプで抽出された正しい格要素対の数}}{\text{プロトタイプで抽出された全格要素対の数}}$$

その結果を表 2に示す。適合率は拡張対応法を組み込む前よりも低くなっているが、再現率は大幅に上がった。しかし総合的には組み込んだ後の方が良いので、拡張対応法は有効と言える。

人手では抽出できたがアルゴリズムでは抽出で

きなかったものは、下記に示すものである。

- 一つの格要素内に複数の未知語がある場合
  （例　日：“オンラインＭＡ装置、
  オフラインＭＡ装置”
  英：“on-line MA equipment,
  off-line MA equipment”）
- 一つ文の中に格要素全体で対応が取れているもの（例では下線の部分）と格要素全体で対応の取れないものが同時に複数ある場合
  （例　日：“特にことわりのない限り、
  大セミナー室と中セミナー室は”
  英：“Unless otherwise specified,
  both the Main and Medium Seminar
  Rooms”）
- 日本語の解析失敗が原因である場合

一方、人手では抽出していないがアルゴリズムでは抽出してしまったものは、下記に示すものである。

- 記号と記号の対応が取れてしまった場合
  （例　日：“、”
  英：“and”）
- 英文の格要素内にそれと対応する日本文にはない情報（文脈や世界知識等からくる情報（例では下線の部分））が含まれている場合
  （例　日：“ハードディスク等”
  英：“hard disk and other related equipment”）
- 上記の逆で、日本文にあった情報（例では下線の部分）が英文では省略されている場合
  （例　日：“監視カメラ装置”
  英：“Monitor camera”）
- 日本語の解析失敗が原因である場合

上記のうち今後解決できそうな問題は、記号と記号の対応が取れてしまった場合（数件）ぐらいで、残りの問題はこの方法では簡単には解決できない。したがって、この方法では、表 2の拡張対応法を取り入れた結果以上にするのは困難である。

表 2：再現率、適合率の変化

拡張対応法	再現率	適合率
なし	$\frac{280}{619}$ = 45.2%	$\frac{280}{280}$ = 100%
あり	$\frac{580}{619}$ = 93.7%	$\frac{580}{620}$ = 93.5%

4. おわりに

本稿では、言語情報と解析情報を利用して対訳コーパスから対応する表現対の自動抽出を行う方法を提案した。この方法を用いると、入手が容易な比較的小規模な対訳コーパスからでも、未知語が取得できる。また、用言だけの対応関係を除いて、特に専門用語の形を決めているわけではないので、辞書に依存したある決まった形の品詞列以外も抽出できる。これは、我々が提案した分野適応型翻訳機構[7]の単語対応事例に関するプロトタイプに2章で述べた拡張対応法を組み込むことで実現されている。本手法を評価したところ、再現率93.7%、適合率93.5%の結果が得れらた。

本稿では人手で理想訳の構文構造を作成しているが、理想訳のほうは構文構造を利用しなくても、本稿で提案した方法が扱えるように検討中である。また、今後は対象文が単文から複文、重文等のより複雑な文に対しても適用できるように検討する予定である。

参考文献

[1] 井ノ上，野垣：対訳テキストを用いた日英対訳辞書の自動作成，電子情報通信学会，NLC93-39，pp.65-72，1993

[2] 森，長尾：ｎグラム統計によるコーパスからの未知語抽出，情報処理学会，95-NL-108，pp.7-12，1995

[3] 石本，長尾：対訳文章を用いた専門用語対訳辞書の自動作成－訳語対応における両立不可能性を考慮した手法について－，言語処理学会第１回年次大会，pp.217-220，1995

[4] 山本，坂本：対訳コーパスを用いた専門用語対訳辞書の作成，情報処理学会，93-NL-94，pp.85-92，1993

[5] A.Kumano，H.Hirakawa：Building an MT Dictionary form Parallel Texts Based on Linguistic and Statistical Information，Proc of COLING-94，pp.76-81，1994

[6] 池原，宮崎，白井，林：言語における話者の認識と多段翻訳方式，情報処理学会論文誌，Vol.28，No.12，pp.1269-1279，1987

[7] S.Yamada，H.Nakaiwa，K.Ogura，S.Ikehara：A Method of Automatically Adapting a MT System to Different Domains，Proc of TMI-95，303-310，1995